UNIVERSIDAD DEL PACÍFICO | Ingeniería Empresarial
Facultad de Ingeniería

CARACTERIZACIÓN DE LA DINÁMICA EPIDEMIOLÓGICA Y NIVEL DE CONOCIMIENTO SOBRE EL VIH/SIDA EN LOS HABITANTES DEL PERÚ

Tesis presentada para optar al Título profesional de Ingeniero Empresarial

Presentado por

Paulo José Alejandro Aybar Flores

Asesor: Alvaro Gustavo Talavera Lopez
0000-0002-2193-4270

Lima, octubre 2021

Resumen

Los estudios de análisis epidemiológico de enfermedades como el VIH/SIDA han adquirido una gran importancia y significado en el diseño y gestión de la salud pública alrededor del mundo ya que, mediante estos, se puede optar por estrategias y medidas para controlar y, ulteriormente, reducir a largo plazo la epidemia. De esta manera, esta investigación se ha realizado con el propósito de caracterizar la dinámica epidemiológica del VIH/SIDA, mediante un modelo determinístico S-I-R de Cadenas de Markov, y el nivel de conocimiento sobre la epidemia en hombres y mujeres en el Perú, a través de modelos de Machine Learning para predicción y agrupamiento, fundamentado en la evaluación de su relación con determinantes estructurales sociales e información de vigilancia sanitaria nacional. Las bases de datos secundarias para el presente estudio fueron extraídas de la Dirección Epidemiológica del Perú, el Programa Conjunto de las Naciones Unidas sobre el VIH/SIDA, el Banco Mundial y el Instituto Nacional de Estadística e Informática del Perú. En consideración de los hallazgos obtenidos, se establece que existen diferencias en las tasas de infectividad y mortalidad del virus entre los períodos de estudio, el comportamiento estacionario de las tendencias evolutivas de la epidemia en cada año de análisis sigue las premisas teóricas de desarrollo epidemiológico de una enfermedad a nivel poblacional y las proyecciones estocásticas del VIH/SIDA en el país evidencian la necesidad de estrategias estatales de intervención y control sanitario con un enfoque integrado y holístico a fin de frenar la epidemia. Asimismo, se establece la conexión entre ciertos factores socio-demográficos, económicos y de salud con el nivel de conocimiento sobre las formas de prevención y rechazo de ideas erróneas sobre la transmisión del VIH/SIDA en el Perú. Adicionalmente, la capacidad de predicción del nivel de conocimiento sobre la epidemia se estimó entre los valores de 59.47 % y 64.30 % para los modelos considerados, siendo el modelo de Random Forest (64.30 %) el que mostró el mejor desempeño. Finalmente, a partir de la aplicación del SOM o la red neuronal de Kohonen, se pudo distinguir la conformación concreta de 5 conglomerados poblacionales que mostraron perfiles únicos que podrían usarse para identificar características estructurales subyacentes al nivel de conocimiento sobre VIH/SIDA en el país. Investigaciones futuras deberán considerar técnicas y/o métodos suplementarios a los presentados previamente que permitan analizar la situación epidemiológica del VIH/SIDA en el Perú e impulsar nuevos espacios para la investigación científica bajo distintas configuraciones o contextos a raíz de una escasa literatura nacional referente al tema y la presencia de dificultades y/o complejidades técnicas en el área de estudio.

Palabras Clave: VIH/SIDA, DGE, ONUSIDA, ENDES, Cadenas de Markov, Determinantes sociales, Regresión cuasi-binomial, Machine Learning, Análisis de conglomerados, SOM, Perú.

Abstract

Epidemiological analysis studies of diseases such as HIV/AIDS have acquired great importance and significance in the design and management of public health around the world since, through these, it is possible to choose strategies and measures to control and, subsequently, reduce the epidemic in the long-term. Thus, this research has been carried out with the purpose of characterizing the epidemiological dynamics of HIV/AIDS, through a deterministic SIR model of Markov Chains, and the level of knowledge about the epidemic in men and women in Peru, by way of Machine Learning models for prediction and clustering, based on the evaluation of their relationship with social structural determinants and national health surveillance information. The secondary databases for the present study were extracted from the Center for Disease Control and Prevention of Peru, the Joint United Nations Program on HIV/AIDS, the World Bank and the National Institute of Statistics and Informatics of Peru. In consideration of the findings obtained, it is established that there are differences in the infectivity and mortality rates of the virus between the study periods, the stationary behavior of the evolutionary trends of the epidemic in each year of analysis follows the theoretical premises of epidemiological development of a disease at the population level and the stochastic projections of HIV/AIDS in the country show the need for state intervention and health control strategies with an integrated and holistic approach in order to curb the epidemic. Likewise, the relationship between certain socio-demographic, economic and health factors and the level of knowledge about the forms of prevention and rejection of erroneous ideas about the transmission of HIV/AIDS in Peru is confirmed. Additionally, the ability to predict the level of knowledge about the epidemic was estimated between the values of 59.47 % and 64.30 % for the models considered, being the Random Forest model (64.30 %) the one that showed the best performance. Finally, from the application of the SOM or Kohonen's neural network, it was possible to distinguish the specific conformation of 5 population clusters that showed unique profiles that could be used to identify structural characteristics underlying the level of knowledge about HIV/AIDS in the country. Future research should consider supplementary techniques and/or methods to those previously presented that allow analyzing the epidemiological situation of HIV/AIDS in Peru and promoting new spaces for scientific research under different configurations or contexts as a result of a scarce national literature on the subject and the presence of technical difficulties and/or complexities in the study area.

Keywords: HIV/AIDS, CDC, UNAIDS, DHS, Markov Chains, Social determinants, Quasi-binomial Regression, Machine Learning, Cluster Analysis, SOM, Peru.

Dedicatoria

Esta disertación está humildemente dedicada a mi querida madre Jenny Flores, cuyo amor, apoyo y aliento inquebrantables han enriquecido mi alma y me han inspirado a continuar y completar esta investigación. Sin su enorme sacrificio personal, orientación sin fin y su amor incondicional, nunca me hubiera convertido en el individuo que soy hoy.

Había prometido enorgullecer a mi madre con el logro de esta monumental meta académica y espero haber cumplido ese juramento, teniendo la oportunidad de compartir la celebración y el éxito de este acontecimiento junto a ella.

Agradecimientos

Un número significativo de personas ha contribuido de diversas formas para permitirme alcanzar el éxito en la realización de esta tesis.

Me gustaría expresar mi más sincero agradecimiento a mi asesor, el Dr. Alvaro Talavera, quien me acompañó, motivó y me brindó la orientación necesaria para completar esta investigación.

Agradezco las inmensas contribuciones del Dr. Walter Aliaga, el Dr. Luciano Stucchi y el Dr. Akram Hernández en mi búsqueda para producir un trabajo exitoso y el cumplimiento de mis objetivos académicos.

Doy gracias a Dios por bendecirme con las fuerzas necesarias para emprender y llevar a cabo este proyecto, sin desistir o flaquear en ningún momento.

Finalmente, mi gratitud va a mi familia y amistades por su apoyo, cariño e inspiración constante durante todo este proceso académico.

Índice general

Índice de figuras

Índice de tablas

Glosario

VIH Virus de Inmunodeficiencia Humana

SIDA Síndrome de Inmunodeficiencia Adquiridad

OMS Organización Mundial de la Salud

ENDES Encuesta Demográfica y de Salud Familiar

ART/TAR Terapia AntirRetroviral

ONUSIDA Programa Conjunto de las Naciones Unidas sobre el VIH/SIDA

HSH Hombres que tienen Sexo con Hombres

DGE Dirección General de Epidemiología del Perú

ITS Infección de Transmisión Sexual

SOM Self-Organizing Map

ETS Enfermedad de Transmisión Sexual

TARGA Tratamiento Antirretroviral de Gran Actividad

PVV/PVVIH Personas que viven con el VIH/SIDA

ONU Organización de las Naciones Unidas

S Estado Susceptible

I Estado de Infección por VIH/SIDA

M/R Estado de fallecido o retirado por infección

S.T. Sin Tratamiento contra el VIH/SIDA

INEI Instituto Nacional de Estadística e Informática del Perú

Capítulo 1

Introducción

1.1. Consideraciones preliminares

El VIH/SIDA (virus de la inmunodeficiencia humana y síndrome de inmunodeficiencia adquirida, respectivamente) es una de las enfermedades infecciosas más devastadoras en la historia de la humanidad desde su descubrimiento en 1981. Aproximadamente 78 millones de personas han contraído la infección y han fallecido unos 35 millones de individuos por enfermedades asociadas al VIH/SIDA desde el inicio de la epidemia a nivel mundial desde la década de los años 80 (Boza, 2016). En la actualidad, la lucha contra el VIH/SIDA es considerada como un aspecto crucial y prioritario dentro de las políticas de salud pública a nivel mundial a raíz de los asoladores efectos de la enfermedad por su alto potencial de propagación y elevada letalidad si es que no se cuenta con tratamientos y medidas contenciosas (OMS, 2016). La diseminación del VIH a lo largo del mundo se originó en distintos contextos y circunstancias mediante numerosos mecanismos de contagio como la transmisión vertical, contaminación a través de transfusiones sanguíneas y/o utilización de material médico infectado, contacto por vía sexual; entre otros (Soto, 2004). Estos hechos, aunados con la acción gubernamental de cada nación, han determinado la preocupante variedad de escenarios que han caracterizado la problemática que esta enfermedad representa y la necesidad de estudiar la naturaleza y el desarrollo de esta. Este acuciamiento por analizar la situación epidemiológica del VIH/SIDA ha sido abordado mediante profusas investigaciones que se han realizado al respecto, conforme la enfermedad evoluciona en diferentes niveles de enfoque de estudio (Teva *et al.*, 2012; APMG Health, 2019).

Gran parte de la literatura de investigación al respecto del VIH/SIDA, en términos generales, está fundamentada o elaborada sobre la base de modelos matemáticos de infección y enfermedad, los cuales resultan significativos en: su aplicación en la construcción de distintas metodologías y formas de abordar estas problemáticas, su uso como herramientas de analítica para la estimación de parámetros epidemiológicos y su seguimiento como guías de la información para el mejoramiento del entendimiento epidemiológico y en el planeamiento de programas de control. En el caso de los estudios acerca del VIH/SIDA, Mukandavire *et al.* (2009) expresan que el modelamiento matemático y las simulaciones computacionales de los esfuerzos sobre el control epidemiológico de la enfermedad se han convertido en herramientas

imperativas para la evaluación de diferentes políticas a nivel de estado, de la evolución de la salud a nivel poblacional e de intervenciones en sectores gubernamentales afines al de salud.

Sin embargo, pese a los avances en el mundo de la medicina y de las ciencias de la salud en los últimos veinte años, a saber: la reducción del número de nuevas infecciones por el VIH en todo el mundo de 3,4 millones en 1996 a 1,8 millones en 2017 y la disminución por la expansión del tratamiento del 34 % de las muertes por causas relacionadas con el SIDA entre 2000 y 2017 (El Fondo Mundial, 2019); el VIH y el SIDA continúan siendo una amenaza sanitaria para numerosos países alrededor del mundo, teniendo implicancias en el diseño y ejecución de las políticas públicas y para el desarrollo cotidiano de diversos grupos poblacionales en riesgo: a través de las consecuencias económicas y sociales adversas del VIH/SIDA en las personas y los hogares, la epidemia también crea desafíos para la política social traducidos en la disminución de la capacidad estatal en sectores como la salud, educación, gasto social, financiamiento y productividad de una nación; al mismo tiempo, las repercusiones fiscales (impacto en la administración pública, el gasto público y los ingresos públicos) van mucho más allá de lo que generalmente se subsume en los costos o la respuesta a la epidemia y el aumento en los costos y la pérdida en la eficiencia de los servicios públicos descata la necesidad de un planeamiento y desarrollo apropiado de las respuestas gubernamentales ante el virus para mitigar los efectos estructurales que pueda ocasionar (Haacker, 2004).

La actual cobertura de servicios es inadecuada y el ritmo de su expansión es demasiado lento para alcanzar las metas mundiales (a fines de 2014, de los 37 millones de personas infectadas por el VIH en todo el mundo, 17 millones no conocían su estado serológico, y 22 millones de personas no tenían acceso a los tratamientos antirretrovíricos) y el éxito en la respuesta mundial frente al VIH no se distribuye de manera pareja ni equitativa (las violaciones de los derechos humanos, junto con la violencia de género, la estigmatización y la discriminación ampliamente generalizadas, continúan obstruyendo el acceso a los servicios de salud, en particular para los niños, los adolescentes, las mujeres jóvenes y las poblaciones vulnerables como las minorías LGBT, entre otras) en algunos países y regiones (OMS, 2016). Frente a ello, el análisis epidemiológico de la infección y de la enfermedad (que se traduce en estudios prospectivos, las tendencias de la evolución y patrones de las dinámicas de infección e influencias que diversos indicadores tengan sobre estas) permite una mejor toma de decisiones a nivel gubernamental en materia de salud pública/social y las acciones que se derivan de los hallazgos y evaluaciones de los resultados de este mismo análisis se establecen como los ejes principales de los esfuerzos contra la epidemia y sus estragos.

Los estudios de análisis epidemiológico de enfermedades como el VIH han adquirido una gran importancia y significado en el diseño y gestión de las políticas de salud públicas alrededor del mundo ya que, mediante la evaluación y monitoreo de la infección a nivel poblacional, se puede analizar las tendencias que adopta la enfermedad en el tiempo y los factores que poseen influencia sobre esta a fin de optar por estrategias y medidas para controlarla y, ulteriormente, erradicarla a largo plazo (OPS, 2002).

Asimismo, a través de estos mismos análisis, difundir el conocimiento y la conciencia sobre

el VIH/SIDA resulta en una de las estrategias clave que se utilizan en la prevención y el control de la epidemia en todo el mundo. Los conocimientos inadecuados y las prácticas de riesgo son los principales obstáculos para prevenir la propagación del virus (Alhasawi *et al.*, 2019). En muchos países, las enfermedades de transmisión sexual (ETS) y los embarazos no planificados se observan con frecuencia entre los adolescentes. Los jóvenes comenzaron a tener relaciones sexuales con una o varias parejas sexuales de forma indiscriminada, y esto facilitó la propagación de las ETS y el VIH. Por lo tanto, los adolescentes en general tienen un mayor riesgo de contraer el VIH a través de la transmisión sexual. Por ello, se torna aún más importante la necesidad de comprender el conocimiento y la actitud de los jóvenes hacia el VIH/SIDA y los esfuerzos públicos para un enfoque personalizado de control y prevención de enfermedades a través de programas de educación y concienciación fundamentales. Los esfuerzos exitosos de control de enfermedades dependen de comprender tanto la distribución como la frecuencia de los comportamientos de salud y medir el conocimiento del público en general sobre el VIH / SIDA y las asociaciones de sus conocimientos y actitudes con diferentes factores sociodemográficos (Janahi *et al.*, 2016).

De esta manera, este trabajo consta de tres etapas: la primera propone un modelo probabilístico y la aplicación de herramientas estadísticas de proyección para determinar la evolución del VIH/SIDA en el Perú permitiendo una evaluación del efecto de las políticas y gestión de salud relacionados al virus y la enfermedad por parte del estado peruano. Esta etapa permitirá conocer el contexto epidemiológico del VIH/SIDA en el Perú y los diferentes cursos de evolución que el virus pueda tomar dependiendo de la gestión pública de la epidemia por parte del estado peruano; estos hallazgos proveerán un fundamento para pesquisar sobre el nivel de comprensión y entendimiento sobre la transmisión y dinámica del VIH en la población juvenil del país que expliquen las tendencias detectadas en las curvas de infección.

La segunda etapa de este trabajo formula modelos de regresión binomial, incorporando dentro de su construcción la estructura compleja del diseño muestral correspondiente a los datos, que estudian a la población de hombres y mujeres jóvenes y jóvenes adultos en territorio nacional considerados en la Encuesta Demográfica y de Salud Familiar - ENDES del año 2019, considerando que es el año con la información más reciente de la población adolescente y joven adulta del país al momento de realizar la presente investigación, utilizando como variables predictoras a diversos factores demográficos, económicos y sociales a fin de determinar aquellos que posean una asociación significativa e influencien al nivel de conocimiento que los entrevistados puedan tener acerca del VIH/SIDA. A su vez, se formulan modelos matemáticos lineales (regresión logística múltiple) y no lineales (redes neuronales artificiales, árboles de decisión, random forest y algoritmo k-NN) utilizando como variables regresoras a factores que caracterizan a los hogares peruanos que forman parte de la muestra, con independencia de las características individuales de grupos poblacionales específicos, para estimar la probabilidad del nivel de conocimiento sobre el virus y la enfermedad. Este estudio proveerá de información al gobierno sobre aquellos determinantes de la salud que influyen directamente en el conocimiento y percepción que los individuos tengan en relación con el VIH/SIDA en el país para una mejor toma de decisiones y la obtención de resultados más eficientes y be-

neficiosos para las principales poblaciones de riesgo y la formulación de políticas de salud y sexualidad. Esta etapa concede la posibilidad de evidenciar qué factores estructurales son los que influencian directamente sobre el conocimiento del VIH/SIDA por parte de la población juvenil en territorio nacional (aquellos en lo que se puede intervenir de manera eficiente para generar un cambio a nivel conductual y educativo de los adolescentes y jóvenes del país a fin de mejorar la situación epidemiológica del virus) y ser capaces de clasificar el perfil de conocimiento de un individuo en base a estos atributos socio-demográficos, económicos y de salud.

Finalmente, la tercera etapa en el proyecto desarrolla un análisis de conglomerados exploratorio a partir de la base de datos de hogares de hombres y mujeres en el Perú (Encuesta Demográfica y de Salud Familiar - ENDES) para determinar agrupamientos (clusters) del nivel de conocimiento sobre el VIH/SIDA que los individuos tengan a lo largo del territorio nacional. Este estudio le permitirá al estado peruano y los organismos descentralizados de control formular regímenes de concientización y control epidemiológico más acordes a las condiciones demográficas, económicas, sociales y sanitarias de la población peruana y la capacidad técnica del aparato estatal.

1.2. Alcance

Los estudios de análisis epidemiológico de enfermedades como el VIH han adquirido una gran importancia y significado en el diseño y gestión de las políticas de salud publicas alrededor del mundo ya que, mediante la evaluación y monitoreo de la infección a nivel poblacional, se pueden analizar las tendencias que adopta la enfermedad en el tiempo y los factores que poseen influencia sobre esta a fin de optar por estrategias y medidas para controlarla y, ulteriormente, reducirla a largo plazo. Asimismo, actualmente existe una escasa literatura a nivel nacional que trate sobre el VIH/SIDA a través de un enfoque de machine learning y/o determinístico, por lo que este estudio resulta novedoso y un punto de partida en el campo de la investigación epidemiológica del virus en el Perú. De esta manera, este trabajo se ha realizado con el propósito de estudiar la evolución prospectiva de los efectos del VIH/SIDA en el tiempo en la población peruana (referidos a la tasa de susceptibles o población en riesgo de contraer el virus, tasa de infectados con VIH/SIDA y aquellos fallecidos producto de la enfermedad), la efectividad de la política de salud establecida para un horizonte de tiempo determinado proyectándolo al futuro y la influencia que ciertos factores socio-demográficos, económicos y de salud ejercen sobre el comportamiento y conocimiento de adolescentes y jóvenes adultos sobre el virus y la enfermedad a lo largo del país. Por consiguiente, esta investigación aporta resultados útiles en la gestión de las políticas de salud y en el análisis de los efectos de la epidemia del VIH en el Perú para orientar y mejorar las medidas sanitarias que el gobierno peruano planee implementar como esfuerzos contra el avance del VIH/SIDA en el país.

1.3. Objetivo de la tesis

El objetivo general del estudio es caracterizar la dinámica epidemiológica del VIH/SIDA y el nivel de conocimiento sobre la epidemia en hombres y mujeres en el Perú mediante la evaluación de su relación con determinantes estructurales sociales e información de vigilancia sanitaria nacional.

De manera complementaria, el objetivo específico a alcanzar dentro de la caracterización de la dinámica epidemiológica del VIH/SIDA en el país es el siguiente:

1. Diseñar y evaluar un modelo probabilístico basado en Cadenas de Markov para la estimación determinística de la situación epidemiológica y la evolución prospectiva del VIH/SIDA en la población peruana a nivel nacional posterior a cambios significativos en materia de salud pública relacionada a la epidemia.

Por otro lado, los objetivos específicos subordinados a la caracterización del nivel de conocimiento sobre la epidemia en hombres y mujeres en el Perú son:

1. Identificación de aquellos determinantes de la salud (factores demográficos, económicos y sociales) en el territorio peruano que poseen una influencia empírica en el nivel de conocimiento sobre el VIH/SIDA en la población de adolescentes y jóvenes adultos en el país en el 2019.
2. Definición del modelo computacional que proporcione la mejor bondad de ajuste y precisión para la estimación y clasificación del nivel de conocimiento del VIH/SIDA en la población juvenil en el Perú mediante una comparación de técnicas paramétricas y no paramétricas.
3. Evaluación del análisis de conglomerados como una herramienta potencialmente útil para definir empíricamente asociaciones resultantes y facilitar la evaluación de medidas por parte del estado al identificar los tipos de patrones y/o comportamientos que siguen los hombres y mujeres que residen en el país basados en factores demográficos, económicos, sociales y el nivel de conocimiento que posean acerca del VIH/SIDA; cuantificando y localizando aglomeraciones a lo largo del país.

1.3.1. Contribución de la investigación

El análisis y cumplimiento de los objetivos plasmados anteriormente resultarán ser útiles y relevantes para los gestores de políticas públicas (principalmente, sanitarias) en el Perú para reforzar, justificar, dirigir y apoyar en el planeamiento y ejecución de mejoras y nuevas medidas en los esfuerzos referidos a contrarrestar la epidemia del VIH/SIDA en el país, ya que, en la medidad en que sean alcanzados, podrán proveer información acerca de la evolución del virus y la enfermedad en el tiempo, permitirán identificar y emplear modelos para el pronóstico de aspectos relevantes que caracterizan la dinámica infecciosa de la misma y posibilitarán identificar potenciales focos de atención de actuales y nuevas políticas de salud que

se basan en aquellos indicadores/determinantes que prueban tener una conexión con el nivel de conocimiento que los individuos puedan tener sobre el VIH/SIDA, representando así una herramienta para la evaluación del rol del gobierno peruano en relación a este problema de salud pública y para posibles formas de optimizar e incentivar un eficiente control y reducción sobre este en el futuro.

1.4. Estructura de la tesis

Para alcanzar los objetivos del proyecto, se consideran las siguientes secciones que formarán parte de la organización del presente trabajo de investigación:

1. Antecedentes y consideraciones epidemiológicas del estudio.
2. Marco teórico (Bases teóricas y perspectivas teóricas-contextualización del VIH/SIDA).
3. Estudio del modelo de evolución probabilística del VIH/SIDA.
4. Determinantes del conocimiento del VIH/SIDA.
5. Análisis de conglomerados sociales del VIH/SIDA.
6. Conclusiones y recomendaciones.
7. Anexos.

En el mismo sentido, la estructura metodológica a emplear a lo largo de la presente tesis se evidencia en la Figura 1.1 exhibida a continuación.

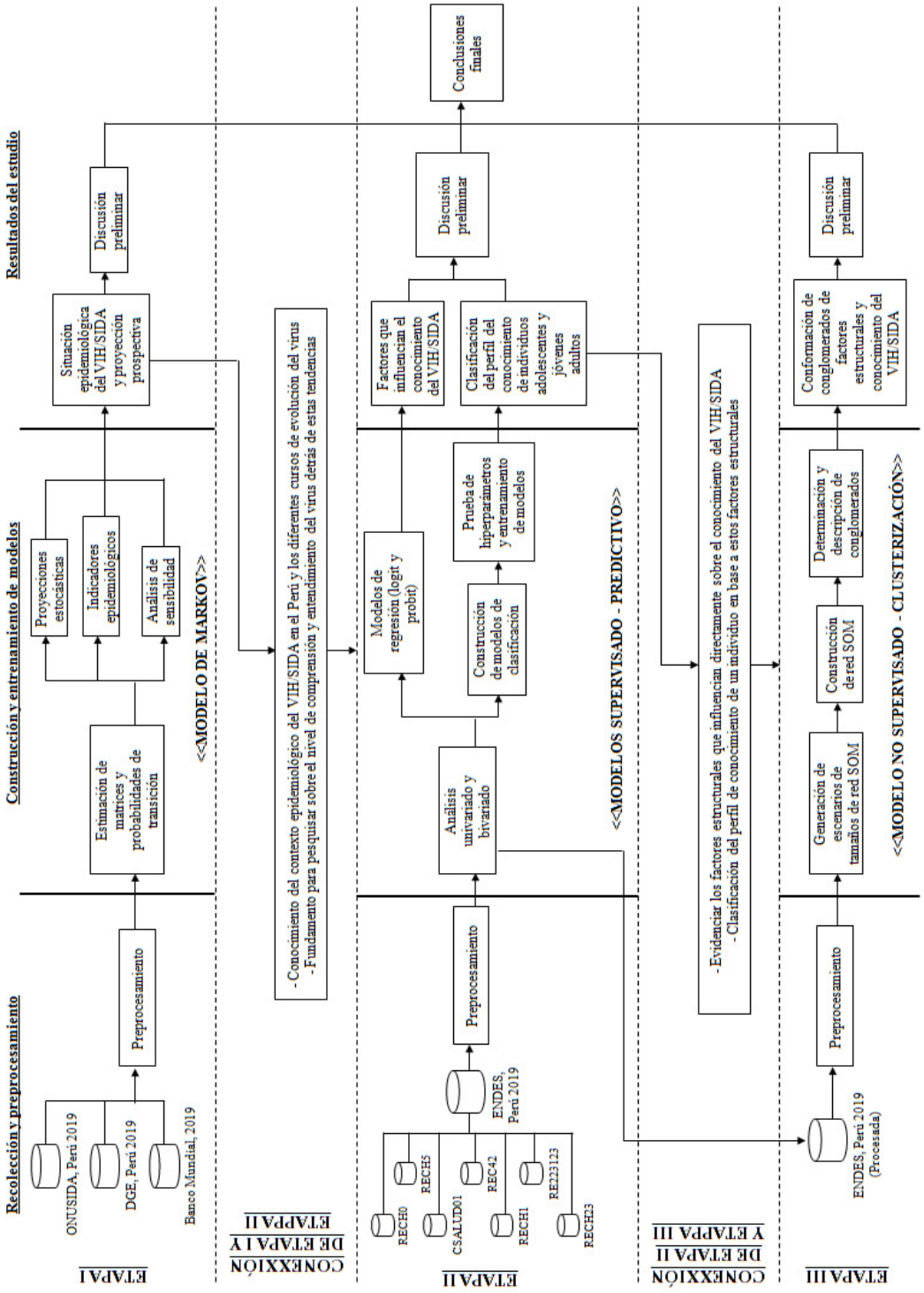


Figura 1.1: Diagrama de metodología para la presente investigación. Fuente: Elaboración propia.

Capítulo 2

El VIH/SIDA: Perspectivas teóricas de la epidemia

Como Fajardo-Ortiz *et al.* (2017) señalan, el VIH/SIDA es una de las enfermedades infecciosas más estudiadas, con más de 260,000 artículos (que mencionan el tema) que figuran en GOPubMed y más de 42,000 artículos (que mencionan el VIH/SIDA en el título) en la Web de la Ciencia, que abarca más de 30 años de investigación científica. Sus hallazgos indican que el VIH/SIDA es estudiado por una pluralidad de disciplinas biomédicas como epidemiología, virología, inmunología o desarrollo de fármacos y disciplinas no biomédicas como las ciencias sociales y las humanidades. Todas las disciplinas biomédicas que trabajan sobre el VIH/SIDA dependen en gran medida de un sólido consenso científico, que explica la manifestación clínica del virus y la enfermedad en términos de las interacciones del virus con las células del sistema inmunitario, el comportamiento y la demografía de las células del sistema inmunitario y, lo más importante, la interacción del virus con la maquinaria biomolecular de las células huésped.

Basándonos en el precedente de investigación provisto por Fajardo-Ortiz *et al.* (2017) y de la significancia de los estudios relacionados al VIH/SIDA, el presente capítulo se centra en la definición clínica y distribución epidemiológica del VIH/SIDA y los diversos enfoques adoptados para el modelamiento del virus y la enfermedad en el tiempo.

2.1. Contextualización de la epidemia

Prevenir la propagación del VIH/SIDA y trabajar eficazmente con las personas ya infectadas requiere una comprensión básica del problema y de los pasos necesarios para abordarlo. Una comprensión básica del problema incluye: conocimiento básico del VIH y cómo causa enfermedades relacionadas con el SIDA, conocimiento de cómo se transmite el VIH y los comportamientos que facilitan la transmisión, conocimiento sobre la situación clínica de pacientes infectados y conocimiento de la terapia, diagnóstico y prevención del VIH/SIDA disponibles.

2.1.1. Definición del VIH/SIDA

Sims (2009) afirma que el virus de la inmunodeficiencia humana (VIH) es la causa de una de las pandemias humanas más destructivas en la historia registrada. Desde su primer reconocimiento en 1981, ha matado a, aproximadamente, 35 millones de personas a nivel mundial

(Boza, 2016). Sims (2009) expresa que a menudo hay confusión entre los términos SIDA y VIH. El síndrome de inmunodeficiencia adquirida (SIDA) es un conjunto de síntomas que ocurren en la etapa final de una infección causada por el virus de inmunodeficiencia humana (VIH) (Sims, 2009). El virus de la inmunodeficiencia humana (VIH) es un virus que ataca a las células inmunes llamadas células CD4, que son un tipo de célula T. Estos son glóbulos blancos que se mueven alrededor del cuerpo, detectando fallas y anomalías en las células, así como infecciones. Cuando el VIH ataca e infiltra estas células, reduce la capacidad del cuerpo para combatir otras enfermedades. Sims (2009) determina que se transmite por contacto con ciertos fluidos corporales de una persona con VIH, más comúnmente durante el sexo sin protección (sexo sin condón o medicamento contra el VIH para prevenir o tratar el VIH), o al compartir equipos de drogas inyectables.

Sims (2009) manifiesta que el SIDA ocurre cuando el virus ha destruido el sistema inmune, dejando al paciente altamente susceptible a otras infecciones que amenazan la vida. Sin tratamiento, es probable que la infección por VIH se convierta en SIDA a medida que el sistema inmunitario se debilita gradualmente. Sin embargo, los avances en ART significan que un número cada vez menor de personas progresa a esta etapa.

Sims (2009) precisa que se considera que una persona con VIH ha progresado a SIDA cuando:

1. El número de sus células CD4 cae por debajo de 200 células por milímetro cúbico de sangre (200 células/mm^3). (En alguien con un sistema inmunitario sano, los recuentos de CD4 están entre 500 y 1.600 células/mm^3).
2. Desarrollan una o más infecciones oportunistas independientemente de su recuento de CD4.

Sin medicamentos contra el VIH, las personas con SIDA generalmente sobreviven unos 3 años (Sims, 2009). Una vez que alguien tiene una peligrosa enfermedad oportunista, la esperanza de vida sin tratamiento cae a aproximadamente 1 año. No obstante, los fármacos antirretrovirales aún pueden ayudar a los individuos en esta etapa de la infección por el VIH, e incluso puede salvarles la vida (Sims, 2009).

2.1.2. Cuadro clínico de un paciente infectado con VIH/SIDA

Con respecto a los mecanismos de transmisión, HIV.gob (2019) establece que el VIH se encuentra en fluidos específicos del cuerpo humano. Existen formas muy específicas en que el VIH se puede transmitir a través de fluidos corporales, como las siguientes:

1. **Durante el contacto sexual**: Se puede contraer el VIH a través del sexo anal, oral o vaginal.
2. **Durante el embarazo, el parto o la lactancia**: Los bebés pueden contraer el VIH a través del contacto que tienen con los fluidos corporales de su madre, incluidos los fluidos amnióticos y la sangre, durante el embarazo y el parto y post-parto.

3. **Como resultado del uso de drogas inyectables**: Las agujas o medicamentos que están contaminados con sangre infectada con VIH pueden transmitir el virus directamente a su cuerpo.

4. **Como resultado de la exposición ocupacional**: Los trabajadores de la salud tienen el mayor riesgo de este tipo de transmisión del VIH porque pueden entrar en contacto con sangre infectada u otros fluidos a través de pinchazos o cortes.

5. **Como resultado de una transfusión de sangre con sangre infectada o un trasplante de órgano de un donante infectado**: Este método de transmisión es extremadamente raro debido a los requisitos de detección.

Con respecto a la sintomatología de un paciente infectado con VIH/SIDA, HIV.gob (2019) formula que el virus a veces puede hacer que las personas se sientan enfermas, pero la mayoría de los síntomas y enfermedades graves del VIH provienen de las infecciones oportunistas que atacan el sistema inmunitario dañado. También es importante reconocer que algunos síntomas del VIH son similares a las enfermedades comunes, como la gripe o las infecciones respiratorias.

A continuación se presentan etapas de la sintomatología del VIH en pacientes detectados:

(a) **Etapa 1 (infección aguda por VIH)**: Dentro de las 2 a 4 semanas posteriores a la infección con el VIH, aproximadamente dos tercios de las personas tendrán una enfermedad similar a la gripe. Esta es la respuesta natural del cuerpo a la infección por VIH. Los síntomas seudogripales pueden incluir: fiebre, escalofríos, erupción, sudores nocturnos, dolores musculares, dolor de garganta, fatiga, ganglios linfáticos inflamados y úlceras en la boca. Estos síntomas pueden durar desde unos pocos días hasta varias semanas.

(b) **Etapa 2 (latencia clínica)**: En esta etapa, el virus todavía se multiplica, pero a niveles muy reducidos. Los individuos en esta fase pueden no sentirse enfermos o presentar alguna sintomatología. Esta etapa también se denomina como infección crónica por VIH. Sin tratamiento contra el virus, las personas pueden permanecer en este estadio durante 10 o 15 años, pero algunas otras pueden llegar a este estado más rápido. Sin embargo, si el infectado asume el tratamiento contra el VIH de forma diaria, exactamente como se lo recetaron y obtiene y mantiene una carga viral indetectable, este puede proteger su salud de forma eficaz y prevenir la transmisión a otros.

(c) **Etapa 3 (SIDA)**: Si el paciente tiene VIH y no está en tratamiento contra el virus, eventualmente este debilitará el sistema inmunitario de su cuerpo y progresará a SIDA (síndrome de inmunodeficiencia adquirida). Esta es la etapa tardía de la infección por VIH. Los síntomas del SIDA pueden incluir: pérdida de peso rápida, fiebre recurrente o sudores nocturnos profusos, cansancio extremo e inexplicable, hinchazón prolongada de las glándulas linfáticas en las axilas, la ingle o el cuello, diarrea que dura más de una semana, llagas en la boca, ano o genitales, neumonía crónica, manchas rojas, marrones, rosadas o violáceas sobre o debajo de la piel o dentro de la boca, nariz o párpados y/o pérdida de memoria, depresión y otros trastornos neurológicos.

2.1.3. Distribución epidemiológica del VIH/SIDA en el mundo y América Latina

2.1.3.1. Distribución epidemiológica del VIH/SIDA a nivel mundial

ONUSIDA (2019) informa que el progreso general contra la epidemia de VIH se mide a través del cálculo de estimaciones de nuevas infecciones por VIH y muertes por causas relacionadas con el SIDA. Las tendencias en todo el mundo siguen siendo similares: ha habido un progreso global constante en la reducción de las muertes relacionadas con el SIDA en la última década, y un progreso más gradual en la reducción de nuevas infecciones por VIH.

Por otra parte, ONUSIDA (2019) afirma que el número anual de muertes por enfermedades relacionadas con el SIDA entre las personas que viven con el VIH a nivel mundial se ha reducido de un máximo de 1,7 millones en 2004 a 770 000 en 2018, como lo muestra la Fig. 2.1. Desde el 2010, la mortalidad relacionada con el SIDA ha disminuido en un 33 %. La disminución mundial de las muertes ha sido impulsada en gran medida por el progreso en África oriental y meridional, que alberga al 54 % de las personas del mundo que viven con el VIH. Las muertes relacionadas con el SIDA en las regiones de Europa oriental y Asia central y Oriente Medio y África del Norte han aumentado un 5 % y un 9 %, respectivamente, durante el período de ocho años.

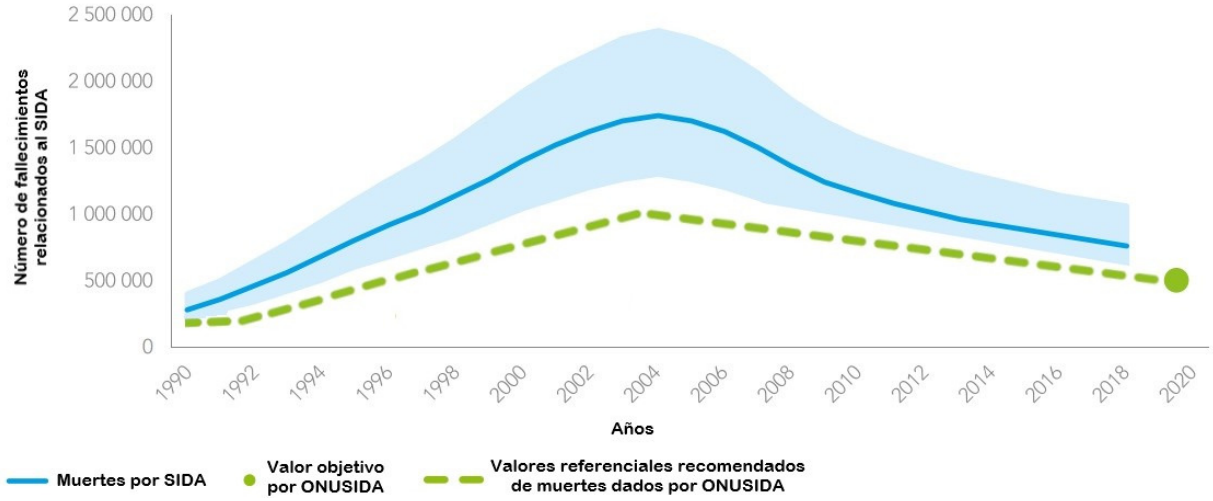


Figura 2.1: Número de muertes relacionadas con el SIDA a nivel mundial durante el período 1990–2018 y el objetivo 2020 (la sombra de color celeste representa a los intervalos de confianza que oscilan sobre la cantidad de muertes relacionadas con la enfermedad SIDA). Fuente: ONUSIDA, 2019.

Por otro lado, ONUSIDA (2019) reporta que el número de nuevas infecciones por el VIH en todo el mundo continuó disminuyendo gradualmente en 2018. El número anual de nuevas infecciones desde 2010 ha disminuido de 2,1 millones a 1,7 millones en 2018, una reducción del 16 %, como lo muestra la Fig. 2.2. La reducción de nuevas infecciones por el VIH entre 2010 y 2018 fue más fuerte en África oriental y meridional (disminución del 28 %). También hubo avances en el Caribe (disminución del 16 %), África occidental y central (disminución del 13 %), Europa occidental y central y América del Norte (disminución del 12 %), y Asia y

el Pacífico (9 %). Sin embargo, el número anual de nuevas infecciones por VIH ha aumentado en Europa oriental y Asia central (aumento del 29 %), Oriente Medio y África del Norte (aumento del 10 %) y América Latina (aumento del 7 %).

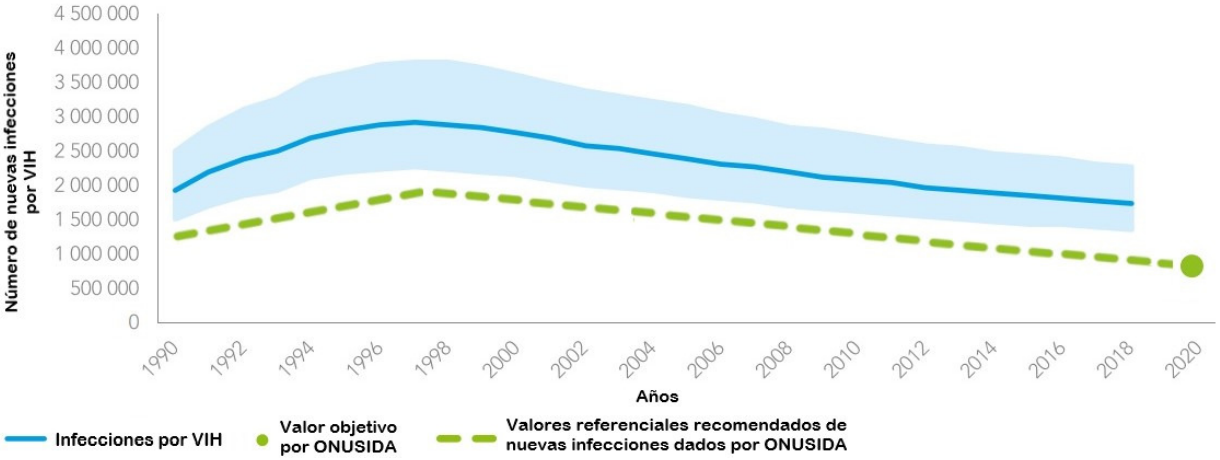


Figura 2.2: Número de nuevas infecciones por VIH, global, 1990–2018 y objetivo 2020. Fuente: ONUSIDA, 2019.

A su vez, ONUSIDA (2019) y sus socios han desarrollado métricas de transición epidémica como medidas complementarias que los países pueden utilizar para seguir mejor su progreso hacia el fin del SIDA como una amenaza para la salud pública.

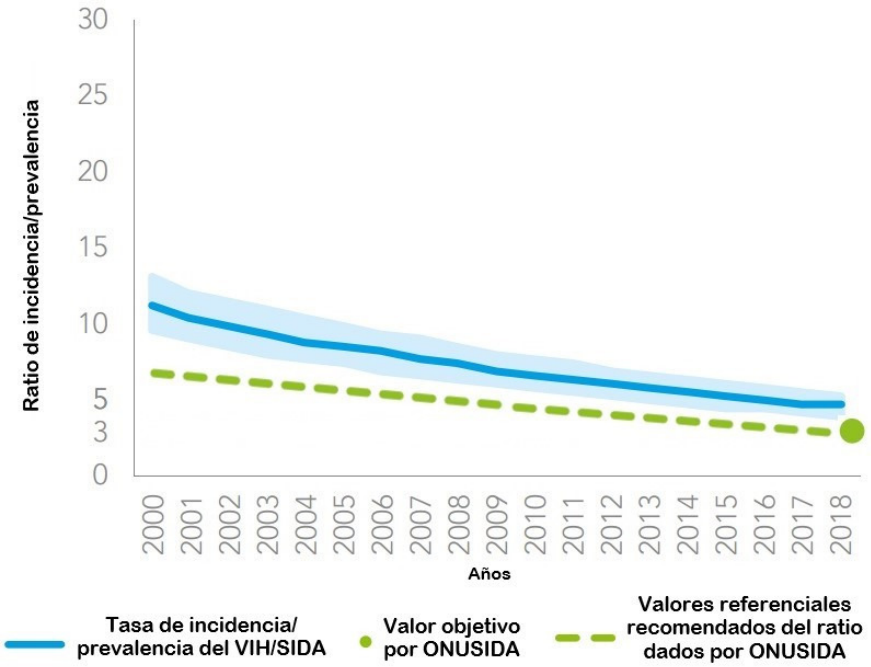


Figura 2.3: Proporción de nuevas infecciones por personas que viven con el VIH, global, 2000–2018 y objetivo 2020 (la línea verde representa el objetivo a largo plazo propuesto por ONUSIDA que el ratio debe alcanzar). Fuente: ONUSIDA, 2019.

Una de esas métricas, el ratio de incidencia-prevalencia, utiliza el número de nuevas infecciones por VIH y el número de personas que viven con el VIH dentro de una población para

producir el inverso del tiempo promedio que una persona vive con el VIH en una epidemia que permanece estable durante muchos años. La relación global de incidencia-prevalencia ha disminuido del 11,2 % en 2000 al 6,6 % en 2010 al 4,6 % en 2018, como lo muestra la Fig. 2.3, lo que refuerza la conclusión de que se han logrado importantes avances contra la epidemia.

2.1.3.2. Distribución epidemiológica del VIH/SIDA en Latinoamérica

En el caso de América Latina, según ONUSIDA (2019), la respuesta al VIH se financia principalmente con recursos nacionales. Sin embargo, no ha habido suficiente inversión estatal por parte de los gobiernos latinoamericanos en programación para poblaciones clave, incluida la expansión de los servicios de prevención para hombres homosexuales y otros hombres que tienen sexo con hombres (HSH), trabajadoras sexuales y personas transgénero: para el 2015, el nivel de inversión gubernamental o gasto público promedio destinado a los esfuerzos contra el VIH/SIDA dentro del sector de salud en la región como porcentaje del PIB ascendió a, aproximadamente, 1.2 % (un incremento en 0.20 % a comparación del 2014), siendo superado notablemente por el gasto público en recursos contra el VIH de regiones como Europa Occidental y América del Norte (con una inversión de 5.8 %), Asia y Pacífico (con una inversión de 2.1 %) y África Occidental y Central (con una inversión de 4 %) (UNAIDS, 2015). En el mismo sentido, América Latina enfrenta desafíos adicionales, que incluyen niveles de migración que aumentan dramáticamente debido a la incertidumbre sociopolítica (ONUSIDA, 2019).

ONUSIDA (2019) estima que 100 000 personas contrajeron el VIH en América Latina en 2018, un aumento del 7 % en comparación con 2010. El número anual de muertes relacionadas con el SIDA en la región disminuyó en un 14 % entre 2010 y 2018, con un estimado de 35 000 vidas perdidas por causas relacionadas con el SIDA en 2018. La relación incidencia-prevalencia de la región continúa disminución, alcanzando 5.4 % en 2018, pero un mayor progreso es requerido para alcanzar el punto de referencia de transición epidémica de 3.0 %. Finalmente, aproximadamente la mitad de los países de la región experimentó un aumento en la incidencia entre 2010 y 2018, y se produjeron los mayores aumentos en Brasil (21 %), Costa Rica (21 %), el Estado Plurinacional de Bolivia (22 %) y Chile (34 %). Al mismo tiempo, se registraron descensos visibles en El Salvador (-48 %), Nicaragua (-29 %) y Colombia (-22 %).

Finalmente, ONUSIDA (2019) expresa que el número anual de muertes relacionadas con el SIDA en la región disminuyó en un 14 % entre 2010 y 2018, con un estimado de 35 000 vidas perdidas por causas relacionadas con el SIDA en 2018. La relación incidencia-prevalencia de la región continúa disminución, alcanzando 5.4 % en 2018, pero se necesita más progreso para alcanzar el punto de referencia de transición epidémica de 3.0 %.

2.1.4. Distribución epidemiológica del VIH/SIDA en el Perú

En el caso peruano, el boletín sobre VIH-SIDA diseñado por el DGE (2020b), recopila los principales indicadores de la epidemia en el país. Anualmente, se presenta el análisis de las notificaciones de casos, de transmisión madre-hijo, de sida y de un conjunto de indicadores de gestión en prevención, diagnóstico y tratamiento.

De acuerdo con el boletín VIH/SIDA 2018 (DGE, 2020b), desde el año 1983 en que se reportó el primer caso de sida en el país, hasta el 31 de diciembre de 2018 se han notificado un total de 120389 casos de infección por VIH, de los cuales 43072 se encuentran en estadio SIDA. En la Fig. 2.4, se observa incluidos en la curva de casos de infección VIH a todos los estadios de la infección (incluso el estadio SIDA).

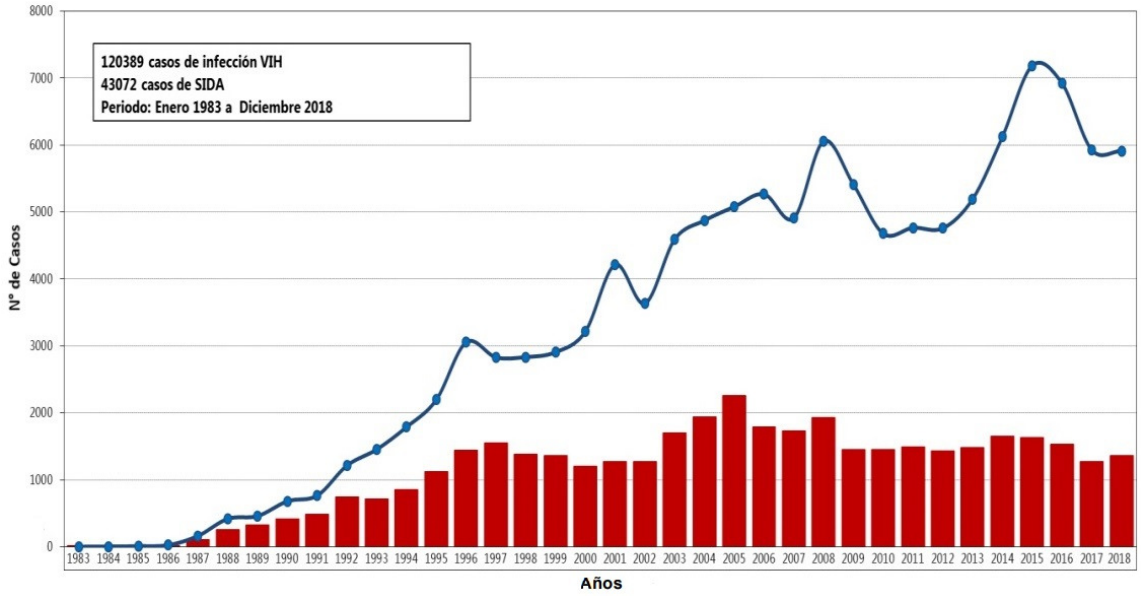


Figura 2.4: Casos de infección por VIH y casos de SIDA notificados, según año de diagnóstico en Perú, 1983-2018. Fuente: DGE Perú, 2018.

En el mismo sentido, DGE (2020b) reporta que en el 2018 hasta diciembre la relación es 3.7 hombres por una mujer en los casos diagnosticados de infección VIH y para los casos SIDA es 4.5 hombres por una mujer, como se puede observar en la Fig. 2.5 y la Fig. 2.6, respectivamente.

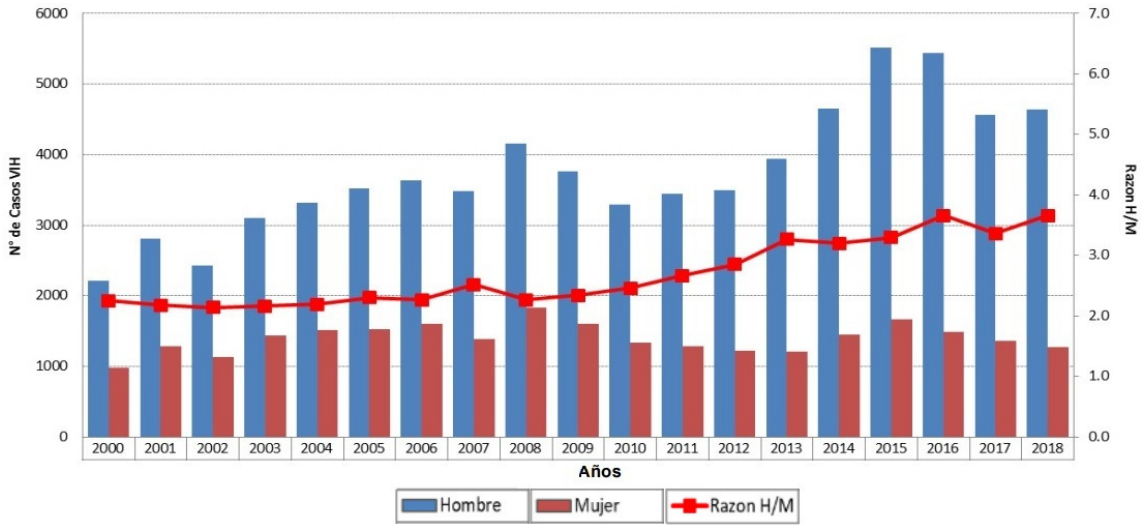


Figura 2.5: Casos de infección por VIH notificados según sexo y razón hombre/mujer, Perú 2000-2018. Fuente: DGE Perú, 2018.

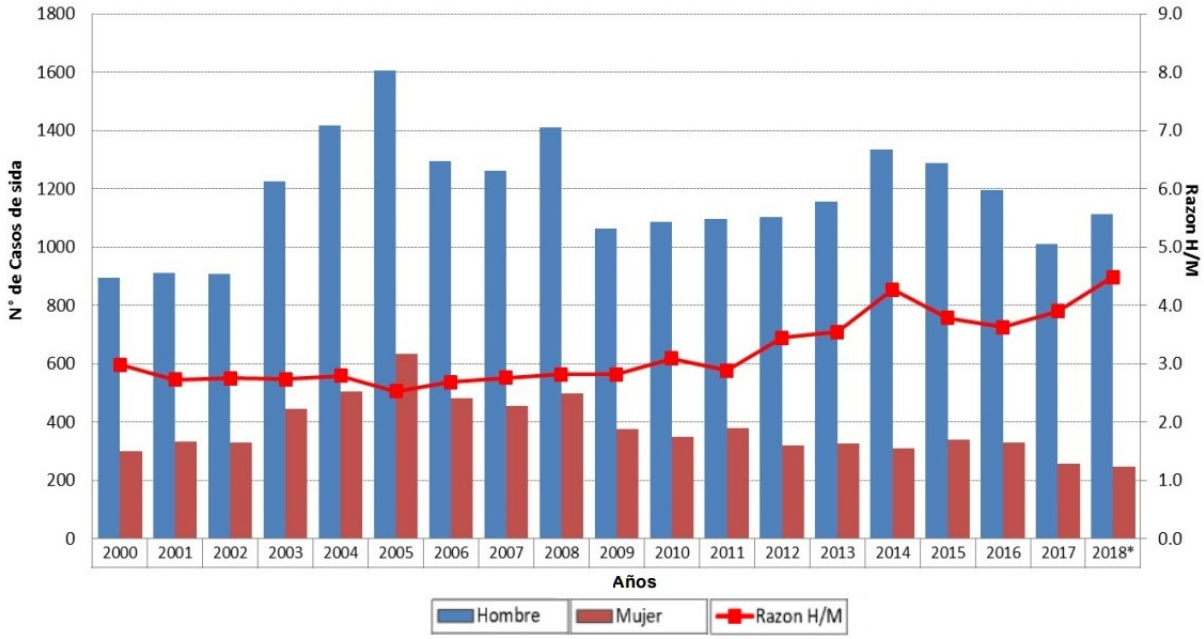


Figura 2.6: Casos de sida notificados según sexo y razón hombre/mujer, Perú 2000-2018. Fuente: DGE Perú, 2018.

Por otra parte, DGE (2020b) señala que del total de casos notificados en VIH en el período de 1983 a diciembre 2018 la vía de transmisión más frecuente es la vía sexual con 97.58 %, seguido del 1.98 % por transmisión madre-niño (vertical) y 0.44 % vía parenteral, como se aprecia en la Fig. 2.7.

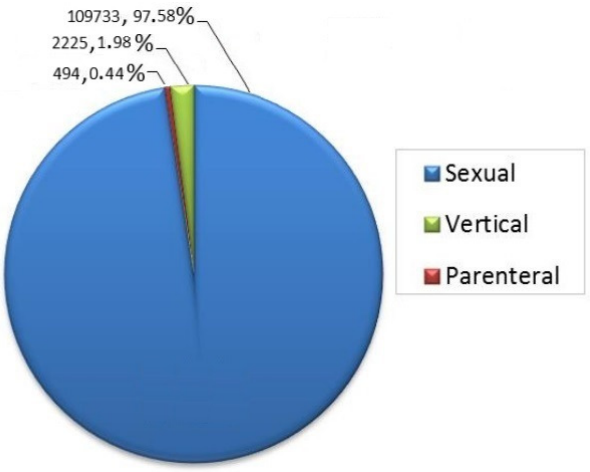


Figura 2.7: Vía de transmisión en casos de VIH acumulados, Perú 1983-2018. Fuente: DGE Perú, 2018.

A su vez, DGE (2020b) indica que los casos de VIH de Lima y Callao sumados a los casos de Loreto, La Libertad, Arequipa, Ica y Lambayeque; representan el 80.7 % de todos los casos de VIH notificados en el período 2000 a 2018 como lo muestra la Fig. 2.8. Para los casos de Sida notificados, el 80.2 % de ellos se concentran en Lima, Callao, Loreto, Ica y Arequipa como lo muestra la Fig. 2.9.

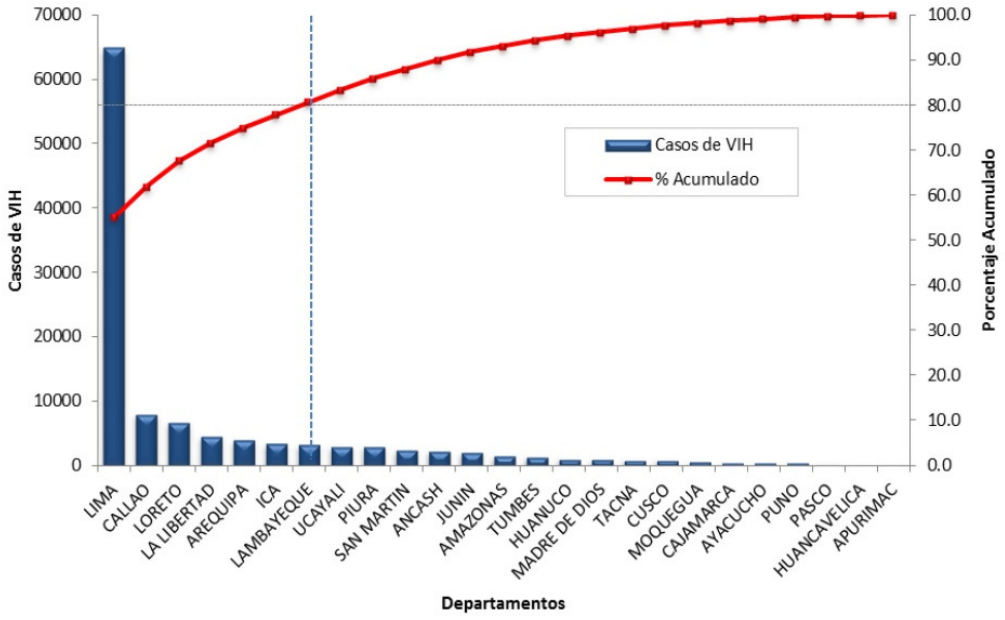


Figura 2.8: Frecuencia acumulada de casos de VIH notificados por departamento, Perú 2000-2018. Fuente: DGE Perú, 2018.

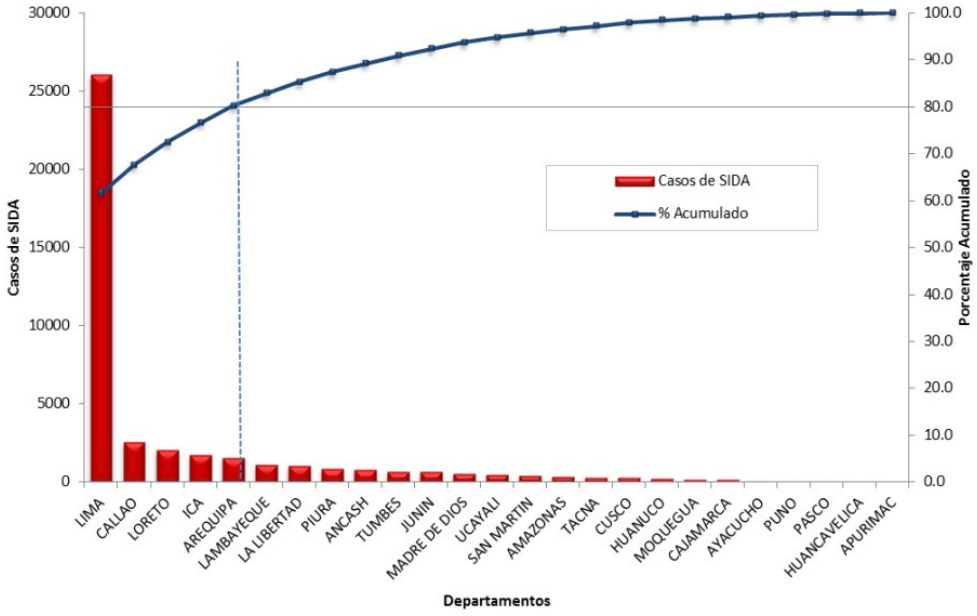


Figura 2.9: Frecuencia acumulada de casos de sida notificados por departamento, Perú 2000-2018. Fuente: DGE Perú, 2018.

Finalmente, la Defensoría del Pueblo (2009) afirma que, aunque exista un reconocimiento a las acciones hasta hoy realizadas por parte del estado, el balance de la respuesta multisectorial a la epidemia del VIH/SIDA en el país evidencia la ausencia de políticas claras desde los diferentes sectores del Estado.

La Defensoría del Pueblo (2009) precisa que diversas evaluaciones e informes realizados reconocen la necesidad de una respuesta multisectorial orgánica en la lucha contra el SIDA para

optimizar el uso de recursos y potenciar el accionar de los diferentes actores. Asimismo, esta señala que, en la actualidad, la respuesta inmediata al VIH en el país está determinada, de manera importante, por los compromisos asumidos con el Fondo Global para la ejecución de proyectos sometidos y financiados.

Ello, ciertamente, implica grandes oportunidades, ligadas a la magnitud del financiamiento y a su potencial impacto en caso de ser bien conducidos, pero también involucra grandes amenazas ligadas a la posibilidad de no aprovechar esta inversión si se usa en intervenciones inefectivas o mal implementadas (Defensoría del Pueblo, 2009).

Si bien es el Sector Salud el que ha mostrado mayores avances, la Defensoría del Pueblo (2009) expresa que aún no ha sido posible movilizar una respuesta integral efectiva frente a la epidemia del VIH y SIDA desde este sector, en el cual se han hecho cada vez más evidentes los requerimientos por parte de la población afectada para acceder a mejores condiciones de vida y gozar de una atención integral en salud.

Capítulo 3

Marco Teórico

Este capítulo presenta las bases y antecedentes teóricos sobre el diseño y aplicación de los modelos propuestos en la presente investigación. La revisión del marco teórico incluye: fundamentos sobre teoría estocástica y cadenas de Markov de tiempo discreto, nociones sobre la regresión logística cuasi-binomial de la familia logit y probit, bases sobre modelos paramétricos y no paramétricos de predicción y conceptos sobre las redes neuronales de Kohonen o redes SOM para agrupamiento de datos.

Asimismo, incluye la revisión de literatura relevante y otras aplicaciones de los modelos propuestos en la presente investigación. En primer lugar, se analizará la teoría estocástica basada en cadenas de Markov en el campo de la investigación epidemiológica, explorando sus aplicaciones concretas en el estudio y evaluación del VIH/SIDA. A su vez, se explorarán las conexiones de diferente índole entre los determinantes de salud (factores ambientales/biológicos, económicos y sociales que caracterizan a los individuos y comunidades) y aspectos relevantes sobre el VIH/SIDA basados en modelos de estimación paramétricos y no paramétricos. Finalmente, se examinarán aplicaciones del análisis de conglomerados exploratorio para la identificación a nivel territorial de individuos según condiciones y/o comportamientos relacionados a la salud.

3.1. Bases Teóricas

3.1.1. Modelamiento probabilístico del VIH/SIDA

Esta sección se centra en todos los métodos y modelos estadísticos utilizados en el análisis del modelamiento probabilístico del VIH/SIDA en el Perú. Las técnicas que se discutirán incluyen: procesos de Markov discretos y probabilidades de transición, estimación prospectiva de probabilidades y estados de transición, estimación de métricas epidemiológicas asociadas al VIH/SIDA y análisis de sensibilidad de variables de intervención.

3.1.1.1. Tipos de procesos estocásticos

Dependiendo de las condiciones particulares de cada estudio, la metodología para el análisis de procesos estocásticos difiere (Ocaña-Riola, 2009). Cuando los valores que puede tomar el

proceso son discretos suele hablarse de cadenas de Markov, mientras que el término proceso de Markov suele reservarse para procesos con espacio de estados continuo. Aunque la terminología no está estandarizada, la clasificación general de los modelos markovianos puede esquematizarse como se muestra en la Tabla 3.1.

Tabla 3.1: Clasificación de los modelos de Markov. Fuente: Ocaña-Riola, 2009.

CLASIFICACIÓN DE LOS MODELOS DE MARKOV

		Espacio de Estados	
		Discreto	Continuo
Tiempo de observación	Discreto	Cadena de Markov discreta	Proceso de Markov discreto
	Continuo	Cadena de Markov continua	Proceso de difusión

3.1.1.2. Cadenas de Markov discretas

Bertsekas & Tsitsiklis (2008) precisan que las cadenas de Markov de tiempo discreto son aquellas cadenas en las cuales el estado cambia en ciertos instantes de tiempo discreto, indexado por una variable entera n. La cadena de Markov se describe en términos de sus probabilidades de transición p_{ij}: siempre que el estado sea i, hay una probabilidad p_{ij} de que el siguiente estado sea igual a j. Matemáticamente:

$$p_{ij} = P(X_{n+1} = j | X_n = i), \quad i, j \in S. \tag{3.1}$$

3.1.1.3. Propiedad de Markov

Matemáticamente, asumimos la propiedad de Markov que requiere que:

$$P(X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P(X_{n+1} = j | X_n = i) = p_{ij} \tag{3.2}$$

para todos los tiempos n, todos los estados $i, j \in S$, y todas las secuencias posibles $i_0, ..., i_{n-1}$ de estados anteriores. La ley de probabilidad del siguiente estado X_{n+1} depende del pasado solo a través del valor del estado actual X_n (Bertsekas & Tsitsiklis, 2008).

3.1.1.4. Clasificación de los estados en una cadena de Markov

Sea $A(i)$ el conjunto de estados a los que se puede acceder desde i. El estado i es recurrente si por cada estado j que es accesible desde el estado i, i también es accesible desde j (Bertsekas & Tsitsiklis, 2008).

Un estado es denominado transitorio si no es recurrente: hay estados $j \in A(i)$ de modo que i no es accesible desde j (Bertsekas & Tsitsiklis, 2008).

Por otro lado, un estado i de una cadena de Markov se denomina absorbente si es imposible

alcanzar otro estado fuera del mismo en el tiempo (es decir, $p_{ij} = 1$) (Bertsekas & Tsitsiklis, 2008).

3.1.1.5. Probabilidades de transición en la n-ésima etapa

Bertsekas & Tsitsiklis (2008) señalan que la ley de probabilidad de un estado es capturada por las probabilidades de transición de n pasos, definidas por:

$$r_{ij}(n) = P(X_n = j | X_0 = i). \tag{3.3}$$

Donde $r_{ij}(n)$ es la probabilidad de que el estado después de n períodos de tiempo sea j, dado que el estado actual es i.

Se puede calcular utilizando la siguiente recursión básica, conocida como la ecuación de Chapman-Kolmogorov:

$$r_{ij}(n) = \sum_{k=1}^{m} r_{ij}(n-1)p_{kj}, \quad \text{para } n > 1, \quad \text{y todo } i, j, \tag{3.4}$$

donde m es la cantidad finita de estados en la ecuación incluidos en el conjunto de espacios S.

En el mismo marco, existe otro método adicional para estimar las matrices de probabilidad de transición de n-ésimo paso (P^n) empleando el enfoque de autovalores y autovectores (Bhat, 1984). La matriz de probabilidad de transición P^n puede ser estimada mediante un método de descomposición que requiere autovalores (λ) y sus correspondientes autovectores calculables de forma determinística.

Por lo tanto, se puede estimar computacionalmente utilizando la siguiente ecuación:

$$P^n = Q\Lambda^n Q^{-1}, \tag{3.5}$$

Sea Q una matriz $m \times m$ no singular $(X_1, X_2, \ldots, X_n)$, donde X_i es el autovector derecho que pertenece al autovalor $\lambda_i (i = 1, 2, \ldots, m)$ y Q^{-1} una matriz $m \times m$ no singular, donde Y_i es el autovector izquierdo que pertenece al autovalor $\lambda_i (i = 1, 2, \ldots, m)$. Expresando que Λ es una matriz diagonal de autovalores $\lambda_i (i = 1, 2, \ldots, m)$. Por lo que:

$$\Lambda^n = \begin{pmatrix} \lambda_1^n & \cdots & \cdots \\ \vdots & \ddots & 0 \\ \vdots & 0 & \lambda_m^n \end{pmatrix}. \tag{3.6}$$

3.1.1.6. Comportamiento estacionario en las cadenas de Markov

Como las probabilidades de estado estacionario o estable π_j suman 1, estas forman una distribución de probabilidad en el espacio de estado, llamada distribución estacionaria de la cadena (Bertsekas & Tsitsiklis, 2008). Usando el teorema de probabilidad total, se tiene que:

$$P(X_1 = j) = \sum_{k=1}^{m} P(X_0 = k)p_{kj} = \sum_{k=1}^{m} \pi_k p_{kj} = \pi_j, \tag{3.7}$$

donde la última igualdad se deduce del teorema de convergencia en estado estacionario. Del mismo modo, se obtiene $P(X_n = j) = \pi_j$, para todo n y j.

3.1.1.7. Modelamiento S-I-R en Cadenas de Markov

Uno de los modelos más sencillos para el estudio de enfermedades de forma estocástica establece que cada sujeto de la población puede estar exclusivamente en uno de los siguientes estados discretos en un período t de tiempo: el sujeto susceptible puede permanecer no infectado - estado 0 a estado 0 (S); no obstante, tras el contacto con un infectado desarrollará la infección y podrá contagiar a otros individuos de la población - estado 0 a estado 1 (I); cuando el sujeto infectado, puede morir por causa de la enfermedad - estado 1 a estado 2 (R) (Ocaña-Riola, 2009).

Matricialmente, dichas transiciones descritas entre estados pueden representarse de la siguiente forma:

$$p_{ij} = \begin{pmatrix} p_{00} & p_{01} & 0 \\ 0 & p_{11} & p_{12} \\ 0 & 0 & 1 \end{pmatrix}, \quad \text{donde} \quad \sum_{j=0}^{2} p_{ij} = 1. \tag{3.8}$$

Donde $i, j = 0, 1, 2$, representa las probabilidades de transición de que un individuo esté situado en un estado i y, en el futuro, alcance un estado j.

En ese sentido, para el cálculo de dichas probabilidades de transición, se emplea un método conocido como la estimación por máxima verosimilitud (MLE) (Anderson & Goodman, 1957). La estimación de máxima verosimilitud de las probabilidades de transición se da a través de:

$$p_{ij} = \frac{x_{ij}}{\sum x_{ij}} = \frac{x_{ij}}{N_i}, \quad s.e(p_{ij}) = \sqrt{\frac{p_{ij}(1 - p_{ij})}{N_i}}. \tag{3.9}$$

Donde x_{ij} es el número de individuos en consideración que provienen del estado i y alcanzan el

estado j, N_i es el número de población total en la que se basan las transiciones que comienza desde el estado i hasta j y $s.e$ es el error estándar asociado al cálculo de la probabilidad de transición.

3.1.1.8. Análisis de sensibilidad en cadenas de Markov

El análisis de sensibilidad es una forma importante de cuantificar los efectos de los cambios en estos parámetros sobre el comportamiento de la cadena (Caswell, 2019). La matriz de transición de orden $n \times m$ para la aplicación de control con coeficiente de ponderación s sobre una probabilidad de transición ϵ es:

$$T_{\in} = \begin{pmatrix} p_{00} & \cdots & p_{0n} \\ \vdots & \vdots & (1-s)\epsilon \\ p_{m0} & \cdots & p_{mn}. \end{pmatrix}. \tag{3.10}$$

3.1.2. Determinantes del conocimiento del VIH/SIDA

Esta sección se centra en todos los métodos y modelos estadísticos utilizados en el análisis de los determinantes estructurales de la salud que poseen alguna influencia o impacto sobre el nivel de conocimiento del VIH/SIDA en la población peruana y la predicción de este conocimiento. La discusión de las técnicas que incluyen: modelo de regresión binomial con diseño muestral complejo en encuestas y predicción de clasificaciones en modelos paramétricos y no paramétricos.

3.1.2.1. Modelo de regresión logística para datos de diseño muestral complejo

Mediante la perspectiva de un problema de clasificación binaria, se puede considerar que una población finita $U = \{1, 2, \ldots, N\}$ se divide en $h = 1, 2, \ldots, H$ estratos, cada estrato se divide además en $j = 1, 2, \ldots, n_h$ unidades de muestra primarias (UMP), cada una de las cuales está constituida por $i = 1, 2, \ldots, n_{hj}$ unidades de muestra secundarias (UMS), las cuales comprenden n_{hji} elementos. Los datos observados consisten en n'_{hj} UMS elegidas de n'_h UMP en el estrato h. El número total de la observación viene dado por $n = \sum_{h=1}^{H} \sum_{j=1}^{n'_h} \sum_{i=1}^{n'_{hj}} n_{hji}$. Cada unidad de muestreo tiene un peso muestral asociado dado por el inverso de su probabilidad de inclusión en la muestra, denotado por $w_{hijk} = \frac{1}{\pi_{hijk}}$, para la $hjik$-ésima unidad (Cassy *et al.*, 2016).

Más aún, Y_{hjik} denota la variable de respuesta binaria, $\boldsymbol{x}_{hjik}$ denota la matriz de covariables y $\boldsymbol{\beta}$ denota los coeficientes de regresión. Así, en general, el modelo de regresión logística con diseño complejo viene dado por:

$$\text{logit}\{P(Y_{hjik} = 1|x_{hjik})\} = \ln\left\{\frac{P(Y_{hjik} = 1|x_{hjik})}{(1 - P(Y_{hjik} = 1|x_{hjik}))}\right\} = \boldsymbol{x}'_{hjik}\boldsymbol{\beta}. \tag{3.11}$$

En el mismo sentido, cuando el valor del parámetro de dispersión es mayor a la unidad en escenarios de respuestas binarias (en contraste con el valor de una familia binomial) como en el contexto de muestras obtenidas mediante métodos de diseño complejos (recuentos no enteros producidos por el uso de ponderaciones muestrales diferenciales), se indica que el modelo tiene una dispersión excesiva y que los parámetros del modelo pueden estar subestimados. Por lo tanto, la familia cuasi-binomial es la opción ideal para enfrentar esta situación particular modelando la sobre-dispersión en un problema de regresión logística (R Core Team, 2020).

Adicionalmente, considerando la función de clasificación en una regresión logística (enlace entre los predictores lineales y la media de la variable respuesta), la estimación puede darse mediante dos tipos de formas funcionales: logit, que emplea la distribución logística estándar, y probit, que utiliza la distribución normal estándar (Ariza *et al.*, 2016). Las especificaciones de dichas funciones de clasificaciones vienen dadas por:

$$P(Y=1|X)_{logit} = \frac{1}{1+e^{-(\beta_0+\beta_{ni}X_{ni})}} \quad \Big| \quad P(Y=1|X)_{probit} = \Phi(\beta_0+\beta_{ni}X_{ni}). \tag{3.12}$$

Donde $P(Y=1|X)$ es la probabilidad de ocurrencia de Y, β_0 la constante del modelo, β_i los cambios marginales en las variables explicativas (X_i) para n observaciones y $\Phi(\cdot)$ la función de distribución acumulada normal estándar.

3.1.2.2. Modelos de clasificación paramétricos

Según Bonilla *et al.* (2003), los modelos paramétricos parten de una forma funcional (función de distribución o clasificación) conocida, y reducen el problema a estimar los parámetros de los que depende el modelo que mejor ajusten las observaciones de la muestra.

3.1.2.3. Regresión logística binomial

La regresión logística estima directamente la probabilidad de que ocurra $Y_i(Y_i = 1)$ dados los valores de X_i (Aldás & Uriel, 2017), bajo la siguiente función de distribución logística:

$$Pr(\,Y=1/X) = \frac{1}{1+e^{-(\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\cdots+\beta_nX_{ni})}} = \frac{1}{1+e^{-Y}}\ , \tag{3.13}$$

donde Pr(Y) es la probabilidad de ocurrencia de Y y e es la base del logaritmo natural. β_0 representa los desplazamientos laterales de la función logística, β_i son los coeficientes que ponderan las variables independientes y de los que depende la dispersión de la función y X_i son las variables independientes (Bonilla *et al.*, 2003).

3.1.2.4. Modelos de clasificación no paramétricos

De acuerdo con Bonilla *et al.* (2003), los modelos no paramétricos o métodos de distribución libre tratan de aproximar la función de clasificación mediante el empleo de formas funcionales flexibles. Por tanto, tales modelos permiten reconstruir bajo pocas restricciones la función de

clasificación en todo tipo de situaciones.

3.1.2.5. Redes neuronales artificiales

Según Bonilla *et al.* (2003), las redes neuronales artificiales (RNA) están formadas por un conjunto de procesadores simples altamente interconectados denominados nodos o neuronas, los cuales se organizan en capas (de entrada, oculta y de salida) que permiten el procesamiento de información ante una determinada conexión o sinapsis con otra neurona.

Por tanto, la señal de entrada total y_i a cada una de las q neuronas de la capa intermedia se calculará sumando los valores de entrada ponderados por sus pesos correspondientes. Posteriormente, a dicha entrada se le aplica una función no lineal denominada función de activación, obteniendo de esta forma el valor de salida de cada nodo intermedio, que, a su vez, será transmitido a la neurona de salida a través de la conexión ponderada correspondiente. Donde $F(y_i)$ es la salida de cada nodo intermedio bajo la función de activación F, y es la salida de la red y β_i son las conexiones a la capa de salida,

$$y = \sum_{j=1}^{q} \beta_i F(y_i). \tag{3.14}$$

3.1.2.6. Random Forests

Conforme a Cutler *et al.* (2011), el modelo de random forest es un conjunto basado en árboles de particiones recursivas binarias en el que cada árbol depende de una colección de variables aleatorias. Para un vector aleatorio p-dimensional $X = (X_1, \ldots, X_p)^T$ que representa la entrada de valor real o las variables predictoras y una variable aleatoria Y que representa la respuesta de valor real, asumimos una distribución conjunta desconocida $P_{XY}(X, Y)$. El objetivo es encontrar una función de predicción $f(X)$ para predecir Y. La función de predicción está determinada por una función de pérdida $L(Y, f(X))$ y definida para minimizar el valor esperado de la pérdida:

$$E_{XY}(L(Y, f(X))), \tag{3.15}$$

donde los subíndices denotan el valor esperado con respecto a la distribución conjunta de X e Y y $L(Y, f(X))$ es una medida de lo cerca que $f(X)$ está de Y.

3.1.2.7. Árboles de decisión

Los árboles de decisión son una técnica no paramétrica de clasificación binaria que reúne las características del modelo clásico univariante y las propias de los sistemas multivariantes (Bonilla *et al.*, 2003). Los autores definen que el proceso consiste en dividir sucesivamente la muestra original en submuestras, sirviéndose para ello de reglas univariantes que buscarán aquella variable independiente que permita discriminar mejor la división. Con objeto de

encontrar la mejor regla de división, el algoritmo estudiará cada una de las variables explicativas, analizando puntos de corte para, de este modo, poder elegir aquella que mayor homogeneidad aporte a los nuevos subgrupos, bajo la premisa de minimización de la impureza del nodo. El proceso finaliza cuando resulte imposible realizar una nueva división que mejore la homogeneidad existente.

3.1.2.8. Algoritmo K-Nearest Neighbors

Basándonos en Zapata-Tapasco *et al.* (2014), el método de clasificación basado en los k-vecinos más cercanos se fundamenta en que las propiedades de un dato x de entrada son similares a las de los datos de su vecindad, entonces éste pertenece a la misma clase que la clase más frecuente de sus k vecinos más cercanos. Los datos son de la forma presentada a continuación:

$$(\mathbf{x}_i, c_i) = (c_{i1}, c_{i2}, \ldots, x_{ip}, c_i). \tag{3.16}$$

3.1.2.9. Evaluación de la eficiencia de predicción en modelos

En un problema típico de clasificación de datos, las métricas de evaluación se emplean en dos etapas para obtener una evaluación confiable de la calidad de la aproximación de un modelo: la etapa de entrenamiento (proceso de aprendizaje) y la etapa de prueba (proceso de validación) (Hossin & Sulaiman, 2015).

Considerando diversas métricas de evaluación, se presentan los indicadores más representativos a continuación (Mierswa & Klinkenberg, 2018):

Tabla 3.2: Indicadores de bondad de ajuste en modelos de clasificación.

Indicador	Definición	Cálculo
Precisión	Predicciones correctas entre el total de observaciones evaluadas	Precisión $= \frac{VP+VN}{VP+VN+FN+FP}$
Sensibilidad	Predicciones positivas correctas entre el total de predicciones positivas	Sensibilidad $= \frac{VP}{VP+FN}$
Especificidad	Predicciones negativas correctas entre el total de predicciones negativas	Especificidad $= \frac{VN}{VN+FV}$
Kappa de Cohen	Medida de ajuste del azar en la proporción de la concordancia observada	Kappa $= \frac{po-pe}{1-pe}$
F1-Score (Valor-F)	Determinación del valor único ponderado de la precisión y exhaustividad	$F1-Score = \frac{2TP}{2VP+FP+FN}$
Error de clasificación	Predicciones incorrectas entre el total de observaciones evaluadas	Error $= \frac{FP+FN}{VP+FP+FN+VN}$
Área bajo la curva (AUC)	Capacidad discriminante de clasificación del modelo	Mediante la curva ROC
Valor predicho positivo	Predicciones positivas correctas entre el total de observaciones positivas	VPP $= \frac{VP}{VP+FP}$
Valor predicho negativo	Predicciones negativas correctas entre el total de observaciones negativas	Especificidad $= \frac{VN}{VN+FN}$

Notas. VP: Verdadero positivo, VN: Verdadero negativo, FN: Falso negativo, FP: Falso positivo, po: Precisión observada (po $= \frac{VP+VN}{VP+FP+FN+VN}$) y pe: Precisión esperada (pe $= \frac{(VP+FP)(VP+FN)+(FN+VN)(FP+VN)}{(VP+FP+FN+VN)^2}$).

Fuente: Mierswa & Klinkenberg (2018)

3.1.3. Análisis de conglomerados sociales del VIH/SIDA

Esta sección se centra en todos los métodos y modelos estadísticos utilizados en el análisis de conglomerados sociales basados en características socio-económicas, demográficas y familiares en la población peruana. Esta sección se centra en técnicas de aprendizaje automático como: aprendizaje competitivo, definición de los mapas auto-organizados de Kohonen, arquitectu-

rade las redes SOM, el algoritmo de aprendizaje del mapa y el proceso de aprendizaje en una red SOM.

3.1.3.1. Aprendizaje competitivo

El aprendizaje competitivo es un método que divide los datos en grupos específicos mediante la iteración de una regla de actualización una sola vez cada que llega un nuevo patrón de entrada (Wang, 1977). Este algoritmo de aprendizaje en línea es el más adecuado para entornos cambiantes, ya que pueden cambiar los grupos en línea de acuerdo con las distribuciones cambiantes de los patrones de entrada (Wang, 1977).

3.1.3.2. Redes de Kohonen: mapa auto-organizado de características

Un mapa auto-organizado (SOM) es un tipo especial de redes neuronales que se puede utilizar para agrupar tareas y visualizaciones de datos de alta dimensión. El objetivo principal del SOM es transformar un patrón de señal entrante de dimensión arbitraria en un mapa discreto de una o dos dimensiones, y realizar esta transformación de forma adaptativa de una manera ordenada topológicamente (Haykin, 2009). Esta red representa una estructura de retroalimentación con una sola capa computacional que consta de neuronas dispuestas en filas y columnas. Cada patrón de entrada presentado a la red normalmente consiste en una región localizada o un punto de actividad sobre un fondo silencioso (Haykin, 2009).

3.1.3.3. Arquitectura de la red SOM

Una red SOM consta de neuronas o nodos ubicados en una cuadrícula regular, generalmente bidimensional o tridimensional (Lu, 2018).

Cada nodo de la red contiene un vector modelo, que tiene el mismo número de elementos que el vector de entrada, por lo que si el vector de entrada V de n dimensiones: $V_1, V_2, V_3, \ldots, V_n$, entonces cada nodo contendrá un correspondiente vector de peso X, de n dimensiones: $X_1, X_2, X_3, \ldots, X_n$ (Lu, 2018). Una configuración de alta dimensión puede habilitar la visualización de patrones, pero generalmente no se usan ya que su esta es mucho más problemática.

3.1.3.4. Algoritmo de aprendizaje en una red de Kohonen

La esencia del algoritmo SOM es su capacidad para generalizar: se permite que la neurona ganadora y sus vecinas aprendan, las neuronas vecinas se especializarán gradualmente para representar entradas similares y las representaciones se ordenarán en la red del mapa (Kaski, 1997).

Las neuronas representan la entrada con vectores de referencia $\mathbf{m}_i$, cuyos componentes corresponden a pesos sinápticos. Un vector de referencia está asociado con cada neurona llamada unidad en un entorno más abstracto (Kaski, 1997). La unidad, indexada con c, cuyo vector de referencia está más cerca de la entrada $\mathbf{x}$ es la ganadora de la competencia:

$$c = c(x) = \arg\min_i \{\|\mathbf{x} - \mathbf{m}_i\|^2\}. \tag{3.17}$$

La unidad ganadora y sus vecinas se adaptan para representar la entrada de una mejor forma modificando sus vectores de referencia hacia la entrada actual (Kaski, 1997). La cantidad que aprendan las unidades estará gobernada por un núcleo vecino h, que es una función decreciente de la distancia de las unidades a la unidad ganadora en la red del mapa. Si las ubicaciones de las unidades i y j en la red se indican mediante los vectores bidimensionales $\mathbf{r}_i$ y $\mathbf{r}_k$, respectivamente, entonces $h_{ij}(t) = h(\|\mathbf{r}_i - \mathbf{r}_k\|; t)$, donde t denota el tiempo. Durante el proceso de aprendizaje en el momento t, los vectores de referencia se cambian iterativamente de acuerdo con la siguiente regla de adaptación, donde $\mathbf{x}(t)$ es la entrada en el momento t y $c = c(\mathbf{x}(t))$ es el índice de la unidad ganadora :

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{ci}(t)[\mathbf{x}(t) - \mathbf{m}_i(t)]. \tag{3.18}$$

En la práctica, el núcleo vecino se elige para que sea ancho al comienzo del proceso de aprendizaje para garantizar el orden global del mapa, y tanto su ancho como su altura disminuyen lentamente durante el aprendizaje (Kaski, 1997). El proceso de aprendizaje que consiste en la selección del ganador por la Ecuación 3.17 y la adaptación de los pesos sinápticos por la Ecuación 3.18, se puede modelar con una estructura de red neuronal, en la que las neuronas están acopladas por conexiones inhibitorias.

3.1.3.5. Proceso de funcionamiento de una red de Kohonen

El algoritmo responsable de la formación del mapa auto-organizado procede primero inicializando los pesos sinápticos en la red (Haykin, 2009). Una vez que la red se ha inicializado correctamente, hay tres procesos esenciales involucrados en la formación del mapa auto-organizado, como se resume a continuación:

El primer proceso se conoce como competencia. Para cada patrón de entrada, las neuronas de la red calculan sus valores respectivos de una función discriminante (Haykin, 2009). La neurona particular con el mayor valor de función discriminante se considera ganadora de la competencia.

Seguidamente, se da el proceso de cooperación. La neurona ganadora determina la ubicación espacial de una vecindad topológica de neuronas activadas, proporcionando así la base para la cooperación entre dichas neuronas vecinas (Haykin, 2009).

Por último, se desarrolla el proceso de adaptación sináptica. Este último mecanismo permite a las neuronas en actividad aumentar sus valores individuales de la función discriminante en relación con el patrón de entrada mediante ajustes adecuados aplicados a sus pesos sinápticos (Haykin, 2009).

3.1.3.6. Determinación del número de neuronas en la topología en una red SOM

Matemáticamente, el tamaño del mapa se determina calculando el número de neuronas a partir del número de observaciones en los datos de entrenamiento utilizando una regla empírica sugerida por Kohonen (Tian *et al.*, 2014), representada por la siguiente ecuación:

$$M \approx 5 \times \sqrt{N} \rightarrow M_{neuronas} \approx \sqrt{M}. \tag{3.19}$$

donde M es el número total de nodos sugeridos para la construcción del mapa, que es un número entero cercano al resultado del lado derecho de la ecuación, y N es el número de observaciones del conjunto de datos.

3.1.3.7. Visualización en una red SOM

Se puede establecer que, por lo general, las visualizaciones de SOM se obtienen a partir de matrices de distancia, que contienen las distancias entre unidades vecinas del mapa (Lu, 2018). La densidad de puntos del mapa de salida sigue aproximadamente la función de densidad de probabilidad de los datos; estas distancias son con corta diferencia inversamente proporcionales a la densidad de los datos, por lo que las matrices de distancia realizan la agrupación como un enfoque de búsqueda de modo.

Las matrices de distancia pueden contener todas las distancias entre las unidades de mapa y sus vecinos inmediatos, como la matriz de distancia unificada (Matriz-U), permitiendo discernir cómo las similitudes o disimilitudes de las neuronas vecinas se presentan según la intensidad del color en la SOM, o solo un valor único para cada unidad de mapa, como la mediana de las distancias a los vecinos (Lu, 2018).

3.2. Estado del arte

3.2.1. Aplicaciones estocásticas en el estudio del VIH/SIDA

La aplicación de los modelos de cadenas de Markov a diversos problemas de toma de decisiones y análisis de situaciones dentro de un contexto de salud ocupacional se ha venido desarrollando y abordando desde diversos enfoques de las ciencias de la salud (Beck & Pauker, 1983; Sonnenberg & Beck, 1993; Biritwum & Odoom, 1995).

Los procesos de Markov han resultado ser particularmente útiles para el análisis de coste/efectividad en tratamientos basados en la evolución prospectiva de enfermedades (Welsing *et al.*, 2004) y para el modelamiento de epidemias (Yaesoubi & Cohen, 2011), extendiéndose a otras problemáticas sanitarias relevantes. Así, se puede determinar que los modelos basados en Cadenas de Markov representan una herramienta útil y viable para el planteamiento y análisis prospectivo de aspectos relacionados a la prognosis y estudio de enfermedades e implicancias en el contexto de la administración pública.

A saber, Lee *et al.* (2014) estudiaron la población afroamericana en los Estados Unidos y la población caucásica para predecir las tendencias de la epidemia del VIH/SIDA en ambos casos. El análisis basado en Cadenas de Markov de los autores se utilizó para modelar la progresión de la enfermedad entre las personas vulnerables, las personas infectadas por el VIH y los casos de SIDA para los dos grupos raciales por separado. Los resultados de su estudio implican, primeramente, que las consideraciones etnográficas deberían incorporarse en las intervenciones de prevención, manejo y evaluación de programas de VIH/SIDA para optimizar la asignación de recursos y, por otro lado, que existe un potencial de aplicabilidad del modelo propuesto para predecir la prevalencia del virus y de la enfermedad en otros grupos raciales del país.

Nucita *et al.* (2013) presentaron un modelo fundado en Cadenas de Markov, que explota datos clínicos reales para modelar la evolución de la enfermedad en un solo paciente, junto con una red sexual que modela la propagación de la epidemia en toda la población general (un modelo para evaluar las relaciones entre individuos y, luego, la infectividad variable), para predecir las tendencias epidemiológicas del VIH/SIDA. Los resultados experimentales de los autores basados en simulaciones de diferentes escenarios, mediante el ajuste de varios parámetros como la composición de la población, la definición de la red sexual, el nivel de retención, entre otros; les permitieron concluir que la ampliación del acceso a las pruebas y la terapia antiretroviral afectaría positivamente la evolución epidemiológica y limitaría la propagación del VIH/SIDA.

En el estudio de Rotich (2016), la dinámica del VIH se analizó utilizando un modelo matemático de corte estocástico basado en Cadenas de Markov de tiempo discreto. Se investigaron parámetros demográficos y epidemiológicos que afectan la dinámica de la población en el modelo. Debido a la continuidad de la epidemia de VIH/SIDA en todo el mundo, Rotich (2016) señala que es importante que los formuladores de políticas tomen en cuenta las recomendaciones médicas y científicas, como: estrategias de intervención, el uso de medicamentos retrovirales y la prevención de la transmisión vertical; intensificando estrategias preventivas que pueden reducir los riesgos de la población infectada con VIH/SIDA y la recopilación de datos apropiados para el modelado, predicciones y planificación futura.

Con respecto al modelo de estimaciones de supervivencia dentro de la población infectada, de Vasconcellos *et al.* (2013) aplicaron modelos de Markov multi-estados para analizar los factores asociados a las transiciones entre los diferentes estados de cronicidad del VIH/SIDA y la rentabilidad de varios regímenes antirretrovirales. Los autores concluyeron que este enfoque permite el análisis de las circunstancias que influyen en las transiciones entre estados específicos de un paciente individual y evaluar los cursos de tratamiento más idóneos según cuadro clínico específico.

3.2.2. Asociación entre determinantes de la salud y el VIH/SIDA

El VIH/SIDA y su proliferación, evolución y tratamiento, evaluados desde un marco biomédico y de políticas de salud públicas, se ven influenciados de manera equívoca por numerosos

elementos y agentes endógenos a la población. Ama *et al.* (2015) señalan que estos cofactores pueden ser denominados como determinantes sociales de la salud y que pueden ser definidos como condiciones económicas, sociales y ambientales/biológicas. Asimismo, estos determinantes sociales inciden en otras concausas a nivel nacional: a través de factores condicionantes intermedios y proximales producen diferentes patologías en la población - mala salud de los pobres, el gradiente social de salud y las grandes desigualdades sanitarias, entre otros (Alfaro-Alfaro, 2014).

Es así como el interés de los investigadores por entender las causas por las que las personas con diferentes características socio-económicas experimentan de manera diferente la salud y las enfermedades ha llevado el debate hacia un nuevo enfoque que reconoce que el estado de salud del sujeto está determinado por factores sociales conductuales y estructurales (Tovar-Cuevas & Arrivillaga-Quintero, 2011).

De acuerdo a los hallazgos de la literatura científica, diversos autores han establecido conexiones de diferentes cortes entre los determinantes de salud y la prevalencia del VIH: la variación y cambio en factores macroeconómicos pueden llegar a frenar la proliferación del VIH en el mundo en desarrollo (Chikermane *et al.*, 2016), bajos niveles del ingreso promedio monetario se relacionan con mayores tasas de incidencia del VIH (Ogunmola *et al.*, 2014), la asociación entre la infección por VIH y de ciertos determinantes difieren por zonas geográficas (Bunyasi & Coetzee, 2017), el nivel de pobreza y de empleabilidadse configuran como importantes determinantes de la prevalencia del VIH (Scott & Simon, 2011), factores socioeconómicos, demográficos y culturales evidencian cambios en las tendencias de influencia en el tiempo sobre el VIH (Woldemariame, 2013).

De la misma manera, existen numerosos estudios empíricos de la asociación entre ciertos determinantes de la salud y el conocimiento sobre el VIH/SIDA que un individuo o población pueda poseer: entre mujeres casadas se detectó un fuerte impacto de la educación, acceso a medios de comunicación, residencia, el índice de riqueza y el estado laboral en el conocimiento sobre VIH (Haque *et al.*, 2018); entre HSH, se evidencia que aspectos como bajos niveles de escolaridad, etnicidad no blanca, pertenencia a clases económicas bajas, juventud, el no haberse realizado una prueba de descarte y monogamia sexual presentaron asociación con un bajo nivel de conocimiento sobre el VIH/SIDA (Gomes *et al.*, 2017); entre las mujeres en edad fértil, se descubrió que el nivel de educación es el factor dominante asociado con el conocimiento del VIH (Najmah *et al.*, 2020); en cuanto a la población adolescente y juvenil, aunque tienen un riesgo muy alto de transmisión del VIH durante las relaciones sexuales, están mal informados sobre el VIH y tienen actitudes muy negativas hacia el virus (Pahn *et al.*, 2020).

A pesar de que las conclusiones particulares dadas por los estudios referidos a la literatura científica tomada en consideración se encuentran acotadas al contexto y a las características propias de la población en estudio, la conclusión general deja entrever que existen asociaciones entre determinantes de la salud de los individuos y la prevalencia y nivel de conocimiento del VIH que deben ser analizadas para complementar la evaluación clínica de las enfermedades

(Gala *et al.*, 2007).

3.2.3. Modelamiento predictivo paramétrico y no paramétrico del VIH/SIDA

En la actualidad, los esfuerzos por desarrollar modelos de aprendizaje los cuales pudieran ser capaces de asimilar información de datos acumulados o de gran dimensionalidad y predecir diferentes factores, con una mayor precisión y flexibilidad que los modelos de regresión multivariada convencionales (paramétricos), permitieron la formulación de otro tipo de modelos predictivos que confieren flexibilidad y toma de decisiones no estructuradas (McCulloch & Pitts, 1943), denominándolos como modelos de clasificación no paramétrica.

En el campo de la sanidad pública, una de las preocupaciones subyacentes de los proveedores de servicios de salud es la ampliación del conocimiento sobre el estado e implicancias del VIH (Hailu, 2015). Por tanto, el modelado predictivo es una de las herramientas más eficaces para los hacedores de políticas públicas. Hailu (2015) plantea que los programas de salud no pueden brindar la atención, el tratamiento y el asesoramiento adecuados para el VIH/SIDA sin saber quién está infectado ni el grado de conocimiento sobre el estado de estos que se posea. Esto implica que identificar el mejor modelo predictivo para estos aspectos que influyen o impactan sobre el VIH/SIDA es fundamental.

En tal sentido, Ahlström *et al.* (2019) enfatizan que los algoritmos de aprendizaje automático, un conjunto de herramientas matemáticas que extraen patrones generalizables de grandes conjuntos de datos para hacer predicciones sobre el resultado en casos nuevos o desconocidos, son áreas de investigación en rápido crecimiento que también se han abierto camino en la investigación del VIH; destacando que estos no solo mejorar la capacidad de discriminación, sino que también puede ayudar a identificar a las personas con mayor riesgo de contraer el VIH y con un menor grado de entendimiento acerca del mismo.

Asimismo, Tang *et al.* (2018) encuentran en una tecnología de reciente desarrollo basada en el estudio de máquinas en inteligencia artificial y bases de datos el potencial para la identificación con precisión de enfermedades y condiciones en función de ciertos atributos importantes resultando en herramientas valiosas en el campo médico: el modelamiento predictivo paramétrico y no paramétrico en la minería de datos. Concluyendo que, en tiempos actuales, el estudio de prevención, diagnóstico y tratamiento del VIH/SIDA entró en una nueva fase gracias a estas tendencias en el modelamiento predictivo descubriendo factores potenciales y tratamientos más eficientes para la epidemia (Tang *et al.*, 2018).

En una perspectiva comparativa entre los ambos tipos de modelos, Bao *et al.* (2020) apuntan a que los enfoques convencionales para la predicción del diagnóstico del VIH/ITS (paramétricos) son cuestionables. Por consiguiente, establecen que el uso de enfoques de aprendizaje automático es una tendencia creciente en la investigación del VIH/ITS, dado que estos enfoques pueden incorporar una mayor cantidad de covariables en un gran conjunto de datos, manejar relaciones complejas entre predictores y el resultado, y lograr una alta precisión (Bao *et al.*, 2020).

3.2.4. Análisis exploratorio de conglomerados asociados al VIH/SIDA y factores de salud pública

Yahaya & Kola (2017) puntualizan que la agrupación de datos o proceso de clusterización es una herramienta vital cuando se trata de comprender elementos con características similares en un conjunto de datos.

En ese aspecto, Merzouki *et al.* (2021) explican que la prevalencia e incidencia del VIH/SIDA varían ampliamente entre países y subnacionalmente, lo que destaca la necesidad de utilizar datos y métodos numéricos para adaptar las intervenciones a poblaciones y países específicos. Por ello, los autores que el aprendizaje no supervisado y el análisis de agrupamiento permite encontrar subgrupos ocultos de personas con diferentes factores y potencialmente diferentes niveles de riesgo de tener o adquirir el VIH (Merzouki *et al.*, 2021).

Por otro lado, Yahaya & Kola (2017) exponen que la técnica del clustering se puede aplicar para clasificar la propagación del VIH/SIDA entre la población: la posible clasificación de esta información ayudaría a tomar mejores decisiones sobre qué facción de un país necesita cierta atención e investigar la concentración de la enfermedad con respecto a las zonas geopolíticas. Por otro lado, O'Neill Institute (2019) determina que la detección mediante un análisis de conglomerados complementa la información obtenida de la vigilancia tradicional para identificar y responder más eficaz y rápidamente con recursos más intensivos a los casos de transmisión rápida del VIH.

Teniendo en cuenta prácticas concretas del análisis de conglomerados, Blondeel *et al.* (2021) utilizaron el análisis jerárquico de conglomerados para identificar patrones de conducta sexual y explorar asociaciones con el historial de ITS y el estado serológico del VIH/SIDA; concluyendo que ciertos factores relacionados al comportamiento sexual refuerzan la vulnerabilidad a las ITS y el VIH de ciertas poblaciones en riesgo. En la misma línea, Lang & Belenko (2001) buscaron identificar perfiles de riesgo del VIH/SIDA entre delincuentes y examinaron si existen patrones de conducta que podrían ser objeto de intervenciones adaptadas a las poblaciones criminales, sugiriendo que los esfuerzos de intervención para los infractores por delitos graves relacionados con las drogas deben abordar las conductas de riesgo epidemiológico diferencial.

En investigaciones más contemporáneas y noveles del análisis de clustering, Jagadesh *et al.* (2020), probando un enfoque novedoso para identificar grupos dentro de la salud pública, construyeron mapas auto-organizados e identificaron conglomerados basados en variables demográficas y socioeconómicas; concluyendo que los mapas auto-organizados (SOM) puede guiar los esfuerzos de salud pública para optimizar la prevención y las pruebas del VIH en la Guayana Francesa y otros países en desarrollo.

Tabla 3.3: Literaturas involucradas en la exploración y estudio del VIH/SIDA.

Número	Autor	Técnica	Objetivo	Resultados
1	Beck & Pauker (1983)	Procesos de Markov	Descripción de un modelo de pronóstico médico de propósito general basado en el proceso de Markov	Los procesos de Markov son herramienta matemática que puede utilizarse para generar evaluaciones detalladas y precisas de la esperanza de vida y el estado de salud de los pacientes de forma óptima
2	Sonnenberg & Beck (1993)	Cadenas de Markov	Revisión de la teoría detrás del modelo de pronóstico de Markov y una guía práctica para la construcción de los mismos	Los modelos de Markov son útiles cuando un problema de decisión implica un riesgo continuo y repetitivo en el tiempo, permitiendo una representación más precisa de los entornos clínicos que involucran estos problemas
3	Biritwum & Odoom (1995)	Procesos de Markov	Aplicación del modelado del proceso de Markov al comportamiento de cambio del estado de salud de los bebés	Las transiciones entre la salud y la enfermedad de los bebés, de un mes a otro, pueden modelarse mediante una cadena de Markov para la cual las probabilidades de transición dependen generalmente del tiempo o de la edad
4	Welsing *et al.* (2004)	Modelo de cadena de Markov	Determinar la rentabilidad de las estrategias de tratamiento para los pacientes con artritis reumatoide	Durante el período de 5 años, el efecto esperado sobre la actividad de la enfermedad y los AVAC fue mejor para las estrategias de tratamiento propuestas. Demuestran la eficacia de la aplicación de los modelos de Markov
5	Yaesoubi & Cohen (2011)	Cadenas de Markov de tiempo discreto	Diseñar una clase de modelos matemáticos para la transmisión de enfermedades infecciosas en grandes poblaciones	Estos modelos aplican para generar políticas de salud dinámicas óptimas para controlar la propagación de enfermedades infecciosas y pueden aproximar la propagación de la enfermedad en poblaciones relativamente grandes
6	Lee *et al.* (2014)	Cadenas de Markov	Predicción de las tendencias de la epidemia entre los afroamericanos y los caucásicos por separado en EE.UU.	El estudio muestra que el número de diagnósticos de VIH SIDA y muertes cada año es constante, mientras que el número de personas que viven con el VIH/SIDA entre la población infectada está aumentando
7	Nucita *et al.* (2013)	Cadenas de Markov	Utilizar datos clínicos reales para simular escenarios epidemiológicos para la epidemia del VIH a nivel de distrito	El modelo propuesto es útil para predecir tendencias epidemiológicas y, por lo tanto, para apoyar la toma de decisiones para enfrentar la difusión del VIH/SIDA y otras enfermedades de transmisión sexual

Tabla 3.4: Literaturas involucradas en la exploración y estudio del VIH/SIDA. (Continuación)

Número	Autor	Técnica	Objetivo	Resultados
8	Rotich (2016)	Procesos de cadenas de Markov	Determinar los valores umbral de los parámetros que determinan el gradiente de propagación del VIH	Un aumento de infecciones a través de las relaciones sexuales, la sangre y la transmisión de madre a hijo llevaría a aumentar la población de infectados lo que a su vez aumentaría la población con SIDA
9	de Vasconcellos *et al.* (2013)	Modelos de Markov multi-estados	Analizar la adherencia al tratamiento y los factores socioeconómicos y terapéuticos que pueden afectar la cronicidad entre los pacientes con SIDA	Representa un enfoque poderoso en el estudio de enfermedades crónicas, posibilitando la adopción de intervenciones más individualizadas y eficaces
10	Ama *et al.* (2015)	Modelo de regresión logística	Describir las características de los adultos mayores infectados por el VIH y determinar cómo el estado del VIH se ve influido por factores socioeconómicas y demográficas	El reconocimiento de los adultos mayores como uno de los grupos vulnerables / poblaciones clave que requiere especial atención en temas de VIH y SIDA y otros asuntos relacionados con la salud
11	Alfaro-Alfaro (2014)	-	Definir qué son los determinantes sociales de la salud y las funciones esenciales de la salud pública social en base a ellos	Los determinantes sociales inciden en otras concausas que están relacionadas con indicadores de salud a nivel nacional, produciendo diferentes patologías en la población
12	Tovar-Cuevas & Arrivillaga-Quintero (2011)	Modelo de regresión logística	Determinar la asociación entre la prevalencia de VIH/SIDA y determinantes sociales estructurales en Colombia	Se encontró evidencia estadísticamente significativa (p= 1%, p=5%) de la asociación entre estos determinantes y la prevalencia de VIH en nueve de 21 municipios analizados
13	Chikermane *et al.* (2016)	Análisis de regresión multivariada	Establecer el vínculo entre la educación y las variables socioeconómicas de nivel macro con la prevalencia del VIH en los EEUU, India, Sudáfrica y Rusia	Concluyen en una indicación importante de qué factores a nivel macro pueden abordarse para frenar el VIH en el mundo en desarrollo, proporcionando una indicación importante de la mejor manera de emplear recursos limitados
14	Ogunmola *et al.* (2014)	Modelo de regresión binaria multivariada	Investigar la relación entre el nivel socioeconómico (NSE) y la infección por VIH en Nigeria	Los hallazgos sugirieron que algunos indicadores de NSE están relacionados de manera diferente con la infección por VIH. Las infecciones prevalentes por el VIH se concentran entre las personas de bajos ingresos

Tabla 3.5: Literaturas involucradas en la exploración y estudio del VIH/SIDA. (Continuación)

Número	Autor	Técnica	Objetivo	Resultados
15	Bunyasi & Coetzee (2017)	Análisis de regresión multivariada	Determinar la asociación entre el nivel socioeconómico (NSE) y el VIH en mujeres en edad reproductiva en Sudáfrica (SA)	La asociación entre la infección por el VIH y el nivel socioeconómico difirió según provincias y la medida del NSE y subraya la carga desproporcionadamente mayor entre las mujeres más pobres y con bajo nivel educativo
16	Scott & Simon (2011)	Modelo de auto-regresión vectorial	Examinar la fuerza y dirección de la asociación entre pobreza y VIH/SIDA en Trinidad y Tobago	Los resultados de la prueba sugieren un vínculo bidireccional entre el desempleo femenino y la incidencia del VIH/SIDA
17	Woldemariame (2013)	Modelo de regresión logística binaria	Determinar los factores que impulsan la prevalencia del VIH/SIDA comparando dos períodos diferentes de tiempo en Etiopía	Factores demográficos, socioeconómicos y culturales tienen un impacto en la prevalencia de la epidemia en los períodos considerados
18	Haque *et al.* (2018)	Modelo de regresión logística	Encontrar el nivel de conciencia de VIH y los factores que influyen entre las mujeres casadas en Bangladesh	Aunque una proporción considerable de mujeres tenía un conocimiento y conciencia adecuados sobre el VIH / SIDA, se recomienda implementar programas educativos relacionados con el VIH/SIDA
19	Gomes *et al.* (2017)	Modelo de regresión logística ordinal	Analizar factores de vulnerabilidad social, individual y programática, asociados al bajo conocimiento en VIH/SIDA entre hombres que practican sexo con hombres (HSH)	Es fundamental mejorar el nivel de conocimiento sobre VIH/SIDA entre los jóvenes HSH, con condiciones socioeconómicas desfavorables
20	Najmah *et al.* (2020)	Modelo de regresión logística multivariada	Proporcionar información sobre el papel de determinantes sociales en la adquisición de conocimientos sobre el VIH entre las mujeres en edad fértil	La concienciación e información sobre el VIH entre las mujeres de bajo riesgo de infección deben integrarse con la educación formal, así como en los servicios de salud materna, especialmente en las zonas rurales
21	Pahn *et al.* (2020)	Análisis de regresión logística múltiple	Describir el nivel de conocimientos y actitudes sobre el VIH y sus factores relacionados entre los adolescentes camboyanos	Importancia de informar a los entes públicos y personal escolar sobre la necesidad de un programa de educación sanitaria específico y culturalmente sensible sobre el VIH

Tabla 3.6: Literaturas involucradas en la exploración y estudio del VIH/SIDA. (Continuación)

Número	Autor	Técnica	Objetivo	Resultados
22	Gala *et al.* (2007)	Modelo de infección transmisible	Explicar la adaptación un modelo de enfermedades transmisibles a la dinámica de la adquisición del VIH	Se demostró la aplicabilidad de este modelo y se previó su utilidad para el estudio de los determinantes que intervienen en la adquisición de la infección
23	McCulloch & Pitts (1943)	Lógica neuronal	Describir eventos neuronales y las relaciones entre ellos	Se muestra que muchas elecciones particulares entre posibles supuestos neurofisiológicos son equivalentes
24	Hailu (2015)	Algoritmos de Machine Learning	Comparar el poder de predicción de las diferentes técnicas de minería de datos utilizadas para desarrollar el modelo de predicción de la prueba del VIH	La minería de datos es fundamental para extraer información relevante para la utilización eficaz de los servicios de pruebas del VIH que tienen importancia clínica, comunitaria y de salud pública en todos los niveles
25	Ahlström *et al.* (2019)	Algoritmos de Machine Learning	Examinar datos de registros electrónicos para predecir el estado del VIH mediante algoritmos de aprendizaje automático	Los algoritmos de aprendizaje automático pueden aprender de los datos del registro electrónico y ayudar a predecir el estado del VIH con una precisión alta
26	Tang *et al.* (2018)	Algoritmos de Machine Learning	Utilizar algoritmos de minería de datos para establecer el modelo de identificación de la infección por VIH y comparar su desempeño predictivo	La minería de datos puede ayudar al personal médico a detectar y diagnosticar el SIDA rápidamente a partir de una gran cantidad de información
27	Bao *et al.* (2020)	Algoritmos de Machine Learning	Desarrollar modelos de aprendizaje automático y evaluar su desempeño en la predicción del diagnóstico de VIH e ITS	Los enfoques de aprendizaje automático son ventajosos sobre los modelos de regresión logística multivariable para predecir el diagnóstico de VIH/ITS

Tabla 3.7: Literaturas involucradas en la exploración y estudio del VIH/SIDA. (Continuación)

Número	Autor	Técnica	Objetivo	Resultados
28	Yahaya & Kola (2017)	Análisis de conglomerados	Aplicar el análisis de conglomerados para clasificar la propagación del VIH entre la población de Nigeria	Agrupación exitosa que clasificó la propagación del VIH entre los estados de Nigeria donde las zonas geopolíticas norte-central y sur-sur
29	Merzouki *et al.* (2021)	Análisis de componentes principales y clustering jerárquico	Investigar cómo la heterogeneidad socioconductual en África subsahariana podría explicar la variación de la incidencia del VIH entre países	Se reveló tres grupos de países, cada uno con perfiles socio-conductuales característicos
30	O'Neill Institute (2019)	Análisis de conglomerados	Utilizar la detección de conglomerados para identificar y responder rápidamente a los grupos de transmisión de VIH	Es una estrategia de salud pública para identificar y responder rápidamente a los patrones de transmisión para informar las respuestas de salud pública
31	Blondeel *et al.* (2021)	Análisis de clustering jerárquico	Identificar patrones de conducta sexual en HSH y explorar su relación con diagnósticos de ITS y el estado de VIH+	Otros factores además del comportamiento sexual parecen reforzar la vulnerabilidad a las ITS y al VIH de algunos HSH
32	Lang & Belenko (2001)	Análisis de regresión logística y agrupamiento jerárquico aglomerativo	Identificar perfiles de riesgo del VIH entre los delincuentes y examinar factores y patrones de conducta comunes para intervenciones adaptadas	Dos grupos distinguibles por la alta frecuencia de conductas sexuales que eran en gran parte sin protección o uso de drogas de alta frecuencia, abordando conductas de riesgo diferencial
33	Jagadesh *et al.* (2020)	Mapas auto-organizados (SOM)	Identificar grupos de población que necesitan mayores esfuerzos de salud pública	Los vecindarios con desventajas socioeconómicas siguen siendo focos de infección por el VIH / SIDA

Capítulo 4

Modelamiento probabilístico del VIH/SIDA

En este capítulo, en primera instancia, se definen los años de estudio y sus respectivos indicadores epidemiológicos a considerar como las unidades operacionales (que se configuran como hitos importantes) que van a ser recopiladas y modeladas mediante los métodos a plantear. Adicionalmente, se proveerá una descripción de las fuentes de dichas estructuras de datos consideradas.

Por otro lado, se muestra el análisis de los diversos datos secundarios relacionados con la distribución epidemiológica del VIH/SIDA en el Perú para los años de estudio (1995, 2005, 2011, 2013 y 2018). Se construyeron los procesos de Markov para cada caso de estudio descrito anteriormente, incluyendo los estados y probabilidades de transición necesarias para representar la dinámica epidemiológica del virus y la enfermedad en el país; estimándose, consecuentemente, el comportamiento estacionario de los modelos de Markov a fin de evaluar el progreso prospectivo de la epidemia dependiendo de las condiciones iniciales que caracterizan cada escenario de evaluación. A su vez, se realizó un estudio de cohorte para evaluar proyecciones estocásticas (las tendencias que el virus y la enfermedad asumirían) basadas en las políticas de salud y medidas de contención y tratamiento presentes en cada año de estudio durante un horizonte temporal planteado. Finalmente, se generan estrategias de control y contención del VIH/SIDA mediante un análisis de sensibilidad ofreciendo herramientas de evaluación sobre el desarrollo y evaluación del VIH/SIDA a nivel poblacional.

Todos los modelos, así como varios cálculos matemáticos, se realizaron en el entorno estadístico y de programación R (R Core Team, 2020) mediante el paquete estadístico `markovchain` (Spedicato, 2017) y funciones adicionales acondicionadas y desplegadas en R.

4.1. Bases de datos

Para la aplicación de los modelos y escenarios propuestos, tomaremos en consideración cifras demográficas y de salud en el Perú asociadas a los siguientes períodos anuales de investigación: 1995, 2005, 2011, 2013 y, finalmente, 2018; obtenidas de las bases de datos del Banco Mundial (World Bank Group), el Programa Conjunto de las Naciones Unidas sobre el VIH/SIDA (ONUSIDA) y la Dirección General de Epidemiología del Perú (DGE Perú), respectivamente.

En esta sección, se fundamenta la elección de los años de estudio que responde al interés de

estudiar los efectos de los hitos y medidas importantes de la lucha contra el VIH/SIDA en las proyecciones de progreso y cambios de la epidemia en el país. Además, se destacan las estructuras de datos construidas y su contenido.

4.1.1. Hitos en la respuesta nacional ante el VIH/SIDA

En el caso del año 1995, este simboliza la inauguración de una fuerte planificación subsectorial, crecimiento y organización en torno a servicios de salud y estrategias dirigidas, fundamentalmente, a grupos vulnerables; pero con persistencia de una incipiente multisectorialidad y articulación de los esfuerzos de la sociedad civil en conjunción con el Estado peruano contra el VIH/SIDA (Defensoría del Pueblo, 2009). De la misma manera, en dichas circunstancias, las primeras estrategias de intervención relacionadas a las ETS en el país fueron planteadas y se creó el Programa de Control de Enfermedades de Transmisión Sexual y SIDA (PROCETSS), el cual planteó y estableció de manera preliminar el protocolo de notificación de infección/transmisión del VIH/SIDA en ciudadanos dentro del sector salud (MINSA, 2001a): la notificación por SIDA sería de manera obligatoria a la direcciones de salud especializadas si el cuadro clínico cumplía con la definición de caso del virus y la enfermedad; representando una de las primeras respuestas sólidas del gobierno para la intervención en la epidemia.

En relación al año 2005, el gobierno peruano, a través del Ministerio de Salud, puso en disposición de los proveedores públicos el programa TARGA - Tratamiento Antirretroviral de Gran Actividad (Prieto, 2016) para la masificación de la atención estatal de la epidemia, lo que marcó la implementación del primer tratamiento con fármacos antirretrovirales en el Perú, bajo un programa de salud pública, implicando la adecuación de los sistemas de salud previos, así como la generación y el desarrollo de nuevos instrumentos gerenciales para administrar el programa y la epidemia (MINSA, 2001b); fundando el esfuerzo altamente significativo por disminuir los casos detectados y fallecimientos con una iniciativa novel en el país.

En el mismo sentido, el año 2011 marca el fin de la ejecución y actividad del promulgado Plan Estratégico Nacional Multisectorial 2007–2011 (PEM) para la prevención y control de las ITS y el VIH/SIDA en el país. La implementación de dicho plan fue uno de los objetivos centrales de las acciones desplegadas en respuesta a la epidemia durante los últimos años (Defensoría del Pueblo, 2011). Sin embargo, una evaluación de su implementación revela que fue necesario para la época el refuerzo de estrategias de intervención y sostenibilidad de los avances logrados en años previos posterior a los resultados generados por el plan; advirtiendo la falta de armonización de los planes y acciones sectoriales, la ausencia de espacios para la evaluación periódica de las intervenciones de cada sector y la falta de mecanismos de vigilancia y notificación en materia de VIH y SIDA.

Examinando el año 2013, ONUSIDA propuso, en términos de salud pública, poner fin a la epidemia de VIH hacia el 2030, mediante el logro de las denominadas Metas 90-90-90 (sobre el conocimiento del estado serológico, suministro de tratamiento y presentación de una carga vírica indetectable al 90 %). En línea con dicha directriz, en el Perú, la construcción de la cascada del continuo de la atención del VIH siguió las recomendaciones internacionales para

el monitoreo del cuidado y tratamiento de las personas viviendo con VIH (PVV). Se muestra como uno de los primeros ejercicios para graficar el continuo de la atención del VIH en el Perú y habilitar la revisión, definición y delineación de las intervenciones en salud pública, buscando la mejora de estos indicadores, así como su comparabilidad respecto a otros países de la región (Garcia-Fernandez *et al.*, 2001).

Por último, indagando acerca del año 2018, este período da cuenta de las medidas del gobierno peruano por incluir, dentro de la política y estructura de salud relacionados al VIH/SIDA, a manera de programa sanitario público nacional la profilaxis pre-exposición (PrEP) como opción dentro de un paquete de prevención combinada del VIH (ONUSIDA, 2017). El uso de medicamentos antirretrovirales para la prevención de la infección por el VIH en personas seronegativas se considera crítico en la respuesta contra la epidemia, en especial al ser realizado mediante la asistencia pública y como parte de los servicios del aparato de salud del país para la adecuación de dicho paquete posteriormente en América Latina.

4.1.2. Fuentes y estructura de datos

Para cada período de investigación contemplado en este capítulo, se compiló una estructura de datos que se compone por los siguientes indicadores: número de habitantes en el país (tamaño de la población general), número de nuevos casos registrados de VIH/SIDA en el territorio nacional, cifra registrada de la población prevalente o que vive infectada con VIH/SIDA y número de fallecimientos oficiales a causa del VIH/SIDA (descartando defunciones por comorbilidades o enfermedades secundarias).

Todos los conjuntos de datos se recopilaron bajo una frecuencia anual para cada período descrito anteriormente. En el caso del número de habitantes en el país, dicha variable demográfica fue extraída del repositorio de datos de libre acceso del Banco Mundial, basados en las revisiones de la división de población de la ONU, asociado al Perú (World Bank, 2020). Teniendo presentes la cifra registrada de la población prevalente de VIH/SIDA y el número de muertes asociadas al virus y la enfermedad, estos fueron obtenidos por medio del archivo Spectrum (formato de datos de vigilancia) oficial del 2019 sobre las estimaciones nacionales del VIH para el país por parte de la Secretaría de ONUSIDA (ONUSIDA, 2020). Finalmente, el número de nuevos casos notificados y registrados de infección de VIH y desarrollo de SIDA por parte de individuos fue compendiado del boletín epidemiológico de VIH/SIDA para diciembre del 2018 que el Centro Nacional de Epidemiología, Prevención y Control de Enfermedades (DGE, 2020a) genera, de forma mensual, para evaluar la situación epidemiológica en el Perú. Todas las cifras fueron consideradas y tratadas de manera absoluta/numérica.

4.2. Metodología

El propósito de esta sección es presentar la metodología de investigación necesaria para desarrollar y justificar el estudio del modelamiento probabilístico del VIH/SIDA en el Perú del presente capítulo. Se precisará la aplicabilidad de la teoría o los aspectos vinculados a la literatura para explicar por qué se están utilizando ciertos métodos/técnicas, procedimientos

y criterios para el análisis de los datos y el fundamento académico de las elecciones dadas en cada subsección de los resultados del estudio propuesto.

4.2.1. Modelamiento mediante Cadenas de Markov

Considerando que los procesos de Markov pueden configurarse como una representación más precisa de los entornos clínicos que involucran problemas relacionados a la prognosis y decisiones terapéuticas de tratamientos y cursos de enfermedades (Beck & Pauker, 1983; Sonnenberg & Beck, 1993); para el presente estudio, se establece que los estados discretos del modelo de la Cadena de Markov para analizar la evolución probabilística del VIH/SIDA en el Perú para los años de estudio establecidos, son los siguientes: individuo sano susceptible a ser contagiado del virus (estado 1), paciente infectado con VIH/SIDA (estado 2) y sujeto fallecido a causa de la enfermedad (estado 3).

4.2.2. Estimación de probabilidades de transición

Ocaña-Riola (2009) precisa que, en epidemias humanas, la matriz de transición y las probabilidades que la componen suelen obtenerse a partir de los datos de población, incidencia, prevalencia y mortalidad publicados en los registros oficiales de la comunidad o país en estudio.

Bajo esta premisa, a cada individuo de la población se le asigna inicialmente uno de los estados de transición (a saber: S, I o R) dentro de la dinámica epidemiológica de una enfermedad como el VIH/SIDA para el presente estudio (Olia *et al.*, 2006).

Los individuos susceptibles tienen una propensión a contagiarse o volverse infecciosos considerando cierto período de tiempo, lo que se configura como una tasa de contagio o incidencia dentro del conjunto de personas examinado inicialmente (tasa que analiza el número de nuevos casos de VIH/SIDA en la población general en un período determinado), lo que permite obtener una probabilidad de infección en el tiempo al dividir este número de casos nuevos detectados entre la población total registrada en cierta unidad de tiempo (usualmente medida en años en este tipo de epidemias) o ajustando la tasa de incidencia oficial recopilada de la población por cada millón de habitantes para obtener dicha probabilidad - p_{12} (Olia *et al.*, 2006). En ese mismo sentido, los individuos que permanezcan sanos o sigan siendo susceptibles a poder contagiarse en el futuro configuran una tasa de susceptibilidad que resulta de la división del número de personas que no estuvieron infectadas entre la población total en el período de estudio determinado, lo que se traduce en una probabilidad de riesgo de infección - p_{11} (Olia *et al.*, 2006).

Por otro lado, los individuos infecciosos o que han adquirido el VIH/SIDA pueden permanecer contagiados en el tiempo con variaciones en su carga viral (lo que representa cambios en su estado serológico y el cuadro clínico que pueden enfrentar a causa de los efectos negativos de la enfermedad), lo que se traduce en un conjunto de personas consideradas prevalentes dentro de la población de infectados (ya que siguen presentando características o signos de contagios

por el virus/enfermedad posteriormente a haberlos evaluado en cierto lapso futuro), siendo capaces de calcular una probabilidad de prevalencia basada en una proporción de individuos en una comunidad determinada que permanezcan o sigan infectados con VIH/SIDA en un período dado sobre la población total de infectados oficial; dicho cálculo permite obtener una probabilidad de padecimiento dentro los individuos contagiados - p_{22} (Olia *et al.*, 2006). Sin embargo, no todos aquellos individuos que son portadores del virus o que han desarrollado la enfermedad permanecen vivos en el futuro. Algunas personas infectadas llegan a fallecer por razones asociadas a comorbilidades o complicaciones biológicas de la enfermedad, lo que se traduce en un ratio de mortalidad dentro del conjunto total de infectados al existir un riesgo de fallecimiento asociado al VIH/SIDA una vez que un individuo ha sido contagiado y basado en el desarrollo del SIDA en el organismo; dicha estadística (tasa de individuos occisos entre la población total de portadores de VIH/SIDA) permite definir una probabilidad de muerte dentro de la dinámica de la epidemia - p_{23} (Olia *et al.*, 2006).

4.2.3. Estimación de la matriz de transición del n-ésimo paso P^n a nivel nacional

La matriz de probabilidades de transición entre estados de una cadena de Markov confiere las estimaciones asociadas al traslado de cada estadio del desarrollo del VIH/SIDA a otro en una sola unidad de tiempo (Miller & Childers, 2012). Es así como las probabilidades de transición del n-ésimo paso pueden ser utilizadas para encontrar estimaciones de transición en diferentes momentos discretos (en años). La probabilidad de transición de n pasos para una cadena de Markov viene dada por la ecuación 3.6. La predilección por aquel enfoque se fundamenta en su capacidad de obtener una solución matricial para la estimación a futuro. Para encontrar las propiedades de estabilidad de cada cadena planteada, la resolución del sistema matricial puede darse mediante la elección de coordenadas convenientes, en las que la matriz del sistema es diagonal (Λ) y, por lo tanto, las entradas de la matriz contienen autovectores asociados a autovalores (λ_i) concretos. Por lo tanto, son los autovalores los que determinan la estabilidad de la solución y, por ende, de la cadena en una forma explícita y directa.

4.2.4. Comportamiento estacionario de las matrices de transición del n-ésimo paso P^n a nivel nacional

De manera análoga al procesamiento de las probabilidades de transición para n finitos ciclos markovianos (partiendo de P^1) que serán evaluados, la evolución al futuro de las curvas de análisis de los estados que forman parte de los procesos de Markov planteados dependen de las condiciones y las características iniciales de la población en estudio (Ossa, 2013), observadas y registradas en el cálculo de las probabilidades de transición de una matriz en el horizonte temporal definido; considerando que, para este caso, n representará un marco de tiempo significativamente más extenso y no finito.

Para los períodos de estudio de la situación epidemiológica relacionada al VIH/SIDA en el Perú, el comportamiento estacionario de las matrices de transición de los casos de estudio

está determinado mediante el producto de las matrices iniciales (P^1 para todos los años) por sí misma n veces, dado que n convergerá a un límite que corresponderá al infinito ($n \to \infty$), el cuál dependerá de las condiciones iniciales de cada cadena de Markov y que puede ser cero para ciertos estados, permitiendo que cada estado que forma parte del proceso tenga una probabilidad positiva de estado estable o estacionario de ser ocupados en un futuro.

4.2.5. Proyecciones estocásticas de la evolución del VIH/SIDA en el Perú

Las cadenas de Markov harán un uso discreto del tiempo; dentro de períodos finitos de tiempo t, el modelo progresará a través de incrementos de tiempo fijos, técnicamente llamados ciclos de Markov, los cuales tendrán una duración habitual en términos anuales como intervalo de tiempo clínicamente significativo en infecciones como la discutida en la presente investigación (Mar *et al.*, 2010).

Los acontecimientos se modelizarán como pasos o transiciones de unos estados a otros que se producirán en períodos uniformes de tiempo y con unas probabilidades de transición que dependerán del estado en el que se encuentre el individuo en cada momento mediante un vector fila (Rubio-Terrés & Echeverría, 2006). En otros términos, en cada ciclo, el paciente llevará a cabo una transición de un estado a otro en función de las probabilidades especificadas para el estado y el ciclo en que se encuentre.

Ocaña-Riola (2009) establece que el curso que seguirá la evolución de la epidemia en el tiempo dependerá de las condiciones iniciales de la población estudiada con respecto al virus y la enfermedad. A saber, de la distribución temporal de los individuos en cada uno de los estados del virus y la enfermedad. Asimismo, Ocaña-Riola (2009) indica que, al modificar un vector de prevalencias iniciales (las condiciones iniciales de la situación epidemiológica de un año en análisis determinado), se pueden obtener diferentes simulaciones de la evolución hipotética de la epidemia a nivel nacional. Conocer qué ocurriría en el futuro si la situación actual fuese diferente constituye una importante fuente de información para la toma de decisiones, siendo ésta una de las principales aportaciones de los modelos markovianos a la investigación biomédica.

4.2.6. Análisis de sensibilidad bajo estrategias de control sobre el VIH/SIDA

Debido a la continuidad de la epidemia del VIH/SIDA, es importante que los responsables políticos tengan en cuenta el diseño y ejecución de planes orientados a disminuir los efectos nocivos de la infección en la población sana del país y en aquellos infectados que viven con el virus y/o la enfermedad (Rotich, 2016).

A fin de dilucidar cómo los cambios en la eficacia con la que pueden ser aplicadas estas estrategias en el sistema de salud público y privado pueden impactar en las probabilidades de transición de los estados establecidos que caracterizan la dinámica del VIH/SIDA en diferentes horizontes de tiempo, se pueden realizar proyecciones en base a dichas probabilidades a través del procesamiento de las matrices de transición con alteraciones mediante parámetros que

representan estrategias o políticas gubernamentales (Rotich, 2016).

Las matrices por plantear, según los parámetros a utilizar, son matrices modificadas, aplicables para una dinámica sistema, cuya población entera cambia con el tiempo (Rotich, 2016). Para un sistema conservado, donde la población permanece constante, la suma de elementos de cada columna es igual a uno. En este caso, la diferencia es el valor de la tasa de supervivencia para cada clase.

4.3. Resultados del modelo propuesto

Esta sección congrega el reporte de todos los resultados relacionados al tratamiento de los datos secundarios recolectados, el modelamiento de los indicadores epidemiológicos en el Perú para determinados períodos de investigación a través de transiciones y matrices de probabilidad, el comportamiento estacionario de la epidemia en el país y las estimaciones poblacionales estocásticas a futuro y, por último, la evaluación de diversas estrategias de control y contención de la epidemia a través de un análisis de sensibilidad con diversos parámetros a considerar.

4.3.1. Modelamiento mediante Cadenas de Markov

Como la Ecuación 3.8 describe, el número de individuos dentro de un año específico de estudio se encuentra representado por el vector X_{ij} con estados que forman parte del conjunto S establecido, donde i corresponde al estado inicial y j corresponde al estado final en el que el individuo se situara luego de un espacio de tiempo t. El vector X_{ij} satisface el modelo de cadenas de Markov que posee el conjunto de estado S = 1, 2, 3, descritos anteriormente.

La tabla 4.1 muestra el número de individuos dentro de la población de estudio en los estados 1, 2 y 3 a nivel nacional para los años 1995, 2005, 2011, 2013 y 2018.

Tabla 4.1: Número de individuos en cualquier estado de transición en el Perú para los años de estudio definidos

Año de estudio		**Año 1995**		
Grupos	**Total**	**Susceptible**	**VIH/SIDA**	**Muerte**
Individuos susceptible	24' 000,000	23' 996,688	**3,312**	0
Individuos infectados	68,321	0	65,108	**3,213**
Año de estudio		**Año 2005**		
Grupos	**Total**	**Susceptible**	**VIH/SIDA**	**Muerte**
Individuos susceptible	27' 870,000	27' 862,639	**7,361**	0
Individuos infectados	64,759	0	60,120	**4,639**
Año de estudio		**Año 2011**		
Grupos	**Total**	**Susceptible**	**VIH/SIDA**	**Muerte**
Individuos susceptible	29' 260,000	29' 253,684	**6,316**	0
Individuos infectados	66,855	0	65,337	**1,518**
Año de estudio		**Año 2013**		
Grupos	**Total**	**Susceptible**	**VIH/SIDA**	**Muerte**
Individuos susceptible	29' 770,000	29' 763,217	**6,783**	0
Individuos infectados	70,273	0	68,981	**1,292**
Año de estudio		**Año 2018**		
Grupos	**Total**	**Susceptible**	**VIH/SIDA**	**Muerte**
Individuos susceptible	31' 990,000	31' 980,694	**9,306**	0
Individuos infectados	78,752	0	77,722	**1,030**

Fuente: Elaboración propia.

De la tabla anterior, puede señalarse que el año 1995 es el que presenta la menor cifra oficial

registrada por los organismos de vigilancia sanitarias en el país (con 3,312 casos de VIH/SIDA notificados); sin embargo, este hecho se debe principalmente a la inoperancia de los mecanismos y servicios del estado peruano para una correcta fiscalización de la epidemia (Defensoría del Pueblo, 2009). Situación reflejada efectivamente en el alto número de fallecidos en la época producto de las complicaciones de la enfermedad (3,213 defunciones oficiales), casi equivalente al número de casos. Desestimando los mayores y mejores esfuerzos articulados intersectorialmente en el Perú para la época, el año 2005 fue el período de repunte de todas cifras de vigilancia epidemiológica (número de casos registrados y muertes); traduciéndose en el año en que la epidemia se recrudeció en el territorio nacional. Los años venideros estuvieron marcados por una tendencia de aumento de casos de contagios notificados de VIH/SIDA pero por una disminución del fallecimiento de los pacientes, lo que se entiende como una mejor política de antirretrovirales y tratamientos contra la enfermedad pero un comportamiento sexual riesgoso y no mesurado por parte de la población.

4.3.2. Estimación de probabilidades de transición

Basándonos en la descripción y formulación del método de estimación por máxima verosimilitud (MLE) en la Ecuación 3.9, dicho procedimiento será empleado para determinar las matrices de probabilidades de transición entre los estados i y j del conjunto S, considerando el número de individuos en cualquier estado de transición en el Perú según año de estudio (representados por los vectores X_{ij}).

A continuación, se muestran las estimaciones de las probabilidades de transición (representadas por p_{ij}), el error estándar asociado a la estimación (representado por SE) y los intervalos de confianza de las probabilidades de transición (representadas por I.C. 99 %) para los estados definidos para los modelos de Cadenas de Markov de cada año de estudio:

Tabla 4.2: Estimaciones de probabilidades de transición para el Perú en el año 1995

Parámetros	Estimación	SE	I.C. 99 %
p_{11}	9.998×10^{-1}	0.240×10^{-5}	9.998×10^{-1} - 9.999×10^{-1}
p_{12}	0.200×10^{-3}	0.240×10^{-5}	0.132×10^{-3} - 0.144×10^{-3}
p_{22}	9.529×10^{-1}	0.810×10^{-3}	9.510×10^{-1} - 9.549×10^{-1}
p_{23}	0.471×10^{-1}	0.810×10^{-3}	0.451×10^{-1} - 0.489×10^{-1}

Fuente: Elaboración propia.

Tabla 4.3: Estimaciones de probabilidades de transición para el Perú en el año 2005

Parámetros	Estimación	SE	I.C. 99 %
p_{11}	9.997×10^{-1}	0.310×10^{-5}	9.997×10^{-1} - 9.998×10^{-1}
p_{12}	0.300×10^{-3}	0.310×10^{-5}	0.257×10^{-3} - 0.271×10^{-3}
p_{22}	9.284×10^{-1}	0.101×10^{-2}	9.260×10^{-1} - 9.307×10^{-1}
p_{23}	0.716×10^{-1}	0.101×10^{-2}	0.693×10^{-1} - 0.740×10^{-1}

Fuente: Elaboración propia.

Tabla 4.4: Estimaciones de probabilidades de transición para el Perú en el año 2011

Parámetros	Estimación	SE	I.C. 99 %
p_{11}	9.998×10^{-1}	0.270×10^{-5}	9.997×10^{-1} - 9.998×10^{-1}
p_{12}	0.200×10^{-3}	0.270×10^{-5}	0.210×10^{-3} - 0.222×10^{-3}
p_{22}	9.773×10^{-1}	0.576×10^{-3}	9.760×10^{-1} - 9.786×10^{-1}
p_{23}	0.227×10^{-1}	0.576×10^{-3}	0.214×10^{-1} - 0.235×10^{-1}

Fuente: Elaboración propia.

Tabla 4.5: Estimaciones de probabilidades de transición para el Perú en el año 2013

Parámetros	Estimación	SE	I.C. 99 %
p_{11}	9.998×10^{-1}	0.280×10^{-5}	9.997×10^{-1} - 9.998×10^{-1}
p_{12}	0.200×10^{-3}	0.280×10^{-5}	0.221×10^{-3} - 0.234×10^{-3}
p_{22}	9.816×10^{-1}	0.507×10^{-3}	9.804×10^{-1} - 9.828×10^{-1}
p_{23}	0.184×10^{-1}	0.507×10^{-3}	0.172×10^{-1} - 0.196×10^{-1}

Fuente: Elaboración propia.

Tabla 4.6: Estimaciones de probabilidades de transición para el Perú en el año 2018

Parámetros	Estimación	SE	I.C. 99 %
p_{11}	9.997×10^{-1}	0.300×10^{-5}	9.997×10^{-1} - 9.998×10^{-1}
p_{12}	0.300×10^{-3}	0.300×10^{-5}	0.284×10^{-3} - 0.298×10^{-3}
p_{22}	9.869×10^{-1}	0.405×10^{-3}	9.860×10^{-1} - 9.878×10^{-1}
p_{23}	0.131×10^{-1}	0.405×10^{-3}	0.121×10^{-1} - 0.140×10^{-1}

Fuente: Elaboración propia.

Las estimaciones de las probabilidades de transición mostradas en las tablas anteriores se basan en el valor demográfico y las cantidades de casos y muertes reportadas para cada año como se muestra en la Sección 4.3.1. Así, mediante la estimación por máxima verosimilitud se tiene, para el año 1995 como instancia, que la probabilidad de que un individuo susceptible pueda permanecer sano en el futuro, se expresa como: $p_{11} = \frac{x_{11}}{N_1} = \frac{23'996,688}{24'000,000} = 9{,}998 \times 10^{-1}$. Bajo el mismo procedimiento y considerando los estados de inicio i y fin j para cada caso, las probabilidades fueron obtenidas. Adicionalmente a la determinación de estas, se consigue presentar los errores estándar asociados a cada una de las mismas, siendo bajos en todos los casos (aunque es relevante mencionar que estos aumentan al calcular las probabilidades de que un sujeto continúe infectado con VIH/SIDA p_{22} y un paciente fallezca estando enfermo p_{23}).

En términos comparativos, el año 2018 es el que presenta una mayor propensión a que los individuos lleguen a contagiarse de VIH y desarrollen SIDA en el tiempo (con un valor de probabilidad de 0.00029); contraria a la tendencia del año 1995, en el que las probabilidades de que un individuo se contagie ascendían únicamente a 0.00014, considerando estadísticas oficiales (debido a que, como se remarcó con antelación, existe un desfase y un subregistro no notificado de pacientes seropositivos en aquella época que no fueron atendidos o identificados por el aparato estatal).

Por otra parte, la mayor proporción de occisos dentro de la población prevalente infectada por VIH/SIDA se constata en el año 2005, con un valor de 0.07163, superando ampliamente al resto de probabilidades de transición de otros períodos de investigación, aspecto que coincide con la crisis mundial que se sufrió en dicho año por el aumento desmesurado e insólito de la epidemia. Las cifras de los años posteriores (a saber, 2013 y 2018), exhiben que un mejor tra-

tamiento y acceso masificado a medios de tratamiento disminuyen la proporción de fallecidos y aumentan la calidad de vida de los infectados.

Por lo tanto, se puede concluir de las estimaciones calculadas previamente que las matrices de probabilidad de transición para los estados de la distribución epidemiológica del VIH/SIDA pueden presentarse de la siguiente forma, respectivamente:

$$P_{Peru1995} = \begin{array}{c} \begin{array}{ccc} \text{Susceptible} & \text{VIH/SIDA} & \text{Muerte} \end{array} \\ \begin{bmatrix} 9{,}998 \times 10^{-1} & 0{,}200 \times 10^{-3} & 0 \\ 0 & 9{,}529 \times 10^{-1} & 0{,}471 \times 10^{-1} \\ 0 & 0 & 1 \end{bmatrix} \end{array} \begin{array}{r} \text{Susceptible} \\ \text{VIH/SIDA} \\ \text{Muerte} \end{array}$$

$$P_{Peru2005} = \begin{array}{c} \begin{array}{ccc} \text{Susceptible} & \text{VIH/SIDA} & \text{Muerte} \end{array} \\ \begin{bmatrix} 9{,}997 \times 10^{-1} & 0{,}300 \times 10^{-3} & 0 \\ 0 & 9{,}284 \times 10^{-1} & 0{,}716 \times 10^{-1} \\ 0 & 0 & 1 \end{bmatrix} \end{array} \begin{array}{r} \text{Susceptible} \\ \text{VIH/SIDA} \\ \text{Muerte} \end{array}$$

$$P_{Peru2011} = \begin{array}{c} \begin{array}{ccc} \text{Susceptible} & \text{VIH/SIDA} & \text{Muerte} \end{array} \\ \begin{bmatrix} 9{,}998 \times 10^{-1} & 0{,}200 \times 10^{-3} & 0 \\ 0 & 9{,}773 \times 10^{-1} & 0{,}227 \times 10^{-1} \\ 0 & 0 & 1 \end{bmatrix} \end{array} \begin{array}{r} \text{Susceptible} \\ \text{VIH/SIDA} \\ \text{Muerte} \end{array}$$

$$P_{Peru2013} = \begin{array}{c} \begin{array}{ccc} \text{Susceptible} & \text{VIH/SIDA} & \text{Muerte} \end{array} \\ \begin{bmatrix} 9{,}998 \times 10^{-1} & 0{,}200 \times 10^{-3} & 0 \\ 0 & 9{,}816 \times 10^{-1} & 0{,}184 \times 10^{-1} \\ 0 & 0 & 1 \end{bmatrix} \end{array} \begin{array}{r} \text{Susceptible} \\ \text{VIH/SIDA} \\ \text{Muerte} \end{array}$$

$$P_{Peru2018} = \begin{array}{c} \begin{array}{ccc} \text{Susceptible} & \text{VIH/SIDA} & \text{Muerte} \end{array} \\ \begin{bmatrix} 9{,}997 \times 10^{-1} & 0{,}300 \times 10^{-3} & 0 \\ 0 & 9{,}869 \times 10^{-1} & 0{,}131 \times 10^{-1} \\ 0 & 0 & 1 \end{bmatrix} \end{array} \begin{array}{r} \text{Susceptible} \\ \text{VIH/SIDA} \\ \text{Muerte} \end{array}$$

De la representación matricial de las transiciones de la dinámica epidemiológica de la infección, se puede mencionar que los estados “Susceptible” y “VIH/SIDA” son, en ambos casos, del tipo transciente o transitorio y el estado “Muerte” es del tipo absorbente, ofreciendo algunas

precisiones importantes:

- El tipo de clasificación que poseen los estados mencionados anteriormente es producto de la naturaleza epidemiológica del progreso de la enfermedad en aquellos individuos infectados. Los sujetos que no hayan contraído el virus en un determinado momento pueden permanecer, en un siguiente horizonte temporal, como susceptibles (no contagiados pero en riesgo de poder ser infectados) o pueden pasar a ser portadores del virus del VIH y desarrollar la enfermedad dependiendo de la conducta del individuo.
- En el mismo sentido, aquellos portadores del virus o que se encuentran en el último estadio de este (SIDA) pueden permanecer, en un futuro, infectados dependiendo de las conductas que estos mismos adopten o pueden fallecer a raíz de complicaciones asociadas al virus o la enfermedad. Sin embargo, una vez que el individuo abandone un determinado estado, no puede retornar al mismo en un siguiente horizonte temporal.
- En el caso del estado "Muerte", este puede ser clasificado como un estado del tipo absorbente (una vez accedido este estado, solo puede acceder a sí mismo o sucederse a sí mismo en un próximo horizonte temporal), ya que la probabilidad estimada de acceder a otro tipo de estado es 0. El tipo de clasificación que posee dicho estado es generado por la mortalidad asociada al virus y la enfermedad.
- Finalmente, considerando que los diagramas de los casos de estudio poseen exclusivamente estados del tipo transitorios y absorbentes, se puede concluir que el tipo de cadena de Markov que caracteriza la evolución del VIH/SIDA en el Perú es una cadena absorbente de Markov con espacio de estados finito.

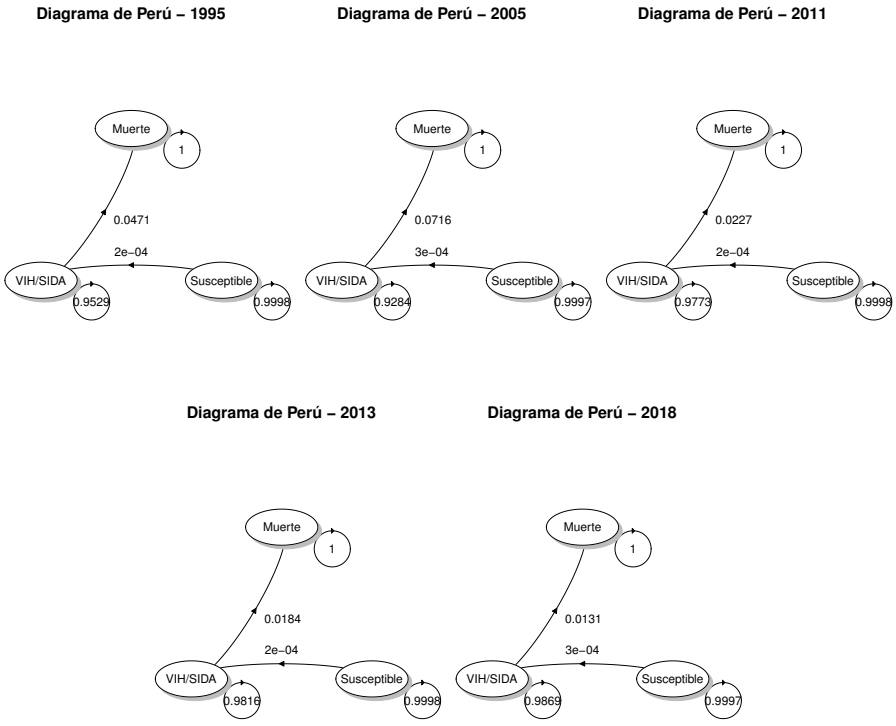


Figura 4.1: Diagramas de probabilidades de transición a nivel nacional por año de estudio. Fuente: Elaboración propia.

Asimismo, Bertsekas & Tsitsiklis (2008) establecen que el gráfico de probabilidades de transición se utilizó para diseñar el modelo de Cadenas de Markov y representar las estimaciones de los estados descritos para cada año de estudio, cuyos nodos son los estados y cuyos arcos son las transiciones calculadas. Al registrar los valores numéricos de p_{ij} cerca de los arcos correspondientes, se puede visualizar todo el modelo de una manera en que pueden hacerse evidentes algunas de sus principales propiedades. La Figura 4.1 muestra la representación gráfica de las probabilidades de transición entre cada estado para los años de estudio del VIH/SIDA a nivel nacional.

4.3.3. Estimación de la matriz de transición del n-ésimo paso P^n a nivel nacional

Dadas las probabilidades de transición de un n=1 paso, es factible calcular o estimar estimaciones de orden superior utilizando el enfoque de descomposición de autovalores y autovectores. Bajo la aplicación de la descomposición, una matriz P_t de probabilidad de transición del primer paso (donde t sea el período a evaluar), se puede expandir mediante la Ecuación 3.5, donde λ es la matriz diagonal de valores propios, I representa la matriz identidad $i \times j$, Q es la matriz cuyas columnas son los vectores propios correspondientes y $Q-1$ es la matriz inversa de Q para cada matriz de transición de los casos de estudio subyacentes.

En ese sentido, las estimaciones de las matrices de transición del n-ésimo paso P^n para la presente investigación pueden calcularse a partir de las matrices de probabilidades de

transición originales o de un n=1 paso, descritas en secciones anteriores de este capítulo, y la determinación de los autovalores y autovectores asociados a cada matriz de transición a través del procedimiento mostrado y descrito en la Sección A del capítulo de Anexos.

Considerando al año 1995 como referencia para la estimación del n-ésimo paso, se tiene la siguiente matriz de probabilidades de transición para la dinámica epidemiológica del VIH/SIDA en el Perú:

$$P_{Peru1995} = \begin{array}{c} \begin{array}{ccc} \text{Susceptible} & \text{VIH/SIDA} & \text{Muerte} \end{array} \\ \begin{bmatrix} 9{,}998 \times 10^{-1} & 0{,}200 \times 10^{-3} & 0 \\ 0 & 9{,}529 \times 10^{-1} 0 & 0{,}471 \times 10^{-1} \\ 0 & 0 & 1 \end{bmatrix} \end{array} \begin{array}{r} \text{Susceptible} \\ \text{VIH/SIDA} \\ \text{Muerte} \end{array}$$

La ecuación determinística, presentada en la Sección A en Anexos, denominada como la ecuación característica de P es el resultado de igualar al polinomio característico de n-ésimo orden de P - det(λI - P) o $|\lambda I - P|$ - a 0 en λ con n raíces para el año 1995 (estas raíces serán conocidas como los valores propios o autovalores de P para el año de estudio).

Al resolver dicha expresión determinística, se pudo obtener 3 cifras que corresponden a los autovalores de la matriz P_{1995}: $\lambda_1 = 1$, $\lambda_2 = 9{,}998 \times 10^{-1}$ y $\lambda_3 = 9{,}530 \times 10^{-1}$. En base a esos *eigenvalues* o autovalores, la matriz de autovectores Q y la matriz inversa Q^{-1} fueron calculadas siguiendo el procedimiento planteado en la Sección 3.1.1.5.

Tomando en cuenta los autovectores y sus inversas para el año 1995, la matriz de transición del n-ésimo paso para ese año será producto de la multiplicación de dichas matrices (Q y Q^{-1}) por la matriz λ^n, la cual tiene la estructura de una matriz diagonal cuyos valores en la diagonal principal serán los 3 autovalores obtenidos con anterioridad, donde n será la unidad de tiempo (por lo genral, años) en el futuro a estimar.

Es así como, en base a estas operaciones, la matriz de probabilidades de transición P^n para la evolución de la epidemia en el Perú en el año 1995 pudo ser estimada con los siguientes valores:

$$P^n_{Peru1995} = \begin{pmatrix} 0{,}999^n & -0{,}003(0{,}953^n - 0{,}999^n) & 1 - 1{,}003(0{,}999^n) + 0{,}003(0{,}953^n) \\ 0 & 0{,}953^n & 1 - 0{,}953^n \\ 0 & 0 & 1 \end{pmatrix}. \quad (4.1)$$

En síntesis, una matriz de transición en n pasos es el resultado de multiplicaciones sucesivas de la misma matriz por ella misma n veces. Basado en $P^n = P^{n-1} \times P$, $n = 1$ daría como resultado la primera matriz de transición $P_{Peru1995}$.

Tomando en consideración el propósito de esta sección, mediante el procedimiento estocástico descrito previamente para el año 1995, se procedió a generalizar este proceso hacia la esti-

mación de las matrices de probabilidades de transición del n-ésimo paso P^n para la dinámica epidemiológica del VIH/SIDA en el Perú para los otros años de análisis planteados, como se muestra a continuación en cada caso.

$$P^n_{Peru2005} = \begin{pmatrix} 0{,}999^n & -0{,}004(0{,}928^n - 0{,}999^n) & 1 - 1{,}004(0{,}999^n) + 0{,}004(0{,}928^n) \\ 0 & 0{,}928^n & 1 - 0{,}928^n \\ 0 & 0 & 1 \end{pmatrix}. \quad (4.2)$$

$$P^n_{Peru2011} = \begin{pmatrix} 0{,}999^n & -0{,}009(0{,}977^n - 0{,}999^n) & 1 - 1{,}009(0{,}999^n) + 0{,}009(0{,}977^n) \\ 0 & 0{,}977^n & 1 - 0{,}977^n \\ 0 & 0 & 1 \end{pmatrix}. \quad (4.3)$$

$$P^n_{Peru2013} = \begin{pmatrix} 0{,}999^n & -0{,}013(0{,}982^n - 0{,}999^n) & 1 - 1{,}013(0{,}999^n) + 0{,}013(0{,}982^n) \\ 0 & 0{,}982^n & 1 - 0{,}982^n \\ 0 & 0 & 1 \end{pmatrix}. \quad (4.4)$$

$$P^n_{Peru2018} = \begin{pmatrix} 0{,}999^n & -0{,}023(0{,}987^n - 0{,}999^n) & 1 - 1{,}023(0{,}999^n) + 0{,}023(0{,}987^n) \\ 0 & 0{,}987^n & 1 - 0{,}987^n \\ 0 & 0 & 1 \end{pmatrix}. \quad (4.5)$$

Como puede evidenciarse, las matrices de transición en cualquier paso de tiempo (n≤1) para cada año de estudio se pueden generar a partir de sus matrices P^n ajustadas. De la misma manera, las cadenas formadas son aperiódicas pero no irreductibles ya que el estado retirado (3: Muerte) es un estado absorbente. Así, las matrices de probabilidades de transición P^n predicen las probabilidades de transición para la epidemia del VIH/SIDA en cualquier paso o instante de tiempo n, lo que permite hacer una evaluación prospectiva de los indicadores o estadísticas epidemiológicas relevantes del VIH/SIDA al generalizar estas probabilidades de transición para predicciones futuras.

En otro orden de ideas, bajo una perspectiva comparativa, se puede establecer que los resultados de las matrices del n-ésimo permiten distinguir aquellos años de estudio que, a futuro, tienen un mejor desempeño epidemiológico considerando los indicadores de la epidemia discutidos con anterioridad. De esa manera, se puede precisar que el año 2018 tendrá una mayor probabilidad a largo plazo de tener la mayor proporción de individuos infectados con VIH/SIDA que sobrevivan en el tiempo o se mantengan en dicho estado como población prevalente, lo que se traduce al mismo tiempo en la menor probabilidad de mortalidad a futuro entre todos los años de análisis (debido a que los resultados de los cálculos exponenciales del

valor de la probabilidad p_{22}=0.987 elevado a un instante de tiempo n serán siempre los mayores entre todos los valores de los 5 años). De igual modo, se puede establecer que el año 2005 es el período que tendrá las estimaciones más desalentadoras relacionadas a las probabilidades de supervivencia de los infectados y mortalidad en la población a nivel nacional, ya que estas probabilidades p_{22}=0.861 y p_{23}=0.139 elevadas a un tiempo n mostrarán las mayores reducciones en población prevalente de contagiados y el mayor número de fallecimientos en el tiempo. Por otro lado, considerando las proyecciones de las tasas de individuos susceptibles a adquirir el VIH/SIDA (o población sana) y aquellos que han sido infectados, se puede definir a los años 1995 y 2011 como aquellos períodos que, a futuro, tendrán las menores tasas o ratios de individuos que adquirieron el VIH/SIDA en algún momento, debido a que las probabilidades p_{12} con valores de $1{,}380 \times 10^{-3}$ y $1{,}980 \times 10^{-3}$, respectivamente, que representan a la transición de individuos sanos a un estado de infección tendrán los menores aumentos a futuro. No obstante, se puede determinar que los años 2005 y 2018 son aquellos lapsos de estudio que presentarán las mayores tendencias de infección dentro de la población susceptible a largo plazo, lo que permite especificar que estos son los períodos en los que cuales habrá una cantidad más elevada de individuos que adquirieron el virus/desarrollaron la enfermedad a futuro, en vista de que las probabilidades p_{12} que reportan cifras de $2{,}840 \times 10^{-3}$ y $2{,}760 \times 10^{-3}$, correspondientemente, indican las mayores propensiones en el tiempo de registrar mayores contagios que en otros años de evaluación.

4.3.4. Comportamiento estacionario de las matrices de transición del n-ésimo paso P^n a nivel nacional

Considerando el producto matricial de las probabilidades de transición en el Perú para el año 1995, la Figura 4.2 muestra la evolución de la epidemia del VIH/SIDA tomando como distribución inicial a dicha matriz de transición y calculando las probabilidades para cada uno de los estados incluidos en la cadena cuando el número de periodos (es decir, n) tiende al infinito.

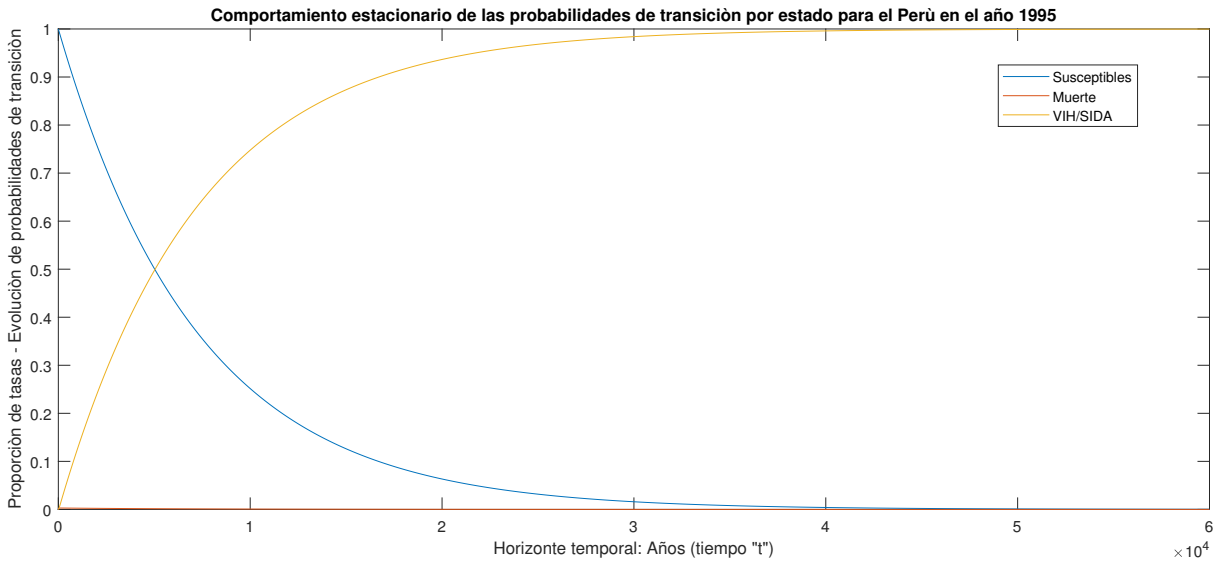


Figura 4.2: Comportamiento estacionario de la distribución epidemiológica en el Perú para el año 1995. Fuente: Elaboración propia.

Como se puede apreciar, la proporción de individuos que se encuentran en el estado "Susceptible" tiende a disminuir en el tiempo (cuanto t crece) conforme a que las probabilidades de fallecer o del estado "Muerte" (como lo muestra la Figura B.1) aumentan progresivamente. Ambas curvas se interceptan cuando están próximas al valor de probabilidad $p = 0{,}5$ y habiendo pasado aproximadamente 2,000 unidades de tiempo o ciclos markovianos a futuro; luego de ese hecho, la curva de "Susceptibles" decrecerá hasta tener un comportamiento asintótico con el eje X aproximándose a 0 en relación con el eje Y y la curva de "Muerte" crezca hasta aproximarse a 1 en relación al eje Y en el tiempo. Asimismo, la particularidad de la evolución del estado estacionario para cada cuadro clínico considerado es que la probabilidad de estar infectado con VIH/SIDA (o alcanzar el estado 2 - Infección) no sufrirá cambios relevantes o apreciables que puedan desprenderse de la figura, solo se detecta un nimia disminución desde el $t = 1$ hasta asumir un comportamiento asintótico con el eje X (de tiempo) cuando este es 0 en el tiempo; esto se debe principalmente a que el valor de probabilidad de transición p_{12} tiene una influencia poco considerable dentro del procesamiento de la matriz, el estado "Susceptible" y "Muerte" impactan en gran magnitud en el transcurso de la epidemia a través del tiempo. Con un mayor énfasis, el gráfico de comportamiento estacionario refleja la propiedad principal de las cadenas de Markov absorbentes, que puntualizan que, a través del tiempo, todo estado i progresará dentro de la matriz de transición hasta llegar al estado absorbente j, el cual no podrá abandonar; de manera equivalente, los estados "Susceptible" y "VIH/SIDA" asumirán diferentes tendencias en el tiempo hasta que el estado "Muerte" sea eventualmente alcanzado y asuma la totalidad de las probabilidades, como se aprecia en la figura correspondiente al año 1995.

Bajo dicha premisa, en las siguientes figuras se muestran los comportamientos estacionarios para los otros años de estudio restantes. De igual modo, se debe mencionar que la escala

de tiempo (t) o el horizonte temporal con el que se han realizado y evaluado los gráficos de comportamiento estacionario se ha estandarizado entre los años de estudio para que pueda apreciarse cómo las curvas de todos los estados asumen un comportamiento próximo a cualquiera de los dos ejes (ya sea el X cuando adopten un valor cercano al 0 o el Y cuando estén próximos al valor 1) de forma comparativa para poder distinguir la tendencia que asumen los estados a largo plazo entre los períodos.

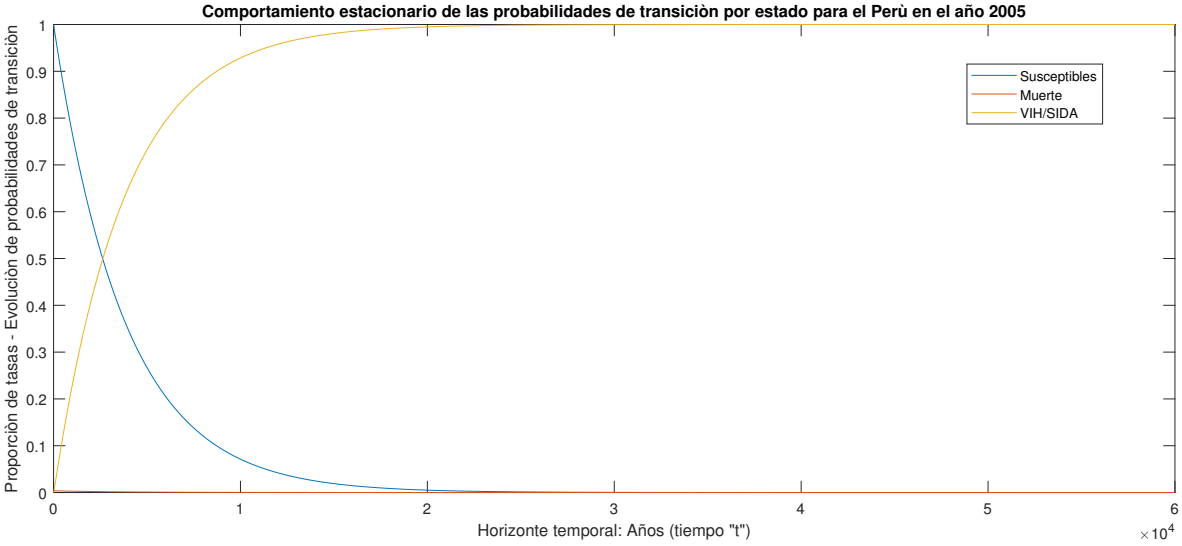


Figura 4.3: Comportamiento estacionario de la distribución epidemiológica en el Perú para el año 2005. Fuente: Elaboración propia.

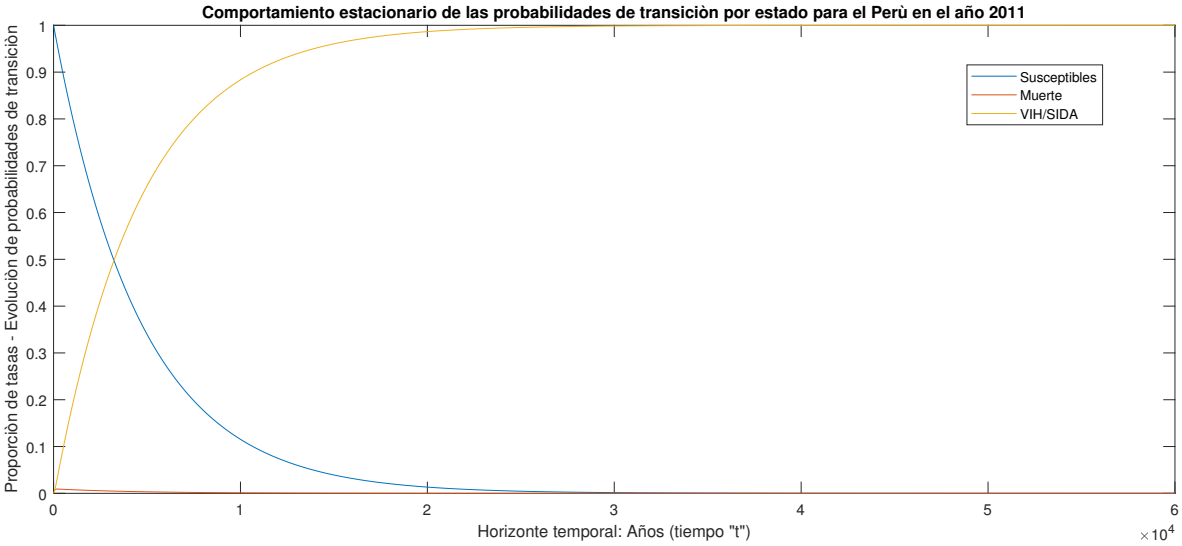


Figura 4.4: Comportamiento estacionario de la distribución epidemiológica en el Perú para el año 2011. Fuente: Elaboración propia.

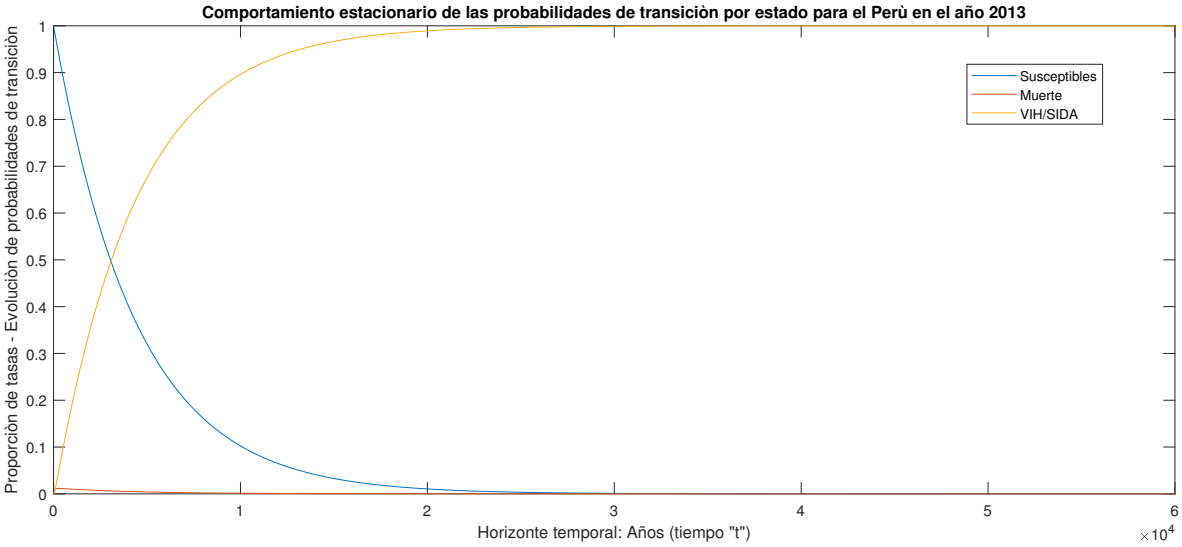


Figura 4.5: Comportamiento estacionario de la distribución epidemiológica en el Perú para el año 2013. Fuente: Elaboración propia.

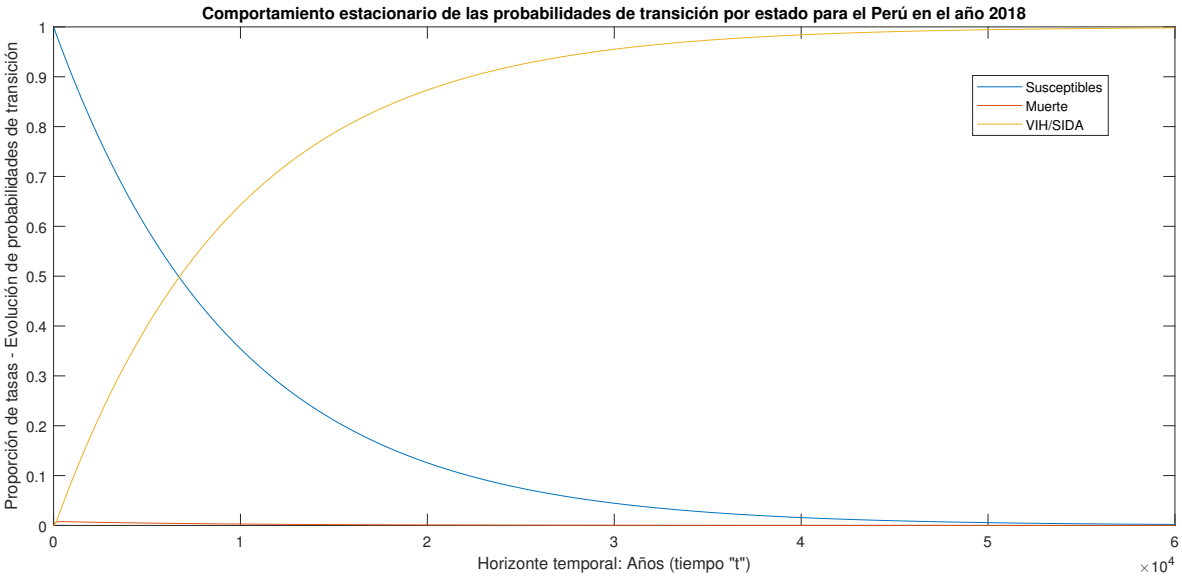


Figura 4.6: Comportamiento estacionario de la distribución epidemiológica en el Perú para el año 2018. Fuente: Elaboración propia.

A partir de las evoluciones en el tiempo anteriormente presentadas del comportamiento estacionario del resto de años de estudio, puede concluirse que todos los períodos siguen la misma propiedad absorbente que el año 1995 - el estado "Susceptible" y "VIH/SIDA" tenderán a 0 en el tiempo y el estado "Muerte" se aproximará a 1 a largo plazo. Todas las intersecciones entre los estados "Susceptible" y "Muerte" pueden ser visualizadas cuando ambos están próximos al valor 0.5 de la probabilidad de transición en los años considerados. Empero, dichas intersecciones serán alcanzadas en marcos de tiempo diferentes en cada escenario: en el caso del 2005, la intersección se presenta antes de las 50,000 unidades de tiempo; para el 2011 y

2013, esta ocurrirá aproximadamente en 60,000 unidades de tiempo; y, finalmente, tomando en cuenta el 2018, la intersección se dará alrededor de las 100,000 unidades t. Estas diferencias simbolizan la efectividad de los tratamientos de salud y la política sanitaria en términos de la epidemia que tienen como fin aumentar la esperanza de vida de los individuos y retrasar los efectos mortales del VIH/SIDA; distinguiéndose la naturaleza del año 2018, el cual, conforme a los indicadores epidemiológicos de la Sección 4.3.1, registra la menor cifra de defunciones.

Por otro lado, en la Sección B del capítulo de Anexos, se presentan las curvas de comportamiento estacionario del estado (3) Muerte para cada año de estudio en mayor escala ofreciendo así una mejor visibilidad de la evolución de este estado en el tiempo y las diferencias apreciables entre los años de análisis que ya fueron discutidas con anterioridad en esta sección. En este caso, de dichas figuras que representan el comportamiento estacionario del estado Muerte, se puede establecer que los años 2013 y 2018 son aquellos que tienen el mayor crecimiento de la tasa de mortalidad entre todos los períodos de evaluación a largo plazo (debido a que el primer año alcanza un crecimiento hasta el $0{,}12 \times 10^{-1}$ y el segundo llega hasta un tasa de $0{,}20 \times 10^{-1}$), los más altos entre todos los años; no obstante, son los años que alcanzan una tendencia asintótica más rápido en el tiempo (2×10^4 y $2{,}5 \times 10^4$, respectivamente), lo que precisa que si bien existen mayor proporción de individuos infectados que fallecen a causa de la enfermedad dados los vectores iniciales establecidos en la simulación, la política de salud en dichos lapsos es lo suficiente eficaz para controlar los fallecimientos en el tiempo. Es relevante especificar que si bien el año 1995 es que el período que presenta el menor crecimiento de las probabilidades de mortalidad de los individuos infectados a largo plazo, es el año en el que la curva demorá más para estabilizarse y alcanzar una tendencia asintótica. Este hecho responde a que, para ese entonces, el Perú así como la mayoría de países de la periferia carecían de tratamientos eficaces contra el virus/enfermedad y, aunado a la incapacidad de poder detectar si un fallecimiento era producto del VIH/SIDA, la estabilización demoraría mucho más comparada con otros años con innovaciones y políticas más efectivas.

4.3.5. Proyecciones estocásticas de la evolución del VIH/SIDA en el Perú

En primera instancia, en el caso de la situación epidemiológica del Perú en el año 1995, la cifra de la población susceptible total ascendía a 23'928,466 de individuos en riesgo de infección, 68,321 pacientes vivos infectados con VIH/SIDA y 3,213 muertes producto de complicaciones médicas asociadas al virus y la enfermedad. Conformando la distribución de vectores para el año de estudio específico, estas cifras se configuran como las condiciones iniciales para el estudio prospectivo del progreso del virus y la enfermedad en el país. El vector de casos prevalentes para el año 1995 será, entonces, representado como v_{1995} = (23'928,466; 68,321; 3,213) y la suma total de estos valores resulta en la población total estudiada en el territorio nacional para dicho periodo de tiempo: 24'000,000 ciudadanos en el Perú.

En el estudio de epidemias poblacionales, la manera más habitual de expresar el vector de prevalencias iniciales es como número de casos por millón de habitantes (Ocaña-Riola, 2009). Es por ello que el vector de condiciones iniciales de casos prevalentes final para el año 1995

es v_{1995} = (997,019; 2,847; 134).

De la misma manera, el mismo procedimiento fue aplicado al resto de años de estudio para determinar los vectores de condiciones iniciales propios de los períodos de investigación.

A continuación, se presenta en la Tabla 4.7 la distribución final de vectores de condiciones iniciales de la prevalencia de casos de la cohorte (el conjunto de individuos a observar durante el tiempo, expresado en millones de habitantes) a emplear en el procesamiento de las matrices de transición por año de estudio establecido en la presente investigación. A su vez, se muestra el horizonte temporal establecido para realizar la evaluación prospectiva de los estadios en la dinámica epidemiológica del VIH/SIDA basada en el cálculo de los ciclos markovianos y las probabilidades de transiciones descritas en la sección anterior para cada período de investigación considerado.

Tabla 4.7: Vectores de condiciones iniciales de casos prevalentes (número de casos por millón de habitantes) y horizonte de proyección del estudio de cohorte para el Perú según año de estudio.

Año	Población susceptible	Casos de VIH/SIDA	Muerte
1995	997019	2847	134
2005	997510	2324	166
2011	997663	2285	52
2013	997596	2361	43
2018	997506	2462	32
Horizonte de proyección		25 años	

Fuente: Elaboración propia.

La tabla 4.8 muestra el resultado de referencia de la matriz de permanencia de la epidemia de VIH/SIDA para el año 1995 en el Perú por estado y año futuro posterior al procesamiento del modelo de Markov dado según los ciclos markovianos establecidos en el estudio, teniendo en cuenta la matriz de transición P^n obtenida previamente y la cohorte (vector de condiciones iniciales) y horizonte temporal establecidos. El ciclo 0 (el año base - 1995) de la matriz contiene las cifras dadas de la cohorte como distribución inicial en la evaluación prospectiva. Una vez finalizado el horizonte temporal, se puede determinar la variación entre la condicional inicial de los estados y el progreso epidemiológico del virus y la enfermedad a fin de valorar la efectividad de las políticas de salud del gobierno y las acciones dirigidas a controlar la epidemia dadas en dicha época.

Tabla 4.8: Evolución del estudio de cohorte según vector de prevalencias iniciales (número de casos por millón de habitantes) para el Perú en el año 1995.

Año	Población susceptible	Casos de VIH/SIDA	Muerte
1996	996881	2851	268
1997	996744	2854	402
1998	996606	2858	536
1999	996469	2861	671
2000	996331	2864	805
2001	996194	2866	940
2002	996056	2869	1075
2003	995919	2872	1210
2004	995781	2874	1345

Tabla 12: Evolución del estudio de cohorte según vector de prevalencias iniciales (número de casos por millón de habitantes) para el Perú en el año 1995 (Continuación)

Año	Población susceptible	Casos de VIH/SIDA	Muerte
2005	995644	2876	1480
2006	995507	2878	1615
2007	995369	2880	1750
2008	995232	2882	1886
2009	995094	2884	2021
2010	994957	2886	2157
2011	994820	2887	2293
2012	994683	2889	2429
2013	994545	2890	2564
2014	994408	2892	2700
2015	994271	2893	2836
2016	994134	2894	2972
2017	993996	2895	3108
2018	993859	2896	3245
2019	993722	2897	3381
2020	993585	2898	3517
Variación final	-0.33 %	1.65 %	1212.31 %

Fuente: Elaboración propia.

Como se puede apreciar mediante el cuadro anterior, la proporción de individuos que pueden pasar, en cada nuevo período o ciclo, de un estado a otro está en función de las probabilidades de transición que se lo permitan. A saber, para obtener la cifra de 2,854 casos de individuos que viven infectados con VIH/SIDA para el segundo ciclo markoviano (el año 1997 dentro de la matriz de permanencia) se deben tomar en cuenta aquellos individuos que provienen del estado N° 01 y pasan al estado N° 02 ligado a la probabilidad de transición p_{12} que fundamenta dicho progreso (996,881 $\times$ 0.000138) y aquellos individuos que permanecen en el estado N° 02 y la probabilidad p_{22} (2,851 $\times$ 0.9529720).

Por otro lado, el procesamiento de las matrices de transición ofrece una perspectiva a futuro de la situación epidemiológica de la población en estudio en el escenario en que las autoridades y entidades gubernamentales decidan no realizar medidas y/o políticas orientadas a la reducción y concientización del virus y de la enfermedad (Salgado, 2015). A saber, bajo esta técnica de análisis se puede generar un contexto conjetural que permite dilucidar la evolución la epidemia si la labor y la gestión del aparato de salud relacionado al VIH/SIDA no experimenta cambio alguno.

En el contexto del estudio, si las herramientas e intervenciones en materia de salud en el año 1995 no sufrían modificaciones favorables y se continuaba en la misma línea de acción y gestión pública, para el año 2020: la proporción de la población susceptible a contraer el virus o desarrollar la enfermedad en el país disminuiría hasta llegar a 993,585 individuos por millón de habitantes en riesgo para el fin de los ciclos markovianos (en otros términos, una variación negativa del 0.33 % en comparación a la situación inicial de los vectores de prevalencia), el número de casos por millón de individuos infectados por VIH/SIDA ascendería a 2,898 (es decir, una variación positiva del 1.65 % en comparación a la situación inicial de los vectores de prevalencia) y, finalmente, el número de decesos por millón de habitantes producto del

VIH/SIDA ascendería a 3,517 (en otros términos, una variación positiva del 1212.31 % en comparación a la situación inicial de los vectores de prevalencia).

Es así como, bajo el procesamiento de los modelos de Markov para el resto de los años especificados, se pueden evaluar las variaciones de la epidemia a futuro. A continuación, se presenta la comparación entre el procesamiento de los primeros y últimos ciclos markovianos dentro de las cadenas de Markov planteadas para los períodos de estudio restantes y el resultado de la variación final en la matriz de permanencia, mostrando la evolución de la infección del VIH/SIDA dentro de la población en el país.

Tabla 4.9: Evolución del estudio de cohorte según vector de prevalencias iniciales (número de casos por millón de habitantes) para el Perú en el año 1995, 2005, 2011, 2013 y 2018

Año	Evolución	Población susceptible	Casos de VIH/SIDA	Muerte
1995	1996	996881	2851	268
	2020	993585	2898	3517
	Variación final	-0.33 %	1.65 %	1212.31 %
2005	2006	997247	2421	322
	2030	990945	3454	5602
	Variación final	-0.63 %	42.67 %	1639.75 %
2011	2012	997663	2285	104
	2036	992292	5418	2289
	Variación final	-0.54 %	137.11 %	2100.96 %
2013	2014	997369	2545	86
	2038	991930	6060	2011
	Variación final	-0.55 %	138.11 %	2238.37 %
2018	2019	997216	2720	64
	2043	990277	7971	1752
	Variación final	-0.70 %	193.05 %	2637.50 %

Fuente: Elaboración propia.

Considerando la tabla anterior, el año 1995 sería el período con los mejores indicadores epidemiológicos a futuro, tomando en cuenta que la reducción en la población susceptible (-0.33 %), el progreso de las nuevas infecciones por VIH/SIDA (1.65 %) y la mortalidad poblacional asociada a la epidemia (1212.31 %) son las menores a futuro. Sin embargo, como ya se estableció previamente, este año presenta un significativo subregistro de casos de infección sin detectar y defunciones que no fueron reconocidas como consecuencia de los efectos de la inmunosupresión producto del virus y la enfermedad. Por lo que, las cifras proyectadas para este año serán tomadas de forma aislada.

Es así como, basándonos en la tabla 4.9, expresando las cifras por millón de habitantes, se puede discriminar que, luego de la evolución a futuro en 25 años o ciclos markovianos, el año 2005 (como vector de condiciones iniciales) tendría la menor proporción de aumento de individuos infectados con VIH/SIDA (42.67 %) y la menor proporción de decesos (1639.75 %) al final del horizonte temporal dispuesto; sin embargo, el 2005 tendría una de las mayores reducciones de la proporción de individuos susceptibles a contraer y desarrollar el VIH/SIDA (-0.63 %), siendo únicamente superado por la reducción de la población susceptible del 2018 (-0.70 %).

En el caso del 2018, además de tener la mayor reducción de población susceptible de los 4 períodos, posee también la mayor proporción de individuos infectados por VIH/SIDA luego de la proyección de las condiciones iniciales a 25 ciclos (193.05 %) y la mayor proporción de fallecimientos (2637.50 %).

Finalmente, el año 2011 posee un mejor resultado futuro que el año 2013 considerando las variaciones de los 3 estados considerados en el proceso de Markov: estado "Susceptible" (-0.54 % y -0.55 %, respectivamente), estado "VIH/SIDA" (137.11 % y 138.11 %, correspondientemente) y estado "Muerte" (2100.96 % y 2238.37 %, para cada año dado).

4.3.6. Análisis de sensibilidad bajo estrategias de control sobre el VIH/SIDA

En esta sección, se presentarán evaluaciones experimentales de estrategias de control bajo un análisis de sensibilidad de una cadena de Markov de manera prospectiva o a futuro.

Para la realización del procesamiento de la cadena, se emplearán los valores de las probabilidades de transición epidemiológicos del Perú en el año 2018 en los escenarios a analizar, puesto que es el período de investigación en el cual diversos tratamientos y medidas estatales nóveles están en vigor y el que cuenta con los registros oficiales más completos a la fecha del estudio. Matricialmente, la dinámica del VIH/SIDA para el año 2018 se representa como:

$$p_{ij} = \begin{pmatrix} p_{00} & p_{01} & 0 \\ 0 & p_{11} & p_{12} \\ 0 & 0 & 1 \end{pmatrix} \rightarrow P_{2018} = \begin{pmatrix} 9{,}997 \times 10^{-1} & 0{,}300 \times 10^{-3} & 0 \\ 0 & 9{,}869 \times 10^{-1} & 0{,}131 \times 10^{-1} \\ 0 & 0 & 1 \end{pmatrix}. \quad (4.6)$$

Es a dicha matriz de transición que se le aplicarán las estrategias de control con determinados coeficientes de ponderación a proponer. Cabe notar que las sumas de los elementos en cada fila en la matriz, luego de la implementación de cada parámetro en el modelo, no sumarán 1 (como la propiedad de las cadenas de Markov puntualiza bajo $\sum p_{ij} = 1$).

A continuación, se definen y presentan los cálculos de las estrategias de control y contención de la epidemia a emplear en la investigación.

4.3.6.1. Estrategia N°01: Control sobre el comportamiento y dinámica sexual de riesgo

El riesgo de contraer el VIH varía mucho según el tipo de exposición o factores conductuales. Las estrategias de comportamiento y dinámica sexual (que engloban: conocimiento de la infección, reducción del estigma, acceso a los servicios, retraso en el inicio de la primera relación sexual, disminución en el número de parejas sexuales, relaciones no monogámicas de riesgo, aumento en la venta o uso de condones para fomentar el sexo con protección, atenuación de la combinación de numerosas actividades de riesgo y disminución en el intercambio de equipos de inyección contaminados) deben lograrse y mantenerse, tanto entre individuos como entre grandes grupos de personas en riesgo, para reducir la posibilidad de transmisión del virus

en la población (Coates *et al.*, 2008). Se debe adoptar un enfoque multinivel que abarque estrategias conductuales; la prevención del VIH debe integrarse con enfoques biomédicos y estructurales, y el tratamiento de la infección.

Una estrategia planteada en base a una intervención sobre factores conductuales de riesgo se traduce en un parámetro α dentro del procesamiento de una cadena de Markov. La eficacia de dicha estrategia determina el tamaño del parámetro, que puede asumir valores entre $0 \leq \alpha \leq 1$. Cuanto mayor sea el valor de α, mejor será el control de la transmisión del VIH y el número de casos de contagio registrados. De igual modo, la representación matricial de la influencia de dicho parámetro sobre las probabilidades de transición es determinada de la siguiente manera:

$$P_{comportamiento} = \begin{pmatrix} p_{11} & p_{12}(1-\alpha) & 0 \\ 0 & p_{22} & p_{23} \\ 0 & 0 & 1 \end{pmatrix} \tag{4.7}$$

Donde puede evidenciarse la ponderación del parámetro α, bajo el nivel de eficacia a ser provisto o alcanzado por el sector salud en cuanto a su aplicación y efectos sobre la probabilidad p_{12} de adquirir la infección al ser anteriormente un individuo sano.

Generando estimaciones de las probabilidades de transición bajo la influencia del parámetro α para diferentes horizontes de tiempo n y bajo distintos niveles de eficacia que la estrategia de control puede asumir, se obtienen los resultados de la Tabla 4.10.

Tabla 4.10: Variación de las probabilidades de transición en el tiempo (n = años) con estrategia de control del parámetro α.

Simulación de las probabilidades por cambios en el parámetro α								
Estado	$S \to S : p_{11}$				$S \to I : p_{12}$			
Tiempo (años)	S.T.[a]	α=0.1	α=0.2	α=0.3	S.T.	α=0.1	α=0.2	α=0.3
n=5	9.985×10^{-1}	9.995×10^{-1}	9.996×10^{-1}	9.996×10^{-1}	1.416×10^{-3}	4.545×10^{-4}	4.040×10^{-4}	3.535×10^{-4}
n=10	9.971×10^{-1}	9.991×10^{-1}	9.992×10^{-1}	9.993×10^{-1}	2.740×10^{-3}	8.798×10^{-4}	7.820×10^{-4}	6.843×10^{-4}
n=15	9.956×10^{-1}	9.986×10^{-1}	9.988×10^{-1}	9.989×10^{-1}	3.977×10^{-3}	1.278×10^{-3}	1.136×10^{-3}	9.940×10^{-4}
n=20	9.942×10^{-1}	9.981×10^{-1}	9.983×10^{-1}	9.985×10^{-1}	5.134×10^{-3}	1.650×10^{-3}	1.467×10^{-3}	1.284×10^{-3}
n=25	9.928×10^{-1}	9.977×10^{-1}	9.979×10^{-1}	9.982×10^{-1}	6.215×10^{-3}	1.999×10^{-3}	1.777×10^{-3}	1.555×10^{-3}

Estado	$I \to I : p_{22}$	$I \to R : p_{23}$
Tiempo	S.T., $\forall\alpha$	S.T., $\forall\alpha$
n=5	9.363×10^{-1}	6.371×10^{-2}
n=10	8.766×10^{-1}	1.234×10^{-1}
n=15	8.208×10^{-1}	1.792×10^{-1}
n=20	7.685×10^{-1}	2.315×10^{-1}
n=25	7.195×10^{-1}	2.805×10^{-1}

Notas. [a] S.T.: Sin aplicación del tratamiento de control en el procesamiento de la cadena. "S": Estado susceptible, "I": Estado infección por VIH/SIDA, "R": Estado muerto del modelo, $\forall\alpha$: Para todos los valores de α.

Fuente: Elaboración propia.

Tomando en cuenta las simulaciones de n = 5, 10, 15, 20 y 25 años, de acuerdo al procedimiento prospectivo empleado en Delgado-Moya & Marrero-Severo (2017), se estima que las

probabilidades p_{22} y p_{23} no sufrirán cambios o alteraciones en el largo plazo tras la aplicación de la estrategia de control y para diferentes niveles de eficacia que esta pueda alcanzar. Es decir, las probabilidades de continuar infectado o viviendo con VIH/SIDA y de fallecer a causa de complicaciones de la infección seguirán siendo las mismas a futuro, por lo que se determina que una estrategia de contención basada en refuerzos conductuales y una dinámica sexual de menor riesgo no genera impacto alguno en las cifras de personas viviendo con VIH (PVVIH) o las probabilidades de mortalidad. Sin embargo, dicha estrategia genera cambios positivos en las probabilidades p_{11} y p_{12}, referidas a la probabilidad de continuar siendo un individuo sano y un sujeto contagiado de VIH/SIDA, lo que refleja un beneficio relativo de este tipo de intervenciones sobre la población. Como se evidencia en la comparación del escenario original sin aplicación del control (S.T.) y los valores obtenidos a través de niveles de eficacia como 0.1, 0.2 y 0.3 en la tabla; en primer lugar, a mayor eficacia provista a la estrategia, mayor será la probabilidad de continuar siendo susceptible o sano en el futuro, lo que resulta en una mejora sobre el modelo sin control; en segundo lugar, la probabilidad de pasar del estado susceptible al de infección por VIH/SIDA disminuye a mayores niveles de eficacia, lo que configura que la estrategia cumple con los objetivos de control de forma parcial. Permite un incremento de la proporción de la población sana a largo plazo y disminuye la transmisión del virus al registrarse menores casos oficiales de infección. No obstante, la estrategia por sí sola carece de los medios para fomentar un escenario ideal experimental de tratamiento de la epidemia, ya que la infección persistirá en la población prevalente de la misma manera y la tasa de mortalidad no será atenuada en ningún horizonte temporal.

4.3.6.2. Estrategia N°02: Tratamiento antirretroviral para individuos seropositivos a indetectables o "recuperados"

La carga viral asociada al VIH se refiere a la cantidad del virus en la sangre de un individuo que vive infectado o que forma parte de la población prevalente relativo a la infección (CATIE, 2018). Si el paciente adopta un régimen antirretroviral como tratamiento contra el virus de manera constante, puede reducir la carga viral a un nivel sumamente bajo para ser detectado por un análisis de sangre. Una vez que su carga viral ha caído por debajo de este nivel, se dice que es indetectable y su función inmune ha sido reestablecida.

Actualmente es muy raro encontrar una persona que se haya curado con éxito y esté libre del VIH. Esto se debe a que el virus del VIH reside en el núcleo celular y utiliza el mecanismo celular para replicarse. Puede permanecer dentro de una célula inactiva de por vida y, por lo tanto, no es fácil curar a un individuo por completo (Rotich, 2016). Sin embargo, si las células del VIH se han reducido a niveles indetectables, entonces podemos decir que el individuo está en la categoría de "recuperado" dentro del cuadro clínico del VIH/SIDA ya que no representa un riesgo de infección y no comparte el mismo perfil sintomatológico de un paciente inmunosuprimido.

La adopción de un régimen antirretroviral adecuado y continuo por parte de pacientes infectados por VIH (que no hayan desarrollado SIDA) fomentado por el aparato estatal con

acceso oportuno a servicios y un seguimiento del tratamiento como estrategia se traduce en un parámetro β dentro del procesamiento de una cadena de Markov. La eficacia de dicha estrategia determina el tamaño del parámetro, que puede asumir valores entre $0 \leq \beta \leq 1$. Cuanto mayor sea el valor de β, mejor será la prognosis de los infectados y la calidad de vida. De igual modo, la representación matricial de la influencia de dicho parámetro sobre las probabilidades de transición es determinada de la siguiente manera:

$$P_{indetectabilidad} = \begin{pmatrix} p_{11} & p_{12} & 0 \\ 0 & p_{22}(1-\beta) & p_{23} \\ 0 & 0 & 1 \end{pmatrix} \tag{4.8}$$

Donde puede evidenciarse la ponderación del parámetro β, bajo el nivel de eficacia a ser provisto o alcanzado por el sector salud en cuanto a su aplicación y efectos sobre la probabilidad p_{22} de estar viviendo infectado con el VIH/SIDA a llegar a ser indetectable y reducir el riesgo de transmisión e inmunosupresión.

Generando estimaciones de las probabilidades de transición bajo la influencia del parámetro β para diferentes horizontes de tiempo n y bajo distintos niveles de eficacia que la estrategia de control puede asumir, se obtienen los resultados de la Tabla 4.11.

Tabla 4.11: Variación de las probabilidades de transición en el tiempo (n = años) con estrategia de control del parámetro λ.

Simulación de las probabilidades por cambios en el parámetro β						
Estado	$S \to S : p_{11}$		$S \to I : p_{12}$			
Tiempo (años)	S.T.[a]	$\forall\beta$	S.T.	β=0.1	β=0.2	β=0.3
n=5	9.985×10^{-1}	9.995×10^{-1}	1.416×10^{-3}	2.492×10^{-3}	3.414×10^{-4}	4.151×10^{-4}
n=10	9.971×10^{-1}	9.990×10^{-1}	2.740×10^{-3}	3.869×10^{-3}	4.460×10^{-4}	4.802×10^{-4}
n=15	9.956×10^{-1}	9.984×10^{-1}	3.977×10^{-3}	4.629×10^{-3}	4.779×10^{-4}	4.902×10^{-4}
n=20	9.942×10^{-1}	9.979×10^{-1}	5.134×10^{-3}	5.048×10^{-3}	4.875×10^{-4}	4.916×10^{-4}
n=25	9.928×10^{-1}	9.974×10^{-1}	6.215×10^{-3}	5.278×10^{-3}	4.903×10^{-4}	4.915×10^{-4}

Estado	$I \to I : p_{22}$				$I \to R : p_{23}$			
Tiempo (años)	S.T.	β=0.1	β=0.2	β=0.3	S.T.	β=0.1	β=0.2	β=0.3
n=5	9.363×10^{-1}	5.529×10^{-1}	3.068×10^{-1}	1.574×10^{-1}	6.371×10^{-2}	5.354×10^{-2}	4.308×10^{-2}	3.565×10^{-2}
n=10	8.766×10^{-1}	3.057×10^{-1}	9.413×10^{-2}	2.476×10^{-2}	1.234×10^{-1}	6.947×10^{-2}	5.629×10^{-2}	4.126×10^{-2}
n=15	8.208×10^{-1}	1.690×10^{-1}	2.888×10^{-2}	3.897×10^{-3}	1.792×10^{-1}	8.061×10^{-2}	6.035×10^{-2}	4.214×10^{-2}
n=20	7.685×10^{-1}	9.343×10^{-2}	8.860×10^{-3}	6.132×10^{-4}	2.315×10^{-1}	8.243×10^{-2}	6.159×10^{-2}	4.228×10^{-2}
n=25	7.195×10^{-1}	5.166×10^{-2}	2.718×10^{-3}	9.600×10^{-5}	2.805×10^{-1}	8.569×10^{-2}	6.198×10^{-2}	4.230×10^{-2}

Notas. [a] S.T.: Sin aplicación del tratamiento de control en el procesamiento de la cadena. "S": Estado susceptible, "I": Estado infección por VIH/SIDA, "R": Estado muerto del modelo, $\forall\beta$: Para todos los valores del parámetro β.

Fuente: Elaboración propia.

Tomando en cuanta las simulaciones de $n = 5, 10, 15, 20$ y 25 años, se estima que la probabilidad p_{11} no sufrirá cambios o alteraciones en el largo plazo tras la aplicación de la estrategia de control y para diferentes niveles de eficacia que esta pueda alcanzar. Es decir, la probabilidad de continuar siendo un individuo sano susceptible de adquirir el VIH/SIDA en algún momento seguirá siendo las mismas a futuro, por lo que se determina que una estrategia de contención

basada en tratamiento antirretroviral para aumentar el número de infectados indetectables o "recuperados" no genera impacto alguno en el número de individuos susceptibles en el país. Sin embargo, dicha estrategia genera cambios positivos en las probabilidades p_{12}, p_{22} y p_{23}, referidas a la probabilidad de ser adquirir la infección, continuar siendo un individuo sano y un sujeto contagiado de VIH/SIDA, lo que refleja un beneficio relativo de este tipo de intervenciones sobre la población. Como se evidencia en la comparación del escenario original sin aplicación del control (S.T.) y los valores obtenidos a través de niveles de eficacia como 0.1, 0.2 y 0.3 en la tabla; en primer lugar, la probabilidad de pasar del estado susceptible al de infección por VIH/SIDA disminuye a mayores niveles de eficacia; en segundo lugar, la estimación de que los individuos viviendo con el VIH/SIDA continúen infectados en el futuro disminuye de forma progresiva; y, por último, la probabilidad de muerte en los pacientes infectados decrece a mayores niveles de eficacia y horizontes temporales. Es así como, la estrategia por sí sola habilita un escenario ideal experimental de tratamiento de la epidemia, ya que el virus no podrá ser transmitido, la población prevalente con VIH/SIDA no representará un riesgo para sí mismos ni para los individuos susceptibles y los pacientes infectados tendrán una esperanza de vida mayor ante las tasas bajas de mortalidad.

4.3.6.3. Estrategia N°03: Control con tratamiento para reducir el riesgo de fallecimiento y aumentar la esperanza de vida

Los fallecimientos entre personas seropositivas han disminuido desde la introducción de la terapia antirretroviral de gran actividad (TARGA) en 1996. El despliegue generalizado de la terapia antirretroviral combinada ha alterado notablemente la historia natural de la infección por VIH (Parashar *et al.*, 2016). La tolerabilidad, la seguridad y la eficacia mejoradas de las terapias, además de la ampliación de las estrategias para promover el acceso y la adherencia al TARGA, han provocado una disminución drástica de la mortalidad por todas las causas de las personas que viven con el VIH (PVVIH) en una variedad de entornos y entre poblaciones diversas. Dados los altos niveles de adherencia, el TARGA suprime el VIH, restaurando la función inmunológica y previniendo la progresión a la muerte.

Una estrategia planteada en la masificación y desarrollo del servicio de terapia antirretroviral en individuos infectados con una alta carga viral o en el estado SIDA se explica en términos de un parámetro λ dentro del procesamiento de una cadena de Markov. La eficacia de dicha estrategia determina el tamaño del parámetro, que puede asumir valores entre $0 \leq \lambda \leq 1$. Cuanto mayor sea el valor de λ, mejor será la proyección de esperanza de vida de los pacientes enfermos y una reducción del riesgo de fallecimiento por complicaciones derivadas de la infección. De igual modo, la representación matricial de la influencia de dicho parámetro sobre las probabilidades de transición es determinada de la siguiente manera:

$$P_{mortalidad} = \begin{pmatrix} p_{11} & p_{12} & 0 \\ 0 & p_{22} & p_{23}(1-\lambda) \\ 0 & 0 & 1 \end{pmatrix} \tag{4.9}$$

Donde puede evidenciarse la ponderación del parámetro λ, bajo el nivel de eficacia a ser provisto o alcanzado por el sector salud en cuanto a su aplicación y efectos sobre la probabilidad p_{23} de reducir la mortalidad de los pacientes infectados con VIH/SIDA y aumentar su esperanza de vida en el largo plazo.

Generando estimaciones de las probabilidades de transición bajo la influencia del parámetro λ para diferentes horizontes de tiempo n y bajo distintos niveles de eficacia que la estrategia de control puede asumir, se obtienen los resultados de la Tabla 4.12.

Tabla 4.12: Variación de las probabilidades de transición en el tiempo (n = años) con estrategia de control del parámetro λ.

Simulación de las probabilidades por cambios en el parámetro λ

Estado	$S \to S : p_{11}$		$S \to I : p_{12}$			
Tiempo (años)	S.T.[a]	$\forall\lambda$	S.T.	λ=0.1	λ=0.2	λ=0.3
n=5	9.985×10^{-1}	9.995×10^{-1}	1.416×10^{-3}	5.063×10^{-4}	5.076×10^{-4}	5.089×10^{-4}
n=10	9.971×10^{-1}	9.990×10^{-1}	2.740×10^{-3}	9.832×10^{-4}	9.889×10^{-4}	9.947×10^{-4}
n=15	9.956×10^{-1}	9.984×10^{-1}	3.977×10^{-3}	1.432×10^{-3}	1.445×10^{-3}	1.458×10^{-3}
n=20	9.942×10^{-1}	9.979×10^{-1}	5.134×10^{-3}	1.856×10^{-3}	1.878×10^{-3}	1.901×10^{-3}
n=25	9.928×10^{-1}	9.974×10^{-1}	6.215×10^{-3}	2.254×10^{-3}	2.288×10^{-3}	2.323×10^{-3}

Estado	$I \to I : p_{22}$				$I \to R : p_{23}$			
Tiempo (años)	S.T.	λ=0.1	λ=0.2	λ=0.3	S.T.	λ=0.1	λ=0.2	λ=0.3
n=5	9.363×10^{-1}	9.425×10^{-1}	9.488×10^{-1}	9.551×10^{-1}	6.371×10^{-2}	5.749×10^{-2}	5.123×10^{-2}	4.495×10^{-2}
n=10	8.766×10^{-1}	8.883×10^{-1}	9.002×10^{-1}	9.121×10^{-1}	1.234×10^{-1}	1.117×10^{-1}	9.984×10^{-2}	8.787×10^{-2}
n=15	8.208×10^{-1}	8.373×10^{-1}	8.540×10^{-1}	8.711×10^{-1}	1.792×10^{-1}	1.627×10^{-1}	1.460×10^{-1}	1.289×10^{-1}
n=20	7.685×10^{-1}	7.891×10^{-1}	8.103×10^{-1}	8.320×10^{-1}	2.315×10^{-1}	2.109×10^{-1}	1.897×10^{-1}	1.680×10^{-1}
n=25	7.195×10^{-1}	7.438×10^{-1}	7.688×10^{-1}	7.945×10^{-1}	2.805×10^{-1}	2.562×10^{-1}	2.312×10^{-1}	2.054×10^{-1}

Notas. [a] S.T.: Sin aplicación del tratamiento de control en el procesamiento de la cadena. "S": Estado susceptible, "I": Estado infección por VIH/SIDA, "R": Estado muerto del modelo, $\forall\lambda$: Para todos los valores del parámetro λ.

Fuente: Elaboración propia.

Tomando en cuanta las simulaciones de $n = 5, 10, 15, 20$ y 25 años, se estima que la probabilidad p_{11} no sufrirá cambios o alteraciones en el largo plazo tras la aplicación de la estrategia de control y para diferentes niveles de eficacia que esta pueda alcanzar. Es decir, la probabilidad de continuar siendo un individuo sano susceptible de adquirir el VIH/SIDA en algún momento seguirá siendo las mismas a futuro, por lo que se determina que una estrategia de contención basada en el tratamiento antirretroviral para reducir la mortalidad en pacientes no genera impacto alguno en el número de individuos susceptibles en el país. Sin embargo, dicha estrategia genera cambios positivos en las probabilidades p_{22} y p_{23}, referidas a continuar siendo un individuo viviendo con VIH (PVVIH) y la probabilidad de fallecer a causa de la infección, lo que refleja un beneficio relativo de este tipo de intervenciones sobre la población; empero, la estrategia genera cambios negativos en la probabilidad p_{12} asociada a la probabilidad de adquirir el virus y ser infectado, mostrando una disyuntiva entre el aumento de casos nuevos a registrar y la disminución de la mortalidad y una mayor proporción de individuos con VIH que pueden vivir más años. Como se evidencia en la comparación del escenario original sin aplicación del control (S.T.) y los valores obtenidos a través de niveles de eficacia como 0.1, 0.2 y 0.3 en la tabla; en primer lugar, a mayor eficacia provista a la estrategia, mayor será

la probabilidad de continuar siendo un individuo infectado con VIH/SIDA o parte de la población prevalente que vive con la infección en el futuro, lo que resulta en una mejora sobre el modelo sin control ya que son más los individuos que pueden alargar su vida; en segundo lugar, la probabilidad de pasar del estado infectado al estado de muerte (según el modelo S-I-R) disminuye de forma considerable a mayores niveles de eficacia, lo que configura que la estrategia cumple con los objetivos de control de la propensión a morir por la enfermedad, pero con un efecto perjudicial sobre la transmisión del virus a personas susceptibles. A mayor eficacia provista a la estrategia, se dará un aumento de la probabilidad de infectarse estando sano conforme pasa el tiempo, lo que conlleva a un número más elevado de casos reportados oficiales a las entidades de salud respectiva en contraste al modelo sin control, que presentan menores tasas de contagio entre la población. Esta tendencia de mayor infección se basa esencialmente en que al reducir las tasas de mortalidad, existen más personas que podrán continuar viviendo infectadas, generando un efecto indirecto sobre las personas sanas que tienen que interactuar en el futuro con un número cada vez más grande de infectados, lo que provoca que aumente el número de personas que contraigan el virus eventualmente. Por ello, la estrategia por sí sola carece de los medios para fomentar un escenario ideal experimental de tratamiento de la epidemia, ya que la infección seguirá propagándose, pese a que los infectados no fallecerán a un ritmo acelerado.

4.3.6.4. Estrategia N°04: Control combinado de tratamientos y políticas de salud

Si bien las medidas de control y contención de la epidemia analizadas en el estudio fueron previamente descritas y sus efectos sobre las probabilidades de transición de la cadena determinados de forma prospectiva, este proceso fue llevado a cabo de manera individual, evaluando cada estrategia de forma aislada y en un contexto controlado. Sin embargo, en el campo de la epidemiología y en el diseño del aparato de salud gubernamental, dichas estrategias se plantean y ejecutan paralelamente, interactuando unas con otras y generando impactos de manera conjunta (ya sean sinérgicos o disímiles) sobre el curso de la infección y el estado de la población.

Es así como, bajo una estrategia de control combinado de tratamientos y políticas de salud (como las definidas en las estrategias N° 01, 02 y 03), la representación matricial de la influencia de los parámetros de intervención sobre las probabilidades de transición es determinada de la siguiente manera:

$$P_{combinado} = \begin{pmatrix} p_{11} & p_{12}(1-\alpha) & 0 \\ 0 & p_{22}(1-\beta) & p_{23}(1-\lambda) \\ 0 & 0 & 1 \end{pmatrix} \tag{4.10}$$

Donde puede evidenciarse la ponderación de los parámetros α, β y λ, bajo el nivel de eficacia a ser provisto o alcanzado por el sector salud en cuanto a su aplicación y efectos sobre las probabilidades p_{11}, p_{12}, p_{22} y p_{23} en el modelo (siendo p_{11} afectada de forma indirecta).

Generando estimaciones de las probabilidades de transición bajo la influencia de los parámetros α, β y λ para diferentes horizontes de tiempo n y bajo distintos niveles de eficacia que la estrategia de control puede asumir, se obtienen los resultados de la Tabla 4.13.

Tabla 4.13: Variación de las probabilidades de transición en el tiempo (n = años) con control combinado de las estrategias descritas anteriormente.

Simulación de las probabilidades por cambios en los parámetros α, β y λ								
Estado	$S \to S : p_{11}$				$S \to I : p_{12}$			
Tiempo (años)	S.T.[a]	α, β, λ=0.1	α, β, λ=0.2	α, β, λ=0.3	S.T.	α, β, λ=0.1	α, β, λ=0.2	α, β, λ=0.3
n=5	9.985×10^{-1}	9.995×10^{-1}	9.996×10^{-1}	9.996×10^{-1}	1.416×10^{-3}	3.741×10^{-4}	2.743×10^{-4}	1.988×10^{-4}
n=10	9.971×10^{-1}	9.991×10^{-1}	9.992×10^{-1}	9.993×10^{-1}	2.740×10^{-3}	5.821×10^{-4}	3.594×10^{-4}	2.307×10^{-4}
n=15	9.956×10^{-1}	9.986×10^{-1}	9.988×10^{-1}	9.989×10^{-1}	3.977×10^{-3}	6.978×10^{-4}	3.858×10^{-4}	2.357×10^{-4}
n=20	9.942×10^{-1}	9.981×10^{-1}	9.983×10^{-1}	9.985×10^{-1}	5.134×10^{-3}	7.619×10^{-4}	3.939×10^{-4}	2.364×10^{-4}
n=25	9.928×10^{-1}	9.977×10^{-1}	9.979×10^{-1}	9.982×10^{-1}	6.215×10^{-3}	7.975×10^{-4}	3.963×10^{-4}	2.365×10^{-4}
Estado	$I \to I : p_{22}$				$I \to R : p_{23}$			
Tiempo (años)	S.T.[a]	α, β, λ=0.1	α, β, λ=0.2	α, β, λ=0.3	S.T.	α, β, λ=0.1	α, β, λ=0.2	α, β, λ=0.3
n=5	9.363×10^{-1}	5.565×10^{-1}	3.109×10^{-1}	1.605×10^{-1}	6.371×10^{-2}	4.720×10^{-2}	3.460×10^{-2}	2.508×10^{-2}
n=10	8.766×10^{-1}	3.097×10^{-1}	9.665×10^{-2}	2.577×10^{-2}	1.234×10^{-1}	7.347×10^{-2}	4.536×10^{-2}	2.911×10^{-2}
n=15	8.208×10^{-1}	1.724×10^{-1}	3.005×10^{-2}	4.136×10^{-3}	1.792×10^{-1}	8.809×10^{-2}	4.871×10^{-2}	2.976×10^{-2}
n=20	7.685×10^{-1}	9.594×10^{-2}	9.342×10^{-3}	6.640×10^{-4}	2.315×10^{-1}	9.622×10^{-2}	4.975×10^{-2}	2.986×10^{-2}
n=25	7.195×10^{-1}	5.340×10^{-2}	2.904×10^{-3}	1.070×10^{-4}	2.805×10^{-1}	1.008×10^{-1}	5.007×10^{-2}	2.988×10^{-2}

Notas. [a] S.T.: Sin aplicación del tratamiento de control en el procesamiento de la cadena. "S": Estado susceptible, "I": Estado infección por VIH/SIDA, "R": Estado muerto del modelo.

Fuente: Elaboración propia.

Tomando en cuenta las simulaciones de n = 5, 10, 15, 20 y 25 años, se determina que una estrategia de control combinado resulta la forma de intervención más eficaz entre todas las anteriormente planteadas, debido a que: se logra un aumento de la población sana susceptible a ser contagiada como efecto indirecto del despliegue de todos los esfuerzos del sector (la probabilidad de individuos que asumen conductas que evitan que adquieran el virus de forma deliberada aumenta con cada horizonte de tiempo dado); se reducen las estimaciones de nuevos casos a registrar por parte de las autoridades sanitarias, dado que el riesgo de transmisión por actividades sexuales que comprometen el bienestar de las personas ha disminuido de forma significativa al asumir esfuerzos de concientización y protección sexual; se presenta una disminución relevante de la población prevalente o que vive infectada por el VIH/SIDA dado que dichos pacientes pasan a tener una carga viral indetectable, por lo que puede precisarse que no representan un riesgo de transmisión y su estado de salud se aproxima a la de una persona sana y no inmunosuprimida (cabe tener en cuenta que dentro del procesamiento del modelo, las personas con carga viral indetectable se consideran fuera del estado "VIH/SIDA", pese a ello no son consideradas personas recuperadas o libres del virus y/o la enfermedad); y, por último, se reducen las probabilidades de transición de fallecer por complicaciones del VIH/SIDA estando infectados mediante la estrategia combinada que les permite tener una esperanza de vida mayor y un cuadro clínico no comprometido.

Los resultados de las simulaciones experimentales permiten dilucidar que la estrategia de control combinado permite la desaceleración de la epidemia en el territorio peruano; logrando reducir todos los indicadores epidemiológicos de reporte de casos y el aumento de la población sana a nivel nacional en un contexto prospectivo.

Capítulo 5

Determinantes del conocimiento del VIH/SIDA

Este capítulo muestra, mediante el sub-análisis complejo de datos de encuesta a través de la aplicación de una regresión logística cuasi-binomial, la identificación de los factores que se encuentran relacionados o que poseen una asociación significativa con el conocimiento adecuado o inadecuado sobre las formas de prevención y rechazo de ideas erróneas sobre la transmisión del VIH/SIDA en la población entrevistada en el Perú dentro del rango etario de 15 a 29 años de edad entre los meses de enero y diciembre del año 2019.

Posteriormente, se desarrollan modelos de clasificación y predicción como la regresión logística multivariada (R.L.M.), redes neuronales artificiales (R.N.A.), el algoritmo k-Nearest-Neighbors (k-NN), el modelo Random Forest (R.F.) y el modelo basado en árboles de decisión (D.T.). Seguidamente, se realizará una comparación entre los mismos a fin de precisar, bajo diversos criterios de bondad de ajuste, qué modelo computacional es el más idóneo para la predicción de la probabilidad de conocimiento y formas de transmisión del VIH/SIDA en la población descrita anteriormente.

En lo que se refiere a los datos empleados en esta etapa, los registros correspondientes a la población objetivo en este estudio forman parte de la unidad de análisis de la "Encuesta Demográfica y de Salud Familiar 2019" (INEI, 2019a). A su vez, el modelo de regresión logística cuasi-binomial se realizó en el software estadístico R (R Core Team, 2020) mediante el paquete estadístico `survey` (Lumley, 2004) y los modelos de clasificación y predicción del nivel de conocimiento de los individuos se realizaron mediante el programa informático para el análisis y minería de datos RapidMiner (Mierswa & Klinkenberg, 2018).

5.1. Bases de datos

Para la aplicación de los modelos propuestos, consideraremos el conjunto secundario de datos basado en la "Encuesta Demográfica y de Salud Familiar 2019" (INEI, 2019a). A fin de analizar el panorama socio-demográfico, económico y familiar más actualizado que describa y caracterice a la población adolescente y joven adulta del Perú, es necesario recabar e investigar los datos más recientes hasta el momento de desarrollar el presente estudio para evaluar las últimas tendencias y evoluciones de estos factores en la población objetivo; es por ello, que la encuesta ENDES del año 2019 se configura como la fuente principal de datos para el

desarrollo de los métodos y técnicas propuestas en este capítulo.

5.1.1. Encuesta Demográfica y de Salud Familiar (ENDES)

La "Encuesta Demográfica y de Salud Familiar ENDES" del año 2019 es una investigación estadística por muestreo que se realiza con la asistencia técnica del Programa Mundial de Encuestas de Demografía y Salud, para obtener información actualizada y efectuar análisis del cambio, tendencias y determinantes de la fecundidad, mortalidad, así como una serie de indicadores de salud materna e infantil y recientemente indicadores de enfermedades no trasmisibles y trasmisibles en el Perú (INEI, 2019a). Dicha encuesta tiene como objetivo principal dotar al país de elementos y aspectos confiables sobre la dinámica demográfica, así como brindar referencias sobre el estado y factores asociados a las enfermedades no trasmisibles y trasmisibles y para la evaluación y formulación de los programas de población y salud familiar en el país.

De manera específica, la base de datos está constituida, en particular para la presente investigación, por una población objetivo de hombres y mujeres de 15 a 29 años que son residentes habituales de viviendas particulares de áreas urbanas y rurales del país que hayan pernoctado la noche anterior, a la encuesta, en la vivienda seleccionada; ofreciendo estimaciones estadísticamente confiables para un nivel de inferencia nacional en el Perú (INEI, 2019a).

En cuanto a las especificaciones estadísticas del muestreo de los datos secundarios, el marco muestral, para la selección de la muestra, lo constituye la información estadística y cartográfica proveniente de los Censos Nacionales XI de Población y VI de Vivienda del año 2007 y la Actualización del Sistema de Focalización de Hogares (SISFOH) 2012-2013. Asimismo, el diseño de la encuesta es uno de muestra compleja, que se caracteriza por ser bietápica, probabilística de tipo equilibrado, estratificada e independiente, a nivel departamental y por las unidades de muestreo de área urbana y rural. El tamaño de la muestra diseñada para dar estimaciones representativas de la ENDES 2019 (anual) es de 36,760 viviendas, correspondiendo 14,760 viviendas al área sede (capitales de departamento y los 43 distritos que conforman la Provincia de Lima), 9,340 viviendas al resto urbano y 12,660 viviendas al área rural.

5.1.2. Preparación de la población y variables de análisis

Los indicadores de estudio esenciales figuran o se encuentran presentes principalmente en la base de datos (BD) de estado y condiciones de salud CSALUD01 de la ENDES. La base de datos en cuestión alberga los datos de los individuos de 15 años a más sobre diversas afecciones, factores de riesgo, percepción y conocimiento de enfermedades no transmisibles y transmisibles, salud mental, entre otras. Además de dicha base de datos, son necesarias bases de datos adicionales que configuran la ENDES como lo son la de Hogar, Personas y Mujeres para unificar, completar y filtrar la base de datos de salud a fin de poder analizar de manera correcta y directa a la encuesta (Hernández-Vásquez & Chacón-Torrico, 2019).

En cuanto al proceso de unificación, existen identificadores o llaves en cada una de las bases de daos que representan o permiten identificar estructuras relacionales y jerárquicas entre las mismas. A través de dichos identificadores, es posible relacionar a los hogares entre sí o a los individuos con sus respectivos hogares. Para la conformación de la llave de los individuos, se empleará el CASEID que se forma de la concatenación del identificador único del hogar y el número de orden del individuo en el hogar (QSNUMERO, HVIDX o HA0 según la base de datos). Es importante conocer estas precisiones para entender los pasos que se toman durante la unión de las bases, teniendo en cuenta que en algunas bases de datos sólo se dispone del HHID y el número de orden del individuo en el hogar (Hernández-Vásquez & Chacón-Torrico, 2019).

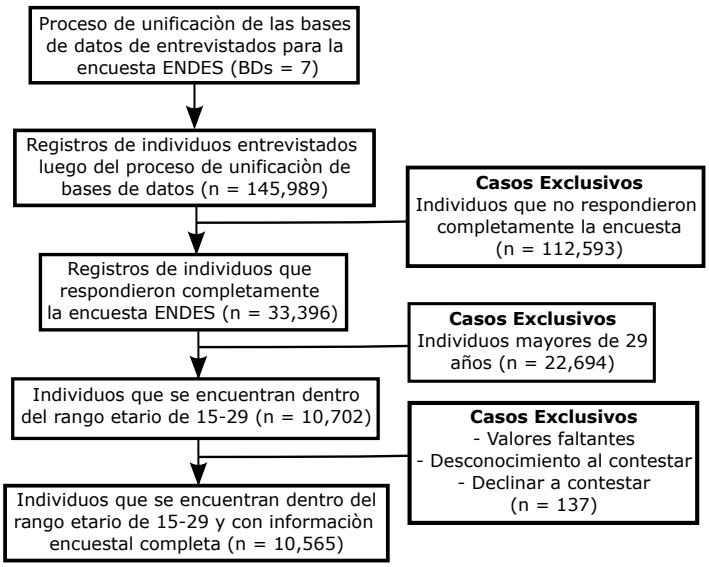


Figura 5.1: Diagrama de flujo del proceso de selección de los individuos entrevistados en la encuesta ENDES 2019 para la presente investigación. Fuente: Elaboración propia.

Considerando la preparación de las variables de análisis en el estudio sigue el proceso definido en la Figura 5.1. Posteriormente al procedimiento de unificación de las bases de datos independientes en un conjunto de datos consolidado y final de la ENDES, se efectúa la etapa de filtrado que consiste en 3 circunstancias de exclusión de datos: en primer lugar, se efectúa una exclusión de aquellos individuos que no respondieron completamente la encuesta de salud, por lo que las cifras de registros de individuos en hogares peruanos entrevistados una vez que se tiene la base de datos consolidada (145,989 individuos) disminuye al valor de 33,396 personas entrevistadas cuya información de salud estuvo debidamente documentada y registrada; en segundo lugar, se realiza la exclusión de aquellos individuos mayores a 29 años ya que el alcance de la investigación contempla aquellos habitantes en territorio nacional que sean adolescentes y jóvenes adultos, por lo que el número de individuos considerados para el estudio decrece a una cifra de 10,702 y, finalmente, es necesario desestimar aquellos casos en donde no se ha obtenido alguna respuesta o el encuestado ha declinado en contestar alguna pregunta determinada dentro del cuestionario de evaluación (habiéndose determinado que la cifra de no contestación o registros faltantes asciende a 137 personas); en razón de la cual,

la cifra final de individuos que cumplen con las características de la población objetivo y que cuentan con información completa alcanza una estimado de 10,565 residentes en el país que se consideran para la presente pesquisa.

Asimismo, se debe tomar en cuenta las especificaciones de diseño muestral dentro del proceso de análisis (como el conglomerado, el estrato y el factor de ponderación) para obtener una adecuada estimación de los indicadores (Hernández-Vásquez & Chacón-Torrico, 2019). La síntesis del proceso de identificación y obtención de la base de datos consolidada de indicadores demográficos y de salud de la ENDES se presenta en la Tabla 5.1.

Las variables de interés para el presente análisis fueron: género de los encuestados (HV104), nivel económico o distribución de riqueza (HV270), región natural habitada actualmente (SHREGION), área de residencia (HV025), rango etario de los encuestados (QS23), nivel educativo más alto alcanzado (HV106), lengua materna o primaria (QS25AA), auto-percepción étnica de los encuestados (Q25BB), si la encuestada ha oído acerca del VIH con anterioridad (QS29A), si la encuestada ha oído acerca del SIDA con anterioridad (QS29B), si el encuestado ha oído acerca del VIH con anterioridad (QS601A), si el encuestado ha oído acerca del SIDA con anterioridad (QS601B), realización de la prueba de descarte del VIH/SIDA (QS603), percepción de menor riesgo de contraer el VIH/SIDA si se tiene una pareja no infectada (QS606), percepción de una persona infectada con VIH/SIDA puede tener una apariencia saludable (QS607), percepción de si se puede adquirir el VIH/SIDA a través de contacto físico (QS608), percepción de si se puede adquirir el VIH/SIDA a través del uso y compartir de utensilios y alimentos (QS610), percepción de menor riesgo de contraer el VIH/SIDA con el uso de preservativos (QS611), estado civil de los encuestados (HV115), si cuentan con una radio en el hogar (HV207), si cuentan con un televisor en el hogar (HV208), si cuentan con internet en el hogar (SH61Q), género del jefe de hogar o familia (HV219) y, por último, la nacionalidad de los encuestados (QH25A).

Tabla 5.1: Variables incluidas en las bases de datos de la ENDES seleccionadas para procesamiento del cuestionario de salud

Utilización/Variable	Contenido	Base de datos (código de la base)
Variables de unión		
HHID	Identificador del hogar	Hogar (RECH0), Personas (RECH1), Vivienda (RECH23), Cuestionario Salud (CSALUD01), Mujer - Antropometria (RECH5)
HVIDX	Número de individuo por hogar	Personas (RECH1)
QSNUMERO	Número de individuos por hogar	Cuestionario Salud (CSALUD01)
HA0	Número de individuos por hogar	Mujer - Antropometria (RECH5)
CASE ID	Identificador de persona	Mujer - Salud y Lactancia (REC42), Mujer - Obstétrica (RE223123)
Variables de filtro		
QSRESINF	Resultado informante	Cuestionario Salud (CSALUD01)
QS23	Edad de los entrevistados	Cuestionario Salud (CSALUD01)
Variables de diseño		
HV001	Número del conglomerado	Hogar (RECH0)
HV022	Estrato	Hogar (RECH0)
PESO15_AMAS[a]	Factor de ponderación	Hogar (RECH0)
Variables de análisis		
HV104	Género	Personas (RECH1)
HV270	Nivel económico	Vivienda (RECH23)

SHREGION	Región Natural	Vivienda (RECH23)
HV025	Área de residencia	Hogar (RECH0)
QS23	Edad	Cuestionario Salud (CSALUD01)
HV106	Nivel educativo	Personas (RECH1)
QS25AA	Lengua materna	Cuestionario Salud (CSALUD01)
QS25BB	Etnicidad	Cuestionario Salud (CSALUD01)
QS29A	Oído acerca del VIH[b]	Cuestionario Salud (CSALUD01)
QS29B	Oído acerca del SIDA[c]	Cuestionario Salud (CSALUD01)
QS601A	Oído acerca del VIH[d]	Cuestionario Salud (CSALUD01)
QS601B	Oído acerca del SIDA[e]	Cuestionario Salud (CSALUD01)
QS603	Prueba de VIH/SIDA	Cuestionario Salud (CSALUD01)
QS606	Riesgo con pareja no infectada	Cuestionario Salud (CSALUD01)
QS607	Persona saludable infectada	Cuestionario Salud (CSALUD01)
QS608	Infección por contacto físico	Cuestionario Salud (CSALUD01)
QS610	Infección a través de utensilios	Cuestionario Salud (CSALUD01)
QS611	Riesgo bajo uso de preservativos	Cuestionario Salud (CSALUD01)
HV115	Estado civil	Personas (RECH1)
HV207	Radio en el hogar	Vivienda (RECH23)
HV208	Televisión en el hogar	Vivienda (RECH23)
SH61Q	Internet en el hogar	Vivienda (RECH23)
HV219	Género del jefe de familia	Vivienda (RECH23)
QH25A	Nacionalidad	Personas (RECH1)

Notas. [a] Para incorporarla en el análisis, se debe dividir entre 1,000,000. [b] Variable para la población femenina. [c] Variable para la población femenina. [d] Variable para la población masculina. [e] Variable para la población masculina.

Fuente: Hernández & Chacón (2019)

5.2. Metodología

El propósito de esta sección es presentar la metodología de investigación necesaria para desarrollar y justificar el estudio de la influencia de los determinantes estructurales de la salud sobre el nivel de conocimiento de VIH/SIDA de la población adolescente y joven adulta del país y la capacidad de predicción de diferentes modelos paramétricos y no paramétricos sobre este nivel de entendimiento de los ciudadanos en el Perú del presente capítulo. Se precisará la aplicabilidad de la teoría o los aspectos vinculados a la literatura para explicar por qué se están utilizando ciertos métodos/técnicas, procedimientos y criterios para el análisis de los datos y el fundamento académico de las elecciones dadas en cada subsección de los resultados del estudio propuesto.

5.2.1. Preparación de datos de factores asociados al conocimiento del VIH/SIDA

A fin de establecer y caracterizar a aquellas variables que cumplirán el rol de dependientes (o de respuesta) e independientes (o regresoras) en el sub-análisis, es necesario asumir una fase de preparación de datos en la que se llevará a cabo la creación de 3 nuevas variables (adicionales al conjunto de variables identificadas previamente en la Sección 5.1.2) que proveerán de mayor sentido a la formulación de los modelos estadísticos a emplear.

En primer lugar, la variable esencial del presente análisis denominada como el "nivel de conocimiento acerca del VIH/SIDA" que un individuo encuestado posea, no se encuentra de forma explícita o directa dentro de la base de datos que se consolidó, sino que se obtiene mediante la

evaluación de las respuestas dadas por los entrevistados a una serie de variables intermedias que permiten determinar la percepción y las ideas acerca de la transmisión y prevención del VIH/SIDA que los individuos puedan tener en torno a dichos temas.

En ese sentido, el INEI define a dicha variable como el conocimiento o entendimiento de las formas de transmisión que evitan que el sujeto incurra en comportamientos de riesgo y conductas discriminatorias hacia otra persona (INEI, 2019b). La encuesta recogió las actitudes de un individuo frente al VIH/SIDA que pueden ser adecuadas o no, de acuerdo al nivel de conocimiento traducido en concepciones y entendimiento que tenga sobre las conductas de riesgo. Es así como este nivel de conocimiento se encuentra basado, por un lado, en la identificación correcta de las formas de prevenir la transmisión sexual del VIH: la persona entrevistada afirma que el riesgo disminuye teniendo una sola pareja sexual fiel y no infectada, usando preservativo en cada relación sexual y reconociendo que un portador del VIH puede aparentar estar saludable. Por otro lado, cuando el individuo rechaza ideas erróneas con respecto a la transmisión del VIH, como que la transmisión del VIH es a través del contacto físico o por vía oral al usar y compartir utensilios alimentarios (INEI, 2019b).

Dicha variable se construye valorando las siguientes interrogantes en la encuesta ENDES: ¿Menor riesgo de adquirir VIH/SIDA con una pareja no infectada? (QS606), ¿Menor riesgo de adquirir VIH/SIDA con el uso de preservativos? (QS611), ¿Una persona de apariencia saludable puede estar infectada con VIH/SIDA? (QS607), ¿Riesgo de adquirir VIH/SIDA mediante contacto físico? (QS608) y ¿Riesgo de adquirir VIH/SIDA mediante el uso y compartir de utensilios? (QS610). Un individuo tiene conocimiento adecuado acerca de las formas correctas y erróneas para prevenir la transmisión sexual del VIH/SIDA si es que contestó con "SÍ" a las tres primeras interrogaciones y con "NO" a las últimas dos. Caso contrario, al responder a cualquier de las cinco interrogantes de forma incorrecta, se considera que la persona tiene un conocimiento inadecuado.

En segundo lugar, como se puede deslindar de las variables recogidas de forma preliminar en la BD de la ENDES, las variables "Oído acerca del VIH" y "Oído acerca del SIDA" se manifiestan de forma independiente y dividas por género del encuestado (las codificaciones QS29A-QS29B en el caso de las mujeres y QS601A-QS601B en el caso de los hombres).

Sin embargo, resulta ineficiente e incoherente, desde una perspectiva de salud pública, conceptuar de forma separada a dichas variables para el análisis estadístico. Dicha separación de ambos conceptos propios de la infección dentro de la encuesta nace por la forma en que la población concibe a la epidemia desde el punto de vista informativo o de concientización pública. Dentro del contexto nacional, las ligeras diferencias históricas distinguibles entre las tendencias del nivel de información referido a la existencia (haber escuchado) del VIH y el SIDA que los ciudadanos del país reportan son resultados esperables en la medida que los medios masivos privados y públicos, desde hace décadas, han informado de forma indiscriminada más sobre el segundo término y también porque en el lenguaje común el término SIDA es el que se usa más para definir y caracterizar el estado de salud o cuadro clínico de una persona infectada con VIH/SIDA en el país (INEI, 2008). Empero, desde la perspectiva

sanitaria estatal, es necesario reconocer que ambos términos no son iguales para asegurar el correcto entendimiento de la epidemia por parte de los ciudadanos. Considerar a ambos conceptos de la infección en una sola pregunta se traduce en la posibilidad de medición tangible del nivel de información concreto con el que cuenta el individuo a fin de que pueda diferenciar al VIH del SIDA y las implicancias que ambos fenómenos acarrean para la seguridad sexual individual y comunitaria (ONUSIDA, 2019). Tanto el virus como la enfermedad se encuentran íntimamente relacionados, pero no poseen la misma definición: el hecho de que la población haya escuchado que el VIH es un tipo de virus que provoca de forma progresiva una inmunodeficiencia total en el cuerpo y que el SIDA es una enfermedad provocada por este virus luego de haber estado contagiados por un período prolongado de tiempo (aprox. 10 a 15 años) y sin haberse sometido a algún tipo de tratamiento antirretroviral que puede ser mortal (ONUSIDA, 2019) permitirá, de forma subsiguiente, determinar la influencia de este conocimiento elemental o esencial de la epidemia en su totalidad sobre la percepción y entendimiento correcto que los individuos tengan sobre formas de prevención y transmisión del VIH/SIDA. Por ello se procede a unificarlas en una sola variable conocida como "Oído acerca del VIH/SIDA".

Finalmente, resulta sugerente en el estudio conocer la influencia que el acceso a medios multimedia de información pueda tener sobre la percepción e ideas acerca del VIH/SIDA que los entrevistados denoten. Al igual que en el caso de la variable "Conocimiento del VIH/SIDA", el acceso a medios multimedia no se expresa de forma explícita a manera de pregunta dentro del cuestionario, sino que se origina del relacionamiento de tres variables contempladas en la base de datos consolidada de la ENDES.

Teniendo en cuenta dicha premisa, la variable "Acceso a medios" se compone de las siguientes variables intermedias: ¿Cuenta con radio en su hogar? (HV207), ¿Cuenta con televisión su hogar? (HV208) y ¿Cuenta con Internet en su hogar? (SH61Q). Se considera que un habitante en territorio nacional cuenta con acceso a medios de comunicación, si es que por lo menos respondió con "SÍ" a una de las tres preguntas.

5.2.2. Análisis estadístico de los factores asociados al conocimiento sobre el VIH/SIDA

Se realizará un análisis univariado y bivariado para obtener una perspectiva y conocimiento sobre los datos. Una vez que se exploren las variables individualmente mediante el análisis univariado, se podrá aprender acerca de la distribución estadística de los factores y analizar los patrones existentes en los mismos (luego de definirlos y sintetizarlos) (Koslowsky, 1979). Para el caso de este análisis univariado, se emplearán las tablas de distribución de frecuencias para identificar la cantidad y porcentajes de individuos en una muestra que cumplen con ciertas categorías de los regresores estudiados. Mediante el análisis bivariado, se podrá aprender acerca de las relaciones entre dos o más variables con el fin de definir las asociaciones significativas o relevantes empíricamente existentes en el conjunto de datos (Koslowsky, 1979). Para el caso del análisis bivariado, se evaluarán las asociaciones entre variables a través del

estadístico Chi-cuadrado; en el caso de que la tabla de contigencia, que compara dos variables categóricas, presente un valor menor a 5 en una de sus celdas, se empleará el estadístico de Fisher para determinar la relación entre los factores en cuestión.

Es así como estos análisis ayudanrán a aprender de qué se tratan los datos y las relaciones subyacentes a estos, creando de esa forma nuevas características o eliminando algunas de los factores existentes que pueden no ser útiles para la investigación. Por lo tanto, el análisis univariado y bivariado serán los pasos necesarios para extraer información importante sobre el conjunto de datos a emplear (Koslowsky, 1979).

5.2.3. Asociación entre los determinantes de la salud y el conocimiento sobre el VIH/SIDA en el Perú

Con el fin de determinar cómo las variables socio-demográficas, económicas y de salud afectan al nivel de conocimiento de los adolescentes y jóvenes adultos en el Perú, se ajustará un modelo de regresión cuasi-binomial multivariante del tipo logit al conjunto de datos extraído de la encuesta ENDES con la probabilidad de tener un nivel de conocimiento adecuado sobre el VIH/SIDA como variable dependiente y el género del encuestado, área de residencia, nivel económico, región natural, rango etario, nivel educativo más alto alcanzado, la auto-percepción étnica de los entrevistados, si han oído acerca del VIH/SIDA con anterioridad, realización de la prueba de descarte del VIH/SIDA, acceso a medios multimedia, género del jefe de hogar y la lengua materna de los individuos como variables independientes o covariables dentro del modelo que representarán los atributos de entrada con los que se desarrollará el modelo.

Las hipótesis para poder aplicar el modelo mencionado anteriormente son sintetizadas de la siguiente manera:

En primer lugar, una regresión logística convencional investiga la relación entre variables de respuesta categórica y un conjunto de variables explicativas. Sin embargo, este enfoque no es válido si los datos provienen de otros diseños de muestra, como: diseños de encuestas complejas con estratificación, agrupamiento y/o ponderación desigual (Anthony, 2002). En estos casos, como en el que ocurre en la encuesta ENDES, se deben aplicar técnicas especializadas para reducir el riesgo de una visión distorsionada de la población encuestada y realizar inferencias estadísticamente válidas para la misma, que incorporen el diseño complejo de la muestra en el análisis de datos (Anthony, 2002). Es así como nace el modelo de regresión binomial o binaria multivariante de diseño de muestreo complejo para realizar estimaciones apropiadas (considerado en el presente estudio).

En segundo lugar, para que dicho modelo de regresión binomial multivariante de diseño de muestreo complejo pueda estar construido bajo una familia cuasi-binomial o cuasi-binaria, se han verificado dos condiciones principales: (a) cuando el valor del parámetro de dispersión ϕ es mayor a la unidad en escenarios de respuestas binarias como en el contexto de muestras obtenidas mediante métodos de diseño complejos (recuentos no enteros producidos por el uso de ponderaciones muestrales diferenciales), se indica que el modelo tiene una dispersión

excesiva y que los parámetros del modelo pueden estar subestimados (R Core Team, 2020) (la familia cuasi-binomial es la opción ideal para enfrentar esta situación particular modelando la sobre-dispersión) y (b) en términos computacionales del uso del software, una regresión con una familia binomial asume que los pesos o la contribución que cada observación hace a la probabilidad ponderada dentro del modelo es un entero; sin embargo, cuando los factores de ponderación resultan ser cifras no enteras (como sucede en la encuesta empleada en este estudio), el procesamiento del modelo no llega a ajustar un resultado a las observaciones (R Core Team, 2020) (la familia cuasi-binomial es la variante que acepta estas contribuciones no enteras y llega a un ajuste del modelo en esta situación).

Los resultados de los coeficientes de regresión, errores estándar, el estadístico de prueba t basado en la prueba ajustada de Wald, el nivel de significancia asociado a las variables, las razones de probabilidades (*odds ratios*) ajustados y el diagnóstico de multicolinealidad (mediante el GVIF, el factor de inflación generalizado de la varianza) serán reportados a fin de dilucidar los resultados estadísticos que el modelo provea para que puedan ser evaluados en función del efecto o influencia que tengan las variables independientes sobre la clase de respuesta (dependiendo tanto del signo del coeficiente como de la magnitud del *odd ratio*) y los niveles de confianza con los cuales se pueda reportar una asociación empírica significativa en este tipo de relaciones.

De igual modo, como prueba de solidez o *robustness check*, se ajustará un modelo de regresión cuasi-binaria multivariante del tipo probit con la misma relación de variable dependiente y covariables o regresores. Una verificación de solidez o *robustness check* da una idea de la confianza que se puede depositar en los resultados del modelo primario (en este caso, el logit), al examinar cómo se comportan ciertas estimaciones de coeficientes centrales de una regresión cuando se modifica las especificaciones de la regresión agregando o eliminando regresores o cuando se cambia la forma funcional con la que se construye el modelo. Si los coeficientes son plausibles y robustos, esto se interpreta comúnmente como evidencia de validez estructural (Lu & White, 2014).

5.2.4. Pre-procesamiento del conjunto de datos para la aplicación de modelos paramétricos y no paramétricos

Referido al tratamiento de la naturaleza de las variables independientes o explicativas, muchos problemas importantes que el aprendizaje automático intenta resolver son de naturaleza binaria (expresión de problemas como clasificaciones binarias para predecir dos categorías o clases). Sin embargo, dependiendo del tipo de modelo que desee aplicarse, un problema central es cómo representar y convertir estas características categóricas discretas como entradas en ciertas técnicas a fin de que estas puedan procesarlas (considerando que algunos modelos exigen valor numéricos en su funcionamiento) y generar resultados a partir de ellas (Seger, 2018).

Considerando los modelos paramétricos y no paramétricos planteados en el estudio, los métodos a emplear para el pre-procesamiento de estos factores categóricos son los siguientes: (a)

método/enfoque del *label encoding* y el (b) método/enfoque de *one-hot encoding*.

En primer lugar, se puede definir al *label encoding* como el método estándar aplicado para el uso de características codificadas, el cual es un enfoque simple que implica convertir cada nivel de una variable en un número que lo representará como entrada y que será procesado por ciertos modelos como una sola variable (Pargent, 2019). Las variables categóricas a convertir pueden ser del tipo nominal u ordinal dependiendo la naturaleza de la variable (cabe destacar que los algoritmos asumirán a dichas etiquetas/números como factores que forman parte de una misma variable) (Pargent, 2019). En segundo lugar, el método de *one-hot encoding* se basa en la binarización de variables categóricas de alta cardinalidad cuando ciertos algoritmos de aprendizaje automático supervisados solo aceptan entradas aritméticas, en el cual para convertir dichas variables categóricas en representaciones binarias, es necesario transformar una variable categórica en un número de variables binarias (que asumen únicamente valores 0 y 1) equivalente a x-1 niveles o categorías que posean; es decir, la dimensionalidad del factor (Garavaglia *et al.*, 1998).

5.2.5. Selección del tamaño de muestra para los conjuntos de entrenamiento y prueba en el modelamiento predictivo

La mayoría de los algoritmos de clasificación tienen uno o más parámetros que se utilizan para controlar la complejidad y desempeño de estos. Para encontrar un conjunto óptimo de parámetros, es necesario dividir los datos en conjuntos de entrenamiento y validación bajo lo que se conoce como el método de retención o el *holdout method* (Xu & Goodacre, 2018a). En este caso, el conjunto de entrenamiento permitirá la construcción de los modelos paramétricos y no paramétricos y la prueba de hiper-parámetros, para estos últimos, y el conjunto de prueba ofrecerá la posibilidad de comparación y la validación de la bondad de ajuste de cada uno de ellos.

Para detectar el comportamiento de un modelo de aprendizaje automático, se necesita utilizar observaciones que no se utilizan en el proceso de entrenamiento. De lo contrario, la evaluación del modelo estaría sesgada. El uso del *holdout method* reduce el riesgo de posibles problemas, como la fuga de datos o el sobreajuste. Por lo tanto, existe la garantía de que el modelo entrenado se generalice bien sobre nuevos datos (Xu & Goodacre, 2018a).

En este sentido, el procedimiento del método de retención tiene como único parámetro de configuración principal al tamaño del conjunto de entrenamiento y el conjunto de prueba. Esto se expresa más comúnmente como un porcentaje entre 0 y 1 para el entrenamiento o para los datos de prueba (Xu & Goodacre, 2018a). Sin embargo, se remarca que no existe un porcentaje de división óptimo o estándar que pueda aplicado a todo conjunto de datos. Se debe elegir un porcentaje de división que cumpla con los objetivos de una investigación con consideraciones que incluyen: costo computacional en el entrenamiento del modelo, costo computacional en la evaluación del modelo, representatividad del conjunto de entrenamiento y representatividad del conjunto de pruebas (Xu & Goodacre, 2018a).

5.2.6. Tratamiento muestral del desbalance del conjunto de entrenamiento

La clasificación desequilibrada implica el desarrollo de modelos predictivos en conjuntos de datos de clasificación que tienen un desequilibrio de clases severo. El desafío de trabajar con conjuntos de datos desequilibrados es que la mayoría de las técnicas de aprendizaje automático ignorarán y, a su vez, tendrán un rendimiento deficiente en la clase minoritaria, siendo normalmente el rendimiento en la clase minoritaria lo más importante dentro del desempeño predictivo de dichos modelos y dentro de una investigación (Chawla *et al.*, 2002).

Considerando el conjunto de entrenamiento definido en la Sección 5.3.4.2, se puede establecer que este sufre un problema de desbalanceo, ya que la clase de interés en el estudio (1 = un individuo que demuestra poseer un nivel de conocimiento adecuado sobre el VIH/SIDA) sólo representa un 34 % del total de observaciones (las categorías de la variable respuesta no están representadas de forma equitativa). Lo que indica que es necesario un tratamiento muestral de los datos con el fin de balancear dicho conjunto y tener un mismo porcentaje de clases en la variable respuesta.

El muestreo de datos proporciona una colección de técnicas que transforman un conjunto de datos de entrenamiento para balancear o equilibrar mejor la distribución de clases. Una vez equilibrados, los algoritmos de aprendizaje automático estándar se pueden entrenar directamente en el conjunto de datos transformado sin ninguna modificación. Esto permite abordar el desafío de la clasificación desequilibrada, incluso con distribuciones de clases severamente desequilibradas, con un método de preparación de datos (Anis & Ali, 2017).

Existen muchos tipos diferentes de métodos de muestreo de datos que se pueden utilizar dependiendo del tipo de problema de investigación que se tenga y las características del conjunto de datos en cuestión. En el presente caso, el método de muestreo elegido a fin de equilibrar las clases de la variable dependiente "Conocimiento del VIH/SIDA" es el SMOTE (*Synthetic Minority Oversampling Technique*), técnica que combina el sobre-muestreo de la clase minoritaria u anormal y el sub-muestreo de la clase mayoritaria; debido a que, entre todos los métodos disponibles de balanceo de datos, es uno de los más recomendados para obtener un mejor desempeño de los clasificadores y capacidad predictiva de los modelos sin sesgo (Chawla *et al.*, 2002; Anis & Ali, 2017).

5.2.7. Construcción y optimización de los modelos paramétricos y no paramétricos de estimación

La construcción de modelos de aprendizaje automático implicará dos pasos: seleccionar un algoritmo de aprendizaje y luego optimizar los hiper-parámetros para maximizar el rendimiento de este (Province, 2015).

Tomando en consideración que, con anterioridad, en el Capítulo 3 se han definido los modelos paramétricos y no paramétricos a emplear en esta investigación, se debe precisar el procedimiento para la optimización del rendimiento de los mismos. A fin de determinar la estructura y composición idónea para los modelos planteados en el estudio, se empleará un proceso de

validación cruzada y optimización por medio del programa informático Rapidminer. La justificación, desde la perspectiva teórica, del uso de este software de minería de datos radica en las conclusiones dadas por Jovanovic *et al.* (2014), quienes señalan que este software puede ser quizás la única herramienta disponible en el mercado que ofrece este tipo de flexibilidad a través de su interfaz gráfica de usuario, sin programación involucrada en el nivel inferior y precisan que RapidMiner ofrece procesos totalmente personalizables que superan con creces los escenarios de aplicaciones simples. Desde la perspectiva empírica en esta investigación, la última afirmación se plasmó a través de la posibilidad que ofreció Rapidminer de llevar a cabo el diseño y desarrollo del procesamiento de datos y modelos de aprendizaje automático que el estudio requirió en un único y consolidado entorno de trabajo, con una interfaz intuitiva, un manejo de conjuntos de datos rápido y variado y con la utilización de operadores computacionales (que representan las aplicaciones estadísticas y de aprendizaje) que simplificaron y agilizaron el ritmo de trabajo al haber puesto en disposición una gama considerable de parámetros y configuraciones en cada etapa de la ejecución de las técnicas fáciles de entender y adaptar, considerando las necesidades del usuario, reduciendo las exigencias analíticas y de conocimiento para emplear la herramienta.

En esta herramienta para el análisis y minería de datos, la arquitectura general para los métodos planteados se basa principalmente en dos operadores: *Optimize Parameters (Grid)* y *Cross Validation* (Mierswa & Klinkenberg, 2018). El proceso de ajuste de modelos (*Grid Search*) se realiza mediante el operador *Optimize Parameters (Grid)*, que ejecuta sub-procesos para todas las combinaciones de valores suministrados y seleccionados de los parámetros y luego entrega los valores óptimos que generan los mejores indicadores de desempeño (Mierswa & Klinkenberg, 2018). De la misma manera, la evaluación de cada combinación de hiper-parámetros y las métricas que lo caracterizan se efectuará a través de un proceso de validación cruzada, que se lleva a cabo en el operador *Cross Validation*. Este operador permite evaluar al modelo en base a la partición de k sub-conjuntos de entrenamiento (k - 1) y validación (1) y así generar los resultados de un algoritmo y mediciones de su rendimiento (Mierswa & Klinkenberg, 2018). Los datos de entrada (el conjunto de entrenamiento original) se dividen en un conjunto de aprendizaje (*training*) y un conjunto de validación (*testing*). El proceso de ajuste se ejecuta separando el conjunto de datos en k porciones diferentes (en este caso, la separación es producto de una validación cruzada de 10x10, lo que indica que los datos iniciales serán divididos aleatoriamente en $k = 10$ sub-muestras o "pliegues" y este proceso de sub-muestreo se repetirá 10 veces) (Mierswa & Klinkenberg, 2018). Luego, cada clasificador se entrenó en k - 1 porciones para cada solución candidata seleccionada por la técnica de ajuste. El conjunto de validación (el pliegue restante de k) se utiliza para probar el modelo (la precisión de la validación se evalúa aplicando el modelo conseguido en el entrenamiento). Así, el rendimiento relacionado a los hiper-parámetros probados en esa iteración puede ser determinados. Este ciclo se repite N veces, dependiendo de las permutaciones que se obtengan por la combinación de parámetros dados. Para guiar el proceso de búsqueda, se utiliza la precisión de validación promedio como valor central. Finalmente, el *Optimize Parameters (Grid)* devolverá el modelo (y, por ende, los mejores valores de parámetros en los que este se

basó) con la mayor precisión y el cual debe ser generalizado al conjunto de prueba desconocido, posterior a esta fase de ajuste (Mierswa & Klinkenberg, 2018).

5.2.8. Comparación de los clasificadores para la predicción del conocimiento sobre el VIH/SIDA

El procedimiento de comparación de modelos en el que se desea seleccionar el mejor algoritmo de clasificación para procesar un conjunto de datos dado consiste en muestrear un conjunto de entrenamiento a partir de los datos totales considerados (un porcentaje de este conjunto total será designado para el afinamiento de modelos), construir varios clasificadores con diferentes algoritmos y parámetros de clasificación en base a este conjunto de entrenamiento y luego generalizar sus desempeños empíricamente en algunos conjuntos de prueba muestreados a partir de los mismos datos (el porcentaje restante de los datos totales luego de haber sido muestreado previamente) (Zhang *et al.*, 2017). Finalmente, se selecciona el clasificador con mayor rendimiento en base a dicho conjunto de prueba/validación luego de generalizar los modelos (aplicarlos a los datos de prueba que poseen una etiqueta desconocida a priori para la técnica) (Zhang *et al.*, 2017). Al estudiar las propiedades de los datos, luego seleccionar un algoritmo de clasificación adecuado y determinar los valores de los parámetros más apropiados y, por último, usarlo para construir el clasificador en una fase de validación corresponde a un proceso analítico lo suficientemente potente para precisar al mejor modelo bajo una serie de métricas o medidas de bondad de ajuste que permiten estas comparaciones (Zhang *et al.*, 2017). Dado que las diferentes métricas capturan diferentes características de un clasificador, según las propiedades de los datos, la elección de las métricas influye en cómo se mide y compara el rendimiento de los algoritmos de clasificación (Zhang *et al.*, 2017); sin embargo, las métricas más comúnes o tradicionales a emplear para equiparar técnicas son: la precisión, la sensibilidad, la especificidad, el error de clasificación, el kappa de Cohen, F1-Score, el valor predicho negativo, el valor predicho positivo y la AUC (*Area Under the Curve*) - métricas descritas con mayor precisión en la Sección 3.1.2.9.

Por otro lado, dentro del desarrollo y comparación de clasificadores, a menudo nos enfrentamos a la tarea de extraer conocimiento de las mismas variables o factores que fueron necesarios para construir dichos métodos que, cuando se afinan correctamente, pueden tener un rendimiento predictivo sumamente significativo (Greenwell *et al.*, 2018). Sin embargo, tener un modelo que predice bien solo resuelve una parte del problema. También es deseable extraer información sobre las relaciones descubiertas por el algoritmo de aprendizaje. A saber, a menudo queremos saber qué predictores, si los hay, son importantes asignando algún tipo de puntuación de importancia variable a cada característica. Una vez que se ha identificado un conjunto de características influyentes, el siguiente paso es resumir la relación funcional entre cada característica, o subconjunto de la misma, y el resultado de interés (Greenwell *et al.*, 2018). Sin embargo, dado que la mayoría de los algoritmos de aprendizaje estadístico son modelos con una estructura sumamente compleja, extraer esta información no siempre es sencillo. Afortunadamente, existe un enfoque conocido como *Variable Importance Measures*, el cual se emplea para poder definir un indicador que cuantifica la fuerza de la dependencia

entre la variable de salida del modelo y una o un conjunto de variables de entrada (Greenwell *et al.*, 2018). Dicho enfoque será realizado mediante el operador *Explain Predictions* del software Rapidminer que provee observaciones estadísticas y visuales para ayudar a comprender la función de cada atributo en la predicción, empleando valores de correlación local para especificar el papel de cada atributo (este papel puede apoyar o contradecir la predicción) en el pronóstico de un valor particular relacionado con una sola muestra en los datos y así generar un ranking de relevancia de los factores considerados, en el cual valores positivos indican que el atributo apoya las predicciones correctas y negativos sugieren que la característica los contradice (Mierswa & Klinkenberg, 2018).

5.3. Resultados de los modelos propuestos

Esta sección congrega el reporte de todos los resultados relacionados a la preparación de los factores socio-demográficos, económicos y de salud para su aplicación en el análisis estadístico, la asociación empírica entre los cofactores independientes y la variable respuesta, los resultados de la regresión binomial logit y probit para estudiar la asociación de estas características con el nivel de conocimiento sobre el VIH/SIDA que los individuos puedan poseer y la comparación de diversos modelos paramétricos y no paramétricos para obtener el algoritmo con el mejor desempeño para la clasificación del conocimiento sobre el virus y la enfermedad.

5.3.1. Preparación de datos de factores asociados al conocimiento del VIH/SIDA

En base a las interrogantes planteadas en la encuesta ENDES sobre la percepción y el entendimiento de los ciudadanos sobre las formas de prevención y rechazo de ideas erróneas de transmisión y dinámica del VIH/SIDA en la Sección 5.2.1, la distribución de las respuestas a dichas variables intermedias (sobre las que se construirá la nueva variable dependiente) y la precisión del número de individuos sobre el nivel de conocimiento sobre la epidemia (la nueva variable final), tanto adecuado como inadecuado, se presentan en la Tabla 5.2.

Tabla 5.2: Construcción de la variable: Conocimiento adecuado del VIH/SIDA.

Variable	**No adecuado**	**Adecuado**
Menor riesgo con el uso de preservativos	2199 (20.80 %)	8366 (79.20 %)
Menor riesgo con una pareja no infectada	1884 (17.80 %)	8681 (82.20 %)
Un individuo de apariencia sana puede estar infectado	2269 (21.50 %)	8296 (78.50 %)
Se puede adquirir el virus mediante contacto físico	2555 (24.20 %)	8010 (75.80 %)
Se puede adquirir el virus al usar y compartir utensilios	4072 (38.50 %)	6493 (61.50 %)
Nivel de conocimiento	6989 (66.20 %)	3576 (33.80 %)
Total de encuestados	N = 10,565 N(%) = 100.00 %	

Fuente: Elaboración propia.

En la tabla anterior, se puede observar que si bien la tendencia general de los resultados por cada pregunta permite dilucidar que los individuos responden de forma adecuada y demuestran nociones correctas sobre dichos aspectos relacionados al VIH/SIDA de manera individual; al unificar las variables intermedias y los datos que las incluyen, se puede establecer que el nivel de entendimiento global no resultó adecuado o correcto (33.80 %) considerando que los ciudadanos no respondieron de forma precisa a cada interrogante formulada sobre la epidemia.

Por otro lado, analizando las variables intermedias que corresponden a la recepción de información (a través de cualquier canal o medio), que se traduce en haber escuchado con anterioridad algún tipo de dato, por parte de los ciudadanos sobre el VIH y sobre el SIDA (por separado) bajo los patrones de contestación expresados en la Sección 5.2.1, la distribución de las respuestas dadas por los entrevistados y de la nueva variable que representa la agrupación de haber escuchado información sobre el VIH y el SIDA se muestran en la Tabla 5.3.

Tabla 5.3: Construcción de la variable: Oído hablar acerca del VIH/SIDA.

Variable	**NO**	**SÍ**
Oído hablar sobre el VIH	1381 (13.10 %)	9184 (86.90 %)
Oído hablar sobre el SIDA	16 (0.20 %)	10549 (99.80 %)
Oído hablar sobre el VIH/SIDA	1397 (13.20 %)	9168 (86.80 %)
Total de encuestados	N = 10,565 N(%) = 100.00 %	

Fuente: Elaboración propia.

La tabla previa, que se muestra líneas arriba, reafirma lo expresado por (INEI, 2008). Los ciudadanos mostraron un nivel de información de ambos aspectos de la epidemia (el VIH y el SIDA) elevado, considerando que ambos reportan más del 85 % de tasa de respuesta por parte de los individuos encuestados. No obstante, el nivel de información sobre el SIDA (99.80 %) superó de forma relevante al nivel del VIH (86.90 %) a nivel nacional (una diferencia de casi 15 % a favor del primero). Es por ello que la nueva variable generada que versa sobre la unificación de haber escuchado información tanto sobre el VIH como el SIDA revela una tendencia significativemente elevada (86.80 %) de contacto con información sobre la epidemia por parte de los habitantes del país pero evidencia el impacto de aquellos que no han escuchado sobre el VIH a priori.

Finalmente, tomando en cuenta a las variables intermedias que representan el acceso a ciertos medios de comunicación e información en el país como la radio, la televisión y el internet, la distribución de respuesta de las variables mencionadas y de la nueva variable que define el acceso a medios de comunicación en general por parte de la ciudadania se exhibe en la Tabla 5.4.

Tabla 5.4: Construcción de la variable: Acceso a medios multimedia.

Variable	**NO**	**SÍ**
Radio	3884 (36.80 %)	6681 (63.20 %)
Televisión	2005 (19.00 %)	8560 (81.00 %)
Internet	8712 (82.50 %)	1853 (17.50 %)
Acceso a medios multimedia	1115 (10.60 %)	9450 (89.40 %)
Total de encuestados	N = 10,565 N(%) = 100.00 %	

Fuente: Elaboración propia.

En el caso de la tabla presentada anteriormente, se puede determinar que los individuos tienen un acceso más proliferado de medios tradicionales de comunicación e información como lo son la radio (63.20 %) y la televisión (siendo este último el de mayor presencia - 81.00 %). Empero, el Internet como medio presenta la menor tasa de respuesta por parte de los pobladores, por lo que este medio resulta insuficiente dentro de la sociedad para proveer información de distintas

índoles por su baja presencia. Asimismo, considerando que la variable final que representa el acceso general a medios de comunicación por parte de la población se establece que los ciudadanos registran un nivel alto de respuestra frente a la variable (89.40 %) al poseer al menos uno de los tres medios presentados.

Tabla 5.5: Definición operacional de los factores socio-demográficos, económicos y familiares empleados en el estudio.

Variable	Definición	Escala de medición	Tipo de variable	Niveles
Conocimiento	Conocimiento/Entendimiento sobre el VIH/SIDA	Cualitativa Nominal	Dependiente	0. NO 1. SÍ
Género	Género del encuestado	Cualitativa Nominal	Independiente	0. Femenino 1. Masculino
Área de residencia	Contexto cultural específico en el que los encuestados residen	Cualitativa Nominal	Independiente	0. Rural 1. Urbana
Nivel educativo	Nivel de estudio más alto alcanzado por el encuestado	Cualitativa Ordinal	Independiente	0. Sin educación 1. Primaria 2. Secundaria 3. Superior
Estado civil	Relaciones familiares y/o matrimoniales del encuestado	Cualitativa Nominal	Independiente	0. Soltero 1. Casado/Conviviente 2. Divorciado/Separado/Viudo
Nacionalidad	Pertenencia del encuestado al ordenamiento del estado peruano	Cualitativa Nominal	Independiente	0. Extranjero 1. Peruano
Etnicidad	Etnicidad que el encuestado identifica según sus costumbres y antepasados	Cualitativa Nominal	Independiente	0. Origen Nativo[a] 1. Afroperuano 2. Blanco 3. Mestizo 4. Otro/No precisa
Lengua Materna	Idioma o lengua que el encuestado aprendió en sus primeros años de vida	Cualitativa Nominal	Independiente	0. Lengua nativa[b] 1. Castellano 2. Lengua extranjera
Oído sobre VIH/SIDA	Contacto con información acerca del VIH/SIDA	Cualitativa Nominal	Independiente	0. NO 1. SÍ
Prueba de VIH/SIDA	Realización de la prueba de descarte en los últimos 12 meses	Cualitativa Nominal	Independiente	0. NO 1. SÍ
Rango etario	Rango etario al que pertenece el encuestado	Cualitativa Ordinal	Independiente	0. 15-20 años 1. 20-24 años 2. 25-29 años
Región natural	Región de respuesta a la que el individuo pertenece	Cualitativa Nominal	Independiente	0. Lima Metropolitana[c] 1. Resto Costa 2. Sierra 3. Selva
Acceso a medios	Capacidad de acceso a medios multimedia[d]	Cualitativa Nominal	Independiente	0. NO 1. SÍ
Género del jefe de hogar	Género del jefe de familia en el hogar del encuestado	Cualitativa Nominal	Independiente	0. Femenino 1. Masculino
Nivel económico	Grupo poblacional de bienestar o de riqueza al que el encuestado pertenece	Cualitativa Ordinal	Independiente	0. Muy pobre 1. Pobre 2. Medio 3. Rico 4. Muy rico

Notas. [a] Quechua, aimara, nativo de la Amazonía, perteneciente o parte de otro pueblo indígena u originario. [b] Quechua o aimara/ lengua originaria de la Selva u otra lengua nativa. [c] Comprende la provincia de Lima y la Provincia Constitucional del Callao. [d] Medios multimedia: Comprende acceso a radio, televisión o internet.

Fuente: Elaboración propia.

Es así como, basándonos en la información estadística recopilada del INEI y la preparación de datos antedicha, las variables identificadas para la indagación de factores que poseen una influencia sobre el conocimiento adecuado sobre formas de transmisión e ideas erróneas sobre el VIH/SIDA se encuentran descritas en la Tabla 5.5.

5.3.2. Análisis estadístico de los factores asociados al conocimiento sobre el VIH/SIDA

La Tabla 5.6 muestra los resultados del nivel de conocimiento sobre el VIH/SIDA de la población entrevistada entre 15-29 años según características de la muestra recopilada en la base de datos de la ENDES.

De una muestra de 10,565 individuos, existe un contraste significativo en la distribución del nivel de conocimiento sobre el virus y la enfermedad entre la población objetivo, estimando que solo 3,576 (33.80 %) encuestados contaban con un nivel de conocimiento adecuado (cálculo determinado en la Sección 5.3.1) y el resto de las personas (6,989 - 66.20 %) tenía nociones o percepciones incorrectas acerca de la adquisición y mecanismo de los mismos.

Tomando en cuenta los factores socio-demográficos, económicos y de salud se pueden señalar a continuación diferencias en términos de desconocimiento sobre la epidemia, tanto producto de las estimaciones porcentuales muestrales (que ignoran el diseño complejo de la recolección de datos) y las estimaciones ponderadas (que incluyen los pesos y la estructura del marco muestral), apreciables según el nivel de conocimiento de los habitantes del territorio nacional.

En el caso del "Género", la proporción muestral de hombres (69.80 %) supera a la de mujeres (63.90 %) en la dimensión de desconocimiento; lo que refleja que los varones jóvenes en el país conocen en una menor tendencia las formas de prevención y cuidado frente al VIH/SIDA.

En cuanto al "Área de residencia", en las zonas rurales (76.60 %) existe mayor desconocimiento que en las urbanas (61.20 %); lo que refleja que en las zonas urbanas existe una mayor propensión de los individuos a poseer mejores ideas y percepciones sobre el virus y la enfermedad.

Examinando el "Nivel económico", se puede observar una tendencia inversamente proporcional entre la distribución de ingresos de los individuos y el nivel de conocimiento que estos puedan tener; se reconoce que la proporción muestral de ciudadanos con desconocimiento sobre el VIH/SIDA disminuye a medida que el nivel económico mejora progresivamente: el nivel muy pobre resulta ser la categoría con mayor desconocimiento (78.60 %), seguida por el nivel pobre (66.40 %), consecutivamente figura el nivel medio (60.70 %), subsiguientemente se encuentra el nivel rico (53.50 %) y, finalmente, el nivel muy rico resulta ser el de menor desconocimiento (51.70 %).

Observando a la variable "Región", la Sierra es la que presenta una mayor tasa de desconocimiento (72.00 %), seguida por la Selva (67.50 %) y el resto de la Costa (61.00 %); en tanto Lima Metropolitana es la región que presenta el menor desconocimiento (58.60 %).

Considerando el "Rango etario", aquellos entre 15-20 años son quienes tienen un menor en-

tendimiento (71.70 %), seguidos por aquellos entre 21-24 años (65.20 %); entre tanto aquellos con 25-29 años tienen un menor nivel de desconocimiento (63.20 %).

En el caso del "Nivel educativo", aquellos sin educación registran la mayor proporción de desconocimiento (93.50 %), seguida por la incomprensión de los individuos que tienen como máximo instrucción primaria (87.40 %) y la inopia de las personas con educación secundaria como mayor nivel educativo (69.70 %); mientras que aquellos con educación superior o mayor poseen la menor estimación de desconocimiento (51.60 %).

Contemplando la "Etnicidad", aquellos que se identifican con otra etnia (a saber, asiática) o no logran determinar a cuál raza pertenecen con precisión son quienes tienen mayor desconocimiento en la muestra (74.60 %); en tanto los mestizos son quienes tienen una menor predisposición a poseer un nivel inadecuado de conocimiento (58.70 %).

Examinando a la variable "Oído acerca del VIH/SIDA", aquellos que no han oído ni del virus ni de la enfermedad son los que tienen un mayor desconocimiento en proporción (84.60 %); mientras que aquellos que sí han escuchado de ambos tienen un menor nivel de desconexión con el conocimiento sobre la epidemia (63.30 %).

Considerando la "Prueba de VIH/SIDA", aquellos que no se han realizado el descarte son los que presentan mayor desconocimiento (68.30 %) y, caso contrario, quien se hicieron el descarte presentan el nivel de conocimiento más alto (59.90 %).

Observando el "Estado civil", aquellos casados o convivientes son los que presentan más desconocimiento en proporción (66.50 %), seguido por la incomprensión de los individuos que registran un estado civil soltero (66.10 %); en tanto los divorciados, separados y/o viudos son quienes presentan el menor desconocimiento (63.30 %).

En el caso del "Acceso a medios", quienes no poseen ningún medio informativo son aquellos con mayor desconocimiento (77.60 %) y, en contraste, quienes poseen algún medio multimedia presentan un menor nivel inadecuado sobre inopia del VIH/SIDA (64.80 %).

En cuanto al "Género del jefe de familia", los hogares precedidos por varones son los que poseen menos conocimiento en proporción (67.20 %); en tanto familias dirigidas por una mujer son las que poseen un menor nivel de desconocimiento (63.50 %).

Considerando la "Lengua materna", aquellos que tengan a una lengua extranjera como lenguaje primario son quienes poseen el menor entendimiento sobre la epidemia (90.00 %), seguidos por la incomprensión de los ciudadanos que tienen como lengua materna a una lengua originaria o nativa del país (77.90 %); mientras que quienes conocen al castellano como lengua originaria poseen un menor desconocimiento (63.60 %).

Finalmente, contemplando la "Nacionalidad", los ciudadanos reconocidos legalmente como peruanos son quienes presentan un mayor desconocimiento (66.20 %); en tanto los ciudadanos extranjeros son los que presentan una menor proporción de incomprensión sobre el VIH/SIDA (59.40 %).

Se debe tener en cuenta que siempre existirán diferencias entre el n muestral y el n ponderado dentro del cálculo estadístico de encuestas que fueron desarrolladas mediante diseños complejos muestrales, como sucede en el caso de la encuesta de este estudio. La ENDES incluye factores de ponderación que modifican el n muestral, por lo que, por medio de la inclusión de estos factores, se puede obtener una proporción ponderada, producto del diseño y de las especificaciones muestrales de la encuesta, denominada como valores estimados. En ese marco, dentro de la cuarta columna de la Tabla 5.6 (Valor estimado) se reportan los valores estimados (en términos porcentuales) de las categorías o niveles de cada variable independiente considerada en el estudio. Dichos valores estimados son los porcentajes o proporciones ponderadas estimadas promedio de la categoría/nivel subpoblacional de una variable dentro de la población general, dando un referente a nivel inferencial de la tendencia de distribución de los individuos según característica seleccionada como subgrupo. Son estimaciones estadísticas puntuales de las proporciones de subgrupos o subpoblaciones especificados a través de un conjunto de niveles/categorías en una variable luego de aplicar factores de ponderación.

Por otra parte, tomando en cuenta los resultados de las pruebas de independencia del estadístico χ^2 de Pearson, se estableció que las variables "Estado civil" ($\chi^2 = 3.087$, p-valor=0.499) y "Nacionalidad" ($\chi^2 = 4.047$, p-valor=0.275) no mostraron una asociación o relación estadística significativa con la variable respuesta "Conocimiento del VIH/SIDA" ya que ambos p-valores mostrados anteriormente superan los niveles de confianza considerados para el estudio, por lo que se procedió a desestimarlas o descartarlas del subsecuente modelo de regresión logística. Empero, existe suficiente evidencia estadística según los niveles de significancia establecidos para rechazar la hipótesis nula de que las variables restantes poseen independencia en relación con la variable dependiente, por lo que se puede afirmar que existe una asociación de dichas variables independientes (género, área de residencia, nivel económico, región natural, rango etario, nivel educativo, etnicidad, oído acerca del VIH/SIDA, prueba de descarte del VIH/SIDA, acceso a medios multimedia, género del jefe de hogar y lengua materna) con la variable respuesta dentro del análisis bivariado.

Tabla 5.6: Análisis de datos de los factores socio-demográficos, económicos y de salud según nivel de conocimiento sobre el VIH/SIDA.

	No adecuado (n = 6,989)		**Adecuado (n = 3,576)**			
Variable	**n muestral**[b]	**n ponderado**[c]	**n muestral**[b]	**n ponderado**[c]	**Valor estimado**[d]	**Prueba de χ^2**
Género						p<0.01 ***
Femenino	4142 (63.90 %)	30.79 % (0.0067)	2343 (36.10 %)	19.97 % (0.0061)	39.34 %	
Masculino	2847 (69.80 %)	33.30 % (0.0076)	1233 (30.20 %)	15.94 % (0.0063)	32.37 %	
Área de residencia						p<0.01 ***
Rural	2605 (76.60 %)	13.73 % (0.0036)	795 (23.40 %)	4.28 % (0.0023)	23.75 %	
Urbana	4384 (61.20 %)	50.36 % (0.0081)	2781 (38.80 %)	31.63 % (0.0076)	38.58 %	
Nivel económico						p<0.01 ***
Muy pobre	2534 (78.60 %)	14.36 % (0.0041)	691 (21.40 %)	3.91 % (0.0022)	21.39 %	
Pobre	2062 (66.10 %)	17.16 % (0.0056)	1058 (33.90 %)	8.17 (0.0039)	32.25 %	
Medio	1252 (60.70 %)	14.11 % (0.0057)	809 (39.30 %)	8.97 % (0.0047)	38.86 %	
Rico	732 (53.50 %)	10.94 % (0.0055)	636 (46.50 %)	8.49 % (0.0045)	43.69 %	
Muy rico	409 (51.70 %)	7.54 % (0.0048)	382 (48.30 %)	6.38 % (0.0045)	45.86 %	
Región natural						p<0.01 ***

	No adecuado (n = 6,989)		Adecuado (n = 3,576			
Variable	**n muestral**[b]	**n ponderado**[c]	**n muestral**[b]	**n ponderado**[c]	**Valor estimado**[d]	**Prueba de χ^2**
Lima Metropolitana	697 (58.60 %)	21.74 % (0.0082)	492 (41.40 %)	15.42 % (0.0068)	41.49 %	
Resto Costa	1859 (61.00 %)	16.41 % (0.0051)	1191 (39.00 %)	9.62 % (0.0040)	36.96 %	
Sierra	2566 (72.00 %)	16.66 % (0.0054)	996 (28.00 %)	6.77 % (0.0030)	28.89 %	
Selva	1867 (67.50 %)	9.28 % (0.0037)	897 (32.50 %)	4.10 % (0.0021)	30.64 %	
Rango etario						p<0.01 ***
15-20 años	2072 (71.70 %)	22.91 % (0.0069)	817 (28.30 %)	10.32 % (0.0051)	31.05 %	
21-24 años	2182 (65.20 %)	21.66 % (0.0063)	1164 (34.80 %)	12.22 % (0.0052)	36.07 %	
25-29 años	2735 (63.20 %)	19.52 % (0.0054)	1595 (36.80 %)	13.37 % (0.0048)	40.66 %	
Nivel educativo						p<0.01 ***
Sin educación	29 (93.50 %)	0.17 % (0.0005)	2 (6.50 %)	0.01 % (0.0001)	6.20 %	
Primaria	924 (87.40 %)	5.52 % (0.0028)	133 (12.60 %)	9.93 % (0.0013)	15.24 %	
Secundaria	4413 (69.70 %)	40.80 % (0.0078)	1920 (30.30 %)	19.53 % (0.0062)	32.38 %	
Superior	1623 (51.60 %)	17.61 % (0.0062)	1521 (48.40 %)	15.37 % (0.0056)	46.61 %	
Etnicidad						p<0.01 ***
Origen nativo	2520 (70.40 %)	17.39 % (0.0055)	1059 (29.60 %)	8.38 % (0.0043)	32.53 %	
Afroperuano	868 (73.00 %)	8.88 % (0.0041)	321 (27.00 %)	3.38 % (0.0027)	27.57 %	
Blanco	466 (70.20 %)	4.88 % (0.0031)	198 (29.80 %)	2.07 % (0.0022)	29.82 %	
Mestizo	2568 (58.70 %)	27.53 % (0.0071)	1805 (41.30 %)	19.99 % (0.0067)	42.07 %	
Otro/No sabe	567 (74.60 %)	5.41 % (0.0037)	193 (25.40 %)	2.08 % (0.0026)	27.81 %	
Oído del VIH/SIDA						p<0.01 ***
NO	1182 (84.60 %)	7.84 % (0.0035)	215 (15.40 %)	1.66 % (0.0018)	17.51 %	
SÍ	5807 (63.30 %)	56.25 % (0.0075)	3361 (36.70 %)	34.25 % (0.0076)	37.84 %	
Prueba del VIH/SIDA						p<0.01 ***
NO	5338 (68.30 %)	50.41 % (0.0077)	2473 (31.70 %)	26.22 % (0.0070)	34.22 %	
SÍ	1651 (59.90 %)	13.68 % (0.0049)	1103 (40.10 %)	9.69 % (0.0044)	41.46 %	
Estado civil						0.499 N.S.
Soltero	2930 (66.10 %)	33.95 % (0.0075)	1501 (33.90 %)	18.54 % (0.0065)	35.33 %	
Casado/Conviviente	3603 (66.50 %)	26.71 % (0.0062)	1811 (33.50 %)	15.19 % (0.0053)	36.26 %	
Divorciado/Separado/Viudo	456 (63.30 %)	3.44 % (0.0026)	264 (36.70 %)	2.17 % (0.0021)	38.73 %	
Acceso a medios						p<0.01 ***
NO	865 (77.60 %)	5.29 % (0.0029)	250 (22.40 %)	1.72 % (0.0078)	24.49 %	
SÍ	6124 (64.80 %)	58.80 % (0.0018)	3326 (35.20 %)	34.19 % (0.0075)	36.77 %	
Género del jefe de hogar						p=0.014 **
Femenino	1861 (63.50 %)	17.40 % (0.0055)	1068 (36.50 %)	11.00 % (0.0048)	38.74 %	
Masculino	5128 (67.20 %)	46.69 % (0.0075)	2508 (32.80 %)	24.90 % (0.0070)	34.79 %	
Lengua materna						p<0.01 ***
Lengua nativa	1447 (77.90 %)	7.76 % (0.0034)	411 (22.10 %)	2.59 % (0.0023)	25.04 %	
Castellano	5533 (63.60 %)	56.25 % (0.0079)	3164 (36.40 %)	33.31 % (0.0075)	37.19 %	
Lengua extranjera	9(90.00 %)	0.09 % (0.0004)	1 (10.00 %)	0.01 % (0.0001)	9.61 %	
Nacionalidad						p=0.275 N.S.
Extranjero	76 (59.40 %)	0.02 % (0.0021)	52 (40.60 %)	0.017 % (0.0019)	41.91 %	
Peruano	6913 (66.20 %)	65.77 % (0.0078)	3523 (33.80 %)	34.193 % (0.0041)	35.76 %	

Notas. *valores significativos p <0.10; **valores muy significativos p <0.05; ***valores altamente significativos p <0.01. N.S.: No significativo estadísticamente. [b] Se calculan la frecuencia y el porcentaje de las observaciones no ponderadas. [c] Se calculan las proporciones y el error estándar de las observaciones ponderadas. [d] Proporción ponderada de las variables considerando las especificaciones muestrales de la base de datos.

Fuente: Elaboración propia.

5.3.3. Asociación entre los determinantes de la salud y el conocimiento sobre el VIH/SIDA en el Perú

A fin de identificar los determinantes estructurales de la salud subyacentes que tienen un efecto significativo en el nivel de conocimiento sobre el VIH/SIDA de la población adolescente y joven adulta en el Perú, se desarrolló un modelo estadístico a fin de alcanzar este objetivo. Considerando la naturaleza binaria de la variable respuesta en este estudio, tener un nivel

de conocimiento sobre el VIH/SIDA adecuado o incorrecto, se empleará un modelo conocido como la regresión cuasi-binomial multivariante del tipo logit para los datos proporcionados.

En esta subsección, el resultado de la regresión cuasi-binomial logit empleada para medir la relación entre el nivel de conocimiento sobre el VIH/SIDA y las variables independientes clave, descritas anteriormente, se analizará con un nivel de significancia del 10 %, 5 % y 1 %, según corresponda. Los resultados de esta regresión se reportan en la Tabla 5.7. De igual modo, como prueba de solidez o *robustness check*, se ajustó un modelo de regresión cuasi-binomial multivariantge del tipo probit con la misma relación de variable dependiente y covariables y regresores. Los resultados de dicho modelo se reportan en la Tabla 5.8.

5.3.3.1. Resultados de la regresión cuasi-binomial logit para la asociación de factores socio-demográficos, económicos y familiares sobre el conocimiento y prevención del VIH/SIDA

La siguiente tabla identifica y presenta todas las características de la población objetivo que predicen de forma concreta, según los niveles de significancia establecidos, la probabilidad que un individuo en el Perú tenga un nivel de conocimiento adecuado o correcto acerca de la adquisición, formas de transmisión y percepción sobre el VIH/SIDA.

El género es una predictor significativo del nivel de conocimiento entre los adolescentes y jóvenes adultos del país ($p<0.01$). Ser del género masculino se encuentra correlacionado negativamente con la probabilidad de tener un entendimiento adecuado y correcto sobre el VIH/SIDA (β=-0.334). Sujetos del género masculino son 0.7 veces menos probables (I.C: 0.62-0.82) de poseer un conocimiento adecuado sobre el virus y la enfermedad que aquellas personas del género femenino.

El área de residencia no se encuentra correlacionada con la probabilidad de que el individuo tenga un nivel de conocimiento adecuado sobre el VIH/SIDA. Considerando el p-valor ligado al resultado del estadístico t del test de Wald ajustado ($p=0.415$), se concluye que la hipótesis nula no es rechazada; por lo que dicha variable no es significativa dentro del modelo.

Aquellos adolescentes y jóvenes adultos que forman parte de los niveles económicos (o están dentro de niveles de distribución de ingresos) pobre, medio, rico y muy rico en el país son 1.35 (I.C: 1.08-1.68), 1.60 (I.C:1.22-2.08), 1.75 (I.C: 1.31-2.32) y 1.85 (1.35-2.55) veces más probables, correspondientemente, de poseer un nivel de conocimiento adecuado acerca de del virus y la enfermedad en comparación con aquellos individuos que registran un nivel económico muy pobre. Los coeficientes de regresión asociados a las categorías de nivel económico demuestran una asociación positiva con la probabilidad de poseer conocimiento y dichas variables resultan ser significativas dentro del modelo según los niveles de α dados.

Existe una propensión en aquellos individuos que habitan en la región de la Sierra del Perú de no poseer un nivel de conocimiento correcto acerca de la interacción con el VIH/SIDA y sus formas de transmisión (están negativamente correlacionados con la variable respuesta - β=-0.178). Ellos son 0.84 veces (I.C: 0.69-1.02) menos probables de poseer un buen nivel

de conocimiento en comparación con aquellas personas que habitan en la región de Lima Metropolitana (p=0.078). Cabe remarcar que quienes habiten en algún departamento de la Costa peruana o la Selva no poseen una asociación empírica con la probabilidad de tener conocimiento, no resultan significativas en el modelo (p=0.838 y p=0.777, respectivamente).

Aquellas personas que se encuentren dentro del rango etario de 25 a 29 años poseen una correlación positiva con la variable dependiente (β=0.285), ya que son 1.33 veces (I.C:1.11-1.60) más probables de tener un entendimiento adecuado sobre el virus y la enfermedad que quienes tienen entre 15 a 20 años. Sin embargo, pertenecer al rango etario de 21 a 24 años no demuestra una relación significativa con la probabilidad de tener un conocimiento adecuado (p=0.329).

Los resultados en cuanto al nivel educativo más alto alcanzado demuestran que las personas con educación secundaria y superior a más poseen mayores probabilidades de contar con un nivel de conocimiento adecuado o correcto que aquellos individuos sin ningún tipo de educación acreditada, 6.76 (I.C: 1-17-38.98) y 9.00 (1.56-51.96) veces más correspondientemente. Se configura una predisposición de aquellos individuos con un nivel de educación relativamente alto de llegar a tener un buen nivel de discernimiento y entendimiento del VIH/SIDA. Debe señalarse que poseer educación primaria no permite predicir la probabilidad de éxito en la regresión logística, ya que estadísticamente resulta no significativa en base al p-valor de la prueba de Wald ajustada (p=0.169).

Los adolescentes y jóvenes adultos que se auto-identifiquen como afroperuanos u de otra etnia (no considerada en la encuesta o que no logren precisar con exactitud a qué etnicidad pertenecen) son menos probables de contar con un nivel de conocimiento apropiado sobre el virus y la enfermedad que aquellos que se auto-perciben como indígenas o de origen nativo en el país, 0.72 (I.C: 0.56-0.93) y 0.68 (I.C: 0.50-0.93) veces menos, respectivamente. Es necesario recalcar que auto-concebirse étnicamente como blanco o racialmente mestizo no denotan una asociación con la probabilidad de éxito de la variable "Conocimiento del VIH/SIDA", estadísticamente ambas categorías son no significativas dentro del modelo de regresión (p=0.173 y p=0.351, correspondientemente).

El haber escuchado algún tipo de información relativa al VIH/SIDA tiene un impacto significativo en la determinación del nivel de conocimiento de estos que un adolescente o jóven adulto pueda tener (p<0.01). Dicha correlación resulta ser positiva según el valor del coeficiente asociado a la característica en cuestión (β=0.635), por lo que puede establecer que quienes hayan escuchado información de distinta índole sobre el virus y la enfermedad son 1.89 (I.C: 1.47-2.42) veces más probables de tener un nivel de conocimiento adecuado que aquellos que no han entrado en contacto con información relevante sobre la epidemia.

Los hallazgos sugieren que la realización de la prueba de descarte del VIH/SIDA por parte de los adolescentes y jóvenes adultos es un predictor independiente del nivel de conocimiento sobre el VIH/SIDA (p=0.019). Dicha característica de salud se encuentra ligada positivamente a la tenencia de un nivel de entendimiento correcto o adecuado por parte del indi-

viduo (β=0.174). Puntualizando que aquellos que se han realizado el examen de detección del VIH/SIDA son 1.19 (I.C: 1.03-1.38) veces más probables de poseer un buen nivel de conocimiento que aquellos que no se han realizado la prueba.

El acceso a medios multimedia de información no se encuentra relacionado con la probabilidad de que un adolescente o joven adulto tenga un nivel de conocimiento adecuado sobre el VIH/SIDA. Considerando el p-valor ligado al resultado del estadístico t de la prueba de Wald ajustado (p=0.634), se concluye que la hipótesis nula no es rechazada; por lo que dicha variable no es significativa dentro del modelo.

De la misma manera, el género del jefe de hogar o familia no se encuentra relacionado con la probabilidad de que un adolescente o joven adulto tenga un nivel de conocimiento adecuado sobre el VIH/SIDA. Considerando el p-valor ligado al resultado del estadístico t de la prueba de Wald ajustado (p=0.223), se concluye que la hipótesis nula no es rechazada; por lo que dicha variable no es significativa dentro del modelo.

Finalmente, los adolescentes y jóvenes adultos que tienen al castellano como lengua primaria o materna son 1.26 (I.C: 1.00-1.60) veces más probables de contar con un nivel de conocimiento apropiado acerca del VIH/SIDA en comparación a aquellos que poseen un dialecto o lengua nativa del Perú como primer idioma, ya que la correlación de esta característica sociodemográfica con la variable dependiente es positiva (β=0.233) y significativa (p=0.052). Por el contrario, el hablar una lengua extranjera en el Perú está asociado negativamente con la posesión de un nivel de entendimiento correcto del virus y la enfermedad (β=-1.915) pero de forma significativa (p=0.082). Aquellos individuos que hablen un idioma extranjero son 0.15 veces (I.C: 0.02-1.28) menos probables de poseer un nivel de conocimiento adecuado que aquellas personas que hablan una lengua nativa en el país.

Considerando las varianzas de los coeficientes de regresión estimados, se puede establecer que no existe un problema de multicolinealidad en el modelo, ya que analizando los valores del factor generalizado de inflación de la varianza (o GVIF, en inglés), todos los índices calculados de las variables independientes son menores a 5, justificando la conclusión de no multicolinealidad (Fox & Monette, 1992).

Tabla 5.7: Resultado del análisis multivariado sobre la asociación entre factores sociodemográficos, económicos y de salud y conocimiento sobre el VIH/SIDA.

Variable	β	Error Est.	t	p-valor	O.R.a (I.C. 95 %)	GVIF[b]
Intercepto	-3.595	0.921	-3.903	<0.01 ***	0.03 (0.00 - 0.17)	-
Género						1.028
Femenino (REF)	-	-	-	-	-	
Masculino	-0.334	0.070	-4.751	p<0.01 ***	0.72 (0.62-0.82)	
Área de residencia						1.431
Rural (REF)	-	-	-	-	-	
Urbana	-0.076	0.093	-0.815	p=0.415 N.S.	0.93 (0.77-1.11)	
Nivel económico						1.170
Muy pobre (REF)	-	-	-	-	-	
Pobre	0.298	0.112	2.649	p=0.008 ***	1.35 (1.08-1.68)	
Medio	0.467	0.135	3.466	p<0.01 ***	1.60 (1.22-2.08)	
Rico	0.558	0.145	3.843	p<0.01 ***	1.75 (1.31-2.32)	

Variable	β	Error Est.	t	p-valor	O.R.a (I.C. 95 %)	GVIF
Muy rico	0.617	0.163	3.774	p<0.01 ***	1.85 (1.35-2.55)	
Región natural						1.121
Lima Metropolitana (REF)	-	-	-	-	-	
Resto Costa	-0.019	0.093	-0.204	p=0.838 N.S.	0.98 (0.82-1.18)	
Sierra	-0.178	0.101	-1.761	p=0.078 *	0.84 (0.69-1.02)	
Selva	-0.030	0.106	-0.284	p=0.777 N.S.	0.97 (0.79-1.19)	
Rango etario						1.083
15-20 años (REF)	-	-	-	-	-	
21-24 años	0.089	0.090	0.977	p=0.329 N.S.	1.09 (0.92-1.30)	
25-29 años	0.285	0.094	3.020	p=0.003 ***	1.33 (1.11-1.60)	
Nivel educativo						1.097
Sin educación (REF)	-	-	-	-	-	
Primaria	1.249	0.909	1.375	p=0.169 N.S.	3.49 (0.59-20.69)	
Secundaria	1.911	0.894	2.137	p=0.033 **	6.76 (1.17-38.98)	
Superior	2.198	0.894	2.457	p=0.014 ***	9.00 (1.56-51.96)	
Etnicidad						1.086
Origen nativo (REF)	-	-	-	-	-	
Afroperuano	-0.328	0.131	-2.502	p=0.012 **	0.72 (0.56-0.93)	
Blanco	-0.207	0.152	-1.364	p=0.173 N.S.	0.81 (0.60-1.09)	
Mestizo	0.092	0.098	0.933	p=0.351 N.S.	1.10 (0.90-1.33)	
Otro	-0.387	0.160	-2.420	p=0.016 **	0.68 (0.50-0.93)	
Oído del VIH/SIDA						1.067
NO (REF)	-	-	-	-	-	
SÍ	0.635	0.127	4.992	p<0.01 ***	1.89 (1.47-2.42)	
Prueba del VIH/SIDA						1.040
NO (REF)	-	-	-	-	-	
SÍ	0.174	0.074	2.352	p=0.019 **	1.19 (1.03-1.38)	
Acceso a medios						1.116
NO (REF)	-	-	-	-	-	
SÍ	0.063	0.131	0.476	p=0.634 N.S.	1.06 (0.82-1.38)	
Género del jefe de hogar						1.060
Femenino (REF)	-	-	-	-	-	
Masculino	-0.091	0.075	-1.220	p=0.223 N.S.	0.91 (0.79-1.06)	
Lengua materna						1.130
Lengua nativa (REF)	-	-	-	-	-	
Castellano	0.233	0.120	1.942	p=0.052 *	1.26 (1.00-1.60)	
Lengua extranjera	-1.915	1.102	-1.738	p=0.082 *	0.15 (0.02-1.28)	

Ajuste del modelo: 0.05 (Pseudo R^2 de McFadden), 0.01 (Pseudo R^2 de Cragg-Uhler).

Notas. *valores significativos p <0.10; **valores muy significativos p <0.05; ***valores altamente significativos p <0.01. N.S.: No significativo estadísticamente. REF: Nivel de referencia del factor. [b] *Generalized Variance Inflation Factor.*

Fuente: Elaboración propia.

5.3.3.2. Resultados de la regresión cuasi-binomial probit para la asociación de factores socio-demográficos, económicos y familiares sobre el conocimiento y prevención del VIH/SIDA

Como se había señalado, de forma paralela al ajuste de un modelo de regresión cuasi-binomial logístico o logit, se realizó la construcción de un modelo de regresión cuasi-binomial probit a manera de prueba de robustez/solidez de los resultados obtenidos en cuanto al tipo de asociaciones que las covariables mantienen con la variable dependiente en el estudio a fin de demostrar que estos no están sesgados o condicionados a la elección de la función de enlace o distribución empleada en los modelos.

Los determinantes del nivel de conocimiento sobre el VIH/SIDA en un adolescente o joven adulto, obtenidos mediante la regresión probit, se muestran en la Tabla 5.8.

El modelo de regresión probit produce resultados similares a los obtenidos empleando el modelo logit de la tabla 5.7. La siguiente tabla muestra que el género masculino, la región natural de la Sierra, la auto-percepción afroperuana y otra etnia/imprecisión de la misma y aquellos que tienen alguna lengua extranjera como lengua primaria están correlacionados negativamente con la probabilidad de tener un nivel de conocimiento adecuado. Mientras que todos los niveles económicos, el grupo etario 25-29 años, educación secundaria y superior a más, el haber escuchado información acerca del virus y la enfermedad, la realización de la prueba de detección y aquellos que tienen al castellano como lengua materna están correlacionados positivamente con la probabilidad de éxito de la variable dependiente. En el mismo sentido, las variables que no tuvieron una asociación significativa en el modelo logit exhiben el mismo comportamiento en el modelo probit.

En cuanto a otros indicadores, el modelo probit tampoco sufre del problema de multicolinealidad, ya que los valores de los GVIF ligados a las variables independientes son menores a 5, lo que justifica dicha conclusión.

Tabla 5.8: Resultado del análisis multivariado sobre la asociación entre factores sociodemográficos, económicos y de salud y conocimiento sobre el VIH/SIDA.

Variable	β	Error Est.	t	p-valor	GVIF[b]
Intercepto	-2.067	0.444	-4.658	p<0.01 ***	-
Género					1.029
Femenino (REF)	-	-	-	-	
Masculino	-0.203	0.043	-4.774	p<0.01 ***	
Área de residencia					1.449
Rural (REF)	-	-	-	-	
Urbana	-0.049	0.056	-0.868	p=0.385 N.S.	
Nivel económico					1.176
Muy pobre (REF)	-	-	-	-	
Pobre	0.180	0.067	2.681	p=0.007 ***	
Medio	0.284	0.081	3.491	p<0.01 ***	
Rico	0.341	0.088	3.875	p<0.01 ***	
Muy rico	0.378	0.099	3.803	p<0.01 ***	
Región natural					1.124
Lima Metropolitana (REF)	-	-	-	-	
Resto Costa	-0.013	0.057	-0.232	p=0.816 N.S.	
Sierra	-0.110	0.062	-1.782	p=0.075 *	
Selva	-0.021	0.065	-0.321	p=0.748 N.S.	
Rango etario					1.082
-20 años (REF)	-	-	-	-	
21-24 años	0.052	0.054	0.969	p=0.333 N.S.	
25-29 años	0.172	0.057	3.022	p=0.003 ***	
Nivel educativo					1.102
Sin educación (REF)	-	-	-	-	
Primaria	0.667	0.434	1.538	p=0.124 N.S.	
Secundaria	1.055	0.425	2.481	p=0.013 **	
Superior	1.234	0.425	2.899	p=0.004 ***	
Etnicidad					1.089
Origen nativo (REF)	-	-	-	-	
Afroperuano	-0.197	0.079	-2.507	p=0.012 **	

Variable	β	Error Est.	t	p-valor	GVIF
Blanco	-0.133	0.091	-1.467	p=0.142 N.S.	
Mestizo	0.054	0.060	0.909	p=0.363 N.S.	
Otro	-0.239	0.095	-2.507	p=0.012 **	
Oído del VIH/SIDA					1.080
NO (REF)	-	-	-	-	
SÍ	0.372	0.072	5.196	p<0.01 ***	
Prueba del VIH/SIDA					1.041
NO (REF)	-	-	-	-	
SÍ	0.108	0.045	2.397	p=0.017 **	
Acceso a medios					1.119
NO (REF)	-	-	-	-	
SÍ	0.039	0.077	0.503	p=0.615 N.S.	
Género del jefe de hogar					1.061
Femenino (REF)	-	-	-	-	
Masculino	-0.056	0.046	-1.236	p=0.217 N.S.	
Lengua materna					1.137
Lengua nativa (REF)	-	-	-	-	
Castellano	0.142	0.071	2.005	p=0.045 *	
Lengua extranjera	-1.146	0.562	-2.037	p=0.042 *	

Ajuste del modelo: 0.06 (Pseudo R^2 de McFadden), 0.01 (Pseudo R^2 de Cragg-Uhler).

Notas. *valores significativos p <0.10; **valores muy significativos p <0.05; ***valores altamente significativos p <0.01. N.S.: No significativo estadísticamente. REF: Nivel de referencia del factor. [b] *Generalized Variance Inflation Factor.*

Fuente: Elaboración propia.

5.3.4. Modelamiento predictivo paramétrico y no paramétrico del nivel de conocimiento del VIH/SIDA en el Perú

Esta sección analiza el rendimiento de una serie de algoritmos de clasificación para detectar y realizar predicciones sobre la clasificación del nivel de conocimiento sobre el VIH/SIDA de un adolescente o joven adulto en territorio nacional, teniendo en cuenta el escenario de datos desequilibrados, cuya propiedad principal de este tipo de problema de clasificación es que los ejemplos de una clase superan significativamente en número a los ejemplos de la otra clase minoritaria que representa el concepto más importante que se quiere clasificar o estimar (Lopez *et al.*, 2013), que caracteriza al conjunto de datos disponible y las características socio-demográficas, económicas y de salud recopiladas por la encuesta ENDES y que resultaron tener una asociación empírica verificada con la variable dependiente. Los métodos probados incluyen regresión binomial (R.B.), random forest (R.F.), redes neuronales artificiales (R.N.A.), el algoritmo k-NN (k-NN) y el árbol de decisión (A.D.). Bajo un pre-procesamiento de las variables independientes y la optimización de los modelos propuestos, empleando el método de validación cruzada de 10 iteraciones, los algoritmos fueron validados y las medidas de bondad de ajuste fueron calculados y presentados, introduciendo adicionalmente medidas de influencia de los atributos tomados en cuenta dentro de los modelos para la predicción de las clases positivas en la investigación.

5.3.4.1. Pre-procesamiento del conjunto de datos para la aplicación de modelos paramétricos y no paramétricos

Dentro del diseño y despliegue de los modelos de predicción propuestos, se considera principalmente que la variable dependiente o respuesta será el "Nivel de conocimiento sobre el VIH/SIDA" en la investigación, que responde a las percepciones, ideas de transmisión e interacción del VIH/SIDA en la cotidianidad de los adolescentes y jóvenes adultos del Perú. El tipo de variable del principal resultado del modelo es de naturaleza cualitativa binaria, lo que quiere decir que se compone por dos clases o niveles que los individuos pueden adoptar: el nivel de conocimiento adecuado sobre el virus y la enfermedad (clase positiva) y nivel de desconocimiento identificable (clase negativa). La recopilación y preparación de la base de datos para esta investigación se precisan en la Sección 5.1.2. De igual modo, las descripciones y definiciones de los datos se incluyen en la Sección 5.6.

Además, la predicción y clasificación prospectiva del nivel de conocimiento sobre el VIH/SIDA de los ciudadanos se soporta o apoya en una serie de características socio-demográficas, económicas y de salud, que incluyen: "Género", "Área de residencia", "Nivel educativo", "Región natural", "Nivel económico" (quintiles de bienestar), "Rango etario", "Etnicidad", "Ha oído acerca del VIH/SIDA con anterioridad", "Prueba de detección del VIH/SIDA", "Acceso a medios multimedia de información", "Lengua materna" y, finalmente, "Género del jefe de hogar o familia" - variables categóricas recopiladas en base a las respuestas de los sujetos encuestados en la ENDES descritas en la Sección 5.6.

Tomando en cuenta la naturaleza de la variable dependiente y de los factores regresores o covariables descrita previamente, el método de conversión a emplear dependerá directamente del tipo de modelo a construir y la forma en que este opera los atributos de entrada a suministrar dentro de su estructura.

De los métodos paramétricos y no paramétricos propuestos, los modelos de regresión binomial, random forest y árbol de decisiones son aquellas técnicas que se benefician de la inclusión de variables categóricas bajo el *label encoding* para la construcción y desarrollo de su estructura. La Tabla 5.9 a continuación muestra la aplicación del método de *label encoding* para la conversión de las variables categóricas presentadas anteriormente. Como puede notarse, la dimensionalidad de las variables independientes o regresoras no cambiará: seguirán siendo 12 variables las que cumplirán un rol de influencia sobre la variable dependiente; lo que varía, a través del método de *label encoding*, es la forma en que los métodos aceptarán los valores de dichos factores, que serán etiquetas o representaciones nominales luego de la conversión de los niveles que representaban datos cualitativos.

Tabla 5.9: Resultado de la conversión de variables categóricas independientes en valores de entrada a través del método de *label encoding*.

Variable	Niveles de categorías	Conversión numérica
Género	Femenino	0
	Masculino	1
Área de residencia	Rural	0

Variable	Niveles de categorías	Conversión numérica
	Urbana	1
Nivel económico	Muy pobre	0
	Pobre	1
	Medio	2
	Rico	3
	Muy rico	4
Región natural	Lima Metropolitana	0
	Resto Costa	1
	Sierra	2
	Selva	3
Rango etario	15-20 años	0
	21-24 años	1
	25-29 años	2
Nivel educativo	Sin educación	0
	Primaria	1
	Secundaria	2
	Superior	3
Etnicidad	Origen nativo	0
	Afroperuano	1
	Blanco	2
	Mestizo	3
	Otro/No sabe	4
Oído del VIH/SIDA	NO	0
	SÍ	1
Prueba del VIH/SIDA	NO	0
	SÍ	1
Acceso a medios	NO	0
	SÍ	1
Género del jefe de hogar	Femenino	0
	Masculino	1
Lengua materna	Lengua nativa	0
	Castellano	1
	Lengua extranjera	2

Fuente: Elaboración propia.

Por otro lado, en la presente investigación, modelos como redes neuronales artificiales y el k-Nearest Neighbors con métodos que reciben como entrada únicamente a valores numéricos. Se optó por la representación binaria de variables categóricas para facilitar la reducción futura de las variables en los modelos que lo requieran y minimizar el impacto en la estructura del modelo (referida a la complejidad que puede asumir un algoritmo por las variables dadas) (Garavaglia *et al.*, 1998). La Tabla 5.10 a continuación muestra la aplicación del método de *one-hot encoding* para la conversión de las variables categóricas presentadas anteriormente. A diferencia de la Tabla 5.9, la dimensionalidad de las variables independientes o regresoras aumentó: previamente eran 12 covariables; sin embargo, debido a que existen métodos que no admiten etiquetas y solo funcionan gracias a valores numéricos, fue necesario aumentar la cantidad de variables con las que se trabajó a un número de 24, precisando que los valores de cada nueva variable estarán en términos binarios y la representación de las observaciones estará en función del número de nuevas variables creadas de cada regresor original.

Tabla 5.10: Resultado de la conversión de variables categóricas independientes en valores de entrada a través del método de *one-hot encoding*.

Variable	Niveles de categorías	Transformación binaria	Codificación binaria[a]
Género	Femenino Masculino	Genero_M	1
Área de residencia	Rural Urbana	Area_Urbana	1
Nivel económico	Muy pobre Pobre Medio Rico Muy rico	Nivel_Pobre Nivel_Medio Nivel_Rico Nivel_Muy_Rico	1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1
Región natural	Lima Metropolitana Resto Costa Sierra Selva	Region_Costa Region_Selva Region_Sierra	1 0 0 0 1 0 0 0 1
Rango etario	15-20 años 21-24 años 25-29 años	Rango_21_20 Rango_25_29	1 0 0 1
Nivel educativo	Sin educación Primaria Secundaria Superior	Nivel_Primaria Nivel_Secundaria Nivel_Superior	1 0 0 0 1 0 0 0 1
Etnicidad	Origen nativo Afroperuano Blanco Mestizo Otro/No sabe	Etnicidad_Afroperuano Etnicidad_Blanco Etnicidad_Mestizo Etnicidad_Otro	1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1
Oído del VIH/SIDA	NO SÍ	Oído_SI	1
Prueba del VIH/SIDA	NO SÍ	Prueba_SI	1
Acceso a medios	NO SÍ	Acesso_SI	1
Género del jefe de hogar	Femenino Masculino	Genero_Jefe_M	1
Lengua materna	Lengua nativa Castellano Lengua extranjera	Lengua_Castellano Lengua_Extranjera	1 0 0 1

Notas. [a] La transformación binaria resultará en la creación de n-1 variables (donde n es el número de niveles que existen); en dicho caso, los valores en la codificación binaria le asignan al nivel o la categoría no considerada en la transformación una representación de 0(s) binaria (en el caso de Género, la categoría "Femenino" será reprentada con un valor de 0; en el caso del nivel económico, la categoría "Muy poobre" será representada con un valor de 0 0 0 0, etc.)

Fuente: Elaboración propia.

5.3.4.2. Selección del tamaño de muestra para los conjuntos de entrenamiento y prueba en el modelamiento predictivo

Bajo la premisa de Xu & Goodacre (2018a), para propiciar una buena generalización y ajuste de los algoritmos propuestos para la predicción del nivel de conocimiento sobre VIH/SIDA de un individuo, se divide el conjunto original de datos, compuesto por 10,565 observaciones, en dos nuevos sub-conjuntos o sub-muestras de datos.

Tabla 5.11: Distribución del tamaño de datos en los conjuntos de entrenamiento y prueba

Datos	Conocimiento adecuado	Conocimiento no adecuado	Total	Porcentaje
Conjunto de entrenamiento	3218	6291	9509	90 %
Conjunto de prueba	358	698	1056	10 %
Total	**3576**	**6989**	**10565**	**100 %**

Fuente: Elaboración propia.

La tabla 5.11 muestra la distribución del tamaño de los conjuntos de entrenamiento y prueba/validación bajo una regla de división "90/10" para la presente investigación: el primero de estos conjuntos agrupará al 90 % de los datos (congregando 9,509 observaciones) y el segundo agrupará al 10 % de los mismos (agremiando 1,056 registros); considerando que el tamaño de observaciones en la muestra (>1000 o 10,000, en algunos casos) fomenta el uso de este ratio de división (Racz *et al.*, 2021), que esta distribución de datos responde a su amplio y fomentado uso para el tratamiento de datos en el campo de la investigación médica al obtener desempeños estables y eficientes a partir de esta (Fredriksson & Glandberger, 2020; Raghavan, 2020) y, de igual manera, este patrón de división es el resultado de la preparación previa del conjunto de datos a través de una partición bajo el procedimiento de validación cruzada de 10-cortes (*10-fold cross validation*) aplicada en estos para el desarrollo de los modelos planteados; ya que responde a la correspondencia del 10 % del total de los datos para testear a los algoritmos y, dentro del 90 % restante correspondiente al conjunto de entrenamiento, una fracción del 10 % para la validación de las pruebas de hiper-parámetros propuestos (Fredriksson & Glandberger, 2020; Xu & Goodacre, 2018b; Berrar, 2018).

5.3.4.3. Tratamiento muestral del desbalance del conjunto de entrenamiento

La tabla 5.12 muestra la distribución de las observaciones de forma comparativa entre el conjunto de entrenamiento original sin cambios y el conjunto balanceado generado mediante el método SMOTE. Bajo el tratamiento muestral, se daría un sobre-muestreo de la clase minoritaria (conocimiento adecuado), elevando dicha cifra de 3,218 individuos a 6,436 individuos y de la clase mayoritaria (conocimiento no adecuado), de una cifra de 6,291 a 6,436 personas. El nuevo conjunto de entrenamiento estaría compuesto, finalmente, por 12,872 registros con clases o categorías balanceadas (50 % de los datos divididos entre ambas).

Tabla 5.12: Aplicación del tratamiento muestral para equilibrar los datos del conjunto de entrenamiento para los modelos paramétricos y no paramétricos

Método de muestreo	Conocimiento adecuado	Conocimiento no adecuado	Total	Procedimiento
Sin muestreo	3218	6291	9509	Datos originales
SMOTE[a]	6436	6436	12872	Muestra de minoría sintética

Notas. [a] Synthetic Minority Oversampling Technique.

Fuente: Elaboración propia.

5.3.4.4. Construcción y optimización de los modelos paramétricos y no paramétricos de estimación

Se ha establecido con antelación, en la introducción de la Sección 5.3.4, que los modelos de clasificación a utilizar son la regresión binomial, redes neuronales artificiales, árboles de decisión, algoritmo k-NN y el random forest. El rendimiento de estos algoritmos de clasificación dependerá de su dominio específico.

Para cada algoritmo que fue evaluado, los parámetros tomados en cuenta y sus definiciones se listan en la Tabla 5.13. En el caso de la regresión binomial (R.B.), no se incluyó ningún hiper-parámetro para el desarrollo del modelo ya que este únicamente asume una función de enlace o distribución logit entre la variable respuesta y las covariables, careciendo de parámetros que modifiquen la capacidad de predicción de la técnica y que afecten su rendimiento (Schratz *et al.*, 2018). En cuanto a los modelos no paramétricos, se establece que: el modelo de Redes Neuronales Artificiales (R.N.A.) posee parámetros de interés en cuanto a la topología de la red (el número de capas ocultas y las neuronas que componen las capas en cuestión) y medidas que modifican el aprendizaje de la misma (los ciclos de entrenamiento, la tasa de aprendizaje en las ponderaciones de cada neurona y el momentum dentro de la actualización de los pesos), el algoritmo k-NN presenta parámetros que describen el número de vecinos a considerar a ser incluidos en la predicción y las medidas de distancia que se asumen para apoyar dicha clasificación, el modelo Random Forest (R.F.) especifica el número de árboles a ser considerados durante el proceso de predicción y los aspectos relacionados a los criterios de división y formación de los árboles durante el proceso y, por último, la técnica de Árbol de decisión (A.D.) se encuentra configurada por los criterios de división que guiarán el crecimiento del árbol y la composición de los nodos que influencian el rendimiento predictivo del modelo.

Tabla 5.13: Definición y tipos de hiper-parámetros para los modelos paramétricos y no paramétricos

Modelo/Algoritmo	Hiper-parámetro	Tipo	Definición
R.B.	-	-	-
R.N.A.	Número de capas ocultas	Entero	Describe la cantidad de capas ocultas dentro de la red
	Número de neuronas	Entero	Describe la cantidad de neuronas dentro de las capas ocultas
	Ciclos de entrenamiento	Entero	Especifica el número de ciclos usados para el entrenamiento de la red
	Tasa de aprendizaje	Real	Define el costo que tiene el gradiente en la actualización de un peso
	Momentum	Real	Define la fracción de la actualización de peso anterior a la actual
k-NN	k número de vecinos	Entero	Describe al número de vecinos más cercanos a incluir en el proceso
	Voto ponderado	Nominal	Determina si los valores de distancia intervienen en la predicción
	Tipos de medida	Nominal	Describe la medida elegida para encontrar los vecinos más cercanos
R.F.	Número de árboles	Entero	Especifica el número de árboles aleatorios a generar
	Criterio de división	Nominal	Define el criterio sobre el que se elegirán los atributos para dividir
	Profundidad máxima	Entero	Describe el tamaño del árbol de decisiones en capas
	Estrategia de voto	Nominal	Especifica la estrategia de predicción en caso de discrepancias
	Ganancia mínima	Real	Describe la ganancia de umbral en un nodo antes de dividirlo
	Tamaño mínimo de hoja	Entero	Determina el mínimo de hojas para dividir un nodo interno
	Tamaño para división	Entero	Determina el tamaño mínimo de un nodo interno para su división
	Alternativas de pre-poda	Entero	Ajusta el número de nodos alternativos probados para dividir
A.D.	Criterio de división	Nominal	Define el criterio sobre el que se elegirán los atributos para dividir
	Profundidad máxima	Entero	Describe el tamaño del árbol de decisiones en capas
	Nivel de confianza	Real	Especifica el nivel para el cálculo del error pesimista de la poda

Modelo/Algoritmo	Hiper-parámetro	Tipo	Definición
	Ganancia mínima	Real	Describe la ganancia de umbral en un nodo antes de dividirlo
	Tamaño mínimo de hoja	Entero	Determina el mínimo de hojas para dividir un nodo interno
	Tamaño para división	Entero	Determina el tamaño mínimo de un nodo interno para su división
	Alternativas de pre-poda	Entero	Ajusta el número de nodos alternativos probados para dividir

Notas. R.B.: Regresión binomial, **R.N.A.**: Redes neuronales artificiales, **k-NN**: Algoritmo k-Nearest Neighbors, **R.F.**: Random forest, **A.D.**: Árbol de decisión.

Fuente: Elaboración propia.

Posteriormente al constreñimiento de los hiper-parámetros a utilizar en cada algoritmo, un proceso de ajuste de modelos (*Grid Search*) y validación cruzada (*Cross Validation*) fue llevado a cabo para analizar la diferencia de distintas iteraciones de ajuste. La arquitectura de ajuste realizada en el software Rapidminer se muestra en la Figura 5.2.

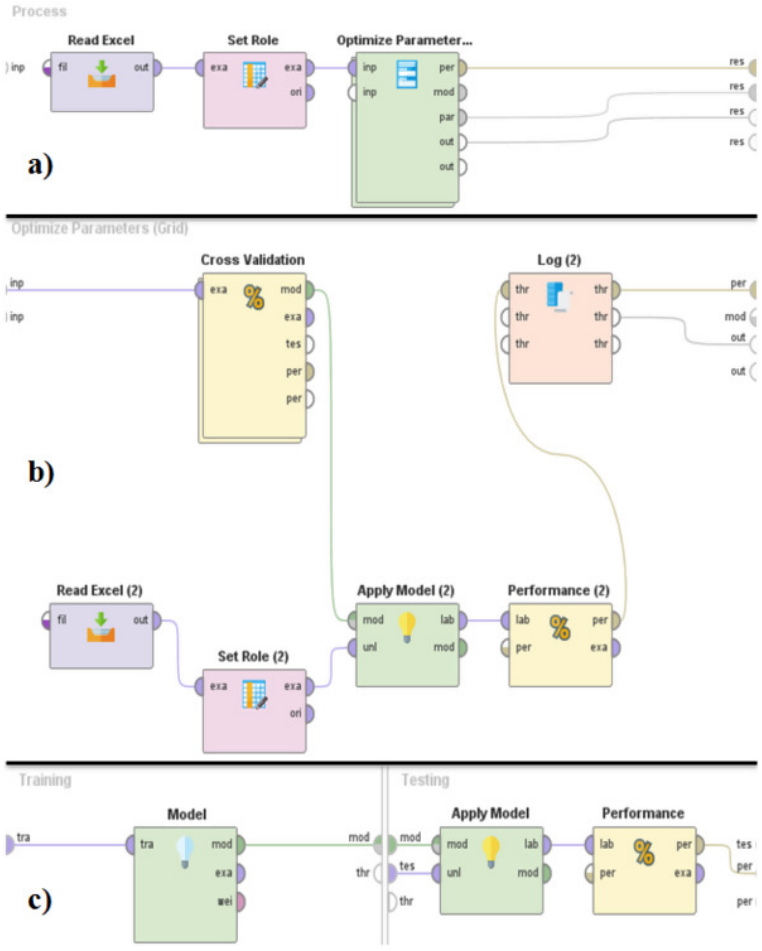


Figura 5.2: Arquitectura genérica de ajuste de hiper-parámetros para los modelos paramétricos y no paramétricos. (a) Hace referencia al proceso de ajuste de parámetros con *Optimize Parameters (Grid)*. (b) Hace referencia al proceso de validación cruzada en la evaluación de modelos con *Cross Validation*. (c) Muestra el proceso de entrenamiento y validación dentro de la validación cruzada y generación de métricas de rendimiento. Fuente: Elaboración propia.

En base a la selección de hiper-parámetros disponibles a optimizar y el procedimiento de validación cruzada presentados anteriormente, las alternativas de prueba disponibles para los hiper-parámetros por cada modelo y el valor seleccionado que optimiza cada uno de ellos,

luego del ajuste de parámetros y la validación cruzada, están dados en la Tabla 5.14.

Tabla 5.14: Pruebas de hiper-parámetros y selección de los valores óptimos para los modelos paramétricos y no paramétricos

Hiper-parámetro	Alternativas de prueba	Selección óptima[a]
	R.B.	
-	-	-
	R.N.A	
Número de capas ocultas	1	**1**
Número de neuronas	10 - 20 - 30 - 40 - 50 - 60 - 70 - 80 - 90 - 100	**20**
Ciclos de entrenamiento	50 - 60 - 70 - 80 - 90 - 100 - 110 - 120 - 130 - 140 - 150	**70**
Tasa de aprendizaje	0.0001 - 0.001 - 0.01 - 0.1	**0.1**
Momentum	0.5 - 0.6 - 0.7 - 0.8 - 0.9	**0.9**
	k-NN	
k número de vecinos	2 - 3 - 4 - 5 - 6	**3**
Voto ponderado	SÍ - NO	**NO**
Tipos de medida	Mixed Measures - Numerical Measures	**Mixed Measures (M.E.D.)**[b]
	R.F.	
Número de árboles	10 - 20 - 30 - 40 - 50 - 60 -70 - 80 - 90 - 100	**40**
Criterio de división	Information gain - Gain ratio	**Information gain**
Profundidad máxima	5 - 10 - 15 - 20	**15**
Estrategia de voto	Confidence vote - Majority vote	**Majority vote**
Ganancia mínima	0.001 - 0.01 - 0.1	**0.01**
Tamaño mínimo de hoja	1 - 2 - 3 - 4 - 5 - 6	**1**
Tamaño para división	1 - 2 - 3 - 4 - 5 - 6	**4**
Alternativas de pre-poda	1 - 2 - 3 - 4 - 5 - 6	**1**
	A.D.	
Criterio de división	Information gain - Gain ratio	**Information gain**
Profundidad máxima	5 - 10 - 15 - 20	**15**
Nivel de confianza	0.1 - 0.2 - 0.3 - 0.4 - 0.5	**0.4**
Ganancia mínima	0.001 - 0.01 - 0.1	**0.01**
Tamaño mínimo de hoja	1 - 2 - 3 - 4 - 5 - 6	**5**
Tamaño para división	1 - 2 - 3 - 4 - 5 - 6	**5**
Alternativas de pre-poda	1 - 2 - 3 - 4 - 5 - 6	**5**

Notas. R.B.: Regresión binomial, **R.N.A.**: Redes neuronales artificiales, **k-NN**: Algoritmo k-Nearest Neighbors, **R.F.**: Random forest, **A.D.**: Árbol de decisión. [a] Los hiper-parámetros óptimos considerados y evaluados fueron aquellos que cumplieron con la condición de no generar problemas de ajuste (*overfitting* o *underfitting*) dentro de los modelos paramétricos y no paramétricos y que generan los resultados con mejor bondad de ajuste durante la construcción de los modelos computacionales. [b] M.E.D.: *Mixed Euclidean Distance.*

Fuente: Elaboración propia.

En base a la anterior tabla, puede señalarse lo siguiente: la regresión binomial se realiza ajustando los datos de entrenamiento y la precisión balanceada (promedio de la validación cruzada) será aquella que represente el modelo sin posibilidad de que sea optimizada; en el caso de las redes neuronales artificiales, una sola capa oculta, 20 neuronas incluidas en esta, 70 ciclos de entrenamiento, una tasa de aprendizaje de 0.1 y un valor de momentum de 0.9 proveen el mejor rendimiento de este modelo no paramétrico tomando en cuenta las combinaciones propuestas; en relación al algoritmo k-NN, son 3 vecinos y una distancia euclídea los que generan el mejor desempeño condicionado a los límites de ajuste adecuados; en el caso del modelo Random Forest, 40 árboles, el criterio *Information Gain* para la división de atributos, una profundidad máxima de 15 para cada árbol, una estrategia de voto por mayoría, una ganancia mínima de 0.01 para nodo, el tamaño de división de nodos de 4 y 1 alternativa de pre-poda para los mismos; por último, en el caso del Árbol de decisión, el criterio de división

Information Gain para los atributos, una profundidad máxima de 15 del árbol, un nivel de confianza de 0.4, una ganancia mínima de 0.01 para cada nodo, el tamaño mínimo 5 de hojas, 5 como el tamaño de división y 5 alternativas de pre-poda permiten obtener el modelo con el mejor ajuste. Dichos modelos con tales especificaciones pueden ser generalizados al conjunto de prueba, obtenido en la Sección 5.3.4.2, para determinar de forma comparativa cuál es el que mejor permite realizar predicciones y clasificaciones sobre el nivel de conocimiento sobre el VIH/SIDA de adolescentes y jóvenes adultos en territorio nacional.

5.3.4.5. Comparación de los clasificadores para la predicción del conocimiento sobre el VIH/SIDA

A continuación, se reportan los resultados de los modelos paramétricos y no paramétricos construidos en la etapa de entrenamiento luego del proceso de afinación e identificación de los hiper-parámetros más idóneos para cada algoritmo propuesto en el estudio (regresión logística, redes neuronales artificiales, árboles de decisión, random forest y algoritmo k-nearest neighbors) en la Tabla 5.15.

Tabla 5.15: Comparación de los indicadores de desempeño de los modelos paramétricos y no paramétricos en el conjunto de entrenamiento

Algoritmo[a]		**R.B.**		**R.N.A**		**k-NN**		**R.F.**		**A.D.**	
		Predicción		**Predicción**		**Predicción**		**Predicción**		**Predicción**	
Matriz de confusión[b]		**SÍ**	**NO**	**SÍ**	**NO**	**SÍ**	**NO**	**SÍ**	**NO**	**SÍ**	**NO**
Observado	**SÍ**	3934	2502	3517	2919	2287	4149	3429	3007	3361	3075
	NO	2381	4055	1587	4849	732	5704	1587	4849	1553	4883
		%	+/-[c]	%	+/-	%	+/-	%	+/-	%	+/-
Precisión		62.06	1.27	64.99*	1.88	62.08	0.72	64.31	1.10	64.05	2.08
Sensibilidad		61.12*	1.38	54.64	16.00	35.53	1.36	53.28	6.52	52.22	6.32
Especificidad		63.00	1.94	75.34	17.76	88.63*	0.95	75.34	6.33	75.87	5.73
Error de clasificación		37.94	1.27	35.01*	1.88	37.92	0.72	35.69	1.10	35.95	2.08
Kappa de Cohen[d]		0.241	0.03	0.300*	0.04	0.242	0.01	0.286	0.02	0.281	0.04
F1 score		61.70*	1.22	60.05	5.91	48.37	1.33	59.68	3.38	59.05	3.76
Valor predicho positivo		62.31	1.16	68.91	5.56	75.75*	1.27	68.36	1.19	68.38	3.24
Valor predicho negativo		61.84	1.21	62.42*	5.74	57.89	1.31	61.72	1.32	61.38	3.27
AUC		66.30	0.02	72.60*	0.01	67.40	0.01	71.60	0.01	68.80	0.02

Notas. R.B.: Regresión binomial, **R.N.A.**: Redes neuronales artificiales, **k-NN**: Algoritmo k-Nearest Neighbors, **R.F.**: Random forest, **A.D.**: Árbol de decisión. [b] En el cálculo de la matriz de confusión, se considera "SÍ" a aquellos individuos con un nivel de conocimiento adecuado sobre el VIH/SIDA y "NO" a aquellos individuos con un nivel de conocimiento no adecuado sobre el VIH/SIDA. [c] Límites de la desviación estándar promedio de la predicción. [d] El resultado se presenta en términos absolutos. *: Mejor valor según indicador.

Fuente: Elaboración propia.

La Tabla 5.15 reporta, en primer lugar, las matrices de confusión asociadas a cada modelo. Se observa que la mayor identificación de la cantidad de verdaderos positivos fue por parte del modelo de regresión binomial (3,934 concordancias entre clases y predicciones) y la menor identificación se dio por parte del algoritmo k-nearest neighbors (2,287 concordancias). En la misma perspectiva, la mayor identificación de la cantidad de verdaderos negativos fue conseguida mediante el modelo k-NN (5,704 concordancias) y la menor identificación fue lograda por la regresión logística binomial (4,055 concordancias).

En cuanto a los indicadores de bondad de ajuste considerados en el presente estudio, se

establece lo siguiente: se subraya el hecho de que los valores de las métricas de comparación para el conjunto de entrenamiento son medias o cifras promedio de los resultados de cada iteración dentro de la etapa de la validación cruzada aplicada para la determinación de los mejores hiper-parámetros en cada algoritmo; fueron 10 repeticiones en total efectuadas para seleccionar los valores óptimos de las alternativas de prueba incluidas, por lo que las medidas de desempeño promedio y sus desviaciones estándar se reportan en la Tabla 5.15.

La precisión ligada al pronóstico de clasificaciones del nivel de conocimiento adecuado sobre el VIH/SIDA se encuentra entre los valores promedios de 62.06 % (+/- 1.27) y 64.99 % (+/- 1.88) generados en el conjunto de entrenamiento, considerando que el valor de precisión más alto fue alcanzado por el modelo de redes neuronales artificiales (R.N.A.) - una precisión de 64.99 %, lo que indica que dicho modelo es el que mejor para realizar predicciones del nivel de conocimiento en la población adolescente y joven adulta del país. De manera análoga, se puede señalar que dicho modelo es el que ostenta la menor tasa o porcentaje de error de clasificación entre los 5 algoritmos propuestos, con un valor promedio de 35.01 % (+/- 1.88).

En cuanto a la Sensibilidad, se puede puntualizar que el modelo de regresión logística tiene el porcentaje o proporción de predicciones positivas correctas entre el total de predicciones positivas más alto entre todos los modelos empleados (61.12 %, con una desviación estándar de +/- 1.38), lo que significa que es el algoritmo con mayor capacidad o precisión para identificar y pronosticar los casos de individuos con un nivel de conocimiento correcto o apropiado sobre el VIH/SIDA. Asimismo, el modelo con el menor desempeño tomando en cuenta a este indicador es el algoritmo k-NN, con una sensibilidad del 35.53 % (+/- 1.36).

Analizando a la Especificidad, se señala que el algoritmo k-NN es el modelo con el mejor desempeño en cuanto a esta métrica de bondad de ajuste, con un valor que asciende al 88.63 % (+/- 0.95), configurando a esta técnica como la que mejor pronostica los casos de individuos que poseen un nivel de conocimiento inadecuado o inexacto. En el mismo sentido, se determina que el modelo con el menor valor de especificidad es la regresión binomial, con una cifra que asciende al 63.00 % (+/- 1.94).

Por otro lado, el modelo que tuvo el mejor rendimiento tomando en consideración a los valores predichos positivos fue el algoritmo k-NN (con 75.75 % y +/- 1.27) y para los valores predichos negativos fue el modelo de redes neuronales artificiales (con 62.42 % y +/- 5.74). El primero identificó acertadamente, en mayor proporción, a los casos de conocimiento adecuado sobre el VIH/SIDA reales y el segundo a los casos de desconocimiento sobre el virus y la enfermedad reales. En el mismo orden de ideas, el modelo que tuvo el menor rendimiento registrado para el primer indicador fue el árbol de decisiones (con 62.31 % y +/- 1.16) y para el segundo fue el algoritmo k-NN (con 57.89 % y +/- 1.31).

Desde otro punto de vista, examinando métricas como el Kappa de Cohen y el F1-Score, se puede definir que el modelo de redes neuronales artificiales es aquel que presenta el mejor desempeño para la primera métrica y la regresión binomial presenta el mejor rendimiento para la segunda, con valores de 0.300 (+/- 0.04) y 61.70 % (+/- 1.22), correspondientemente.

Enfocándonos en el área debajo de la curva (AUC), entre los 5 modelos de paramétricos y no paramétricos empleados en este estudio, la curva de las redes neuronales artificiales muestran el valor de AUC más alto (una cifra de 72.60 % y +/- 0.01), lo que indica que es el algoritmo con la mejor capacidad discriminatoria para clasificar a las personas con un nivel de conocimiento adecuado y sujetos con un nivel notorio de desconocimiento sobre el VIH/SIDA, entre los algoritmos de la etapa de entrenamiento.

Con un importante énfasis, el factor más crítico que afecta el éxito de un aprendizaje automático o *machine learning* es el proceso de entrenamiento y prueba (Uçar *et al.*, 2020), como se mencionó el Sección 5.2. El propósito del entrenamiento fue determinar cuántos datos se aprenden para la precisión de los hiper-parámetros en cuestión, como en la Sección 5.3.4.4, revelándose ciertas métricas de desempeño asociadas a esta selección de parámetros y estructura de algoritmos descritas líneas arriba (indicadores referenciales para el proceso de construcción, empero no comparables). Sin embargo, una vez realizado el proceso de entrenamiento, la validación del comportamiento final del modelo de aprendizaje automático se efectúa mediante la generalización de este a los datos de prueba: los resultados obtenidos aquí representan el rendimiento concluyente e imparcial del modelo y no se pueden modificar más, son estos valores los que vislumbran la capacidad predictiva de los algoritmos y el potencial comparativo en el área de estudio.

Es bajo esa premisa que, a continuación, se reportan los resultados de los modelos paramétricos y no paramétricos construidos en la etapa de entrenamiento y aplicados al conjunto de prueba o validación para evaluar la capacidad de generalización y predicción de los algoritmos (regresión logística, redes neuronales artificiales, árboles de decisión, random forest y algoritmo k-nearest neighbors) en la Tabla 5.16.

Tabla 5.16: Comparación de los indicadores de desempeño de los modelos paramétricos y no paramétricos en el conjunto de prueba

Algoritmo[a]		**R.B.**		**R.N.A**		**k-NN**		**R.F.**		**A.D.**	
		Predicción		**Predicción**		**Predicción**		**Predicción**		**Predicción**	
Matriz de confusión[b]		**SÍ**	**NO**	**SÍ**	**NO**	**SÍ**	**NO**	**SÍ**	**NO**	**SÍ**	**NO**
Observado	**SÍ**	177	181	149	209	78	280	180	178	136	222
	NO	223	475	172	526	148	550	199	499	158	540
			%		%		%		%		%
Precisión			61.74		63.92		59.47		64.30*		64.02
Sensibilidad			49.44		41.62		21.79		50.28*		37.99
Especificidad			68.05		75.36		78.80*		71.49		77.36
Error de clasificación			38.26		36.08		40.53		35.70*		35.98
Kappa de Cohen[c]			0.170		0.174		0.006		0.215 *		0.161
F1 score			46.70		43.89		26.71		48.85*		41.72
Valor predicho positivo			44.25		46.42		34.51		47.49*		46.26
Valor predicho negativo			72.41		71.56		66.27		73.71*		70.87
AUC			62.90*		62.90*		48.90		61.20		60.20

Notas. R.B.: Regresión binomial, **R.N.A.**: Redes neuronales artificiales, **k-NN**: Algoritmo k-Nearest Neighbors, **R.F.**: Random forest, **A.D.**: Árbol de decisión. [b] En el cálculo de la matriz de confusión, se considera "SÍ" a aquellos individuos con un nivel de conocimiento adecuado sobre el VIH/SIDA y "NO" a aquellos individuos con un nivel de conocimiento no adecuado sobre el VIH/SIDA. [c] El resultado se presenta en términos absolutos. *: Mejor valor según indicador.

Fuente: Elaboración propia.

La tabla reporta, en primer lugar, las matrices de confusión (estructuras que contraponen las clases reales del conjunto de datos y las predicciones dadas por un algoritmo) asociadas a cada modelo. Se observa que la mayor identificación de la cantidad de verdaderos positivos (observaciones pertenecientes a la clase positiva y que fueron predichas como tal) fue por parte del modelo de Random Forest (178 concordancias entre clases y predicciones) y la menor identificación se dio por parte del algoritmo k-nearest neighbors (78 concordancias). En la misma perspectiva, la mayor identificación de la cantidad de verdaderos negativos (observaciones pertenecientes a la clase negativa y que fueron pronosticadas como tal) fue conseguida mediante el modelo k-NN (550 concordancias) y la menor identificación fue lograda por la regresión logística binomial (475 concordancias).

En cuanto a los indicadores de bondad de ajuste considerados en el presente estudio, se establece lo siguiente.

La precisión ligada al pronóstico de clasificaciones del nivel de conocimiento adecuado sobre el VIH/SIDA se encuentra entre los límites moderados reportados en el ámbito de las ciencias de la salud y la investigación médica (Bitew *et al.*, 2020; Amusa *et al.*, 2020; Adegbosin *et al.*, 2020; Talukder & Ahammed, 2020), con valores entre 59.47 % y 64.30 % generados en el conjunto de validación, considerando que el valor de precisión más alto (64.30 %) fue alcanzado por el modelo de Random Forest (R.F.), lo que indica que dicho modelo es el que mejor para realizar predicciones del nivel de conocimiento en la población adolescente y jovén adulta del país. De manera análoga, se puede señalar que dicho modelo es el que ostenta la menor tasa o porcentaje de error de clasificación entre los 5 algoritmos propuestos, con un valor de 35.70 % (ya que el error de clasificación se calcula mediante la resta de la unidad y la precisión).

En cuanto a la Sensibilidad, se puede puntualizar que el modelo de Random Forest tiene el porcentaje o proporción de predicciones positivas correctas entre el total de predicciones positivas más alto entre todos los modelos empleados (50.28 %), lo que significa que es el algoritmo con mayor capacidad o precisión para identificar y pronosticar los casos de individuos con un nivel de conocimiento correcto o apropiado sobre el VIH/SIDA (la clase positiva en la investigación). Asimismo, el modelo con el menor desempeño tomando en cuenta a este indicador es el algoritmo k-NN, con una sensibilidad del 21.79 %.

Analizando a la Especificidad (proporción o ratio de predicciones negativas correctas entre el total de predicciones negativas), se señala que el algoritmo k-NN es el modelo con el mejor desempeño en cuanto a esta métrica de bondad de ajuste, con un valor que asciende al 78.80 %, configurando a esta técnica como la que mejor pronostica los casos de individuos que poseen un nivel de conocimiento inadecuado o inexacto. En el mismo sentido, se determina que el modelo con el menor valor de especificidad es la regresión binomial, con una cifra que asciende al 68.05 %. Cabe recalcar que, pese a que el algoritmo k-NN exhibe la mejor capacidad de determinación de los casos negativos, el objetivo principal de la construcción de los modelos de clasificación es generar una forma de predecir los casos positivos, aquellas personas con un nivel de conocimiento correcto, de forma eficiente para el diseño y despliegue

de políticas públicas, por lo que este indicador no influye de manera considerable en la elección del algoritmo idóneo.

Por otro lado, el modelo que tuvo el mejor rendimiento tomando en consideración a los valores predichos positivos (predicciones positivas correctas entre el total de observaciones positivas) y valores predichos negativos (predicciones negativas correctas entre el total de observaciones negativas) fue el modelo de Random Forest. Dicho algoritmo identificó acertadamente al 47.49 % de los casos de conocimiento adecuado sobre el VIH/SIDA reales y 73.71 % de los casos de desconocimiento sobre el virus y la enfermedad reales, elementos que dan indicios de que el modelo de Random Forest es relativamente mejor para predecir casos reales positivos y negativos. En el mismo orden de ideas, el modelo que tuvo el menor rendimiento registrado considerando ambos indicadores fue el algoritmo k-NN, con valores de 34.51 % y 66.27 % para los valores predichos positivos y negativos, respectivamente.

Desde otro punto de vista, examinando métricas como el Kappa de Cohen y el F1-Score, se puede definir que el modelo de Random Forest es aquel que presenta el mejor desempeño en ambos indicadores, con valores de 0.215 y 48.85 %, correspondientemente. Estos resultados sugieren, primeramente, que el Random Forest posee la mayor proporción de comprensión entre la predicción y las clasificaciones reales en el conjunto de datos (cumpliendo el supuesto de que el K de Cohen sea menor o igual a la unidad), con un grado de comprensión justo (al estar contenido entre los rangos de 0.21 y 0.40) de acuerdo con la escala de Landis & Koch (1977). Del mismo modo, los hallazgos plantean que el R.F. es el modelo que funciona mejor en la clase positiva entre todos los construidos para la investigación y el que más se acerca al mejor ajuste entre precisión y sensibilidad (F1-Score = 1). En contraste, el modelo que configura el menor rendimiento en cuanto al Kappa de Cohen y el F1-Score es el algoritmo k-NN con un valor de K de 0.006 (con una comprensión cercana a la nulidad) y un valor-F de 26.71 %.

Enfocándonos en el área debajo de la curva (AUC), entre los 5 modelos de paramétricos y no paramétricos empleados en este estudio, las curvas de los modelos de regresión binomial y de redes neuronales artificiales muestran el valor de AUC más alto (una cifra de 62.90 % para ambos casos), lo que indica que son los mejores para clasificar los casos de conocimiento apropiado y marcado desconocimiento sobre el VIH/SIDA, entre los algoritmos. Sin embargo, es necesario remarcar que, pese a que el valor de AUC para los modelos de regresión binomial y redes neuronales sea el mayor entre las técnicas consideradas, el modelo de Random Forest ofrece el segundo mejor valor para dicho indicador (siendo concretamente similar al valor de 62.90 %, con una diferencia nimia de 1.70 % a favor del primero) y, en la misma línea, para niveles altos de especificidad, este mismo modelo obtiene niveles moderados de sensibilidad: un aspecto deseable dentro de la capacidad predictiva de un algoritmo, a diferencia de la compensación que hacen la regresión binomial y la R.N.A. donde se asumen altas tasas de falsos positivos para poder realizar predicciones correctas en cuanto a la sensibilidad; aspecto desproporcionado que influencia fuertemente la obtención de un mayor valor de AUC por parte de estos modelos en contraste con el R.F., debido a que raramente se exploraría y

trabajaría en las zonas de la curva ROC en donde estos modelos generan su mejor desempeño y desestiman el ajuste de estos como mencionan Lobo *et al.* (2008).

Es así como los indicadores de clasificación presentan su mejor desempeño (exceptuando el valor de la AUC, que resulta aún favorable de forma comparativa) en el modelo de Random Forest. Dicho algoritmo muestra un mayor poder predictivo en contraste con los otros modelos paramétricos y no paramétricos incluidos en este estudio.

Finalmente, es necesario comprobar la importancia de las características socio-demográficas, económicas y de salud consideradas en el estudio con respecto a su poder predictivo e influencia dentro del desempeño de los algoritmos propuestos basándonos en el procedimiento que fue descrito en la Sección 5.2.8.

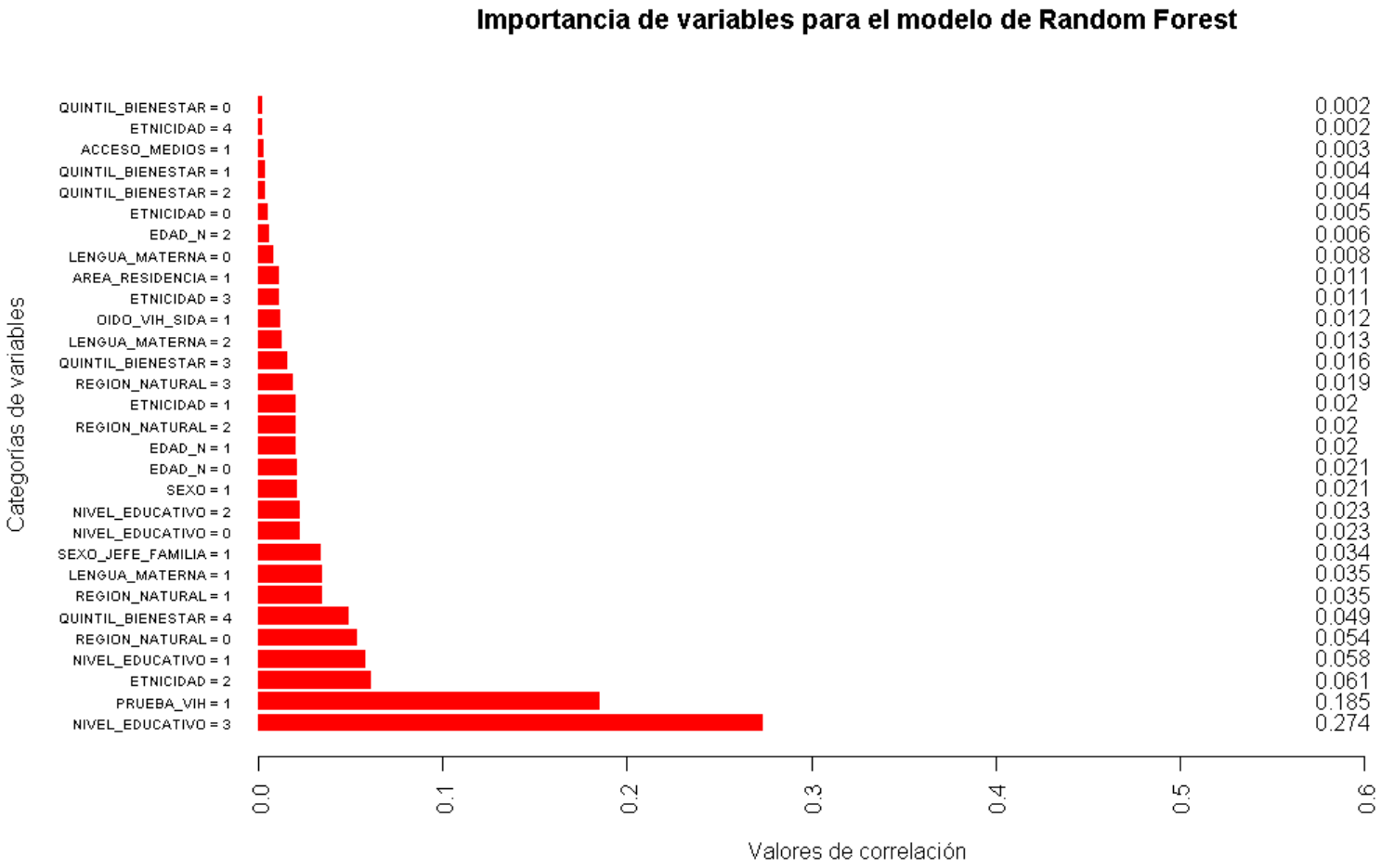


Figura 5.3: Medidas de influencia de las variables o regresores independientes en la predicción del nivel de conocimiento sobre el VIH/SIDA en la población peruana en el modelo de Random Forest. Fuente: Elaboración propia.

La Figura 5.3, presentada anteriormente, muestra las medidas de influencia, calculadas mediante el operador *Explain Predictions* del software Rapidminer, de los niveles/categorías de todas las variables empleadas en el desarrollo del modelo Random Forest, considerando que fue el algoritmo que presentó el mejor desempeño predictivo para estimar la variable respues-

ta (conocimiento del VIH/SIDA) con una precisión final del 64.30 %. De la misma manera, la sección C del capítulo de Anexos muestra los valores de correlación que representan la importancia o el impacto que tienen cada nivel o categoría de los covariables considerados en el estudio para el resto de modelos paramétricos y no paramétricos que registraron menores capacidades o potenciales de clasificación del nivel de entendimiento sobre la epidemia en los adolescentes y jóvenes adultos del país.

Debe destacarse que estos valores resultan una forma efectiva de establecer el grado de influencia de las características sobre la predicción (ya sea que apoyen favorablemente la formación de una predicción correcta de la variable respuesta o contradigan el resultado de la estimación que pueda obtenerse dada una técnica propuesta) del conocimiento adecuado sobre el VIH/SIDA en los individuos en todas las simulaciones del proceso de validación cruzada al considerar que las cifras que hacen referencia a la correlación local (representando su magnitud a través de barras rojas con sus valores en el eje X de los gráficos) son producto de la identificación y cálculo de los pesos o ponderaciones de la importancia local de los atributos (variables) bajo una relación de vecindad que se forma con la variable de salida en cada modelo (Mierswa & Klinkenberg, 2018), tomando en cuenta que valores positivos indican que los atributos apoyan a una correcta estimación y valores negativos señalan atributos contradictorios relacionados con las predicciones correctas. Asimismo, aunque la relación entre atributos y predicciones puede ser muy no lineal a nivel global, la relación lineal local por atributo es lo suficientemente poderosa para explicar las predicciones (Mierswa & Klinkenberg, 2018).

Tomando en cuenta los modelos generados y el conjunto de datos de entrenamiento y validación durante el proceso de validación cruzada, se puede establecer que todas las influencias de las categorías incluidas en los algoritmos son mayores a cero (> 0) en todos los casos, lo que establece que ninguna categoría o nivel de alguna variable contradice la formación de una predicción correcta dentro de los algoritmos (Mierswa & Klinkenberg, 2018). De la misma manera, se debe precisar que es poco probable que aquellas categorías de variables que poseen una correlación local menor a 0.01 % contribuyan a la estimación correcta del nivel de conocimiento sobre el VIH/SIDA en la población objetivo (como en el caso del factor quintil_bienestar=3 del modelo de Regresión Logística con un valor de correlación local de 3×10^{-4}), por lo que prescindir de ellas, si el investigador lo considera necesario, no perjudicará el desempeño de los modelos - no obstante, conservarlas tampoco afecta en algo la capacidad de predictiva de los mismos (Mierswa & Klinkenberg, 2018). Aquellos niveles que superen este valor (> 0.01) serán considerados como empíricamente positivos (Mierswa & Klinkenberg, 2018), por lo que se especifica que estos apoyan a las predicciones positivas en las observaciones relacionadas al nivel de conocimiento adecuado sobre el VIH/SIDA en los adolescentes y jóvenes adultos en el Perú.

De la misma manera, pese a que los valores de correlación local varían entre cada modelo desarrollado (como las figuras permiten apreciar, los gráficos están bajo la misma escala en el eje X pero presentan diferentes magnitudes de cada categoría de variable) siendo el modelo de Regresión Logística la que presenta las mayores correlaciones positivas y el modelo de Redes

Neuronales Artificiales las menores entre todos los algoritmos propuestos (considerando aun así que todos poseen, en amplia mayoría, categorías o niveles positivos con correlaciones altas que apoyan a la formulación de predicciones correctas en el estudio), se puede precisar que los niveles con correlaciones locales entre 0.01 y 40.00 % son lo más factibles y beneficiosos de emplear dentro del modelamiento de las técnicas ya que producen buenos desempeños y permiten estimar la variable respuesta (Mierswa & Klinkenberg, 2018). Es así como la mayoría de las categorías en todos los modelos resultan ser atributos con correlaciones locales positivas: en el caso del Decision Tree, desde el factor edad=0 hasta el nivel_educativo=3; en el caso del algoritmo k-NN, desde la lengua_materna=3 hasta el factor quintil_bienestar=0; considerando el modelo de Random Forest, desde el quintil_bienestar=0 hasta el nivel_educativo=3; analizando el modelo de Regresión Logística, desde el factor quintil_bienestar=3 hasta el quintil_bienestar=1; y, finalmente, tomando en cuenta al modelo de Redes Neuronales Artificiales, desde el factor quintil_bienestar=0 hasta el sexo_jefe_familia=1. Si bien, en el caso de la Regresión Logística, el correlación del factor nivel_educativo=0 tiene un valor mayor a 40.00 % (con un valor de 56.70 %) y podría decirse que presenta una alta correlación positiva con la variable respuesta, en estas situaciones un valor de correlación como este puede ser un indicador de información que un algoritmo puede perder en el momento de la predicción y se sugiere la eliminación del factor en ciertos casos (Mierswa & Klinkenberg, 2018) (no obstante, si el problema de predicción es simple puede obtenerse un mejor modelo cuando se conserva la categoría).

A su vez, puede señalarse que existe una similitud de las variables que encabezan el ranking de los valores de correlación local asociadas a ellas y que juegan el papel más importante en la formación de las predicciones correctas del nivel de conocimiento sobre el VIH/SIDA en los gráficos de influencia de factores de los algoritmos (aspecto que se produce cuando se detecta la repetición apreciable de un mismo factor en las primeras posiciones de correlación local en todos los modelos - en este caso, se repiten categorías o la presencia mayoritaria de ciertas variables en todos los modelos en las primeras posiciones de correlaciones positivas locales): los cofactores "Nivel económico" (quintiles de riqueza), "Lengua materna", "Nivel educativo", "Región natural", "Etnicidad" y "Prueba de detección del VIH/SIDA" son los que figuraron de manera reiterativa a lo largo de las 10 primeras posiciones en la comparación de influencia de atributos en los 5 modelos. Con mayor precisión, niveles como realización efectiva de la prueba de descarte del VIH/SIDA (Prueba = 1), nivel educativo muy rico (Quintil = 4), nivel educativo superior a más (Nivel educativo = 3) y lengua materna castellana (Lengua materna = 1). Otros factores importantes que se deben examinar son otras categorías de las covariables como etnia afroperuana (Etnicidad = 2), las regiones naturales en las que habitan los individuos (Región natural = 0, 1 y 2), el área de residencia urbana (Área = 1) y género del jefe de familia masculino (Género de jefe de familia = 1).

Capítulo 6

Análisis de conglomerados sociales del VIH/SIDA

Este capítulo exhibe, mediante el análisis y asociación de diferentes determinantes de la salud que describen a una muestra entrevistada de individuos entre 15 y 29 años en el Perú a través de un mapa auto-organizado (SOM, en inglés), la identificación y caracterización de conglomerados o agrupaciones de encuestados en torno, principalmente, al nivel de conocimiento sobre las formas de prevención y rechazo de ideas erróneas sobre la transmisión del VIH/SIDA que posean y a los atributos que componen al conjunto de ciudadanos considerados.

Seguidamente, en base a los clusters distinguidos producto de la red SOM o mapa de Kohonen, se procede a ubicar geográficamente a los individuos que forman parte de la base de datos a lo largo del plano nacional, bajo cortes regionales, mediante coordenadas de latitud y longitud a fin de generar mapas de calor que permitan reconocer las concentraciones a nivel departamental más notorias del entendimiento sobre el VIH/SIDA de las personas y atributos relacionados a dicha región y evaluar el estado de la concientización sobre la epidemia en el país.

En lo que se refiere a los datos empleados en esta etapa, los registros correspondientes a las características socio-demográficas, económicas y de salud de los adolescentes y jóvenes adultos y su geolocalización dentro del territorio nacional forman parte de la unidad de análisis de la "Encuesta Demográfica y de Salud Familiar 2019" (INEI, 2019a) y los datos geográficos del plano nacional y los límites regionales del Perú obtenidos a través del Instituto Geográfico Nacional (MINDEF - IGN, 2020). A su vez, el modelo de Mapa Auto-organizado (SOM) para la agrupación de individuos con atributos similares dentro del conjunto de encuestados se realizó mediante el sistema de cómputo numérico MATLAB (MATLAB, 2010) y la representación geográfica y concentración de conglomerados en el Perú se efectuó a través del lenguaje de programación Python (Van Rossum & Drake, 2009).

6.1. Bases de datos

Para la aplicación del modelo de clusterización y creación de mapas de calor, consideraremos dos fuentes de datos relevantes para el estudio: (a) el conjunto secundario de datos unificado, filtrado y acabado de adolescentes y jóvenes adultos entre 15 y 29 años obtenido de la preparación y análisis de regresión logística del Capítulo 5 basado en la "Encuesta Demográfica

y de Salud Familiar 2019" (INEI, 2019a) y (b) los datos geográficos del plano nacional y los límites regionales del Perú obtenidos a través del Instituto Geográfico Nacional (MINDEF - IGN, 2020).

6.1.1. Encuesta Demográfica y de Salud Familiar (ENDES)

La presente investigación subyace a un sub-análisis de la Encuesta Demográfica y de Salud Familiar (ENDES) llevada a cabo en el Perú en el año 2019 por parte de Instituto Nacional de Estadística e Informática en el territorio nacional (INEI, 2019a).

La encuesta ENDES es un producto estadístico de diseño muestral complejo compuesta por una muestra de 36,760 viviendas (correspondiendo 14,760 viviendas al área sede, 9,340 al resto urbano y 12,660 al área rural) y por 33,396 mujeres y hombres entrevistados de 15 y más años en edad oficial reproductiva. Sin embargo, la muestra a evaluar en el análisis de conglomerados es la que corresponde a los adolescentes y jóvenes adultos entre 15 y 29 años que hayan respondido completamente a la encuesta de salud y la general, lo que representa a 10,565 individuos.

Además de las especificaciones de diseño muestral necesarias consideradas en la Sección 5.1.2; se requiere extraer del conjunto original de la ENDES aquellas variables de análisis referidas a la ubicación geográfica de los encuestados para poder habilitar el procesamiento cartográfico de los resultados del clustering. La síntesis de dicho proceso de extracción para obtener el conjunto de datos a emplear se presenta a continuación en la Tabla 6.1.

Tabla 6.1: Variables incluidas en el conjunto de datos de la ENDES para la identificación de conglomeración y ubicación geográfica

Utilización/Variable	Contenido	Base de datos (código de la base)
Variables de análisis		
NIVEL CONOCIMIENTO	Nivel de conocimiento sobre VIH/SIDA	ENDES 15-29[a]
SEXO	Género	ENDES 15-29
QUINTIL BIENESTAR	Nivel económico	ENDES 15-29
REGION NATURAL	Región Natural	ENDES 15-29
AREA RESIDENCIA	Área de residencia	ENDES 15-29
EDAD N	Rango etario	ENDES 15-29
NIVEL EDUCATIVO	Nivel educativo	ENDES 15-29
LENGUA MATERNA	Lengua materna	ENDES 15-29
ETNICIDAD	Etnicidad	ENDES 15-29
OIDO VIH SIDA	Ha oído acerca del VIH/SIDA	ENDES 15-29
PRUEBA VIH	Prueba de VIH/SIDA	ENDES 15-29
ACCESO MEDIOS	Radio en el hogar	ENDES 15-29
SEXO JEFE FAMILIA	Género del jefe de familia	ENDES 15-29
Variables geográficas		
ubigeo	Ubigeo del encuestado	Hogar (RECH0)
longitudx	Coordenada de longitud del hogar	Hogar (RECH0)
latitudy	Coordenada de latitud del hogar	Hogar (RECH0)
HV024	Departamento de la vivienda entrevistada	Hogar (RECH0)

Notas. [a] Base de datos obtenida en el Capítulo 5.

Fuente: Elaboración propia.

Las variables de interés para el análisis de conglomerados son: el nivel de conocimiento sobre

las formas de transmisión y naturaleza del VIH/SIDA que los encuestados posean (NIVEL CONOCIMIENTO), género de los encuestados (SEXO), nivel económico o distribución de riqueza (QUINTIL BIENESTAR), región natural habitada actualmente (REGION NATURAL), área de residencia (AREA RESIDENCIA), rango etario de los encuestados (EDAD N), nivel educativo más alto alcanzado (NIVEL EDUCATIVO), lengua materna o primaria (LENGUA MATERNA), auto-percepción étnica de los encuestados (ETNICIDAD), si el encuestado ha oído acerca del VIH/SIDA con anterioridad (OÍDO VIH SIDA), realización de la prueba de descarte del VIH/SIDA (PRUEBA VIH), si el encuestado cuenta con acceso a medios de comunicación multimedia en sus hogares (ACCESO MEDIOS) y, por último, género del jefe de hogar o familia (SEXO JEFE FAMILIA).

Las variables de interés para el análisis geográfico como las detalla la Tabla 6.1 son: el código de ubicación geográfico o ubigeo que emplea el INEI para codificar las circunscripciones territoriales del Perú (ubigeo), la medida cartográfica de longitud que forman parte de las coordenadas territoriales del hogar del encuestado (longitudx), la medida cartográfica de latitud que forman parte de las coordenadas territoriales del hogar del encuestado (latitudy) y el departamento en donde se encuentra la vivienda entrevistada (HV024).

6.1.2. Datos cartográficos gubernamentales del Perú

Los experimentos de clasificación de este capítulo se llevaron a cabo mediante los datos cartográficos nacional y regionales de Perú. Los datos tabulares se utilizaron en el análisis como base de datos de las regiones. Existen dos tipos de mapas que se utilizarán en el análisis: (a) datos poligonales del contorno del plano del territorio nacional del Perú y (b) datos poligonales del mapa regional o departamental del país.

En cuanto a los datos poligonales del contorno del plano territorial, este mapa físico de Perú contiene datos poligonales de los bordes territoriales y la forma del país (MINDEF - IGN, 2020).

Por otro lado, tomando en consideración los datos poligonales del mapa regional o departamental del país, el mapa regional contiene datos poligonales de las fronteras administrativas de los departamentos y su tabla de atributos. Este archivo de datos es en realidad un mapa basado para efectuar análisis temáticos (MINDEF - IGN, 2020).

6.2. Metodología

El propósito de esta sección es presentar la metodología de investigación necesaria para desarrollar y justificar el estudio del conglomerados sociales del VIH/SIDA en el Perú del presente capítulo. Se precisará la aplicabilidad de la teoría o los aspectos vinculados a la literatura para explicar por qué se están utilizando ciertos métodos/técnicas, procedimientos y criterios para el análisis de los datos y el fundamento académico de las elecciones dadas en cada subsección de los resultados del estudio propuesto.

6.2.1. Preparación de variables de estudio para el análisis de conglomerados

Considerando el diseño y aplicación de la red SOM para el presente estudio, las variables de entrada o independientes a utilizar en el modelo son la siguientes: "Nivel de conocimiento sobre el VIH/SIDA", "Género", "Área de residencia", "Nivel educativo", "Región natural", "Nivel económico", "Rango etario", "Etnicidad", "Ha oído acerca del VIH/SIDA con anterioridad", "Prueba de detección del VIH/SIDA", "Acceso a medios multimedia de información", "Lengua materna" y, finalmente, "Género del jefe de hogar o familia". Todos los factores a ser suministrados a la red de Kohonen son de naturaleza o tipo categórica o no métrica, debido a que están compuestas por un número limitado de valores que se expresan en categorías o niveles nominales.

Sin embargo, como ocurre en el caso de ciertos algoritmos no paramétricos o de *machine learning*, el mapa auto-organizado de Kohonen no admite variables que no sean del tipo numérico o que no estén expresadas en esos términos, ya que no es capaz de procesarlas a fin de generar un resultado. Esto significa que los datos categóricos deben codificarse o pasar por una conversión antes de que puedan ser usados para ajustar y evaluar los resultados de un modelo.

Principalmente, se debe tomar en cuenta que la red SOM está diseñada para representar datos en los que la magnitud de los valores tiene significado. En tal caso, los datos categóricos se pueden incorporar asignándolos a datos numéricos: esto se hace etiquetando cada muestra con un número, asegurándose de que las etiquetas numéricas estén bajo un sentido lógico y no se asignen arbitrariamente (Clark, 2018).

Es por este precedente que, con el fin de emplear las variables mencionadas anteriormente en la construcción de conglomerados a partir de un mapa auto-organizado, sustituimos las etiquetas nominales de los factores por codificaciones numéricas que representarán a los niveles que las definen.

6.2.2. Estadísticas descriptivas de los factores involucrados en el análisis de conglomerados

Para obtener información sobre los factores involucrados en la generación y análisis de conglomerados a través de una red SOM, el análisis univariado a llevar a cabo hace referencia a la exploración de datos cuantitativos y cualitativos que proporcionarán descripciones de dichas variables individuales que resultan importantes o relevantes para su uso en pruebas más avanzadas y que facilitarán la delimitación posterior de la estructura y desarrollo de los subsiguientes análisis bivariados y multivariados a realizar en la investigación (Sandilands, 2013).

La aplicación de un análisis cuantitativo relativamente simple pero fundamental como el univariado será necesaria para resumir o describir una variable a la vez en todos los casos, facilitando la interpretación de los datos y la comprensión de cómo se distribuirán dentro de nuestra muestra de estudio (Sandilands, 2013). Permitiendo filtrar los factores y evaluar si

estos cumplen con los supuestos o criterios requeridos para desarrollar un modelo complejo como el mapa de Kohonen. El tipo principal de análisis univariado a emplear en este capítulo serán los cálculos a través de frecuencias de las variables: la observación de eventos o resultados en los datos, así como el número de veces que se produce una situación en particular (Sandilands, 2013).

6.2.3. Parámetros de entrenamiento de la red SOM

En primera instancia, para entrenar una red SOM se deben definir dos aspectos importantes: las relaciones topológicas dentro de la estructura del mapa y los criterios de entrenamiento para inicializar y desarrollar el modelo.

Tomando en consideración el primer aspecto, en la red SOM clásica, el número de neuronas y sus relaciones topológicas están fijas desde el principio. Hay cuatro cuestiones que deben decidirse: el número de neuronas, las dimensiones de la cuadrícula del mapa, la red del mapa y la forma (Vesanto *et al.*, 2000).

El número de neuronas generalmente debe seleccionarse lo más grande posible, con el tamaño de la vecindad controlando el suavizamiento y generalización del mapeo (Vesanto *et al.*, 2000). Sin embargo, este no sufre considerablemente incluso cuando el número de neuronas excede el número de vectores de entrada, si se selecciona apropiadamente el tamaño de la vecindad. Por otro lado, a medida que aumenta el tamaño del mapa, la fase de entrenamiento se vuelve computacionalmente impracticable para la mayoría de las aplicaciones. Del mismo modo, la forma de la cuadrícula del mapa debe corresponder a la forma de la variedad de datos (Vesanto *et al.*, 2000). Para el mapa en forma de hoja predeterminado (*sheet shaped*), se recomienda que la longitud del lado a lo largo de una dimensión sea más larga que las otras, para que el mapa pueda orientarse correctamente. La elección de la estructura y el tamaño del mapa está relacionada tanto con el tipo de problema como con la elección subjetiva del usuario (Vesanto *et al.*, 2000). Por lo general, se recomienda el uso de la disposición hexagonal. Así, los mapas se vuelven más suaves y agradables a la vista.

De la misma manera, los criterios de inicialización y desarrollo de un mapa de Kohonen son los siguientes: parámetro de inicialización del proceso de aprendizaje de la red, las fases de entrenamiento del mapa, la medición del desempeño durante el entrenamiento y la cantidad de iteraciones que determinarán la duración del aprendizaje en el modelo.

En la inicialización del proceso de aprendizaje, los vectores de peso se pueden incorporar de diversas formas. Antes del entrenamiento, se dan valores iniciales a los vectores prototipo. La red SOM es robusta con respecto a la inicialización, pero si se logra correctamente, permite que el algoritmo converja más rápido en una buena solución (Akinduko & Mirkes, 2012). El criterio a seguir para la inicialización del mapa en cuestión está dado por la inicialización lineal, la cual se realiza seleccionando una malla de puntos del cubo d-dimensional min-max de los datos de entrenamiento (Akinduko & Mirkes, 2012). Los ejes de la malla son los autovectores (que se pueden calcular usando el procedimiento de Gram-Schmidt) correspondientes a los m

valores propios más grandes de los datos de entrenamiento (m es la dimensión de la cuadrícula del mapa) (Akinduko & Mirkes, 2012).

Por otra parte, el entrenamiento de una red se realiza en dos fases: entrenamiento aproximado con un radio (inicial) de vecindario grande y una tasa (inicial) de aprendizaje grande y, posteriormente, un ajuste fino con un radio y una tasa de aprendizaje pequeños. Para dicho fin, el entrenamiento a más idóneo a aplicar en una red neuronal es el algoritmo de reglas de aprendizaje de peso y sesgo con actualizaciones por lotes. La elección se deriva del hecho de que este algoritmo permite cálculos mucho más veloces que otros los tipos de aprendizaje, independiente de la dimensión de los datos de entrada y la topología de la red, y los resultados suelen ser buenos o incluso mejores que los demás (Akinduko & Mirkes, 2012). Basándonose en que la información de las ponderaciones anteriores está en la BMU (*Best Matching Unit*).

A su vez, considerando el desempeño y confianza en la red de Kohonen durante la etapa de entrenamiento, uno de los dos objetivos principales de los mapas auto-organizados es cuantificar el espacio de datos en un número finito de los centroides (o vectores de código), conocido como cuantificación vectorial (de Bodt *et al.*, 2002), para la determinación del desempeño del mapa durante el entrenamiento. La medición de este desempeño se realizará bajo un indicador que consiste en la distancia al cuadrado entre un dato observado x_i y su centroide correspondiente (más cercano) conocido como el error de cuantificación cuadrático o MSE. La suma de este error de cuantificación sobre todos los datos conduce a la distorsión o suma de cuadrados intraclase. Al medir el MSE, se puede evaluar la mejora o el aprendizaje del algoritmo. Para rastrear el desarrollo del entrenamiento del algoritmo, se calcula el Error Cuadrático Medio (MSE) después de cada época. El MSE es el error cuadrado entre cada punto de datos y su neurona más cercana (la mejor neurona coincidente), sumado para todos los puntos de datos dividido por el número total de datos (Montzka, 2018). Para el SOM, el cálculo del MSE se obtiene de la siguiente expresión matemática:

$$MSE = \frac{1}{N}\sum_{j=1}^{N} \min_{k}(x_j - W_k)^2 \tag{6.1}$$

donde W_k es el vector de peso para la neurona k con la distancia mínima al punto de datos, x_j es un punto de dato j y N es el número total de puntos de datos ($1 \leq j \leq N$). El MSE aporta una buena descripción teórica sobre el proceso de aprendizaje (Dvorský, 2018).

Evaluando el parámetro referido al número de iteraciones o reproducciones que medirá la duración del entrenamiento en una red SOM, se puede señalar que la duración de este del mapa se mide en épocas: una época corresponde a una pasada a través de los datos provistos (Liu *et al.*, 2018). El número de épocas es directamente proporcional a la relación entre el número de unidades de mapa (m) y el número de muestras de datos (n), expresada como $\frac{m}{n}$. Para cada época, todos los vectores (o secuencias) de entrenamiento se presentan cada uno una vez en un orden aleatorio diferente con la red y los valores de peso y sesgo actualizados después de cada presentación individual (Liu *et al.*, 2018). El proceso de entrenamiento de

una red se repite hasta que se alcanza el número especificado de épocas.

6.2.4. Evaluación de la topología óptima para la red SOM

Una red neuronal de Kohonen consta de varias neuronas. Cada neurona está representada por un vector de peso que tiene la misma dimensión de los datos de entrenamiento (Tian *et al.*, 2014). Las neuronas se organizan de acuerdo con su similitud, donde las neuronas con los vectores de peso similares se agrupan como vecinas. Esta relación de vecindad describe la estructura del mapa, que refleja la relación en los datos de entrenamiento (Tian *et al.*, 2014).

En ese sentido, el número de neuronas debe decidirse antes de que la red pueda comenzar a procesar los datos de entrada (Valova *et al.*, 2013). Dado que la red SOM aprende sin una señal de enseñanza, no se asume ningún conocimiento a priori del conjunto de datos. Esto significa que no tenemos información sobre cuántos puntos de datos hay o cómo se pueden distribuir estos puntos de datos (Valova *et al.*, 2013). Esto, a su vez, significa que no hay información disponible sobre cuántas neuronas podrían ser óptimas.

Para obtener el número de neuronas recomendado $M_{neuronas}$ con el que se inicializaría la red, se debe aplicar una raíz cuadrada al valor M y aproximarlo al entero más cercano (Tian *et al.*, 2014) en la Ecuación 3.19. No obstante, elegir el número correcto de nodos puede resultar un proceso complicado. Demasiados nodos aumentan el cálculo y es posible que no logren una reducción/simplificación de datos suficiente, pero muy pocos nodos pueden no proporcionar un ajuste suficiente a los datos (Xia, 2017). El mínimo a tener presente es una cantidad de dos nodos, lo que conduce a una clasificación binaria, es decir, los vectores de entrada se asignarán a uno de los dos nodos. La elección idónea del número de neuronas dependerá de un proceso de prueba y evaluación, en donde se generará el mapa SOM y se evaluará de forma visual y descriptiva la calidad de la separación realizada por el modelo a fin de precisar una topología óptima.

6.2.5. Análisis de los resultados del mapa auto-organizado de Kohonen

En la primera etapa del proceso de análisis de agrupamiento, los individuos se asignarán al azar a la matriz que se generó automáticamente (Akçapınar *et al.*, 2014). Luego, la distancia entre el siguiente individuo y los primeros designados se calculará y el siguiente individuo se asignará a la celda más cercana en la matriz. Este proceso continuará hasta que todas las observaciones sean designadas a un grupo individual o nodo. Cada nodo será un vector con un número de elementos que coincide con la cantidad de regresores correspondiente al número de características o variables. Después del entrenamiento, será posible identificar las *Best Matching Units* (BMU) en el mapa. Cada celda hexagonal representará una neurona y el número en una celda será el número de veces que la neurona se ha convertido en una BMU o el número de aciertos en una red.

De la misma manera, luego de la construcción de la estructura de la red y la conformación de los nodos o neuronas, el SOM se puede utilizar de manera eficiente en la visualización

de datos. El uso de estas visualizaciones provee una idea de la estructura y las correlaciones de los datos subyacentes (Qiana *et al.*, 2019). Bajo estas separaciones y uniones visualmente apreciables, se pueden investigar los indicios relevantes de los diferentes conglomerados de individuos similares que se pueden formar, después de haber ejecutado el algoritmo SOM en nuestro conjunto de datos (Qiana *et al.*, 2019).

El mapa por sí solo no es de mucha utilidad sin las técnicas de visualización que mejoran las propiedades particulares de los datos subyacentes (Qiana *et al.*, 2019). Estos planos de componentes resultan útiles para interpretar el tipo de muestras que pertenecen a un grupo, comparándolas con una matriz U, que permite evaluar la presencia y distribución de concentraciones de observaciones similares y sus posibles características a través de una presentación de red. Es así como se puede definir a los resultados provistos por el SOM como una herramienta poderosa y eficaz para validar la correlación esperada entre diferentes características de la muestra y para predecir correlaciones desconocidas (Akçapınar *et al.*, 2014).

6.2.6. Identificación y caracterización de conglomerados basados en características socio-demográficas, económicas y de salud

Para lograr una identificación y caracterización de agrupaciones o conglomerados relevantes en base a los datos a suministrar al modelo, la red SOM creará un conjunto de prototipos de vectores que representarán el conjunto de datos y llevará a cabo una topología preservando la proyección de los prototipos desde el espacio de entrada d-dimensional sobre una cuadrícula de menor dimensión (Vesanto & Alhoniemi, 2000). Esta cuadrícula ordenada se podrá utilizar como una superficie de visualización conveniente para mostrar diferentes características del SOM (y, por lo tanto, de los datos), siendo unas de las más importantes: la estructura tentativa del clusters existentes (Vesanto & Alhoniemi, 2000). Por lo tanto, para poder utilizar de manera efectiva la información proporcionada por el SOM, se requerirán métodos para dar buenos candidatos para los clusters o grupos de unidades en el mapa. Se debe enfatizar que el objetivo aquí no es encontrar un agrupamiento óptimo para los datos, sino obtener una buena división de la estructura del grupo de los datos (Vesanto & Alhoniemi, 2000).

Una idea inicial del número de conglomerados en el SOM, así como sus relaciones espaciales, podrá ser adquirida mediante la inspección visual del mapa. Los métodos más utilizados para visualizar la estructura de conglomerados de la SOM son las técnicas de matriz de distancia, especialmente la matriz de distancia unificada (matriz U) (Vesanto & Alhoniemi, 2000). La matriz U mostrará distancias entre vectores prototipo de unidades de mapa vecinos. Debido a que normalmente tienen vectores prototipos similares, la matriz U estará estrechamente relacionada con la medida de enlace único, dicha medida ofrecerá una visualización a color de la relación o asociación entre neuronas y cómo se distribuyen las uniones o agrupaciones de estas tomando en cuenta la intensidad de dichas dependencias (Akçapınar *et al.*, 2014). Generalmente, la escala de colores que reflejará el grado de enlace en la red va desde el color negro hasta el color amarillo (Akçapınar *et al.*, 2014): el primero indica que no existe relación alguna (lejanía) entre los nodos vecinos y el segundo simboliza una asociación muy alta o

fuerte (cercanía) entre las neuronas, como la Figura 6.1 evidencia.

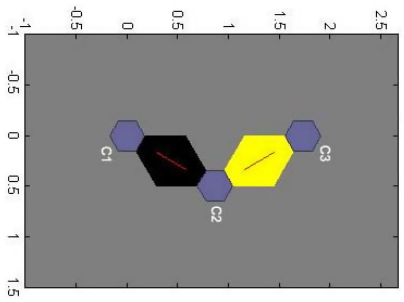


Figura 6.1: Distancias entre neuronas o nodos en una matriz U. Fuente: (Akçapınar *et al.*, 2014).

El entrenamiento de las posiciones de SOM interpolará las unidades de mapa entre los conglomerados y, por lo tanto, oscurecerá los bordes de cada conglomerado identificado. Esta información se podrá utilizar para generar agrupaciones en el SOM mediante el uso de unidades de impacto cero para indicar los límites de dichos clusters (Vesanto & Alhoniemi, 2000). Los grupos resultantes serán rodeados y etiquetados ulteriormente.

Sin embargo, las visualizaciones solo se pueden utilizar para obtener información cualitativa. De ese modo, para producir caracterizaciones, descripciones cuantitativas de las propiedades de los datos, de los grupos visualmente identificados se deberán seleccionar los grupos significativos de unidades de mapa del SOM resultantes de la fase de visualización. Posteriormente, se prepararán descripciones y resultados estadísticos en base a la composición de cada conglomerado según las variables independientes o regresoras (Vesanto & Alhoniemi, 2000).

6.2.7. Análisis de la distribución geográfica de los conglomerados

Con el objetivo de investigar el comportamiento geográfico de los conglomerados identificados en la Sección 6.3.6 basados en los factores demográficos y de salud recogidos de la encuesta ENDES para el 2019, se realizará un análisis espacial exploratorio de la distribución y concentración de individuos bajo la división administrativa del Perú, que consiste en 25 departamentos o regiones en los que el territorio nacional está particionado, tomando en cuenta su pertenencia a un cluster determinado en cada caso.

Los datos de la longitud y latitud ofrecidos en la encuesta ENDES permitirán la ubicación espacial de cada individuo dependiendo los valores de ambas coordenadas a lo largo del mapa del Perú, representando cada observación como un punto dentro del plano territorial a fin de evaluar las dimensiones geográficas de los conglomerados y focos de agrupación o dispersión de personas. A su vez, el ubigeo y el código de departamento son variables que juegan un rol clave dentro del análisis geográfico, debido a que estas serán empleadas para la determinación de las tasas de concentración en cada región del país al facilitar la integración de la información de los clusters con los datos cartográficos gubernamentales.

En el mismo orden de ideas, se procederá a recabar datos cartográficos de los límites nacionales del Perú y las divisiones administrativas de las fronteras de las regiones del estado para

proveerle al estudio los planos geográficos necesarios. Este mapeo físico del país se abstrae de la Infraestructura Nacional de Datos Geoespaciales Fundamentales del Perú, división del Instituto Geográfica Nacional a cargo del Ministerio de Defensa en el país (MINDEF - IGN, 2020). Los datos poligonales y tablas de atributos se presentan en archivos de extensión .shp denominados archivos *shape* que, en esencia, hacen referencia a un formato de almacenamiento que describe espacialmente características vectoriales (puntos, líneas y polígonos) y, junto con los atributos de datos que están vinculados a cada forma, crean la representación de los datos geográficos.

6.3. Resultado del modelo propuesto

Esta sección congrega el reporte de todos los resultados relacionados al proceso de conformación de conglomerados o clusterización de los adolescentes y jóvenes adultos en el Perú en torno al nivel de conocimiento que manejan acerca de las vías de transmisión y naturaleza de la infección del VIH/SIDA y a factores socio-demográficos, económicos y de salud que describen a la muestra de estudio.

Se presenta la preparación de las variables a emplear en el análisis de la red SOM, el proceso de entrenamiento de las redes, la elección de la mejor configuración o tamaño del mapa auto-organizado, la identificación de los clusters existentes dentro de la red SOM a través de la distancia vecinal entre neuronas, caracterización de los mismos junto con la prueba de diferencia de medias por cada atributo considerado entre los conglomerados y el diseño y despliegue de mapas de localización geográfica de los encuestados por conglomerado y mapas de calor de concentración de individuos y variables de interés a fin de evaluar la asociación entre atributos y conocimiento a nivel regional en el país.

6.3.1. Preparación de variables de estudio para el análisis de conglomerados

Basándonos en la información estadística recopilada del INEI y las operaciones efectuadas en el Capítulo 5.3.1, la Tabla 6.2 muestra a los factores, junto con sus definiciones o descripciones, el tipo de variable y dato que representan, el rango de las etiquetas numéricas (el cual va desde 0 hasta $n-1$, siendo n el número de niveles con los que la variable cuenta) y el resultado de la codificación de los niveles nominales, como puede visualizarse a continuación.

Tabla 6.2: Definición y descripción de las variables involucradas en el mapa auto-organizado.

Variable	Definición	Naturaleza	Tipo de variable	Rango	Codificación
Conocimiento	Nivel de conocimiento sobre el VIH/SIDA	Categórica	Independiente	0-1	0. NO 1. SÍ
Género	Género del encuestado	Categórica	Independiente	0-1	0. Femenino 1. Masculino
Área de residencia	Límites zonales de los hogares de los encuestados	Categórica	Independiente	0-1	0. Rural 1. Urbana
Nivel educativo	Nivel de estudio más alto alcanzado por el encuestado	Categórica	Independiente	0-3	0. Sin educación 1. Primaria 2. Secundaria 3. Superior

Variable	Definición	Naturaleza	Tipo de variable	Rango	Codificación
Etnicidad	Etnicidad que el encuestado identifica según sus costumbres y antepasados	Categórica	Independiente	0-4	0. Origen Nativo[a] 1. Afroperuano 2. Blanco 3. Mestizo 4. Otro/No precisa
Lengua Materna	Idioma o lengua que el encuestado aprendió en sus primeros años de vida	Categórica	Independiente	0-2	0. Lengua nativa[b] 1. Castellano 2. Lengua extranjera
Oído sobre VIH/SIDA	Contacto con información acerca del VIH/SIDA	Categórica	Independiente	0-1	0. NO 1. SÍ
Prueba de VIH/SIDA	Realización de la prueba de descarte en los últimos 12 meses	Categórica	Independiente	0-1	0. NO 1. SÍ
Rango etario	Rango etario al que pertenece el encuestado	Categórica	Independiente	0-2	0. 15-20 años 1. 20-24 años 2. 25-29 años
Región natural	Región de respuesta a la que el individuo pertenece	Categórica	Independiente	0-3	0. Lima Metropolitana[c] 1. Resto Costa 2. Sierra 3. Selva
Acceso a medios	Capacidad de acceso a medios multimedia[d]	Categórica	Independiente	0-1	0. NO 1. SÍ
Género del jefe de hogar	Género del jefe de familia en el hogar del encuestado	Categórica	Independiente	0-1	0. Femenino 1. Masculino
Nivel económico	Grupo poblacional de bienestar o de riqueza al que el encuestado pertenece	Categórica	Independiente	0-4	0. Muy pobre 1. Pobre 2. Medio 3. Rico 4. Muy rico

Notas. [a] Quechua, aimara, nativo de la Amazonía, perteneciente o parte de otro pueblo indígena u originario. [b] Quechua o aimara/ lengua originaria de la Selva u otra lengua nativa. [c] Comprende la provincia de Lima y la Provincia Constitucional del Callao. [d] Medios multimedia: Comprende acceso a radio, televisión o internet.

Fuente: Elaboración propia.

Con la codificación exhibida en la tabla anterior se plantea la preparación de los factores socio-demográficos, económicos y de salud para proceder con el análisis de conglomerados; esta permitirá que las 13 variables consideradas preliminarmente puedan ser empleadas en el estudio.

6.3.2. Estadísticas descriptivas de los factores involucrados en el análisis de conglomerados

Posteriormente al proceso de preparación de las variables de entrada, la Tabla 6.3 muestra las características basales descriptivas de la muestra de estudio basada en los individuos entrevistados entre 15-29 años según la base de datos de la ENDES.

La muestra analítica está compuesta por 10,565 individuos que se encontraban dentro del territorio peruano al momento de efectuarse las entrevistas.

Tomando en cuenta los factores socio-demográficos, económicos y de salud en la cohorte de análisis, se pueden señalar los siguientes aspectos.

El 66.20 % de los miembros de la cohorte cuentan con un nivel de conocimiento inadecuado o

inapropiado si se evalúa la percepción e ideas sobre la transmisión y naturaleza del VIH/SIDA que estas personas poseen, lo que refleja que aproximadamente dos tercios del total de entrevistados en la muestra se encuentra en riesgo por la falta de un discernimiento correcto acerca de los efectos y dinámica del virus y la enfermedad; el 61.40 % de los individuos son del género femenino; una proporción considerable de entrevistados vive en un contexto social y zonal urbano dentro de la región o departamento al que pertenezcan, ascendiendo a un valor de 67.80 % en la cohorte; en cuanto al nivel económico o pertenencia a un determinado quintil de bienestar, puede establecerse que el 60 % de los encuestados se encuentra en situación de pobre o muy pobre (30.50 % y 29.50 %, respectivamente), sólo el 20.4 % de individuos denota tener una distribución de ingresos elevada o muy elevada; considerando la región natural a la que los entrevistados pertenecen en el momento de haberse realizado la encuesta, puede establecerse que el 33.70 % de individuos se encuentra residiendo en algún departamento o región de la Sierra del país, el 26.20 % de las personas dentro de la muestra está localizada en la Selva del Perú y el 40.20 % de encuestados restantes se encuentra ubicado en algún departamento de la Costa peruana (resaltando que una proporción de 11.30 % vive en Lima Metropolitana); el 41.00 % de los individuos tiene entre 25 y 29 años de edad simbolizando el rango etario más frecuente en la cohorte, el 31.70 % de los encuestados está dentro del rango etario de 21-24 años y el resto de entrevistados (27.30 %) tiene entre 15 y 20 años; el nivel educativo más alto alcanzado que el grueso de la muestra posee es el grado de secundaria, representando al 59.90 % de las personas, el 29.80 % de los encuestados ostenta un nivel superior o mayor de educación y alrededor del 10.00 % de individuos tiene únicamente el grado de instrucción primaria o sin educación alguna; el 41.40 % de la cohorte se auto-percibe como mestiza dentro del estudio, el 33.90 % indica que posee un origen nativo o forma parte de la cultura indígena nacional, una proporción de 11.30 % se identifica como afro-peruano o miembro de la comunidad negra en el Perú y el resto de entrevistados es de etnia blanca u otra no precisada en el cuestionario de la ENDES (6.30 % y 7.20 %, correspondientemente); considerando si el encuestado ha oído con anterioridad alguna información o contenido acerca del VIH/SIDA, se precisa que el 86.80 % de los encuestados ha escuchado algún tipo de dato sobre el virus y/o la enfermedad; el 73.90 % de los individuos no se ha realizado ningún tipo de prueba de descarte o detección del VIH/SIDA en los 12 meses previos a haberse realizado la presentación del cuestionario; el 89.40 % de los entrevistados asegura que cuenta con al menos un medio multi-media de información (ya sea el radio, la televisión o el internet) en sus hogares; una proporción relevante de los hogares de los individuos dentro de la cohorte de estudio está precedida o jefaturada por un varón, representando un valor de 72.30 %; por último, la lengua materna o primaria más frecuente dentro de la muestra de estudio es el castellano, con 82.30 % de entrevistados que la adquirieron como primer idioma, el resto de encuestados posee una lengua nativa o extranjera como idioma natal (con 17.60 % y 0.10 %, respectivamente).

Tabla 6.3: Análisis de datos de los factores socio-demográficos, económicos y de salud de la encuesta ENDES.

Variable	Frecuencia (n)	Porcentual (%)
Nivel de conocimiento		
NO	6989	66.20 %
SÍ	3576	33.80 %
Género		
Femenino	6485	61.40 %
Masculino	4080	39.60 %
Área de residencia		
Rural	3400	32.20 %
Urbana	7165	67.80 %
Nivel económico		
Muy pobre	3225	30.50 %
Pobre	3120	29.50 %
Medio	2061	19.50 %
Rico	1368	12.90 %
Muy rico	791	7.50 %
Región natural		
Lima Metropolitana	1189	11.30 %
Resto Costa	3050	28.90 %
Sierra	3562	33.70 %
Selva	2764	26.20 %
Rango etario		
15-20 años	2889	27.30 %
21-24 años	3346	31.70 %
25-29 años	4330	41.00 %
Nivel educativo		
Sin educación	31	0.3 %
Primaria	1057	10.00 %
Secundaria	6333	59.90 %
Superior	3144	29.80 %
Etnicidad		
Origen nativo	3579	33.90 %
Afroperuano	1189	11.30 %
Blanco	664	6.30 %
Mestizo	4373	41.40 %
Otro/No sabe	760	7.20 %
Oído del VIH/SIDA		
NO	1397	13.20 %
SÍ	9168	86.80 %
Prueba del VIH/SIDA		
NO	7811	73.90 %
SÍ	2754	26.10 %
Acceso a medios		
NO	1115	10.60 %
SÍ	9450	89.40 %
Género del jefe de hogar		
Femenino	2929	27.70 %
Masculino	7636	72.30 %
Lengua materna		
Lengua nativa	1858	17.60 %
Castellano	8697	82.30 %
Lengua extranjera	10	0.10 %

Fuente: Elaboración propia.

6.3.3. Parámetros de entrenamiento de la red SOM

Bajo en las premisas dadas por (Vesanto *et al.*, 2000) descritas en la Sección 6.2.3, la forma del mapa en este caso será una red hexagonal para la disposición de los nodos y la presentación de los individuos asignadas a cada uno de ellos. De igual manera, la forma de la cuadrícula de la red adoptará el estilo de mapa de hoja predeterminado por las razones citadas previamente.

En el mismo orden de ideas, la Tabla 6.4 muestra a continuación los parámetros seleccionados para el entrenamiento de la red de Kohonen para el presente estudio de conglomerados de individuos basados en factores socio-demográficos, económicos y de salud.

Tabla 6.4: Parámetros de la red de Kohonen para la preparación del modelo en la etapa de entrenamiento.

Componente	Selección
Inicialización	Linear
Entrenamiento	Batch Weight/Bias Rules
Desempeño	Mean Squared Error (mse)
Iteraciones	200 épocas (epochs)

Fuente: Elaboración propia.

El procedimiento de inicialización a emplear en el presente mapa de Kohonen es la inicialización lineal (o *Linear initialization*), donde los vectores de peso se inicializan de manera ordenada a lo largo del subespacio lineal atravesado por los dos vectores propios principales del conjunto de datos de entrada (Akinduko & Mirkes, 2012).

En cuanto a la fase de entrenamiento que será empleada en el modelo, el algoritmo de reglas de aprendizaje de peso y sesgo con actualizaciones por lotes (conocido también como *Batch Weight/Bias Rules algorithm*) fue elegido como la opción a ser aplicada en el diseño y construcción del mapa auto-organizado en base a las variables definidas anteriormente. Este es una variante del aprendizaje en SOM y el sentido elemental detrás de este algoritmo es entrenar al SOM utilizando todo el conjunto de datos de entrenamiento y, por lo tanto, en cada iteración, los pesos de sus neuronas representan la media de las entradas más cercanas (Akinduko & Mirkes, 2012).

Asimismo, la medición del desempeño y confianza por los cuales el entrenamiento de la red estará regido se dará a través del vector compuesto por el error cuadrático medio (MSE, *mean squared error*) para cada época o iteración del entrenamiento.

Finalmente, considerando las épocas (o *epochs*) dentro de una red SOM, la duración del entrenamiento en este caso será de 200 épocas o iteraciones, que es el valor estándar y máximo del software en el que se desarrolla el SOM.

6.3.4. Evaluación de la topología óptima para la red SOM

Bajo la premisa dada por (Tian *et al.*, 2014) en la Sección 6.2.4, el número sugerido de neuronas a través de la regla de Kohonen, puede ser decidido de la siguiente manera, considerando que N es igual a 10,565, ya que es el número de observaciones o individuos presentes en el

conjunto de datos extraído de la ENDES.

$$M \approx 5 \times \sqrt{10,565} \rightarrow M_{neuronas} \approx \sqrt{514} \rightarrow M_{neuronas} = 23. \tag{6.2}$$

Como la Ecuación 6.2 muestra, para nuestros datos, $M_{neuronas}$ es igual a 23, lo que indica que puede emplearse una red de 529 nodos en una configuración de 23x23 neuronas.

Sin embargo, al ser el mapa auto-organizado de Kohonen un tipo especial de red neuronal artificial, el modelo recibe el mismo tratamiento que se le suele dar a la evaluación del número de neuronas, en cuyo caso se hacen variaciones de dicho parámetro con incrementos aritméticos de 5 o 10 unidades entre un valor recomendado y otro, siendo estos valores múltiplos de los números mencionados anteriormente en la mayoría de los casos (Lesinski *et al.*, 2016; Sha, 2006; Koutsoukas *et al.*, 2017; Arifin *et al.*, 2019).

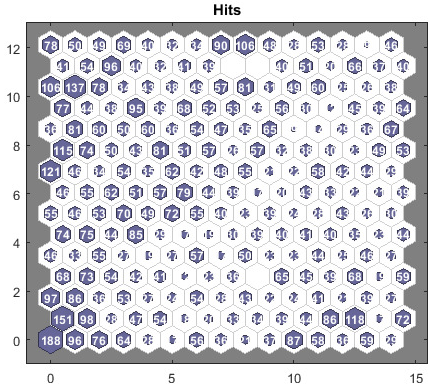
Tomando en consideración el valor de $M_{neuronas}$ de la Ecuación 6.2, el múltiplo de 5 o 10 más cercano a este valor es el 25, por lo que el inicio de la evaluación de la topología óptima para la red SOM se dará con esta cifra. El SOM será entrenado iterativamente hasta que todos los vectores de peso del mapa se agrupan en grupos de acuerdo con su distancia. Cuando finaliza el proceso de aprendizaje, se crea el SOM.

La Figura 6.2c muestra el plano de distribución de observaciones en el mapa con la configuración de 25x25 neuronas y una red de 625 nodos dada líneas arriba. Como puede notarse, en el gráfico no existe una clusterización definible y delimitada de los individuos en los nodos (ya que cada uno de ellos tiene asignado un fracción de individuos de la muestra total que están contenidos en el interior de la neurona), debido a que el grueso de los nodos que poseen observaciones mayoritarias (las que superan la cifra de 25 individuos - límite inferior aproximado al 35 % del valor más alto en el mapa) se ubican en el extremo izquierdo de la cuadrícula con patrón hexagonal hasta la mitad de la red, dimensionándose ineficientemente posibles separaciones entre estas acumulaciones de nodos (representadas por aquellas neuronas con concentraciones de individuos menores a 25).

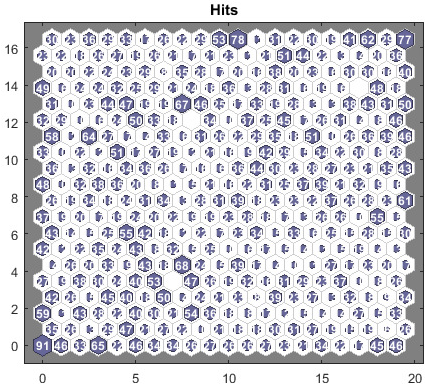
Asimismo, la distribución irregular de los individuos demuestra que el número de neuronas dadas no permite una reducción de los datos en el mapa, ya que existen nodos vacíos (sin ningún individuo miembro) o nodos con una cantidad de agremiados nimia (menores de 10) en zonas sin ninguna interacción/asociación entre neuronas que presuman una conformación de conglomerado o entre las secciones donde la mayor parte de nodos se encuentra. Compaginando a partir de Xia (2017), la configuración de 25x25 no logró una simplificación de datos suficiente ni la posibilidad de esbozar potenciales clusters que puedan conformarse de la concentración de individuos.

Tomando en cuenta que un $M_{neuronas} = 25$ no fue el valor ideal para la construcción del mapa, la Figura 6.2 presenta los resultados de las pruebas de topología con 15, 20 y 30 como valor del parámetro de número de neuronas mediante los gráficos de planos de distribución de observaciones.

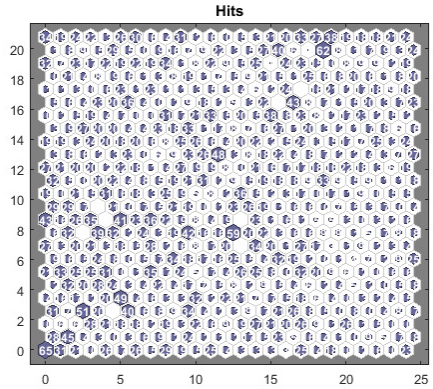
(a) Configuración de 15x15



(b) Configuración de 20x20



(c) Configuración de 25x25



(d) Configuración de 30x30

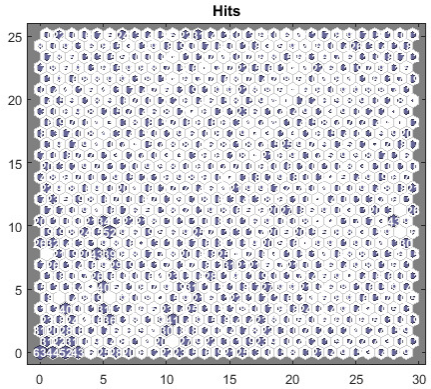


Figura 6.2: Topologías de red evaluadas para el mapa auto-organizado. (a) Red SOM de tamaño 15x15, (b) Red SOM de tamaño 20x20, (c) Red SOM de tamaño 25x25 y (d) Red SOM de tamaño 30x30. Fuente: Elaboración propia.

Como se observa, la red de 225 nodos con una configuración de 15x15 y la red de 900 nodos con una configuración de 30x30 son valores que también fallan en obtener una correcta representación del conjunto de datos mediante el mapa de Kohonen.

En el caso de la primera estructura en mención, no se pueden hacer mayores distinciones entre los focos de concentración de nodos guiados por el número de individuos en su interior, ya que dos particiones (una ubicada desde la esquina superior izquierda hasta llenar la diagonal superior del mapa y la otra en el extremo derecho inferior) con la disposición que muestran dilucidan que existe un sobredimensionamiento de las fracciones reconocidas, posiblemente agrupando en cada una de ellas nodos con comportamientos o características diferentes a los demás que incumplen con una propiedad elemental del proceso de clusterización, que es asegurar la heterogeneidad entre los conglomerados identificados y la homogeneidad dentro de los mismos. No es posible identificar sub-grupos dentro del plano de nodos como se espera de un resultado ideal de clustering, lo que indica que este tamaño de neuronas no generó un ajuste suficiente para los datos.

Por otro lado, analizando la segunda disposición en discusión, esta presenta el problema

opuesto a la anterior configuración: sub-dimensionamiento del conjunto de datos al ser modelado dentro de la red SOM y la incapacidad de poder reducirlos apropiadamente. Existe una dispersión altamente significativa a lo largo del mapa, puesto que se conformaron múltiples neuronas con una cantidad poco representativa o considerable (a saber, igual o menor a 10 miembros en un nodo) en aproximadamente toda el área de la cuadrícula, exceptuando los focos formados en la sección inferior del plano. Las concentraciones de esta zona muestran casos aislados de posibles grupos diferenciados, pero estar rodeados por nodos vacíos o con una escasa cifra de miembros dan cuenta que la topología probada no provee resultados favorables en la conformación de clusters homogéneos.

Finalmente, la prueba de topología de red con una configuración 20x20 y una cuadrícula hexagonal con 400 nodos en total ofrece los resultados más convenientes y adecuados para el estudio. Como se puede apreciar en la Figura 6.2b, la presencia de nodos vacíos o con una baja acumulación de individuos es sumamente reducida, lo que advierte que no existe una dispersión del conjunto de datos en las neuronas. En el mismo orden de ideas, es posible identificar agrupaciones diferenciadas en el plano que ofrecen indicios lógicos de la conformación de potenciales conglomerados en una etapa posterior de evaluación de distancias vecinales ponderadas entre neuronas, estas asociaciones son notorias en las siguientes posiciones: un sub-grupo localizado en la esquina inferior derecha, otro sub-grupo ubicado en la esquina superior derecha, una sub-división en la esquina inferior izquierda, un sub-grupo en la esquina superior izquierda y otra sub-partición en el centro superior del mapa. La capacidad de distinción es otro punto favorable para esta configuración, ya que los nodos menos concentrados (con un número de agremiados entre 10 y 25) asumen un rol implícito de separadores entre las áreas segregadas reconocidas previamente, indicando que este arreglo de nodos fomenta una distinción ordenada y visible entre los focos de concentración de encuestados. Esta configuración ha probado ofrecer una representación adecuada de los datos en la red y evitar problemas de simplificación o ajuste dentro del modelo.

Es por ello que, para la construcción y evaluación de un mapa de Kohonen considerando las variables de entrada y la forma de la cuadrícula y orientación de nodos dadas en la sección anterior, se empleará una configuración de 20 neuronas que generarán un plano de 400 neuronas al ser la topología óptima de red determinada en esta sección.

6.3.5. Análisis de los resultados del mapa auto-organizado de Kohonen

Como vectores de entrada, se emplean a los 10,565 individuos entrevistados en la encuesta ENDES y 13 variables independientes que definen al conjunto de datos (conocimiento sobre VIH/SIDA, género del encuestado, área de residencia, nivel económico - quintil de bienestar, auto-identificación étnica, lengua materna o primaria, si el individuo ha oído con anterioridad acerca del VIH/SIDA, realización de la prueba de descarte del virus y/o enfermedad, rango etario, región natural en la que reside al momento de la encuesta, acceso a medios multimedia de información, género del jefe de familia y nivel/grado educativo más alto alcanzado) en la red de Kohonen como datos de entrenamiento. Los mapas de topología, distribución de

observaciones, asociaciones entre nodos y componentes se producen utilizando el algoritmo descrito en la Sección 3.1.3.4 y representan individuos con propiedades similares que pueden ser agrupados.

La red SOM inicia el análisis con un patrón que incluye la estructura de una red neuronal convencional y es similar a una estructura de plano como cuadrícula que tiene una cierta dimensión al principio. Este dimensionamiento también determina el número de neuronas (en este caso, se seleccionó 20 como el número de neuronas a emplear), según el proceso que se muestra en la Figura 6.3.

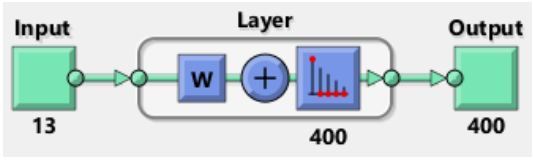


Figura 6.3: Construcción de la red SOM considerando variables de entrada y mapa de salida. Fuente: Elaboración propia.

Como se muestra en la Figura 6.4, los números contenidos en los hexágonos son datos que son absorbidos por cada uno de los nodos de la red neuronal o mapa en (b). Del mismo modo, el mapa muestra que los datos tienden a formar ciertas particiones como las detectadas en el Sección 6.3.4. Además de la distribución de observaciones, se puede puntualizar que la topología generada del SOM es hexagonal como puede notarse en (a). Dicha figura muestra las ubicaciones de las neuronas en la topología a fin de preparar el plano que indicará cuántos de los datos de entrenamiento estarán asociados con cada una de las neuronas (centros de grupos). En detalle, la topología es una cuadrícula de 20 por 20, por lo que se tienen 400 neuronas en total. Cabe resaltar que, según la Figura 6.4b, el número máximo de miembros en un hexágono fue 91, lo que indica que el máximo de datos en un cluster individual es 91. Además, el número mínimo de miembros en un hexágono fue 0, lo que indica que, en estos nodos, no hay datos.

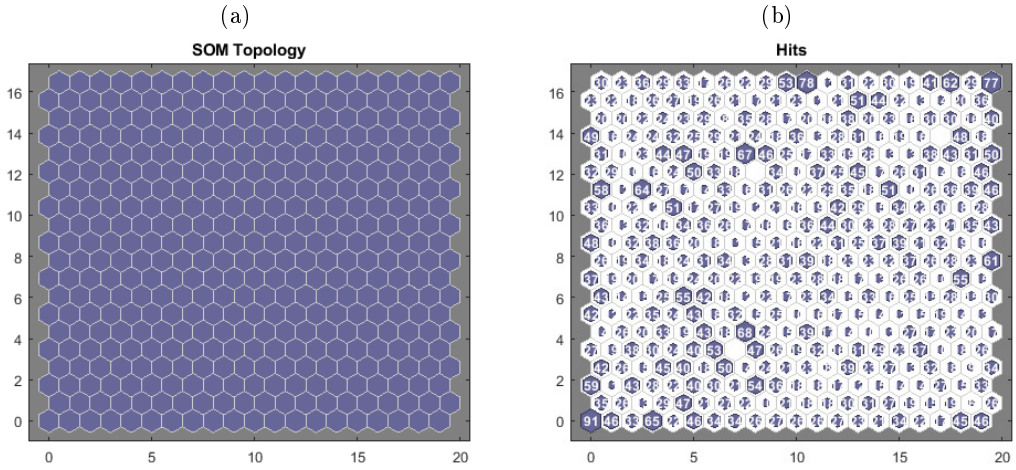


Figura 6.4: (a) Topología final de la red SOM y (b) Plano de distribución de observaciones en las neuronas de la red. Fuente: Elaboración propia.

En otro sentido, la Figura 6.5 hace referencia a la matriz U, que es un medio para encontrar posibles agrupaciones en el conjunto de datos.

El mapa de la matriz U se produjo mapeando la distancia promedio entre cada nodo y sus vecinos más cercanos en el espacio de alta dimensión en el mapa bi-dimensional. Los valores en el mapa de la matriz U reflejan el grado de similitud de cada nodo y sus vecinos. En dicha figura, los hexágonos azules representan las neuronas o nodos. Las áreas de claridad entre las neuronas en el mapa de la matriz U, que se muestran mediante una escala de color próxima a tonalidades amarillas, significan que los hexágonos, así como los puntos de datos de entrenamiento en dichas áreas, están muy cerca unos de otros en el espacio de alta dimensión, es decir, son muy similares o homogéneos entre ellos. Para los puntos de datos de entrenamiento en las regiones o bordes de color rojo o negro, puede indicarse que son muy diferentes entre sí. Como se puede ver en la matriz U a continuación, hay una identificación latente de aproximadamente 5 conglomerados: un cluster que abarca verticalmente todo el extremo izquierdo del gráfico, una partición ubicada en la esquina inferior derecha, una división localizada en la esquina superior derecha, un cluster al lado de la división anterior y otro que va desde los extremos inferiores hasta la zona superior del centro del plano.

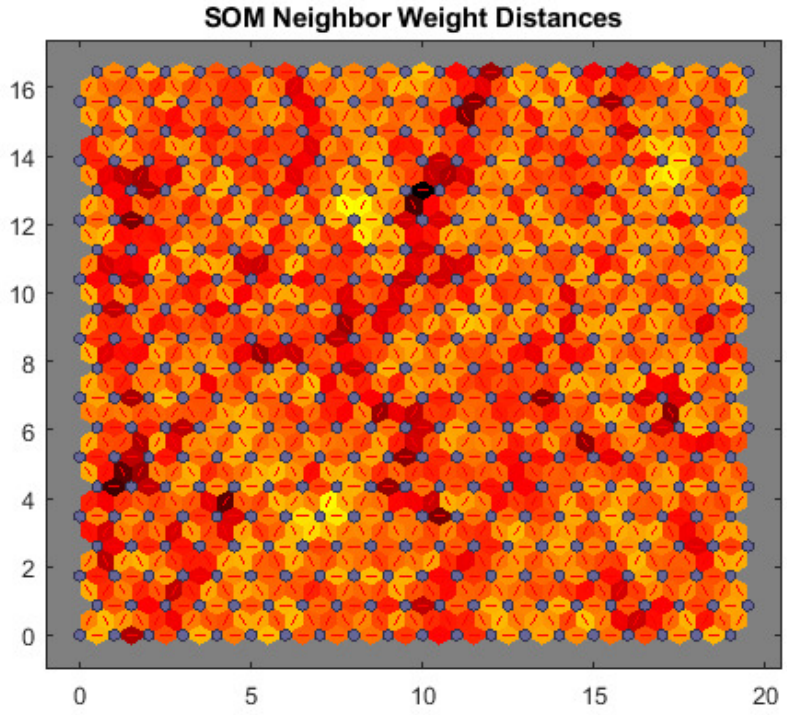


Figura 6.5: Gráfico de distancias vecinales ponderadas entre neuronas de la red SOM. Fuente: Elaboración propia.

En la misma perspectiva, los gráficos de componentes o factores, como se presentan en la Figura 6.6, mapean por separado cada elemento o variable de los pesos del nodo en el mapa bidimensional (pueden ser considerados como versiones sintéticas del SOM, donde cada plano

muestra la distribución de un componente o característica del vector de peso). La escala de color de los mapas de componentes se estableció de manera que, como ocurre en el caso de la matriz U, colores claros cercanos al amarillo representan los valores más altos de integración en el mapa y colores opacos u oscuros reflejan los valores más bajos de cohesión en el plano. La ventaja de este medio de visualización es que la variación de valores en el mapa es significativa, lo que puede ayudar a los usuarios a detectar la correlación entre propiedades o características (Qiana *et al.*, 2019).

Como puede evidenciarse en la Figura 6.6, el número total de planos de componentes que se pueden obtener es igual al número de elementos que contiene la ponderación de un nodo (a saber, el número de variables que tienen los datos de entrenamiento). Así, son 13 los mapas generados a partir de las 13 características demográficas y de salud de nuestros datos de entrenamiento, de los cuales se puede obtener información valiosa sobre los factores a partir de la interpretación de las relaciones entre las diferentes propiedades de los planos. Adicionalmente, las 13 figuras están unidas por posición: en cada una, el hexágono en una determinada posición corresponde a la misma unidad de mapa.

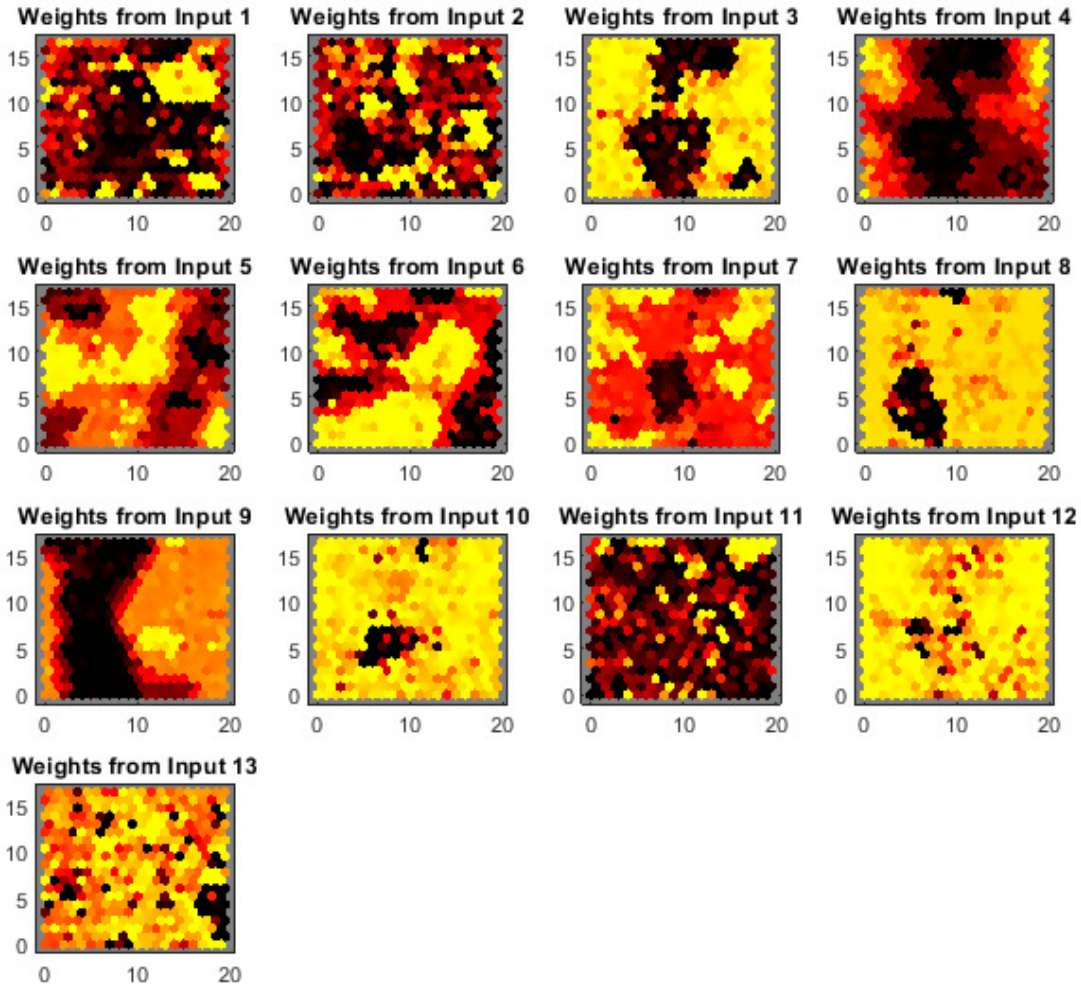


Figura 6.6: Planos de componentes para la conformación de conglomerados en la presente red SOM. Fuente: Elaboración propia.

Bajo un análisis visual de los mapas de componentes, comparando la distribución y el patrón total de la escala de colores (considerando que cada mapa representa una variable) se podría concluir lo siguiente:

- El componente N° 01 ("Nivel de conocimiento sobre el VIH/SIDA") y el N° 11 ("Prueba de descarte del VIH/SIDA") tienen una correlación o asociación directa (positiva) significativa. A medida que el valor del nivel de conocimiento de VIH/SIDA aumenta (desde el lado superior derecho de su matriz), el valor de la realización de la prueba de detección del virus y/o enfermedad aumenta en la misma dirección y casi con el mismo patrón.
- El componente N° 03 ("Área de residencia"), el N° 04 ("Nivel económico/Quintil de bienestar") y N° 09 ("Etnicidad") tienen una correlación o asociación directa (positiva) significativa.
- El componente N° 06 ("Rango etario") y el N° 07 ("Nivel educativo") tienen una correlación o asociación directa (positiva) significativa.
- El componente N° 05 ("Región natural de procedencia") y el N° 06 ("Rango etario") tienen una correlación o asociación inversa (negativa) significativa. A medida que el valor del nivel de la región natural en la que los entrevistados se encuentran en el momento de la encuesta aumenta (desde el lado superior izquierdo de su matriz), el valor del rango etario disminuye en la misma dirección con un patrón diferente.
- El componente N° 08 ("Lengua materna") y el N° 10 ("Ha oído anteriormente acerca del VIH/SIDA") tienen una correlación o asociación inversa (negativa) significativa.
- El componente N° 02 ("Género"), N° 12 ("Acceso a medios multi-media") y N° 13 ("Género del jefe de hogar") no tienen una relación significativa distinguible con los factores.

6.3.6. Identificación y caracterización de conglomerados basados en características socio-demográficas, económicas y de salud

Basándonos en la matriz U resultante de la red SOM de 20 neuronas aplicada a las variables de estudio, se puede distinguir la conformación de conglomerados de individuos considerando las distancias vecinales ponderadas entre las neuronas o nodos del mapa, tal y como la Figura 6.7 muestra a continuación. Los límites en el mapa de conglomerados que se produce se asemejan significativamente a las regiones rojas (o negras) en la matriz U (Figura 6.5), lo que valida los resultados de los clusters.

Como puede abstraerse de la figura, son en total 5 conglomerados o clusters que pueden ser deslindados de los resultados obtenidos mediante la red de Kohonen.

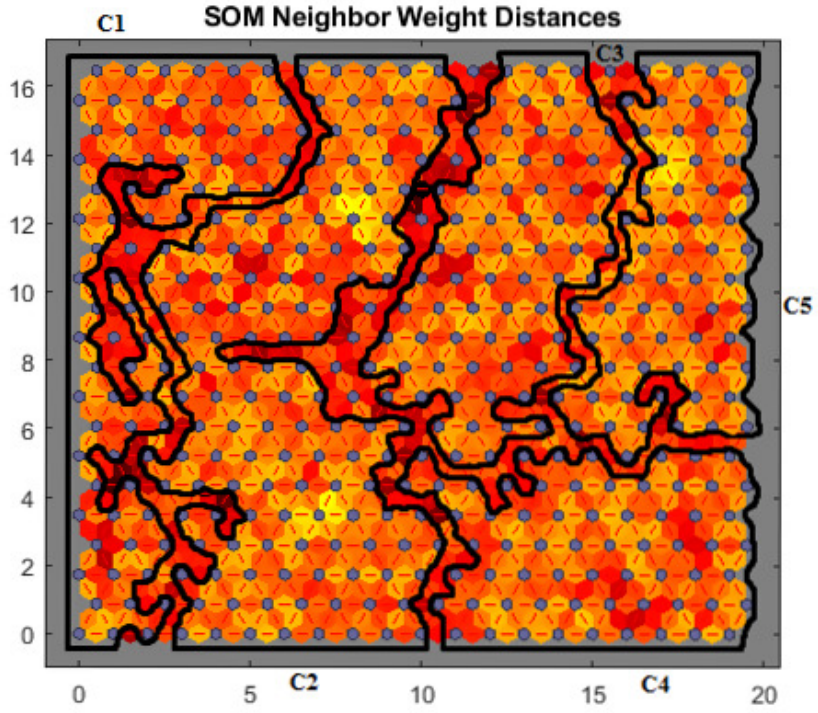


Figura 6.7: Gráfico de identificación de conglomerados presentes en la red SOM. Fuente: Elaboración propia.

Cada uno de estos se encuentra separado o divido del resto de nodos o agrupaciones de individuos en el mapa ya que las conexiones entre clusters son débiles o demuestran que no existe un comportamiento homogéneo entre las observaciones que los conforman que justifiquen la integración o congregación de los mismos (como el gráfico exhibe: las conexiones que existen entre cada uno de los clusters mostrados posee una intensidad de color rojo, lo que sugiere una dependencia débil o diferencias entre los individuos agrupados en cada lado).

En el mismo sentido, la creación y disposición dada de los conglomerados en la red se explica ya que puede evidenciarse que las conexiones de los nodos en cada una de las divisiones de clusters son fuertes o altas entre ellas, debido a que la mayoría de estas asociaciones tienen una intensidad de color amarillo, lo que demuestra que poseen similitudes entre ellos que apoyan la cohesión de las neuronas a fin de conformar un grupo distintivo.

Asimismo, considerando la Figura 6.6 que hace referencia a los componentes que representan a las variables de entrada y las correlaciones entre ellas que construyen el gráfico de distancias vecinales ponderadas, se puede afirmar que ciertas variables o factores dan cuenta de la concentración de las observaciones en sub-grupos a partir de la muestra inicial y permiten dilucidar la conformación de conglomerados dentro de la matriz U.

Con mayor precisión, conforme a la figura, se establece que: el Input 1, que representa a la variable “Nivel de conocimiento sobre el VIH/SIDA”, indica que existe una concentración de individuos en la esquina superior derecha del gráfico de distancias entre nodos y pequeñas

agrupaciones ubicadas en el resto del mapa, reflejando razones para la conformación de los clusters 3 (C3) y 5 (C5); por otro lado, el Input 3, que representa a la variable "Área de residencia" establece que existen agrupaciones notorias en ambos lados de la red, derivando en el fundamento de conglomerado 1 (C1) y parte del resto de clusters (2, 3, 4 y 5) ya que abarca con amplitud el mapa con excepción de la zona central y con conexiones altamente fuertes bajo un color amarillo de asociación; el Input 4, que representa a la variable "Quintil de bienestar", sigue la misma tendencia que el Input 3, ya que forma asociaciones relativamente altas o fuertes entre neuronas en los extremos de la red que fundamentan la conformación de los clusters 1 (C1) y 5 (C5), aunque en una menor proporción, por lo que los grupos formados tienen un menor tamaño en comparación a la variable anterior; en cuanto al Input 5, que representa a la variable "Región natural", este indica que la mayor agrupación o asociación entre neuronas se sitúa aproximadamente en el centro del mapa auto-organizado, lo que refleja que la conformación del conglomerado 2 (C2) está influenciada por este factor; considerando el Input 6, que representa a la variable "Rango etario", este factor denota una concentración en la esquina superior izquierda y en la diagonal de la red, lo que sugiere la conformación de los clusters 2 (C2), 3 (C3) y 1 (C1) por la alta asociación o dependencia (de color amarillo) entre los nodos que componen dichas concentraciones; tomando en cuenta al Input 7, que representa a la variable "Nivel educativo", este sugiere que existe una relación de grado muy alto (con una intensidad de color amarillo) entre las neuronas vecinas de la esquina superior e inferior izquierda y la esquina superior derecha, permitiendo la configuración de los conglomerados 1 (C1) y parte de los conglomerados 3 (C3) y 5 (C5); finalmente, el Input 9, que representa a la variable "Etnicidad", este genera asociaciones relevantes en los extremos del mapa auto-organizado (destacando que las conjunciones del extremo derecho llegan a distribuirse hasta la mitad de la red), reflejando la correcta agrupación de dichas neuronas en el conglomerado 1 (C1) y los conglomerados 3 (C3), 4 (C4) y 5 (C5). No obstante, inputs como el 2 (que representa a la variable "Género") y el 11 (que representa a la variable "Prueba de VIH/SIDA") no poseen ninguna correlación significativa en la construcción de la matriz U, debido a que forma asociaciones totalmente nulas o débiles entre los nodos, lo que indica una falta de cohesión a lo largo de toda la red. Por otro lado, inputs como el 8 (que representa a la variable "Lengua materna"), el 10 (que simboliza a la variable "Oído acerca del VIH/SIDA"), 12 (que representa a la variable "Acceso a medios multi-media") y 13 (que simboliza a la variable "Género del jefe de familia") soportan la tendencia de asociación general entre casi todas las neuronas de la red de Kohonen, contribuyendo a la conformación de los 5 clusters identificados.

De esta manera, la identificación dada en la Figura 6.7 de los 5 clusters es explicada a partir de los cortes diferenciales entre grupos representados a través de las dependencias débiles entre cada conglomerado (las conexiones de color rojo de los nodos o neuronas que pertenecen a grupos disímiles) y los aportes o influencias de la correlación entre cada input o variable de entrada de la red que configuran la forma de los clusters y su distribución a lo largo del mapa auto-organizado.

La Tabla 6.5 mostrada a continuación provee una descripción estadística del perfil socio-

demográfico, económico y de salud separados por conglomerado y el resultado de las pruebas de diferencia de medias (bajo la prueba de chi-cuadrado o la prueba de Fisher dependiendo de la frecuencia n de los datos).

Tabla 6.5: Características socio-demográficas, económicas y de salud de la muestra de estudio identificada por conglomerados.

		Conglomerados										
Variable	**Categorías**	**C1**	%	**C2**	%	**C3**	%	**C4**	%	**C5**	%	Dif. Medias[b]
Conocimiento	0. NO	1143	56.9 %	2628	74.5 %	1137	64.4 %	967	70.8 %	1012	57.6 %	<0.001 ***
	1. SÍ	867	43.1 %	900	25.5 %	628	35.6 %	399	29.2 %	746	42.4 %	
Género	0. Fem	1203	59.9 %	2217	62.8 %	1119	63.4 %	816	59.7 %	1044	59.4 %	<0.05 **
	1. Masc	807	40.1 %	1311	37.2 %	646	36.6 %	550	40.3 %	714	40.6 %	
Área de residencia	0. Rural	60	3 %	1953	55.4 %	868	49.2 %	376	27.5 %	58	3.3 %	<0.001 ***
	1. Urbana	1950	97 %	1575	44.6 %	897	50.8 %	990	72.5 %	1700	96.7 %	
Nivel económico	0. Quintil I	36	1.8 %	1882	53.3 %	850	48.2 %	389	28.5 %	8	0.5 %	<0.001 ***
	1. Quintil II	243	12.1 %	1044	29.6 %	722	40.9 %	925	67.7 %	128	7.3 %	
	2. Quintil III	629	31.3 %	423	12 %	173	9.8 %	52	3.8 %	776	44.1 %	
	3. Quintil IV	679	33.8 %	157	4.5 %	20	1.1 %	0	0 %	507	28.8 %	
	4. Quintil V	423	21 %	22	0.6 %	0	0 %	0	0 %	339	19.3 %	
Región natural	0. Lima	411	20.4 %	28	0.8 %	21	1.2 %	135	9.9 %	594	33.8 %	<0.001 *** [F]
	1. Resto Costa	863	42.9 %	177	5 %	234	13.3 %	674	49.3 %	1089	61.9 %	
	2. Sierra	236	11.7 %	2440	69.2 %	450	25.5 %	300	22 %	75	4.3 %	
	3. Selva	500	24.9 %	883	25 %	1060	60.1 %	257	18.8 %	0	0 %	
Rango etario	0. 15-19	464	23.1 %	979	27.7 %	272	15.4 %	604	44.2 %	548	31.2 %	<0.001 ***
	1. 20-24	600	29.9 %	1092	31 %	550	31.2 %	382	28 %	649	36.9 %	
	2. 25-29	946	47.1 %	1457	41.3 %	943	53.4 %	380	27.8 %	561	31.9 %	
Nivel educativo	0. Sin Ed.	3	0.1 %	15	0.4 %	9	0.5 %	2	0.1 %	1	0.1 %	<0.001 *** [F]
	1. Primaria	39	1.9 %	472	13.4 %	291	16.5 %	167	12.2 %	22	1.3 %	
	2. Secundaria	934	46.5 %	2277	64.5 %	1089	61.7 %	1091	79.9 %	893	50.8 %	
	3. Superior	1034	51.4 %	764	21.7 %	376	21.3 %	106	7.8 %	842	47.9 %	
Lengua materna	0. Nativa	70	3.5 %	1604	45.5 %	105	5.9 %	42	3.1 %	31	1.8 %	<0.001 *** [F]
	1. Castellano	1936	96.3 %	1921	54.5 %	1659	94 %	1323	96.9 %	1726	98.2 %	
	2. Extranjera	4	0.2 %	3	0.1 %	1	0.1 %	1	0.1 %	1	0.1 %	
Etnicidad	0. Nativa	500	24.9 %	3069	87 %	0	0 %	7	0.5 %	0	0 %	<0.001 *** [F]
	1. Afroperuana	516	25.7 %	367	10.4 %	2	0.1 %	260	19 %	0	0 %	
	2. Blanca	201	10 %	89	2.5 %	82	4.6 %	169	12.4 %	51	2.9 %	
	3. Mestiza	717	35.7 %	3	0.1 %	1334	75.6 %	802	58.7 %	1500	85.3 %	
	4. Otra	76	3.8 %	0	0 %	347	19.7 %	128	9.4 %	207	11.8 %	
Oído del VIH/SIDA	0. NO	102	5.1 %	774	21.9 %	245	13.9 %	165	12.1 %	66	3.8 %	<0.001 ***
	1. SÍ	1908	94.9 %	2754	78.1 %	1520	86.1 %	1201	87.9 %	1692	96.2 %	
Prueba de VIH/SIDA	0. NO	1515	75.4 %	2695	76.4 %	1262	71.5 %	1088	79.6 %	1174	66.8 %	<0.001 ***
	1. SÍ	495	24.6 %	833	23.6 %	503	28.5 %	278	20.4 %	584	33.2 %	
Acceso a medios	0. NO	60	3 %	574	16.3 %	295	16.7 %	129	9.4 %	36	2 %	<0.001 ***
	1. SÍ	1950	97 %	2954	83.7 %	1470	83.3 %	1237	90.6 %	1722	98 %	
Género jefe de hogar	0. Fem	656	32.6 %	886	25.1 %	373	21.1 %	444	32.5 %	540	30.7 %	<0.001 ***
	1. Masc	1354	67.4 %	2642	74.9 %	1392	78.9 %	922	67.5 %	1218	69.3 %	
	Sub total[c]	2010	19.3 %	3528	33.8 %	1765	16.9 %	1366	13.1 %	1758	16.9 %	
	Sub total[d]	10,427	99 %									
	Total BD[e]	10,565	100 %									

Notas. *valores significativos p <0.10; **valores muy significativos p <0.05; ***valores altamente significativos p <0.01. N.S.: No significativo estadísticamente. [F]: Simboliza que se realizó la prueba de diferencia de medias de Fisher, en lugar a la prueba de chi-cuadrado. [b] Prueba de diferencia de medias entre conglomerados. [c] Total de individuos que forman parte de cada cluster y su valor porcentual. [d] Subtotal de individuos clusterizados bajo el SOM y el porcentaje que representa de la muestra total del estudio. [e] Muestra total.

Fuente: Elaboración propia.

Ulteriormente a la conformación de los clusters delimitados líneas arriba, se procede a realizar la caracterización de cada conglomerado que consiste en precisar los valores de las categorías o niveles de cada factor o variable tomando en cuenta el tamaño o dimensión de cada grupo (derivado del número de individuos ubicados en cada nodo que componen a los conglomerados).

Del mismo modo, se realiza una prueba de independencia de medias para cada variable comparando los 5 clusters en todos los casos a fin de evaluar estadísticamente si los valores obtenidos en cada categoría de un factor difieren entre los grupos y comprobar efectivamente si los individuos en cada cluster son distintos - asegurar la heterogeneidad entre estos.

Como la Tabla 6.5 señala, el porcentaje de la muestra que pudo ser clusterizado posterior a la aplicación de la red de Kohonen fue del 99 %, con un número de individuos que ascendía a 10,427.

La razón por la cual el resto de los entrevistados no formó parte de los conglomerados deviene principalmente del hecho de que los nodos en los que están ubicados no llegaron a ser agremiados en algún grupo en particular ya que las conexiones que estas neuronas poseían con los clusters eran débiles, lo que evidencia la incapacidad de poder considerar que dichos puntos podían compartir similitudes con otros nodos.

En el mismo orden de ideas, el conglomerado que integra a la mayor cantidad de encuestados es el cluster 2 (C2) con 3,528 personas, representando un 33.8 % del total de la muestra clusterizada; el segundo conglomerado con mayor número de individuos es el cluster 1 (C1) con 2,010 entrevistado, sintetizando el 19.3 % del total de la cohorte; seguidamente, el orden final de conglomerados por cantidad de agrupados viene dado primero por los clusters 3 y 5, que congregan a un 16.9 % en ambos casos (con un número de individuos asociados de 1,765 y 1,758, respectivamente) y el cluster 4, que reúne al 13.1 % de entrevistados (con una cifra que asciende a 1,366 personas). En el mismo sentido, existen diferencias estadísticas marcadas entre la distribución de los individuos en las categorías de cada factor considerado, ya que los resultados de las pruebas de independencia de medias permiten concluir que cada conglomerado es diferente de los demás para todas las variables de estudio, para los niveles de significancia establecidos.

Considerando el "Nivel de conocimiento sobre el VIH/SIDA", puede indicarse que los conglomerados 1 (C1) y 5 (C5) se encuentran caracterizados por una proporción relativamente homogénea de individuos que desconocen aspectos importantes sobre el VIH/SIDA y personas que están conscientes de la naturaleza elemental del virus y la enfermedad; es decir, existe cierto equilibrio entre el desconocimiento y el entendimiento adecuado (56.9 % y 57.6 % de individuos carecen de contenido apropiado sobre el VIH/SIDA, respectivamente para el cluster 1 y 5; 43.1 % y 42.4 % de entrevistados conoce sobre el virus y la enfermedad de una forma adecuada). Por otro lado, el conglomerado 2 (C2) es el que posee una mayor proporción de individuos con desconocimiento (74.5 %), seguido por el conglomerado 4 (C4) con un valor de 70.8 % de falta de entendimiento y el conglomerado 3 (C3), con un porcentaje de 64.4 %.

Con respecto al "Género", existe una tendencia similar entre todos los conglomerados identificados en cuanto a la distribución de los varones y mujeres. En todos los casos, la proporción de mujeres supera a la de los hombres, siendo el conglomerado 4 (C4) el que posee una mayor tasa de mujeres como integrantes del grupo, con un porcentaje de 63.4 %, y el conglomerado 5 (C5) con el menor ratio con un tasa de 59.4 %.

Evaluando al "Área de residencia", se puede determinar que los clusters 1 (C1) y 5 (C5) son aquellos que poseen la mayor proporción de individuos que residen en zonas urbanas, con valores de 97.0 % y 96.7 %, de forma correspondiente. En el mismo sentido, los conglomerados 2 y 3 son los grupos que tienen la menor tasa de individuos que residentes en áreas urbanas, con ratios de 44.6 % y 50.8 %, respectivamente.

Tomando en cuenta al "Nivel económico", puede establecerse que el conglomerado 1 (C1) se compone por individuos con una capacidad adquisitiva media o alta (con 65.1 % de sus integrantes dentro de estas dos categorías). El conglomerado 5 (C5) se caracteriza por ser un cluster conformado por individuos con una posición económica media a muy rica (con el 92.2 % de entrevistados localizados en estos niveles de bienestar). De manera disímil, los clusters 2 (C2), 3 (C3) y 4 (C4) pueden describirse en términos de una población pobre o muy pobre en su mayoría (con proporciones de 82.9 %, 89.1 % y 96.2 %, correspondientemente, de encuestados que pertenecen a estas situaciones económicas).

Considerando a la "Región natural", el conglomerado 1 (C1) y 5 (C5) son clusters que se caracterizan por integrantes que, en su mayoría, se encuentran ubicados en Lima Metropolitana o en alguna región de la Costa peruana, agremiando al 63.3 % y 95.4 % de entrevistados, respectivamente. En la misma perspectiva, el conglomerado 2 (C2) se encuentra compuesto por individuos que están localizados en departamentos de la Sierra del país pluralmente, con una tasa de 69.2 % de encuestados en estas zonas. En el cluster 3 (C3), la región natural más frecuente entre los individuos es la Selva del Perú, con un valor de 60.1 %. Por último, el conglomerado 4 (C4) está caracterizado por entrevistados que se ubican en la Costa en su mayoría (49.3 %) y, en menor proporción, en la Selva y Sierra (18.8 % y 22 %, correspondientemente).

Analizando al "Rango etario", tanto el conglomerado 1 (C1), 2 (C2) y 3 (C3) son clusters que se caracterizan por individuos que están dentro del rango de 25 a 29 años (con tasas de 47.1 %, 41.3 % y 53.4 %, respectivamente). El conglomerado 4 (C4) se compone por una muestra de encuestados más jóvenes, ya que el 44.2 % de estos tiene entre 15 y 19 años. De manera particular, el conglomerado 5 (C5) es un grupo que reúne de manera a homogénea a entrevistados considerando al grupo etario al que pertenecen, ya que representan valores similares entre las 3 categorías (31.2 % para 15-19 años, 36.9 % para 20-24 años y 31.9 % para 25-29 años).

Examinando al "Nivel educativo", el conglomerado 1 (C1) y 5 (C5) son los clusters que agremian a un ratio significativo de individuos que tienen como nivel educativo más alto alcanzado a los grados de instrucción secundaria y superior (que engloba estudios universitarios y de posgrado), con valores como 97.9 % y 98.7 %, de forma correspondiente, resaltando que, en

ambos casos, el nivel de secundaria es el que se presenta con mayor frecuencia. Por otro lado, en los conglomerados 2 (C2), 3 (C3) y 4 (C4), la categoría mayoritaria es el grado de instrucción secundaria, con proporciones de 64.5 %, 61.7 % y 79.8 %, respectivamente.

Considerando a la "Lengua materna", en todos los conglomerados, exceptuando al cluster 2 (C2), el idioma natal o primario aprendido por los entrevistados es el castellano, siendo el conglomerado 5 (C5) con la mayor proporción de encuestados castellanohablantes. Empero, debe reconocerse que el cluster 2 (C2) es aquel grupo que se caracteriza por reunir a individuos que tienen como lengua materna a un dialecto originario o nativo del país y a la lengua castellana, con valores de 45.5 % y 54.5 %, para cada caso.

Con respecto a la "Etnicidad", los conglomerados 3 (C3), 4 (C4) y 5 (C5) están compuestos por individuos que se auto-identifican, en su mayoría, como mestizos (75.6 %, 58.7 % y 85.3 %, correspondientemente). El conglomerado 2 (C2) está integrado por entrevistados que se autoreconocen como nativos o pobladores con un origen indígena en el país (87 %). Por otro lado, el cluster 1 (C1) es un grupo que puede describirse como pluri-étnico, ya que integra diversas etnias en términos similares (mestiza con una proporción de 35.7 %, afroperuana con 25.7 % y nativa con 24.9 %).

Evaluando al factor "Oído acerca del VIH/SIDA", puede establecerse que, en todos los conglomerados identificados, los individuos han entrado en contacto con algún tipo de información o contenido concreto sobre el VIH/SIDA con anterioridad a la entrevista de manera notoriamente elevada, siendo el conglomerado 5 (C5) el que reúne a la mayor proporción de encuestados informados (96.2 %).

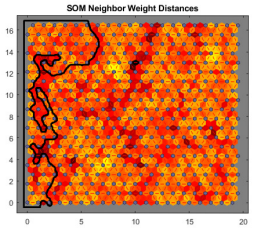
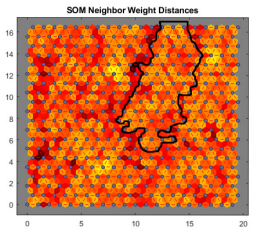
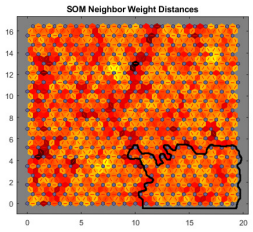
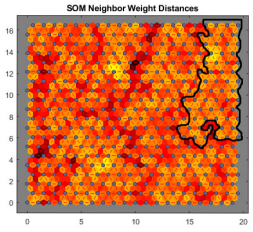
Examinando a la "Prueba de detección del VIH/SIDA", puede determinarse, que en su mayoría, los individuos de todos los conglomerados o clusters no se han realizado una prueba de descarte del VIH/SIDA en los últimos 12 meses antes de contestar la encuesta, resaltando que el conglomerado 4 (C4) es el grupo que tiene el porcentaje más elevado de individuos sin detección del virus o la enfermedad (79.6 %).

Con respecto al "Acceso de medios multi-media", se precisa que, en todos los conglomerados identificados, los entrevistados poseen al menos un medio de comunicación multi-media en sus hogares (sea internet, televisión o radio) en su mayoría, siendo el cluster 5 (C5) el grupo con la mayor tasa de encuestados con medios multi-media (98 %).

Finalmente, considerando al "Género del jefe de hogar", puede señalarse que todos los casos guardan una semejanza en cuanto a la distribución de los géneros de las personas que jefaturan las familias de los entrevistados, revelándose que el género masculino es el que posee la mayor proporción de respuestas, con el conglomerado 3 (C3) el que tiene la mayor tasa (78.9 %).

Sumariamente, la Tabla 6.6 agrupa y exponer las características más relevantes y propias de cada conglomerado identificado según las características socio-demográficas, económicas y de salud, como se visualiza a continuación.

Tabla 6.6: Descripción general de conglomerados basados en dimensión y factores independientes.

Conglomerado	Configuración	Descripción
Cluster 1		Equilibrio entre desconocimiento y entendimiento sobre el VIH/SIDA / Población Femenina / Residencia urbana / Nivel económico medio a rico / Ubicados en Lima Metropolitana o departamentos de la Costa / 25-29 años / Grado de instrucción secundaria a superior / Castellanohablantes / Pluri-étnico / Oído acerca del VIH-SIDA / Baja realización de pruebas de descarte del VIH-SIDA / Acceso a medios multi-media / Jefe varón de familia
Cluster 2		Desconocimiento sobre el VIH/SIDA / Población femenina / Equilibrio en el área de residencia / Nivel económico pobre o muy pobre / Ubicados en la Sierra peruana / 25-29 años / Grado de instrucción secundaria / Castellanohablantes y dialectos originarios / Origen nativo / Oído acerca del VIH-SIDA / Baja realización de pruebas de descarte del VIH-SIDA / Acceso a medios multi-media / Jefe varón de familia
Cluster 3		Desconocimiento sobre el VIH/SIDA / Población femenina / Equilibrio en el área de residencia / Nivel económico pobre o muy pobre / Ubicados en la Selva peruana / 25-29 años / Grado de instrucción secundaria / Castellanohablantes / Mestizos / Oído acerca del VIH-SIDA / Baja realización de pruebas de descarte del VIH-SIDA / Acceso a medios multi-media / Jefe varón de familia
Cluster 4		Desconocimiento sobre el VIH/SIDA / Población femenina / Residencia urbana / Nivel económico pobre o muy pobre / Ubicados en la Costa, mayoritariamente, y seguidamente en la Sierra y Selva de manera similar / 15-19 años / Grado de instrucción secundaria / Castellanohablantes / Mestizos / Oído acerca del VIH-SIDA / Baja realización de pruebas de descarte del VIH-SIDA / Acceso a medios multi-media / Jefe varón de familia
Cluster 5		Equilibrio entre desconocimiento y entendimiento sobre el VIH/SIDA / Población Femenina / Residencia urbana / Nivel económico medio a muy rico / Ubicados en Lima Metropolitana o departamentos de la Costa / Equilibrio entre los rangos etarios / Grado de instrucción secundaria a superior / Castellanohablantes / Mestizos / Oído acerca del VIH-SIDA / Baja realización de pruebas de descarte del VIH-SIDA / Acceso a medios multi-media / Jefe varón de familia

Fuente: Elaboración propia.

6.3.7. Análisis de la distribución geográfica de los conglomerados

Los datos de localización geográfica de cada entrevistado fueron recopilados de la encuesta ENDES y los componentes necesarios para realizar el análisis exploratorio son los que se muestran a continuación en la Tabla 6.7.

Tabla 6.7: Componentes geográficos de la encuesta ENDES.

Variable	Definición	Naturaleza	Rango	Codificación
Ubigeo	Código de ubicación geográfico	Numérico	10101 - 250401	-
longitudx	Coordenada geográfica de longitud del hogar	Numérico	-81.27 - -68.70	-
latitudy	Coordenada geográfica de latitud del hogar	Numérico	-18.07 - -0.97	-
HV024	Departamento en el que viven los encuestados	Categórico	1 - 25	1. Amazonas, 2. Ancash, 3. Apurímac, 4. Arequipa, 5. Ayacucho, 6. Cajamarca, 7. Callao, 8. Cusco, 9. Huancavelica, 10. Huánuco, 11. Ica, 12. Junín, 13. La Libertad, 14. Lambayeque, 15. Lima, 16. Loreto, 17. Madre de Dios, 18. Moquegua, 19. Pasco, 20. Piura, 21. Puno, 22. San Martín, 23. Tacna, 24. Tumbes, 25. Ucayali.

Fuente: Elaboración propia.

Las figuras mostradas subsiguientemente permiten, en primera instancia, visibilizar la disposición de individuos a lo largo del plano nacional y, en segundo lugar, las regiones con las mayores proporciones de entrevistados por conglomerado.

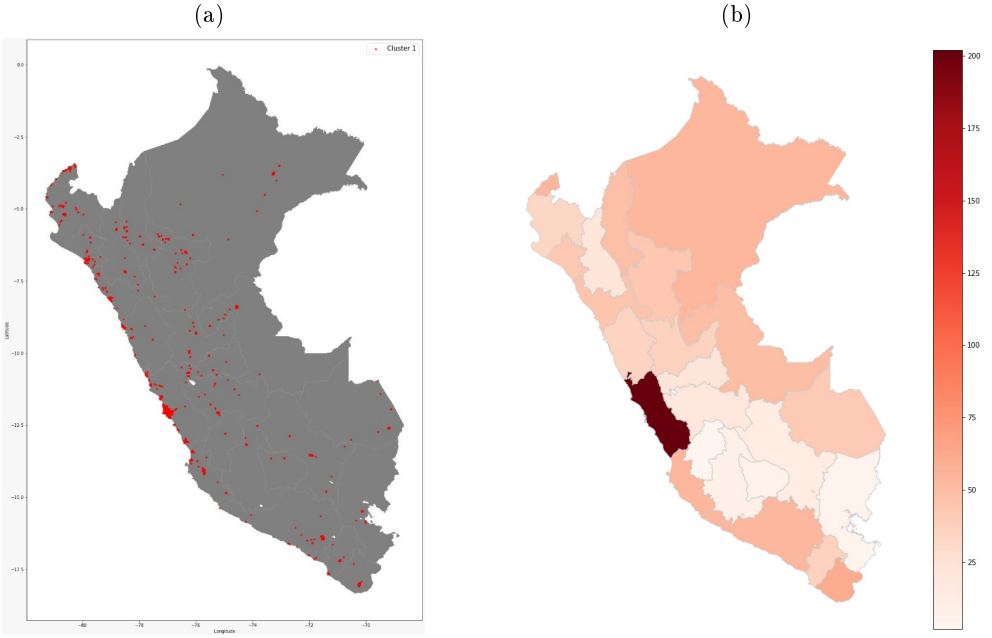


Figura 6.8: (a) Distribución geográfica de los individuos pertenecientes al conglomerado N° 01 y (b) Mapa de calor de la concentración de individuos según regiones en el Perú para el conglomerado N° 01. Fuente: Elaboración propia.

La Figura 6.8 aterriza la caracterización dada al conglomerado 1 (C1) tomando en cuenta a la variable "Región natural". La Figura 6.8a da cuenta que los individuos de este cluster se agrupan a largo de toda la Costa peruana (con énfasis en las regiones norteñas y departamentos del Sur) y con una notoria concentración en Lima Metropolitana y en la Provincia Constitucional del Callao. De manera complementaria, la Figura 6.8b precisa las regiones que poseen

la mayor cantidad de individuos en este cluster. Como se observa, la región de Lima Metropolitana congrega a 413 encuestados (21 % de los miembros de este grupo), convirtiéndola en la región con el mayor foco de asociación. Del mismo modo, las regiones con concentraciones medias del conglomerado son Ica (con 120 personas, 6 % de la cohorte), Callao (117 entrevistados, 6 %), Tacna (109 individuos, 5 %), Loreto (101, 5 %), Tumbes (96, 5 %), Arequipa (92, 5 %), Lambayeque (91, 5 %), Piura (90, 4 %), Ucayali (87, 4 %), Moquegua (86, 4 %), San Martín (84, 4 %) y La Libertad (82, 4 %). Los 13 departamentos mencionados acumulan aproximadamente el 80 % del total del conglomerado (n = 1,568).

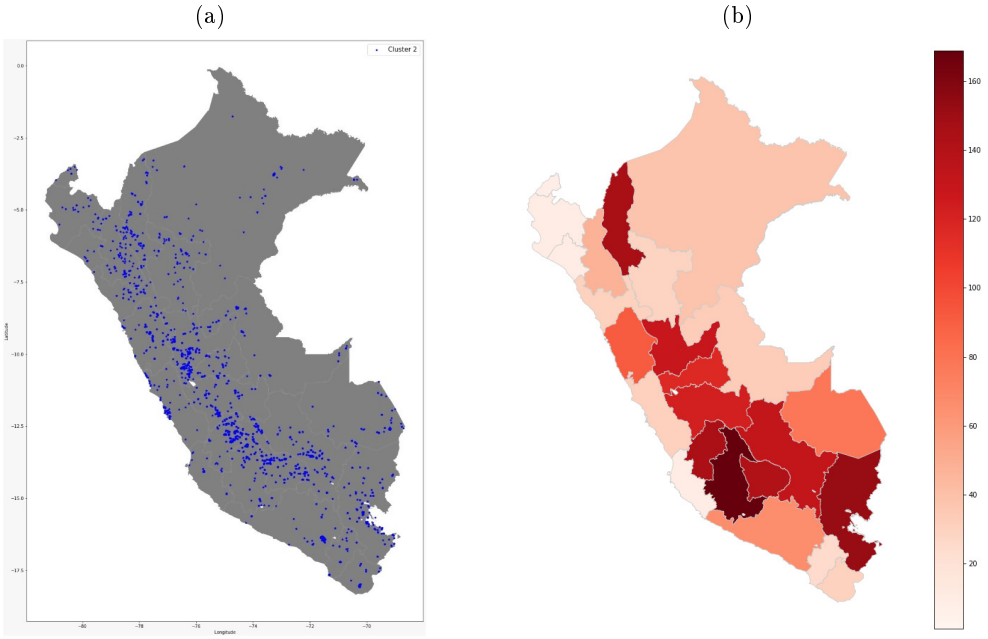


Figura 6.9: (a) Distribución geográfica de los individuos pertenecientes al conglomerado N° 02 y (b) Mapa de calor de la concentración de individuos según regiones en el Perú para el conglomerado N° 02. Fuente: Elaboración propia.

Por su parte, la Figura 6.9a exhibe el comportamiento del conglomerado 2 (C2) en cuando a la ubicación geográfica de sus miembros. Los entrevistados de este grupo se concentran, en su mayoría, en regiones de la Sierra en toda su extensión (desde el sentido norte hacia el sur), con departamentos como Huánuco o Pasco en la zona norte-central de la Sierra, Huancavelica y Ayacucho en la zona central y Puno y la zona altiplánica de Arequipa. En adición, la Figura 6.9b precisa las regiones que poseen la mayor cantidad de individuos en este cluster. Como se observa, en cuanto a los departamentos con las mayores concentraciones, se establece que Ayacucho es la región con la cifra más elevada de individuos reunidos (344), recogiendo al 10 % de personas de la cohorte pertenecientes a este grupo. Seguidamente, zonas como Apurímac y Puno con 316 encuestados (9 %), Huancavelica con 311 individuos (9 %) forman parte de este grupo de regiones con las mayores cifras. Dentro de las proporciones medias en el cluster, Cusco con 254 entrevistados y una concentración de 7 %, Huánuco reuniendo 247 individuos y una tasa de 7 %, Junín con 240 encuestados y un ratio de 7 %, Pasco con 219 personas y una proporción de 6 % encabezan esta lista. Por último, considerando regiones con una

menor cantidad de individuos, se destaca que Madre de Dios (187, 5 %), Áncash (156, 4 %), Amazonas y Arequipa (146, 4 %) integran este conjunto. Los 12 departamentos mencionados acumulan aproximadamente el 80 % del total del conglomerado (n = 2,882).

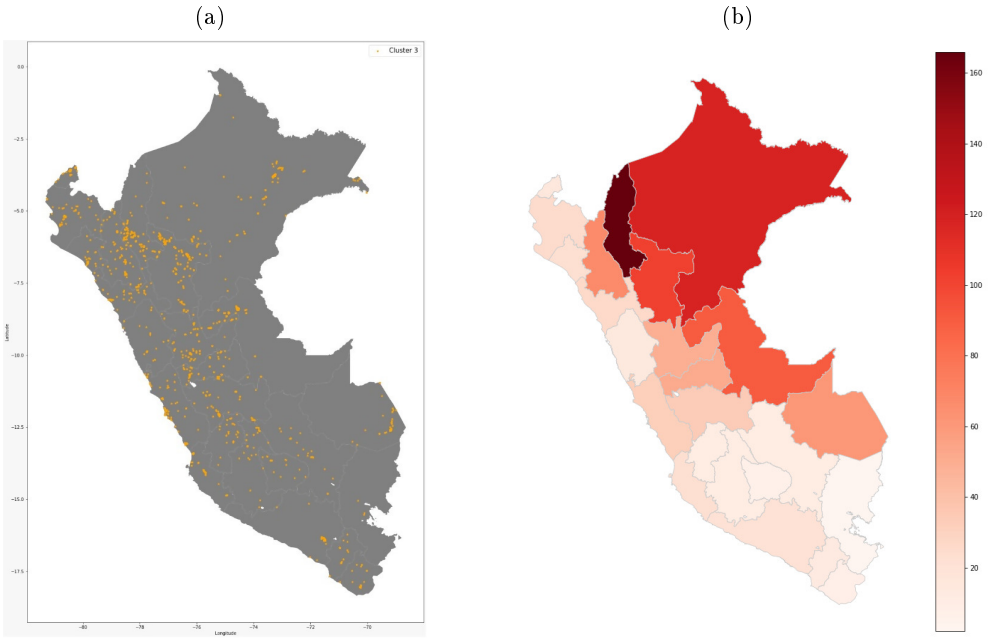


Figura 6.10: (a) Distribución geográfica de los individuos pertenecientes al conglomerado N° 03 y (b) Mapa de calor de la concentración de individuos según regiones en el Perú para el conglomerado N° 03. Fuente: Elaboración propia.

En otro orden de ideas, la Figura 6.10a muestra la tendencia de distribución que asumen los miembros del conglomerado 3 (C3) considerando al factor "Región natural". Los encuestado se agrupan principalmente en regiones de la Amazonía o Selva peruana como, por ejemplo, alrededor de Amazonas, Loreto o San Martín en el norte del país o la ceja de selva en el centro del país con la fracción amazónica de Huánuco o Pasco y la región de Ucayali y en el sur con departamentos como Madre de Dios. En añadidura, la Figura 6.10b precisa las regiones que poseen la mayor cantidad de individuos en este cluster. Como se observa, en cuanto a los departamentos con las mayores concentraciones, se establece que San Martín y Loreto son las regiones con el número más elevado de individuos en este grupo, con cifras como 225 (13 %) y 217 (12 %). Tomando en cuenta las regiones con tasas o proporciones de agremiación medias, se tiene que Ucayali con 178 individuos (10 %), Amazonas con 166 encuestados (9 %), Cajamarca con 116 entrevistados (7 %), Huánuco con 112 personas (6 %) y Madre de Dios con 101 individuos (6 %) conforman a este grupo de regiones con una prevalencia media del cluster. Evaluando los departamentos con las tasas más bajas de agrupamiento que son consideradas como significativas dentro del análisis espacial (ya que forman parte del conjunto de regiones que concentran el 80 % de los miembros del conglomerado), se denota que Pasco con 84 individuos (5 %), Junín reuniendo a 72 personas (4 %), Lima Metropolitana que agrupa a 61 encuestados (3 %) y Piura con 54 sujetos (3 %) pertenecen a esta categoría. Los 11 departamentos mencionados acumulan aproximadamente el 80 % del total del conglomerado

(n = 1,386).

Por otro lado, la Figura 6.11 evidencia el perfil geográfico del conglomerado 4 (C4) descrito en la sección anterior tomando en cuenta las ubicaciones de sus miembros y los focos de densidad de los mismos en el territorio nacional. La Figura 6.11a muestra cómo los integrantes del cluster en cuestión se reparten a lo largo de la costa superior y central en el país, alrededor de los territorios norteños como Piura y Tumbes y trasladándose hacia el centro con departamentos como Lambayeque hasta llegar a Lima Metropolitana e inmediaciones. Igualmente, existe concentraciones relevantes en la Sierra del Perú (se conforman agrupaciones relevantes dentro de regiones como Cajamarca o La Libertad - la fracción altiplánica o que está situada en la serranía). En agregación, la Figura 6.11b precisa las regiones que poseen la mayor cantidad de individuos en este cluster. Puede advertirse que departamentos costeños, entre el norte y centro del Perú, como Piura con 172 integrantes (13 %), Lima con 169 individuos (12 %) y Tumbes con 153 encuestados (11 %) son las regiones que encabezan el conjunto de zonas geográficas que componen al cluster 4 con las tasas más altas de individuos agrupados. Examinando los departamentos con proporciones medias de densidad, se concluye que regiones como Lambayaque con 90 entrevistados (7 %), La Libertad con 87 personas (6 %), Ucayali con 84 individuos (6 %), Cajamarca con 73 agremiados (5 %) e Ica con 60 integrantes (4 %) forman parte de esa fracción de territorios. Por otro lado, Amazonas (58, 4 %), Loreto (53, 4 %), Áncash (49, 4 %) y Madre de Dios (36, 3 %) son los departamentos que constituyen el fraccionamiento de regiones que guardan las proporciones más reducidas en el conglomerado. Los 12 departamentos mencionados acumulan aproximadamente el 80 % del total del conglomerado (n = 1,084).

En suma, la Figura 6.12a expone la organización geográfica de los miembros que conforman al conglomerado 5 (C5) teniendo como premisa elemental a la región natural en la que se encontraban durante el desarrollo de la encuesta. Los entrevistados de este grupo se concentran, mayormente, en departamentos de la Costa peruana. Existen grandes densidades de individuos que integran el cluster que se localizan en regiones norteñas de la Costa como Tumbes dentro del plano territorial. A su vez, departamentos como Lambayaque forman parte del bloque del grupo 5 hasta llegar a la zona costera central para configurar focos de personas en Lima Metropolitana, la Provincia Constitucional del Callao e Ica. En el mismo sentido, las manifestaciones de este conglomerado figuran en la zona costera sur del país, con agremiados en regiones como Tacna o Moquegua (este último, en concentraciones reducidas). En adición, la Figura 6.12b precisa las regiones que poseen la mayor cantidad de individuos en este cluster. Como se observa, la región de Lima Metropolitana congrega a 548 entrevistados (31 % de los miembros de este grupo), convirtiéndola en la región con el mayor foco de concentración. Del mismo modo, las regiones con concentraciones medias del conglomerado son Ica con 188 encuestados (que simbolizan el 11 % del total de la cohorte de C5), Provincia Constitucional del Callao con 179 personas (que representan el 10 % de los miembros del cluster), Lambayaque con 155 individuos reunidos (que alude al 9 % del total de C5), Tumbes con una cifra de 147 de los miembros del grupo (que se le atribuye el 8 % de la cohorte) y, por último, Tacna con 126 entrevistados que la componen (en la que figura el 7 % de los miembros del

conglomerado). Los 6 departamentos mencionados acumulan aproximadamente el 80 % del total del conglomerado (n = 1,343).

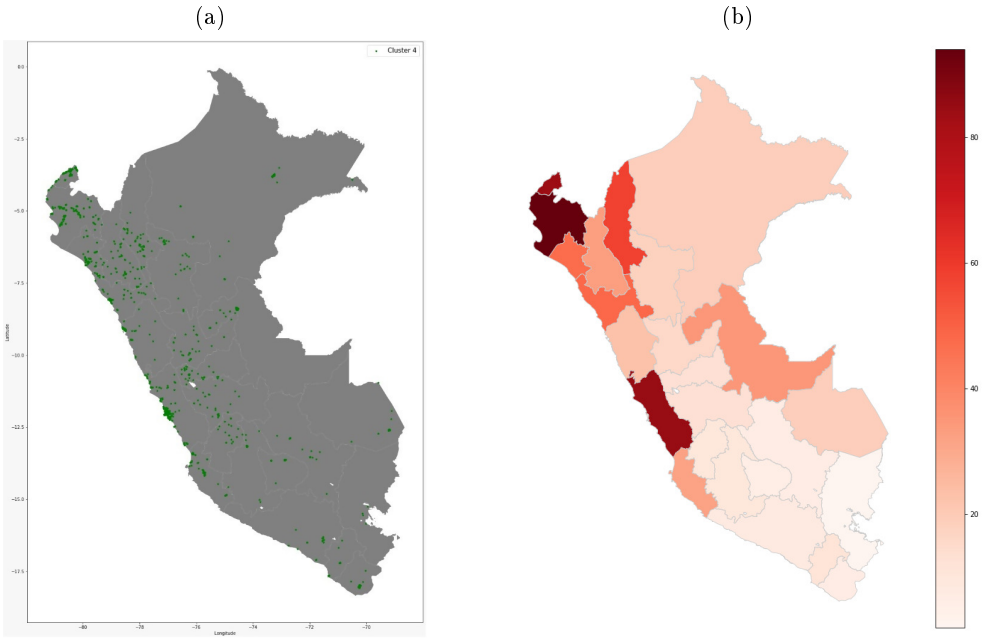


Figura 6.11: (a) Distribución geográfica de los individuos pertenecientes al conglomerado N° 04 y (b) Mapa de calor de la concentración de individuos según regiones en el Perú para el conglomerado N° 04. Fuente: Elaboración propia.

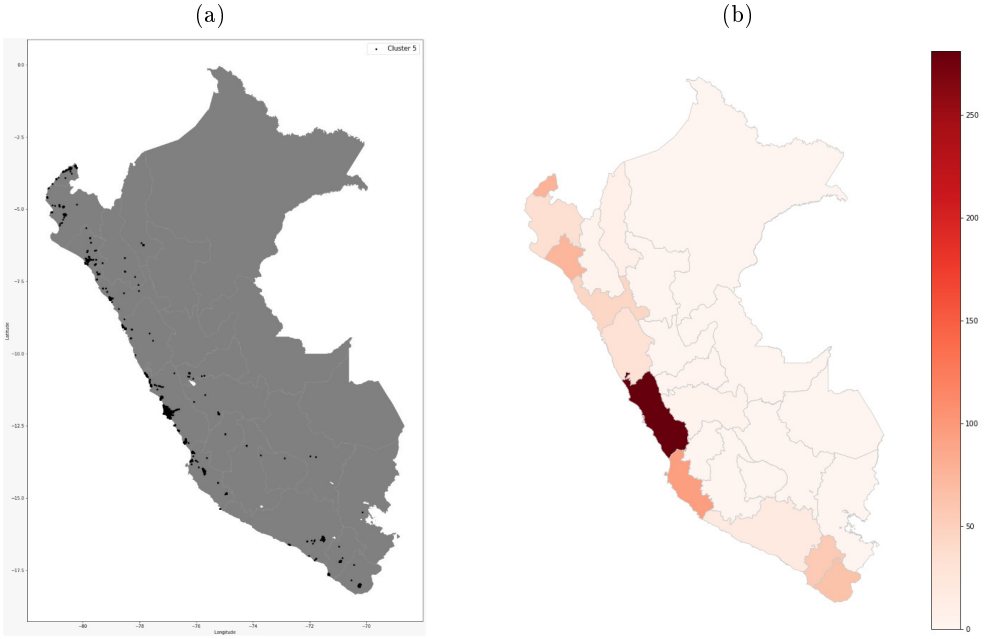


Figura 6.12: (a) Distribución geográfica de los individuos pertenecientes al conglomerado N° 05 y (b) Mapa de calor de la concentración de individuos según regiones en el Perú para el conglomerado N° 05. Fuente: Elaboración propia.

Finalmente, la Tabla 6.8 muestra, por cada conglomerado, el breviario de los departamentos y

el número de individuos que forman parte de ellos, dando cuenta de las regiones que concentran el 80 % de las observaciones para todos los casos.

Tabla 6.8: Distribución de individuos pertenecientes a los conglomerados por departamentos con mayor concentración.

Conglomerados									
C1		**C2**		**C3**		**C4**		**C5**	
Dpto.	n	**Dpto.**	n	**Dpto.**	n	**Dpto.**	n	**Dpto.**	n
Lima	413	Ayacucho	344	San Martín	225	Piura	172	Lima	548
Ica	120	Apurímac	316	Loreto	217	Lima	169	Ica	188
Callao	117	Puno	316	Ucayali	178	Tumbes	153	Callao	179
Tacna	109	Huancavelica	311	Amazonas	166	Lambayeque	90	Lambayeque	155
Loreto	101	Cusco	254	Cajamarca	116	La Libertad	87	Tumbes	147
Tumbes	96	Huánuco	247	Huánuco	112	Ucayali	84	Tacna	126
Arequipa	92	Junín	240	Madre de Dios	101	Cajamarca	73		
Lambayeque	91	Pasco	219	Pasco	84	Ica	60		
Piura	90	Madre de Dios	187	Junín	72	Amazonas	58		
Ucayali	87	Áncash	156	Lima	61	Loreto	53		
Moquegua	86	Amazonas	146	Piura	54	Áncash	49		
San Martín	84	Arequipa	146			Madre de Dios	36		
La Libertad	82								
Total[a]	1,568	**Total**	2,882	**Total**	1,386	**Total**	1,084	**Total**	1,343

Notas. n: frecuencia o número de miembros del cluster, [a] Las distribuciones de individuos mostradas representan el 80 % del total de observaciones registradas por conglomerados.

Fuente: Elaboración propia.

Para una mayor exploración y visualización en detalle de la distribución de los encuestados por conglomerado identificado, se han creado mapas interactivos en formato de páginas web a los que se puede acceder como material suplementario en este capítulo: `https://drive.google.com/drive/folders/1oNwjswr6BBFCtysq4flstlUvuRXi8NlE?usp=sharing`.

Capítulo 7

Conclusiones

Este capítulo presenta las principales conclusiones y recomendaciones extraídas del análisis determinístico de las proyecciones sobre el VIH/SIDA en el Perú y la evaluación de las acciones gubernamentales por parte del estado en cuanto a la contención y control de la epidemia; los determinates socio-demográficos, económicos y de salud que caracterizan a los adolescentes y jóvenes adultos según el nivel de conocimiento que ostentan sobre el virus y la enfermedad a lo largo del país; la clasificación supervisada del nivel de entendimiento que los individuos puedan tener y los hallazgos resultantes de los conglomerados exploratorios basados en determinantes de la salud estructurales que caracterizan a la muestra de estudio.

7.1. Discusión

En esta disertación, el objetivo general fue realizar la caracterización de la dinámica epidemiológica y del nivel de conocimiento sobre el VIH/SIDA en los habitantes del Perú. Para lograr dicho objetivo, a lo largo del presente trabajo, se propusieron y desarrollaron metodologías adecuadas para abordar estos desafíos. En las etapas metodológicas establecidas, que estuvieron motivadas por distintas preguntas de investigación y la naturaleza de los datos disponibles, presentamos tres enfoques amplios cuyo cumplimiento permitió identificar y emplear modelos para el pronóstico de aspectos relevantes de la evolución de la epidemia en el país y la detección de potenciales focos de atención de actuales y nuevas políticas de salud a nivel gubernamental.

Las conclusiones de este trabajo se pueden desagregar según los objetivos específicos establecidos.

En primer lugar, a fin de conocer el contexto epidemiológico del VIH/SIDA en el Perú y los diferentes cursos de evolución que la epidemia pueda asumir dependiendo de la gestión pública por parte del estado en materia de salud, el objetivo específico fue diseñar y evaluar un modelo probabilístico basado en Cadenas de Markov que consiguió la estimación determinística de la situación epidemiológica y el desarrollo prospectivo del VIH/SIDA a nivel nacional basado en cada uno de los cambios más estructurales o significativos de la política de salud relacionada a la epidemia registrados en el Perú.

En ese sentido, mediante los experimentos y simulaciones realizadas, se comprobó que el mo-

delo de Markov adoptado en este trabajo permite representar fielmente la evolución dinámica de la epidemia, y así se logró aproximar un proceso natural real y complejo.

Este estudio de modelización de un sistema tan complejo como el VIH y el SIDA demuestra ser una herramienta poderosa en la investigación epidemiológica que puede ofrecer una comprensión más profunda de la progresión natural de la enfermedad y su transmisión. El análisis que se presenta posee la operatibilidad suficiente para que pueda ser trasladado y aplicado a una variedad adicional de modelos de epidemias basados en agentes víricos e infecciosos más generales teniendo en cuenta las variaciones en los parámetros del modelo para proporcionar trayectorias generales de la enfermedad a través de estados intermedios que puedan alertar respuesta antes de que ocurra un evento adverso.

Asimismo, el modelo es capaz de brindar información sobre la combinación de tratamientos que es más efectiva en cada estado de la dinámica de la epidemia y qué combinación incurre en un beneficio mayor y sobre las variaciones en eficiencia de las intervenciones orientadas a la gestión de la epidemia del VIH a largo plazo.

En segundo lugar, con el propósito de proveer información sobre aquellos factores estructurales sociales que influyen directamente en el conocimiento del VIH/SIDA por parte de la población juvenil en territorio nacional y ser capaces de clasificar el perfil de percepción y entendimiento sobre la epidemia de estos individuos, los objetivos específicos en esta fase metodológica se basan en el ajuste de una técnica estadística multivariante que nos permitió estimar la asociación empírica existente entre los regresores independientes y el nivel de conocimiento sobre la epidemia y la definición de modelos computacionales paramétricos y no paramétricos para la estimación y predicción de dicho conocimiento.

Los hallazgos contribuyen a esta línea de investigación al indicar que es más probable que ciertos factores estructurales (entre ellos: demográficos, sociales, familiares y culturales), más que conductuales y médicos, estén asociados con el nivel de percepción sobre las formas de transmisión y evolución del VIH/SIDA y al confirmar que existe una diferencia marcada de este conocimiento a nivel nacional. En la evaluación de los resultados, se encuentran diferentes factores que ofrecen la posibilidad de modificados con el propósito de diseñar e implementar programas preventivos bajo un enfoque estructural.

Por otro lado, al examinar si los algoritmos paramétricos y no paramétricos podrían aprender de los datos nacionales existentes y ayudar a predecir el nivel de conocimiento sobre el VIH/SIDA, el estudio muestra que tal enfoque es factible y que los algoritmos tienen una precisión relativamente alta para la predicción de la percepción y entendimiento de la dinámica y naturaleza de la epidemia. El enfoque permitió establecer las características más predictivas para el nivel de conocimiento considerando la naturaleza compleja de varios predictores del entendimiento del VIH/SIDA para brindar una comprensión intuitiva de las características clave.

De la misma manera, este estudio demuestra el potencial de los modelos de aprendizaje automático para identificar a los adolescentes y jóvenes adultos que tienen una mayor expo-

sición o una mayor predisposición de aceptar nocisiones incorrectas sobre la transmisión del VIH/SIDA y asumir comportamientos de riesgo bajo esas ideas basado en ciertos regresores. Además, esta metodología facilita la selección de un modelo en función del desempeño con limitaciones de recursos y la estabilidad del desempeño a lo largo del tiempo. La incorporación de dicha información en los algoritmos puede mejorar potencialmente las propiedades discriminatorias en los modelos.

Mediante la experimentación del diseño y ejecución de estos algoritmos de aprendizaje automático, se puede comprobar que dichos modelos pueden identificar relaciones incluso cuando algunos de los datos de entrada son muy complejos, mal definidos y mal estructurados.

En tercer lugar, con la finalidad de desarrollar una evaluación exploratoria para determinar agrupamientos del nivel de conocimiento sobre el VIH/SIDA y factores estructurales que los individuos tengan a lo largo del territorio nacional, el objetivo específico en esta etapa metodológica fue diseñar y desarrollar un método de aprendizaje no supervisado basado en redes neuronales que permitió identificar tipos de patrones y/o comportamientos que siguen los hombres y mujeres que residen en el país, cuantificando y localizando aglomeraciones en un plano geográfico a lo largo del Perú.

La metodología propuesta permitió identificar grupos de adolescentes y jóvenes adultos que comparten características similares; es decir, permite la identificación de los grados de pertenencia de los habitantes del país hacia un perfil y modo de conducta particular basado en factores culturales, económicos, sociales y sanitarios.

La implementación de una red SOM demuestra ser un método eficaz para detectar la fragmentación de las necesidades de los ciudadanos peruanos considerando la heterogeneidad de los conglomerados identificados. De la misma manera, las ventajas del enfoque se relacionan con su forma simple, general y modular; ya que la metodología utilizada permite no solo la identificación de un individuo en un clúster sino también su localización dentro del mismo (su posición relativa basada en la relación del nodo que contiene a la observación con el resto de los individuos).

Asimismo, el mapa de Kohonen se configura como una plataforma eficaz para la visualización de datos de alta dimensión, permitiendo una división eficiente y óptima de los prototipos generados con las observaciones de un conjunto de datos en grupos apreciables y distinguibles. A través de una presentación e interpretación visual aproximada de los clústeres y la ventaja de contar con un reducido costo computacional, este método permite comprender mejor la relación entre las variables independientes consideradas como atributos de entrada y los resultados de la configuración y conformación de los conglomerados, permitiendo determinar los niveles de correlación entre los factores y la forma del mapa, evidenciando su potencial discrimante e interpretativo para dividir y ofrecer información sobre los agrupamientos obtenidos con los datos. De la misma manera, el mapa SOM hace más hincapié en las similitudes locales y distingue claramente los grupos dentro de los datos; ofreciendo una descripción general clara de las relaciones locales entre los datos.

Por otro lado, la red neuronal de Kohonen se perfiló como una poderosa herramienta para el traslado de los resultados estadísticos de los conglomerados identificados hacia una perspectiva o dimensión de análisis geográfico. La investigación, en su carácter exploratorio y descriptivo en su análisis de la ubicación y distribución de los individuos según conglomerado, proporcionó una descripción visual singular de los resultados basados en el conocimiento sobre el VIH/SIDA y diferentes determinantes estructurales a lo largo del Perú, considerando la compleja interacción de estas características socio-económicas, de salud sexual y cultural a nivel de ubigeo que impactan en la prevalencia de dichas covariables a nivel macro relacionadas a la epidemia en toda el área nacional más allá de los factores de riesgo y sexuales individuales que pudieran reportarse.

Estos hallazgos brindan la posibilidad de adecuar intervenciones públicas a las necesidades específicas de la población según los perfiles detectados, así como diseñar la priorización de iniciativas en las localidades (áreas) más vulnerables o comprometidas según las categorías de factores que presentan mayores desafíos en la sociedad, contribuyendo a la identificación y visualización de estos focos a nivel nacional, puesto que el análisis geográfico que pudo desprenderse del mapa de Kohonen permitió encontrar múltiples áreas geográficas por cada conglomerado epidemiológicamente distintas y que pueden caracterizarse por ciertos factores de riesgo sociales y conductuales del VIH/SIDA; confirmando así que el análisis espacial realizado es relevante para comprender dónde y qué poblaciones deben recibir una atención especial en términos de actividades de prevención primaria y secundaria, resumiendo los aspectos geográficos clave de esta epidemia, para ayudar a las partes interesadas adecuadas a diseñar programas de prevención personalizados y asignar recursos de manera más eficaz. Por tanto, los análisis geográfico y espacial que la red SOM expone son potencialmente útiles herramientas completas para investigadores que buscan comenzar un examen de cuestiones de equidad espacial y formuladores de políticas que buscan identificar herramientas para la toma de decisiones con respecto a la asignación de recursos de servicios sociales y de salud.

7.2. Limitaciones y futuras investigaciones

Pese a que la finalidad ulterior de la investigación fue alcanzada, se pudieron identificar ciertas limitaciones en la configuración y aplicación de las técnicas seleccionadas que pueden sentar las bases para su optimización y mejoramiento a través de trabajos futuros.

En un marco general, la disponibilidad restringida de datos sobre el VIH/SIDA en el Perú es una limitación seria en este estudio. La naturaleza de los datos disponibles, en gran medida, adaptó la dirección de esta tesis. La agregación de los datos a nivel nacional hizo imposible realizar un análisis en niveles más detallados o específicos. Asimismo, el estudio y el modelado de la tendencia de la epidemia en los distintos estratos demográficos de la sociedad peruana se vieron obstaculizados por la falta de estratificación de los datos. Se recomienda que, a mediano y largo plazo, los datos se amplien en favor de una mayor disponibilidad para el dominio público después de eliminar todas las identidades de los pacientes y que sean más accesibles para los investigadores.

En un marco más específico, también pueden señalarse limitaciones dentro de cada etapa metodológica propuesta en la introducción de la investigación.

Tomando en consideración el diseño y evaluación de un modelo probabilístico basado en Cadenas de Markov, se puede precisar que un modelado estocástico eficaz se basa en el desarrollo de configuraciones del diseño del algoritmo y la disponibilidad de datos de calidad. Ambos presentaron desafíos o limitantes para el desarrollo de un modelo útil que analizaria la dinámica epidemiológica del VIH/SIDA en el Perú. El país carece de datos en niveles determinados del sistema de salud, apoyándose en valores modelados por agentes internacionales o privados ajenos a la administración pública con diferentes supuestos en la captura de los datos que dificultan su uso para el propósito planteado (a saber: se tuvo que diseñar un estado VIH/SIDA, a diferencia de un estado VIH y otro estado SIDA por separado, al no contar con estadísticas oficiales sobre la población viviendo con SIDA exclusivamente; entre otros inconvenientes relacionados a la manipulación de los datos). Como resultado, la selección de las estadísticas epidemiológicas y los estados de transición que se utilizarían no solo se basó en la idoneidad de los mismos, sino también en la disponibilidad de datos fiables y completos.

No obstante, como trabajo futuro contemplando datos más sólidos y diversificados, se puede validar el modelo desarrollado en otros segmentos demográficos afectados por el VIH/SIDA con una cobertura de muestreo más baja entrenándose para que sea aplicable a epidemias localizadas muy diferentes (a saber: trabajadores sexuales, población homosexual, hombres que tienen sexo con otros hombres, adictos a sustancias químicas, entre otros). Por añadidura, otros estudios futuros pueden considerar estados adicionales al modelo de Markov de tipo S-I-R para distinguir entre las personas que murieron y las que se volvieron indetectables después de un proceso de terapia antirretroviral. En otro contexto, el modelo de Cadena de Markov puede ser utilizado por profesionales de la salud y otros especialistas en modelado para estimar otros indicadores epidemiológicos relevantes de enfermedades infecciosas, así como para generalizar las probabilidades de transición para predicciones futuras para otras epidemias o agentes infecciosos.

En relación a la fase metodológica de identificación de determinantes estructurales de la salud que influyen en el nivel de conocimiento sobre el VIH/SIDA y el modelamiento predictivo del mismo, se puede estableceer que una limitación en el análisis es la medición estática y asociativa del conocimiento sobre la epidemia y los factores socioeconómicos, de salud y familiares inherente al diseño transversal. Una investigación futura debe incluir diseños de tipo longitudinal que puedan medir el nivel actual de conocimiento de un individuo, así como su comportamiento futuro para confirmar la relación temporal y causal. Este tipo de datos permitiría realizar una valoración convincente de los efectos de la percepción y el conocimiento del riesgo sobre el comportamiento.

De la misma manera, la utilización de datos provenientes de fuentes secundarias (como la encuesta ENDES) generó inconvenientes que se traducen en el hecho de que la selección de variables, la calidad de los datos y los indicadores de medición estuvieran fuera del control o determinación del investigador a priori. De manera similar, el estudio puede haber tenido

sesgos de respuesta durante la recopilación de factores de riesgo o sensibles percibidos (i.e. realización de pruebas de detección del VIH/SIDA, auto-percepción étnica, entre otras), aunque esta preocupación es común a la mayoría de los estudios de comportamiento autoinformado. Además, los datos recogidos fueron del año 2019; mientras tanto, el nivel de conocimiento de la población objetivo puede haber cambiado y los resultados presentados previamente pueden no reflejar con exactitud la situación presente del conocimiento entre adolescentes y jóvenes adultos en el Perú. Por lo tanto, los patrones encontrados en este estudio deben ser evaluados por expertos en salud (que tienen experiencia en el dominio del problema) para decidir si son lógicos, prácticos y novedosos para impulsar nuevas direcciones de investigación biológica y clínica. Adicionalmente, se recomienda el uso de covariables adicionales y quizás datos de historia sexual recopilados de diferentes muestras para extraer más inferencias sobre las diferencias asociadas a las características comunitarias o individuales. La identificación de tales diferencias podría mejorar la comprensión de las variaciones en el conocimiento del VIH/sida observadas en la poblacion.

En otra perspectiva, se necesitan estudios futuros adicionales para evaluar más a fondo la utilidad y los efectos de los algoritmos paramétricos y no paramétricos empleados y deberían estar dirigidos a mejorarlos. Esto puede incluir la exploración de otras técnicas y configuraciones de aprendizaje automático para mejorar el rendimiento de los modelos: variaciones en la complejidad y dimensionalidad asociada a la construcción de modelos, modelos que atiendan a la naturaleza de series de tiempo del problema, redefinición o inclusión de nuevos predictores y exploración de interacciones subyacentes en los modelos actuales que puedan explotarse. A su vez, advirtiendo la deficiencia de interpretabilidad y comprensibilidad de las cuales ciertos métodos de aprendizaje automático empleados en esta sección todavía adolecen (por ejemplo, si un método de aprendizaje automático está explotando algunos efectos de interacción y no linealidad, el examen de la importancia de las variables basado en el modelo no puede explicar ni dar cuenta por completo de tales mecanismos predictivos), mejorar el mecanismo de interpretabilidad en la generación de resultados es una dirección crítica y atractiva, pero en gran medida poco estudiada, que puede ser motivo de investigación posterior.

Contemplando la etapa de la metodología que versa sobre la aplicación de una red neuronal o mapa de Kohonen para el agrupamiento social y geográfico de los adolescentes y jovenes adultos en el Perú, se detecta que una limitación radica en el uso de los resultados que encuestas de fuentes secundarias generan. Este tipo de encuestas producen datos transversales destinados a proporcionar una medida del nivel de conocimiento sobre el VIH/SIDA y factores estructurales entre la población general; y por lo tanto, existe una subrepresentación o, incluso, una nula consideración dentro del análisis de poblaciones de alto riesgo o móviles (como trabajadoras sexuales, hombres que tienen relaciones sexuales con hombres, conductores de camiones y trabajadores de temporada); por ello, una ampliación del alcance de estas encuestras con otras fuentes de datos que permitan el mapeo de poblaciones clave permitirán una identificación más precisa de las nociones sobre el VIH/SIDA y puede mejorar aún más nuestra comprensión de la epidemia. Por otro lado, las ubicaciones de muestreo de la encuensta se seleccionan al azar, sobre la densidad de población subyacente dentro del país. En

consecuencia, pueden existir ubicaciones poco muestreadas y con una menor representación de los individuos de dichas zonas. Es por ello que resulta imporante implementar en este tipo de estudios técnicas de muestreo alternativas que sobremuestreen áreas con bajas densidades poblacionales a fin de aumentar la confiabilidad de resultados de encuestas futuras.

En relación al método empleado, una oportunidad de mejora debería centrarse en integrar mejor los campos de visualización de información y reducción de dimensiones. Así como su combinación con otros métodos de minería de datos, como la agrupación pura o las técnicas de escalado multidimensional, para mejorar lols resultados que pueda generar y su confiabilidad. Otros experimentos pueden llevarse a cabo mediante el uso de diferentes vectorizadores en la construcción del mapa SOM y la aplicación de diferentes métricas de distancias, con el fin de investigar su impacto en los resultados de la clasificación (ya que solo se empleó la distancia euclidiana en el estudio). De manera complementaria, los análisis descriptivos previos en el análisis geográfico de la distribución de individuos de acuerdo al conglomerado en el que están ubicados apuntan a la necesidad de un mayor análisis exploratorio de datos espaciales que, en futura instancia, puedan proporcionar una dirección para las políticas públicas y la asignación de recursos sanitarios de acuerdo a medidas de correlación espacial o análisis espacial estadístico.

Anexos

Apéndice A

Matriz de transición del n-ésimo paso P^n

En este anexo se presenta el procedimiento para el cálculo de una matriz de transición del n-ésimo paso P^n considerando el método de descomposición en autovalores y autovectores de una matriz de transición de probabilidades. Este procedimiento se realizará en base al año de estudio 1995 del Capítulo 4.3.1.

El cálculo de la matriz de transición del n-ésimo paso P^n a nivel nacional para el año 1995 se muestra a continuación.

Al resolver la expresión determinística conocida como la ecuación característica de P,

$$|\lambda I - P| = 0,$$

es posible obtener los autovalores $\lambda_1 = 1$, $\lambda_2 = 0{,}999862$ y $\lambda_3 = 0{,}9529720$, con la matriz de autovectores correspondiente al año 1995:

$$Q = \begin{bmatrix} 5{,}774 \times 10^{-1} & 1 & -0{,}294 \times 10^{-2} \\ 5{,}774 \times 10^{-1}0 & 0 & 9{,}999 \times 10^{-1} \\ 5{,}774 \times 10^{-1}0 & 0 & 0 \end{bmatrix}$$

A su vez, basándose en la matriz de autovectores, la inversa de la matriz Q es expresada como:

$$Q^{-1} = \begin{bmatrix} 0 & 0 & 0{,}173 \times 10^{1} \\ 1 & 0{,}2943 \times 10^{-2} & -0{,}100 \times 10^{1} \\ 0 & 0{,}100 \times 10^{1} & -0{,}100 \times 10^{1} \end{bmatrix}$$

Empleando la Ecuación 3.5,

$$P^n = Q\lambda^n Q^{-1},$$

donde λ^n corresponde a:

$$\lambda^n = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0{,}999^n & 0 \\ 0 & 0 & 0{,}953^n \end{bmatrix}$$

Así, la matriz de probabilidades de transición P^n para la evolución de la epidemia en el Perú en el año 1995 pudo ser estimada con los siguientes valores:

$$P^n_{Peru1995} = \begin{pmatrix} 0{,}999^n & -0{,}003(0{,}953^n - 0{,}999^n) & 1 - 1{,}003(0{,}999^n) + 0{,}003(0{,}953^n) \\ 0 & 0{,}953^n & 1 - 0{,}953^n \\ 0 & 0 & 1 \end{pmatrix} . \text{(A.1)}$$

Apéndice B

Comportamiento estacionario del estado Muerte

En este anexo se presentan las figuras que complementan visual y descriptivamente el análisis de las curvas de comportamiento estacionario del estado (3) Muerte que caracteriza la dinámica epidemiológica del VIH/SIDA en el Perú para los años de estudio (1995, 2005, 2011, 2013 y 2018) planteados en el Capítulo 4.3.4 a fin de garantizar una visibilidad más clara y apreciable de la evolución de este estado en el largo plazo y las diferencias relevantes que puedan abstraerse entre los períodos o hitos de análisis.

Apéndice B.1: Comportamiento estacionario del estado Muerte del año 1995

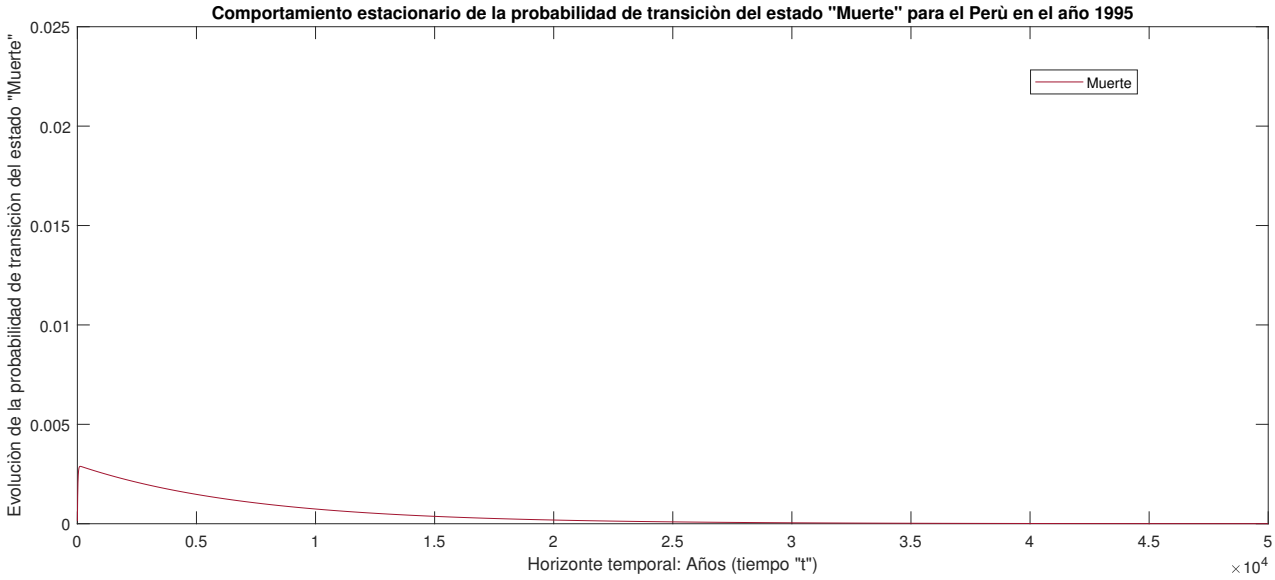


Figura B.1: Comportamiento estacionario del estado "Muerte" en el Perú para el año 1995. Fuente: Elaboración propia.

Apéndice B.2: Comportamiento estacionario del estado Muerte del año 2005

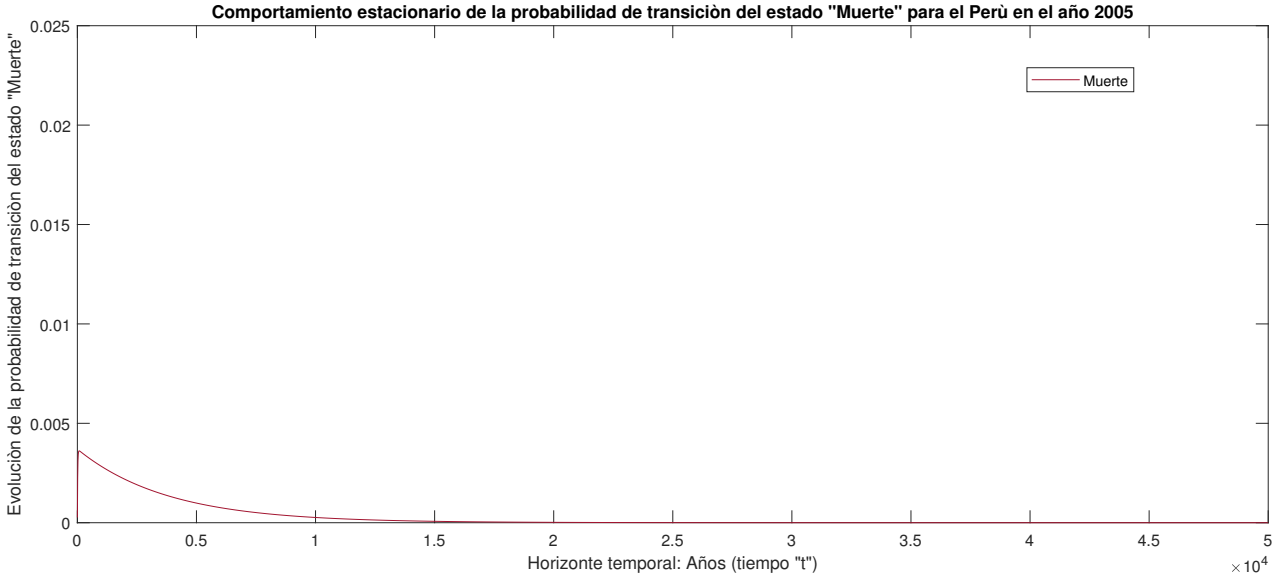


Figura B.2: Comportamiento estacionario del estado "Muerte" en el Perú para el año 2005. Fuente: Elaboración propia.

Apéndice B.3: Comportamiento estacionario del estado Muerte del año 2011

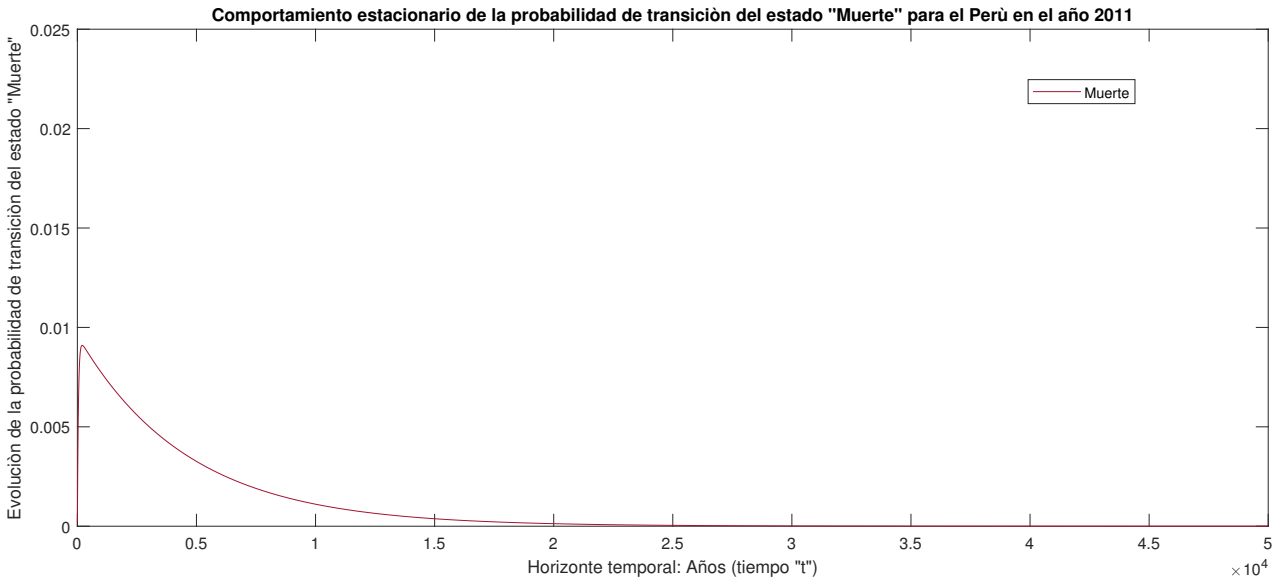


Figura B.3: Comportamiento estacionario del estado "Muerte" en el Perú para el año 2011. Fuente: Elaboración propia.

Apéndice B.4: Comportamiento estacionario del estado Muerte del año 2013

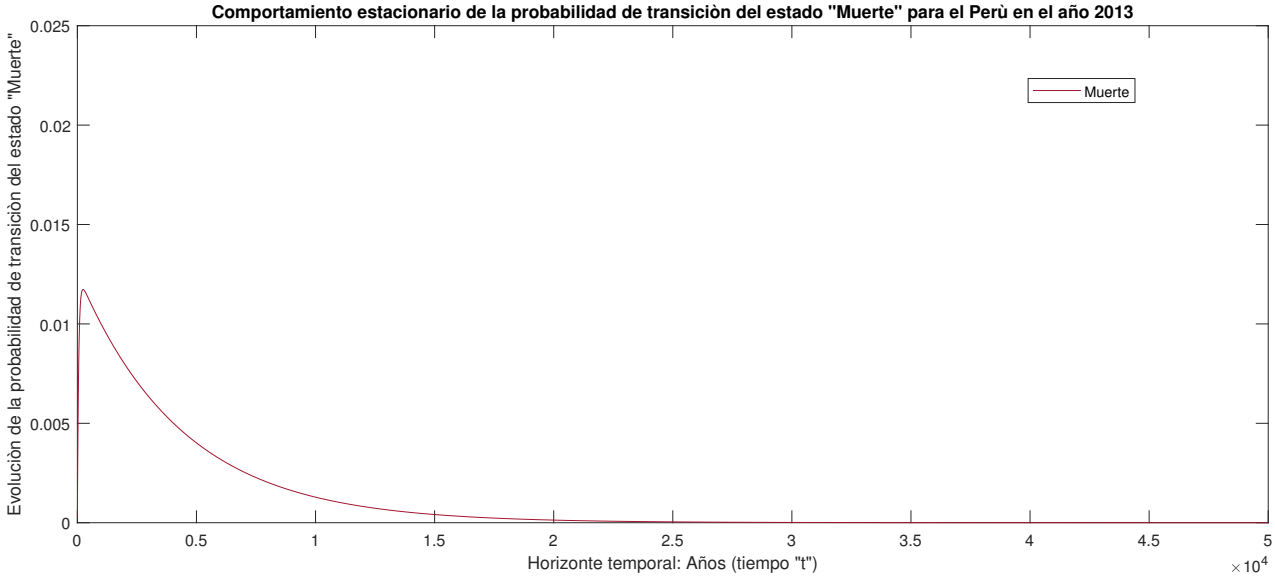


Figura B.4: Comportamiento estacionario del estado "Muerte" en el Perú para el año 2013. Fuente: Elaboración propia.

Apéndice B.5: Comportamiento estacionario del estado Muerte del año 2018

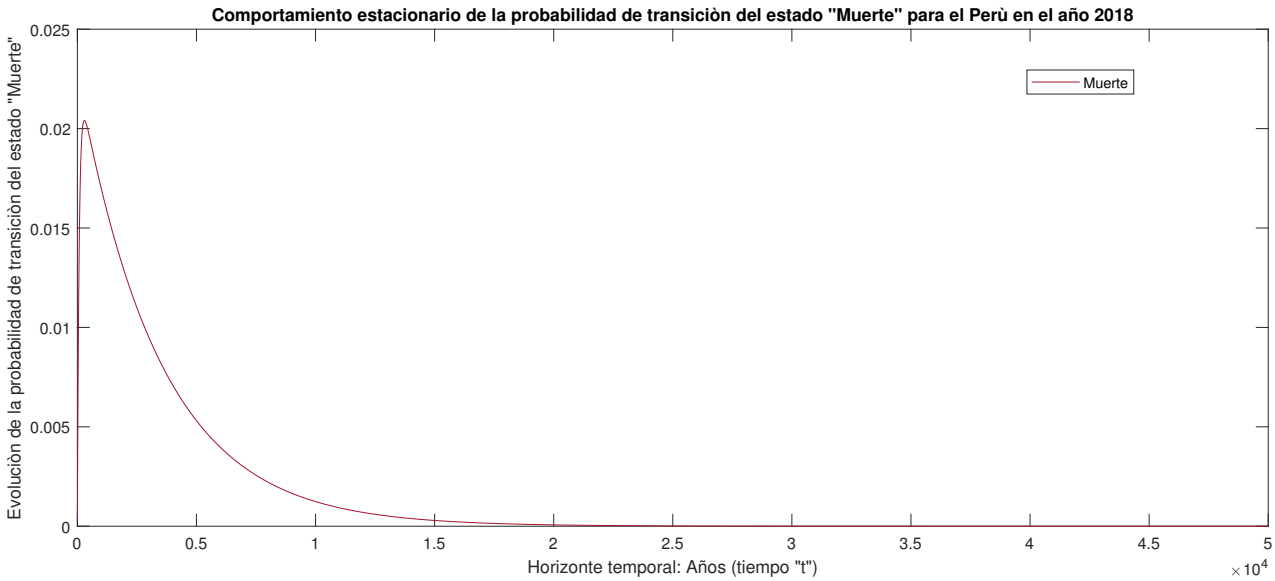


Figura B.5: Comportamiento estacionario del estado "Muerte" en el Perú para el año 2018. Fuente: Elaboración propia.

Apéndice C

Análisis de importancia de variables

En este anexo se presentan los gráficos de importancia de variables, medidos a través de los valores de correlación local de cada categoría o nivel de las covariantes o variables independientes considerados en el estudio, asociados a modelos como Regresión Logística, Decision Trees, Redes Neuronales Artificiales y algoritmo k-NN planteados en el Capítulo 5 que tuvieron un menor desempeño predictivo que el modelo de Random Forest mostrado en el capítulo en cuestión.

Apéndice C.1: Importancia de variables para el modelo de Regresión Logística

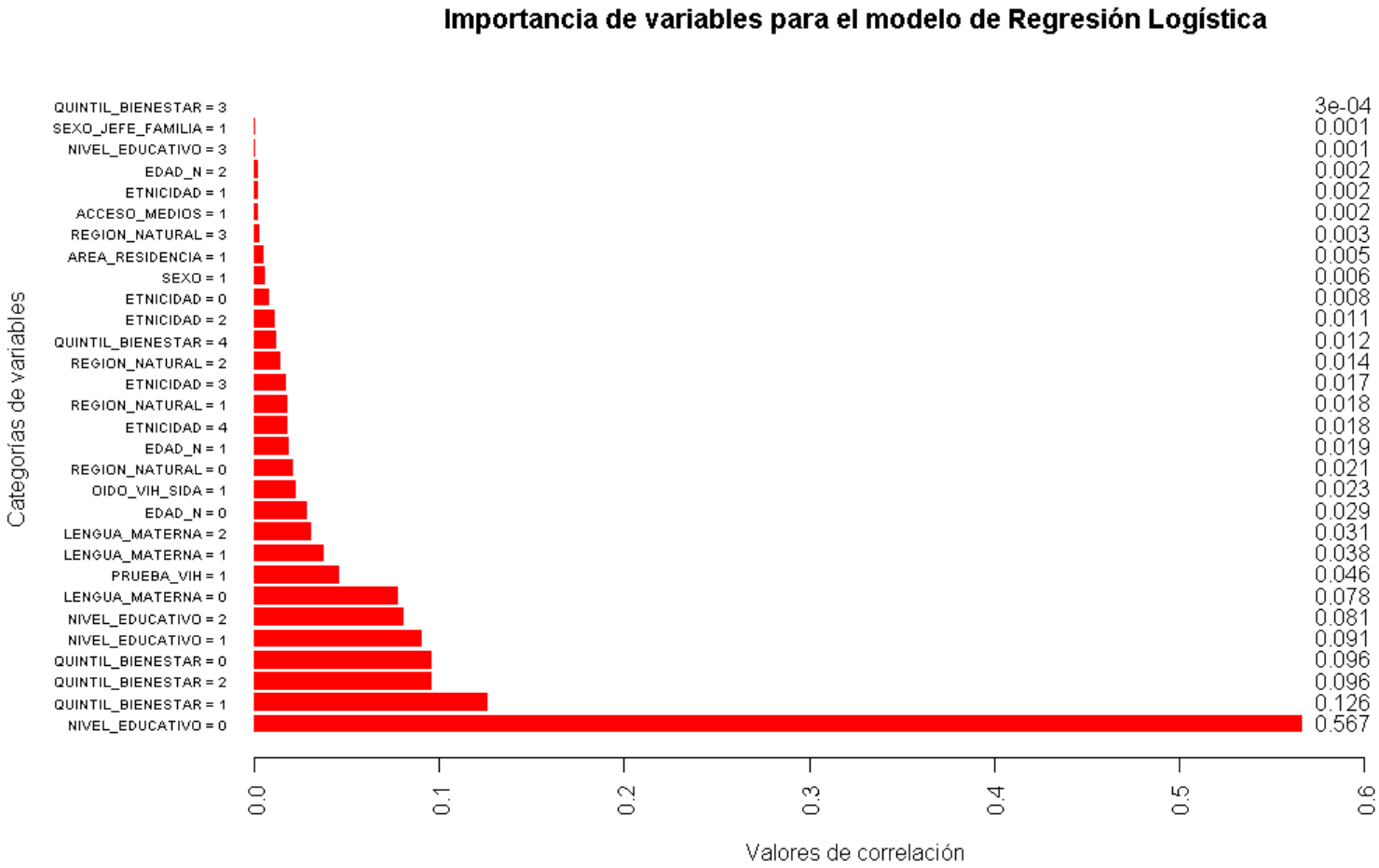


Figura C.1: Medidas de influencia de las variables o regresores independientes en la predicción del nivel de conocimiento sobre el VIH/SIDA en la población peruana en el modelo de Regresión Logística. Fuente: Elaboración propia.

Apéndice C.2: Importancia de variables para el modelo de Decision Trees

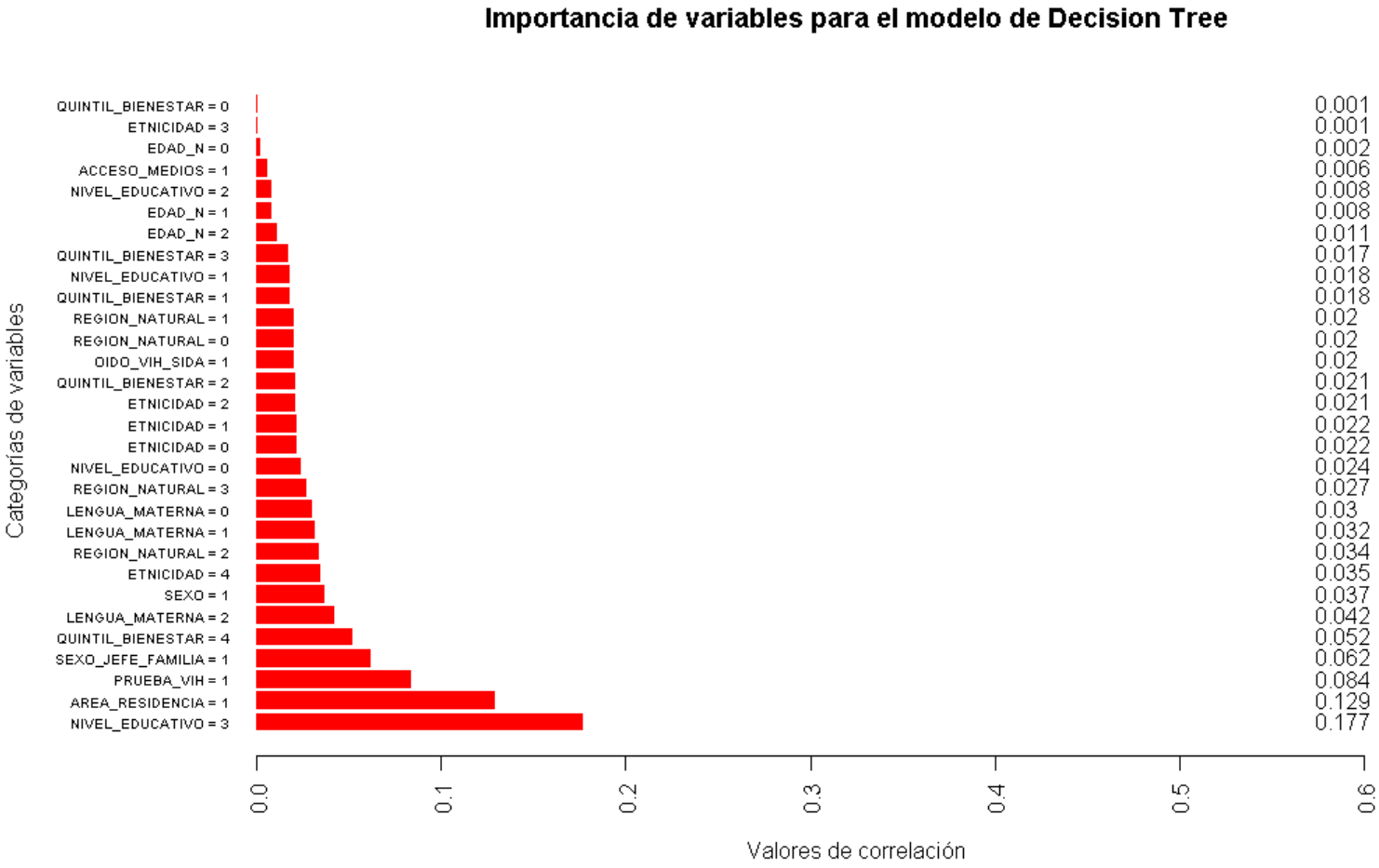


Figura C.2: Medidas de influencia de las variables o regresores independientes en la predicción del nivel de conocimiento sobre el VIH/SIDA en la población peruana en el modelo de Decision Trees. Fuente: Elaboración propia.

Apéndice C.3: Importancia de variables para el modelo de Redes Neuronales Artificiales

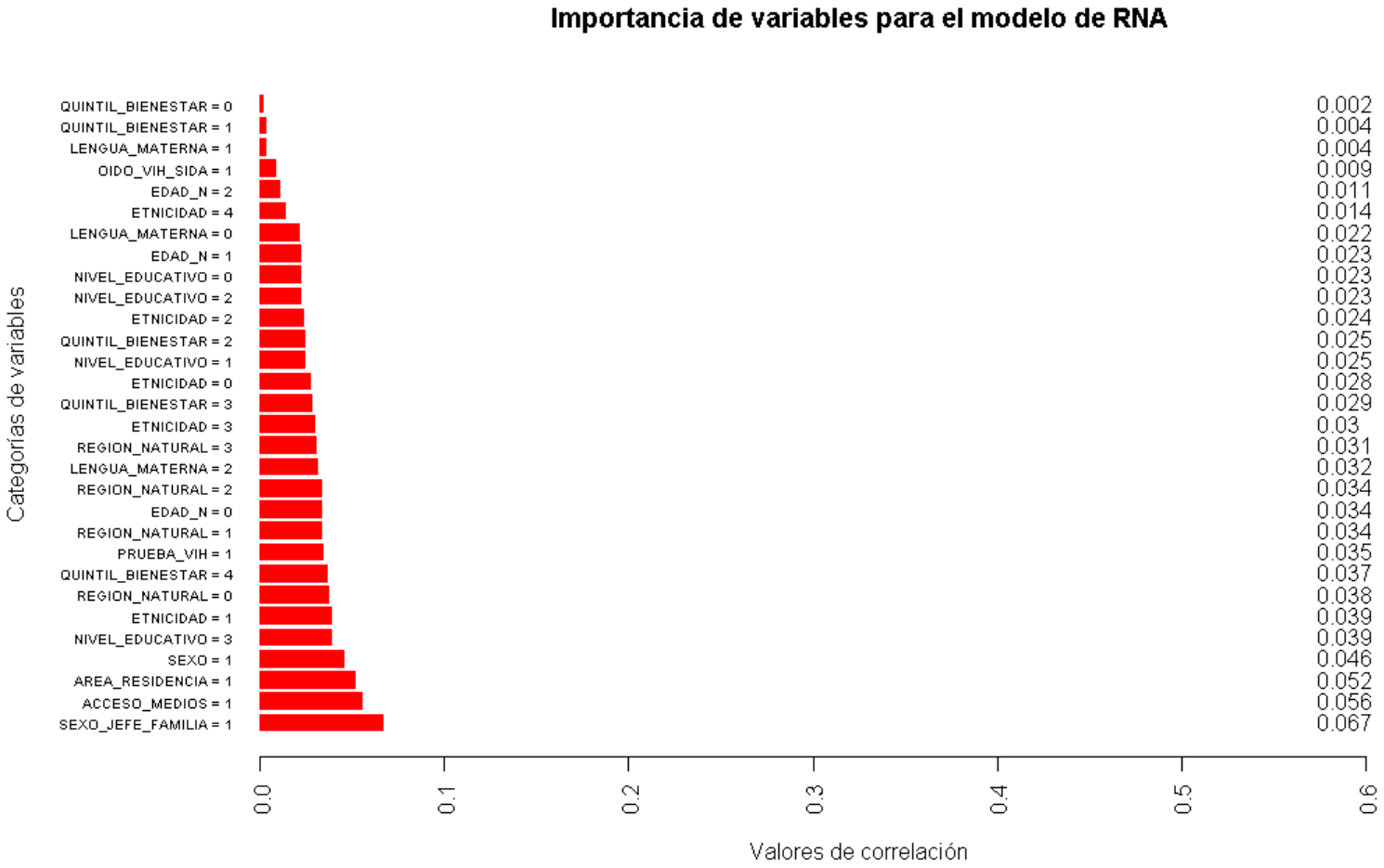


Figura C.3: Medidas de influencia de las variables o regresores independientes en la predicción del nivel de conocimiento sobre el VIH/SIDA en la población peruana en el modelo de Redes Neuronales Artificiales. Fuente: Elaboración propia.

Apéndice C.4: Importancia de variables para el algoritmo k-NN

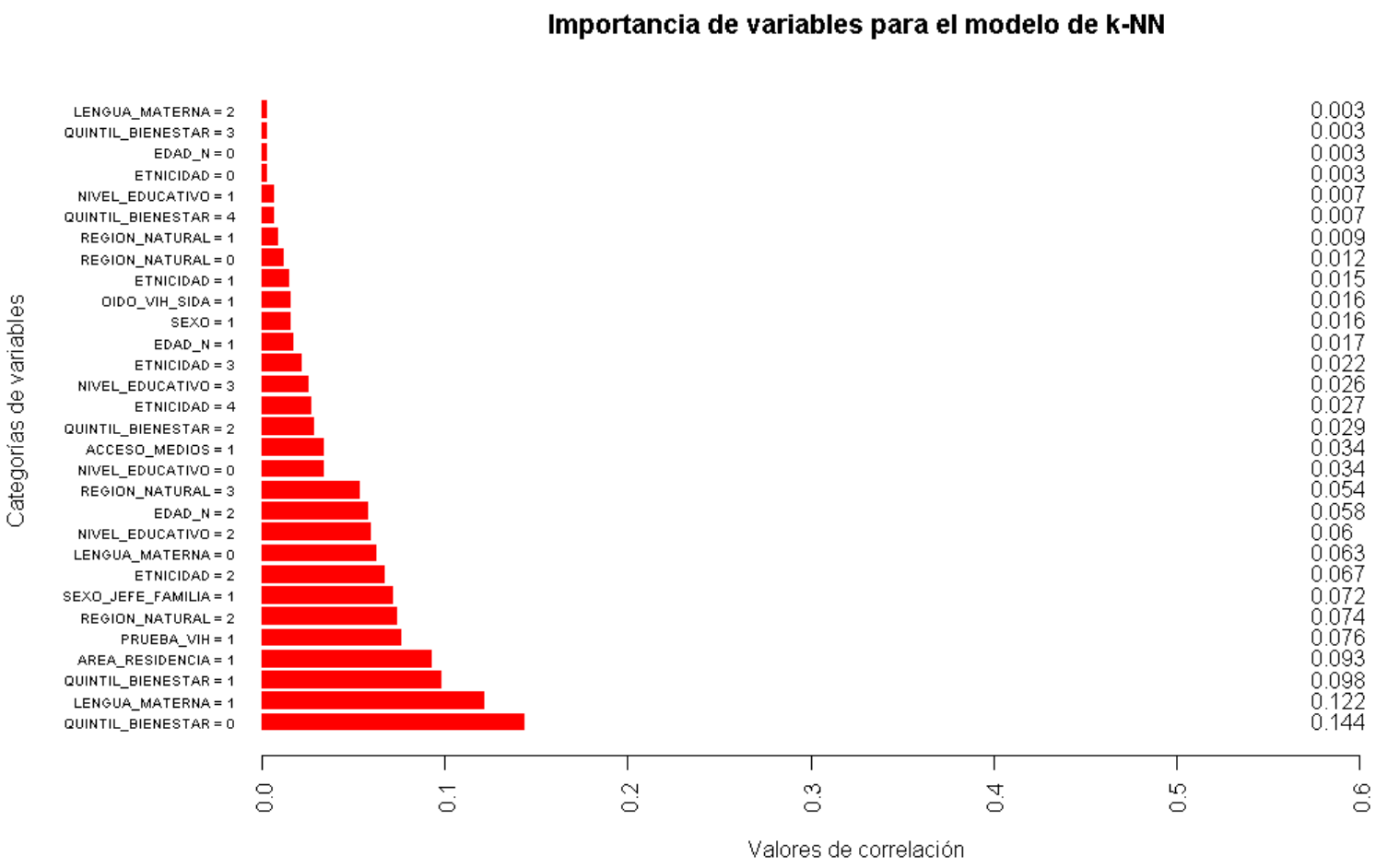


Figura C.4: Medidas de influencia de las variables o regresores independientes en la predicción del nivel de conocimiento sobre el VIH/SIDA en la población peruana en el algoritmo k-NN. Fuente: Elaboración propia.

Bibliografía

1. Adegbosin, A., Stantic, B. & Sun, J. (2020). Efficacy of deep learning methods for predicting under-five mortality in 34 low-income and middle-income countries. *BMJ Open*, **10**(8), e034524.

2. Ahlström, M., Ronit, A., Omland, L., Vedel, S. & Obel, N. (2019). Algorithmic prediction of HIV status using nation-wide electronic registry data. *EClinicalMedicine*, **17**.

3. Akçapınar, G., Altun, A. & Cosgun, E. (2014). *Investigating Students' Interaction Profile in an Online Learning Environment with Clustering* [Conference Session]. 14th IEEE International Conference on Advanced Learning Technologies (ICALT2014), Athens, Greece.
`https://ieeexplore.ieee.org/document/6901411/`.

4. Akinduko, A. & Mirkes, E. (2012). Initialization of Self-Organizing Maps: Principal Components Versus Random Initialization. A Case Study. *arXiv*.

5. Aldás, J. & Uriel, E. (2017). *Análisis multivariante aplicado con R* (2nd ed.). Editorial Paraninfo.

6. Alfaro-Alfaro, N. (2014). Los determinantes sociales de la salud y las funciones esenciales de la salud pública social . *SaludJalisco*, **1**(1), 36–46.

7. Alhasawi, A., Grover, S., Sadek, A., Ashoor, I., Alkhabbaz, I. & Almasri, S. (2019). Assessing HIV/AIDS Knowledge, Awareness, and Attitudes among Senior High School Students in Kuwait. *Med Princ Pract*, **28**, 470–476.

8. Ama, N., Dwivedi, V., Shaibu, S. & Burnette, D. (2015). Socio-Economic and Demographic Determinants of HIV Status among HIV Infected Older Adults (50-64 Years) in Botswana: Evidence From 2013 Botswana AIDS Impact Survey (BAIS IV). *Journal of AIDS & Clinical Research*, **6**, 448.

9. Amusa, L., Bengesai, A. & Khan, H. (2020). Predicting the Vulnerability of Women to Intimate Partner Violence in South Africa: Evidence from Tree-based Machine Learning Techniques. *Journal of Interpersonal Violence*.

10. Anderson, T. & Goodman, L. (1957). Statistical Inference About Markov Chains. *The Annals of Mathematical Statistics*, **28**(1).

11. Anis, M. & Ali, M. (2017). Investigating the Performance of Smote for Class Imbalanced Learning: A Case Study of Credit Scoring Datasets. *European Scientific Journal*, **13**(33).

12. Anthony, A. (2002). *Perfoming logistic regression on survey data with the new SURVEYLOGISTIC procedure*. Proceedings of the 27th Annual SAS Users Group International Conference (SUGI 27), Florida, United States of America. `https://support.sas.com/resources/papers/proceedings/proceedings/sugi27/p258-27.pdf`.

13. APMG Health (2019). *Global summary of findings of an assessment of HIV services packages for key populations in six regions*. `https://www.theglobalfund.org/core_hivservicesforkeypopulationssix regions.pdf`.

14. Arifin, F., Robbani, H., Annisa, T. & Ma'arof, N. (2019). Variations in the Number of Layers and the Number of Neurons in Artificial Neural Networks: Case Study of Pattern Recognition. *Journal of Physics: Conference Series*, **1413**, 012016.

15. Ariza, M., Acosta, K. & Altamar, L. (2016). Aplicación de los Modelos de Respuesta Binaria a los Determinantes de la Demanda de Postgrado en Colombia. *Escenarios*, **14**(1), 7–18.

16. Bao, Y., Medland, N., Fairley, C., Wu, J., Shang, X., Eric, Chow, P., Xu, X., Ge, Z., Zhuang, X. & Zhang, L. (2020). Predicting the diagnosis of HIV and sexually transmitted infections among men who have sex with men using machine learning approaches. *Journal of Infection*, **82**(1), 48–59.

17. Beck, J. & Pauker, S. (1983). The Markov process in medical prognosis. *Med Decis Making*, **3**(4), 419–458.

18. Berrar, D. (2018). Cross-Validation. *Encyclopedia of Bioinformatics and Computational Biology*, **1**, 542–545.

19. Bertsekas, D. & Tsitsiklis, J. (2008). *Introduction to Probability. Athena Scientific optimization and computation series* (2nd ed.). Athena Scientific.

20. Bhat, U. (1984). *Elements of Applied Stochastic Process* (2nd ed.). John Wiley & Sons.

21. Biritwum, R. & Odoom, K. (1995). Application of Markov Process Modelling to Health Status Switching Behaviour of Infants. *International journal of epidemiology*, **24**, 177–182.

22. Bitew, F., Nyarko, S., Potter, L. & Sparks, C. (2020). Machine learning approach for predicting under-five mortality determinants in Ethiopia: evidence from the 2016 Ethiopian Demographic and Health Survey. *Journal of Population Sciences*, **76**(37).

23. Blondeel, K., Dias, S., Furegato, M., Seuc, A., Gama, A., Fuertes, R., Mendão, L., Temmerman, M. & Toskin, I. (2021). Sexual behaviour patterns and STI risk: results of a cluster analysis among men who have sex with men in Portugal. *BMJ Open*, **11**(1), e033290.

24. Bonilla, M., Olmeda, I. & Puertas, R. (2003). Modelos paramétricos y no paramétricos en problemas de credit scoring. *Revista Española de Financiación y Contabilidad*, **32**(118), 833–869.

25. Boza, R. (2016). Orígenes del VIH/SIDA. *Revista Clínica de la Escuela de Medicina UCR-HSJD*, **6**(4), 48–60.

26. Bunyasi, E. & Coetzee, D. (2017). Relationship between socioeconomic status and HIV infection: findings from a survey in the Free State and Western Cape Provinces of South Africa. *BMJ Open*, **7**, e016232.

27. Canadian AIDS Treatment Information Exchange (CATIE) (2018). *The Power of Undetectable: What you need to know about HIV treatment as prevention*. `https://www.catie.ca/sites/default/files/power-undetectable-en.pdf`.

28. Cassy, S., Natário, I. & Martins, R. (2016). Logistic Regression Modelling for Complex Survey Data with an Application for Bed Net Use in Mozambique. *Open Journal of Statistics*, **6**(5), 898–907.

29. Caswell, H. (2019). *Sensitivity Analysis of Discrete Markov Chains. In: Sensitivity Analysis: Matrix Methods in Demography and Ecology. Demographic Research Monographs (A Series of the Max Planck Institute for Demographic Research)* (1st. ed.). Springer.

30. Centro Nacional de Epidemiología, Prevención y Control de Enfermedades (DGE) (2020a). *Vigilancia epidemiológica del VIH/SIDA*. Recuperado de la base de datos en `https://www.dge.gob.pe/epipublic/uploads/vih-sida/vih-sida_202012.pdf`.

31. Centro Nacional de Epidemiología, Prevención y Control de Enfermedades (DGE) (2020b). *Situación epidemiológica del VIH-Sida en el Perú*. `https://www.dge.gob.pe/epipublic/uploads/vih-sida/vih-sida_202012.pdf`.

32. Chawla, N., Bowyer, K., Hall, L. & Kegelmeyer, P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, **16**(1), 321–357.

33. Chikermane, S., Polimeni, J., Burton-Chase, A., Chandrasekara, R. & O'Grady, T. (2016). Effects of Education and other Socioeconomic Variables on HIV Seroprevalence in Russia, India, South Africa and The United States. *Value in Health*, **19**(3), A224.

34. Clark, S. (2018). *Advances in self-organizing maps for spatiotemporal and nonlinear systems* [PhD Thesis, University of New South Wales]. Institutional Repository of the University of New South Wales.

http://unsworks.unsw.edu.au/fapi/datastream/unsworks:52796/SOURCE02?view
=true.

35. Clinical Info - HIV.gob (HIV.gob) (2019). *How Do You Get or Transmit HIV?*. https://www.hiv.gov/hiv-basics/overview/about-hiv-and-aids/how-is-hiv-transmitted.

36. Coates, T., Richter, L. & Caceres, C. (2008). *Behavioural strategies to reduce HIV transmission: how to make them work better*. Organización Mundial de la Salud. https://www.who.int/hiv/events/artprevention/coates.pdf.

37. Cutler, A., Cutler, D. & Stevens, J. (2011). Random Forests. *Machine Learning*, **45**(1), 157–176.

38. de Bodt, E., Cottrell, M. & Verleysen, M. (2002). Statistical tools to assess the reliability of self-organizing maps. *Neural networks : the official journal of the International Neural Network Society*, **15**(8-9), 967–978.

39. de Vasconcellos, R., Emiko, S., Pereira, D., Pavan, F., Rocha, S., Veloso, V., Grinsztejn, B. & Carvalho, M. (2013). Multi-state models for defi ning degrees of chronicity related to HIV-infected patient therapy adherence. *Cad. Saúde Pública*, **29**(4), 801–811.

40. Defensoría del Pueblo (2009). *Informe Defensorial N° 143: Fortaleciendo la respuesta frente a la epidemia del VIH/Sida - Supervisión de los servicios de prevención, atención y tratamiento del VIH/Sida*. https://www.defensoria.gob.pe/wp-content/uploads/2018/05/informe_143.pdf.

41. Defensoría del Pueblo (2011). *Serie Informes de Adjuntía - Informe N° 005-2011-DP/AAE: Fortaleciendo la respuesta frente a la epidemia del VIH/sida: Segunda supervisión de los servicios de prevención, atención y tratamiento del VIH/SIDA*. https://www.defensoria.gob.pe/wp-content/uploads/2020/02/informe-adjuntia-AAE-005-2011.pdf.

42. Delgado-Moya, E. & Marrero-Severo, A. (2017). Modelo Estocástico para la Epidemia del VIH/SIDA. *Revista de Matemática: Teoría y Aplicaciones*, **24**(2), 277–286.

43. Dvorský, J. (2018). *Self-organizing maps and SVD* [Conference Session]. 18th International Workshop on Database and Expert Systems Applications (DEXA 2007), Regensburg, Germany. https://ieeexplore.ieee.org/document/4312874.

44. El Fondo Mundial (2019). *Nota informativa sobre el VIH*. https://www.theglobalfund.org/media/8794/core_hiv_infonote_es.pdf.

45. Fajardo-Ortiz, D., Lopez-Cervantes, M., Duran, L., Dumontier, M., Lara, M., Ochoa, H. & Castano, V. (2017). The emergence and evolution of the research fronts in HIV/AIDS research. *PLoS ONE*, **12**(5), e0178293.

46. Fox, J. & Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, **87**, 178–183.

47. Fredriksson, D. & Glandberger, O. (2020). *Neural Network Regularization for Generalized Heart Arrhythmia Classification* [Master Thesis, Blekinge Institute of Technology]. Institutional Repository of the Blekinge Institute of Technology. `https://www.diva-portal.org/smash/get/diva2:1440369/FULLTEXT01.pdf`.

48. Gala, A., Berdasquera, D., Pérez, J., Pinto, J., Suárez, J., Joanes, J., Sánchez, L., Aragonés, C. & Díaz, M. (2007). Dinámica de adquisición del VIH en su dimensión social, ambiental y cultural. *Revista Cubana de Medicina Tropical*, **59**(2).

49. Garavaglia, S., Dun, A. & Hill, B. (1998). A smart guide to dummy variables: Four applications and a macro.

50. Garcia-Fernandez, L., Novoa, R., Huaman, B. & Benites, C. (2001). Continuo de la atención de personas que viven con VIH y brechas para el logro de las metas 90-90-90 en Perú. *Revista Peruana de Medicina Experimental y Salud Publica*, **35**(3), 491–496.

51. Gomes, R., Ceccato, M., Kerr, L. & aes, M. G. (2017). Fatores associados ao baixo conhecimento sobre HIV/AIDS entre homens que fazem sexo com homens no Brasil. *Cadernos de Saúde Pública*, **33**(10).

52. Greenwell, B., Boehmke, B. & McCarthy, A. (2018). A Simple and Effective Model-Based Variable Importance Measure. *arXiv preprint arXiv:1805.04755*.

53. Haacker, M. (2004). *The Macroeconomics of HIV/AIDS* (1st ed.). International Monetary Fund.

54. Hailu, T. (2015). Comparing Data Mining Techniques in HIV Testing Prediction. *Intelligent Information Management*, **7**(3), 152–179.

55. Haque, A., Hossain, S., Chowdhury, M. & Uddin, J. (2018). Factors associated with knowledge and awareness of HIV/AIDS among married women in Bangladesh: evidence from a nationally representative survey. *Sahara Journal*, **15**(1), 121–127.

56. Haykin, S. (2009). *Neural Networks and Learning Machines* (3rd ed.). Prentice Hall.

57. Hernández-Vásquez, A. & Chacón-Torrico, H. (2019). Manipulación, análisis y visualización de datos de la encuesta demográfica y de salud familiar con el programa R. *Revista Peruana de Medicina Experimental y Salud Publica*, **36**(1), 128–133.

58. Hossin, M. & Sulaiman, M. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, **5**(2), 1–11.

59. Instituto Nacional de Estadística e Informática (INEI) (2008). *Perú: Conocimiento, actitudes y autopercepción de los varones de 15 a 59 años sobre el VIH-SIDA - 2008*. https://proyectos.inei.gob.pe/endes/doc/2008/c.20Varones20VIH-SIDA.pdf.

60. Instituto Nacional de Estadística e Informática (INEI) (2019a). *Perú - Encuesta Demográfica y de Salud Familiar 2019*. Recuperado de la base de datos en http://iinei.inei.gob.pe/microdatos/Consulta_por_Encuesta.asp.

61. Instituto Nacional de Estadística e Informática (INEI) (2019b). *Perú: Enfermedades No Transmisibles y Transmisibles, 2019*. https://proyectos.inei.gob.pe/endes/2019/SALUD/ENFERMEDADES_ENDES_2019.pdf.

62. Jagadesh, S., Combe, M., Couppié, P., Gozlan, R. & Nacher, M. (2020). Mapping priority neighborhoods: A novel approach to cluster identification in HIV/AIDS population. *Research Square*.

63. Janahi, E., Mustafa, S., Alsari, S., Al-Mannai, M. & Farhat, G. (2016). Public knowledge, perceptions, and attitudes towards HIV/AIDS in Bahrain: A cross-sectional study . *The Journal of Infection in Developing Countries*, **10**(9), 1003–1011.

64. Jovanovic, M., Vukicevic, M., Delibašić, B. & Suknovic, M. (2014). *RapidMiner: Data Mining Use Cases and Business Analytics Applications* (1st. ed.). Chapman & Hall.

65. Kaski, S. (1997). Data exploration using self-organizing maps. *Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series*, **82**(57).

66. Koslowsky, M. (1979). Univariate and Multivariate Analysis of Categorical Variables. *Educational and Psychological Measurement*, **39**(4), 747–759.

67. Koutsoukas, A., Monaghan, K., Li, X. & Huan, J. (2017). Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *Journal of Cheminformatics*, **9**(42).

68. Landis, R. & Koch, G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, **33**(1), 159–174.

69. Lang, M. & Belenko, S. (2001). A Cluster Analysis of HIV Risk Among Felony Drug Offenders. *Criminal Justice and Behavior*, **28**(1), 24–61.

70. Lee, S., Ko, J., Tan, X., Patel, I., Balkrishnan, R. & Chang, J. (2014). Markov Chain Modelling Analysis of HIV/AIDS Progression: A Race-based Forecast in the United States. *Indian J Pharm Sci*, **76**(2), 107–115.

71. Lesinski, G., Corns, S. & Dagli, C. (2016). Application of an Artificial Neural Network to Predict Graduation Success at the United States Military Academy. *Procedia Computer Science*, **95**, 375–382.

72. Liu, Y., Sun, J., Yao, Q., Wang, S., Zheng, K. & Liu, Y. (2018). A Scalable Heterogeneous Parallel SOM Based on MPI/CUDA. *Proceedings of Machine Learning Research*, **95**, 264–279.

73. Lobo, J., Jiménez-Valverde, A. & Real, R. (2008). AUC: A misleading measure of the performance of predictive distribution models. *Journal of Global Ecology and Biogeography*, **17**, 145–151.

74. Lopez, V., Fernandez, A., Garcia, S., Palade, V. & Herreraa, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, **250**, 113–141.

75. Lu, W. (2018). *Exploration of Relationships from Texts using SelfOrganizing Maps* [Master Thesis, University of Gävle]. Institutional Repository of the University of Gävle].
`http://www.diva-portal.org/smash/get/diva2:119682/FULLTEXT01.pdf`.

76. Lu, X. & White, H. (2014). Robustness checks and robustness tests in applied economics. *Journal of Econometrics*, **178**(1), 194–206.

77. Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, **9**(1), 1–19. R package verson 2.2.

78. Mar, J., Antoñanzas, F., Pradas, R. & Arrospide, A. (2010). Los modelos de Markov probabilísticos en la evaluación económica de tecnologías sanitarias: una guía práctica. *Gaceta Sanitaria*, **24**(3), 209–214.

79. MATLAB (2010). *version 7.10.0 (R2010a)*. The MathWorks Inc., Natick, Massachusetts.

80. McCulloch, W. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, **5**, 115–133.

81. Merzouki, A., Estill, J., Orel, E., Tal, K. & Keiser, O. (2021). Clusters of sub-Saharan African countries based on sociobehavioural characteristics and associated HIV incidence. *PeerJ*, **9**(1), e10660.

82. Mierswa, I. & Klinkenberg, R. (2018). *RapidMiner Studio (9.1) [Data science, machine learning, predictive analytics]*.
`https://rapidminer.com/`.

83. Miller, S. & Childers, D. (2012). *Probability and Random Processes* (2nd ed.). Academic Press.

84. Ministerio de Defensa - Instituto Geográfica Nacional (MINDEF - IGN) (2020). *Información Geoespacial de límites departamentales del Perú.*. Recuperado de la base de datos en
`http://www.idep.gob.pe/`.

85. Ministerio de Salud (MINSA) (2001a). *Análisis de la Situación Epidemiológica del VIH/SIDA en el Perú - Bases Epidemiológicas para la Prevención y Contro*. http://www.dge.gob.pe/publicaciones/pub_asis/asis19.pdf.

86. Ministerio de Salud (MINSA) (2001b). *Un paso adelante en la lucha contra el SIDA: Los primeros dos años de acceso universal al tratamiento antiretroviral en el Perú*. https://cdn.www.gob.pe/uploads/document/file/419280/un-paso-adelante-en-la-lucha-contra-el-sida-parte-2.pdf.

87. Montzka, T. (2018). *Investigating the Potential of Using SOM on Audit Changed Trades* [Master Thesis, Kth Royal Institute of Technology]. Institutional Repository of the Kth Royal Institute of Technology. https://kth.diva-portal.org/smash/get/diva2:1230333/FULLTEXT01.pdf.

88. Mukandavire, Z., Tchuenche, M., Chiyaka, C. & Musuka, G. (2009). HIV/AIDS and the use of mathematical models in the theoretical assessment of intervention strategies: A review. *Advances in Disease Epidemiology*, **56**(2), 221–241.

89. Najmah, Sari, I., Kumalasari, T., Davies, S. & Andajani, S. (2020). Factors influencing HIV knowledge among women of childbearing age in South Sumatra, Indonesia. *Malaysian Journal of Public Health Medicine*, **20**(1), 150–159.

90. Nucita, A., Bernava, G., Giglio, P., Peroni, M., Bartolo, M., Orlando, S., Marazzi, M. C. & Palombi, L. (2013). *A Markov Chain Based Model to Predict HIV/AIDS Epidemiological Trends* [Conference Session]. International Conference on Model and Data Engineering - MEDI: Model and Data Engineering, Amantea, Italy. https://link.springer.com/chapter/10.10072F978-3-642-41366-7_19.

91. Ocaña-Riola, R. (2009). Modelos de Markov Aplicados a la Investigación en Ciencias de la Salud. *Interciencia*, **34**(3).

92. Ogunmola, O., Oladosu, Y. & Olamoyegun, M. (2014). Relationship between socioeconomic status and HIV infection in a rural tertiary health center. *HIV/AIDS (Auckland, N.Z.)*, **6**, 61–67.

93. Olia, M., Venkataraman, M., Klein, P., Wendland, L. & Brown, M. (2006). Population dynamics of infectious diseases: A discrete time model. *Ecological Modelling*, **198**(1-2), 183–194.

94. O'Neill Institute (2019). *Quick Take: Using Cluster Detection To End The HIV Epidemic*. https://oneill.law.georgetown.edu/oneill-institute-releases-publications-on-hiv-cluster-detection/.

95. Organización Mundial de la Salud (OMS) (2016). *Estrategia Mundial del Sector de la Salud contra el VIH 2016-2021: Hacia el fin del SIDA*. https://www.who.int/hiv/strategy2016-2021/ghss-hiv/es/.

96. Organización Panamericana de la Salud (OPS) (2002). *Módulo de Principios de Epidemiología para el Control de Enfermedades (MOPECE)* (2da ed.). https://www.paho.org/col/dmdocuments/MOPECE1.pdf.

97. Ossa, J. (2013). Matrices de transición y patrones de variabilidad cognitiva. *Universitas Psychologica*, **12**(2), 559–570.

98. Pahn, J., Yang, Y. & Lewis, F. (2020). HIV Knowledge and Attitude and Its Related Factors of Cambodian Adolescents. *Journal of Convergence for Information Technology*, **10**(8), 108–119.

99. Parashar, S., Collins, A., Montaner, J., Hogg, R. & Milloya, M. J. (2016). Reducing Rates Of Preventable HIV/AIDS-Associated Mortality Among People Living With HIV Who Inject Drugs. *Current Opinion in HIV and AIDS*, **11**(5), 507–513.

100. Pargent, F. (2019). *A Benchmark Experiment on How to Encode Categorical Features in Predictive Modeling* [Master Thesis, University of Munich]. Institutional Repository of the University of Munich. https://osf.io/356ed/download.

101. Prieto, L. (2016). *Optimización de las inversiones para la respuesta al HIV en Perú*. Ministerio de Salud. http://www.dge.gob.pe/publicaciones/pub_asis/asis19.pdf.

102. Programa Conjunto de las Naciones Unidades sobre el VIH/SIDA (ONUSIDA) (2017). *Prevención de la infección por el VIH bajo LA LUPA: Un análisis desde la perspectiva del sector de la salud en América Latina y el Caribe*. http://bvs.minsa.gob.pe/local/MINSA/4256.pdf.

103. Programa Conjunto de las Naciones Unidas sobre el VIH/SIDA (ONUSIDA) (2019). *UNAIDS Data 2019*. https://www.unaids.org/sites/default/files/media_asset/2019-UNAIDS-data_en.pdf.

104. Programa Conjunto de las Naciones Unidas sobre el VIH/SIDA (ONUSIDA) (2020). *Archivos sobre las estimaciones nacionales del VIH - Perú*. Recuperado de la base de datos en https://www.unaids.org/es/dataanalysis/datatools/spectrum-epp.

105. Province, B. (2015). *The Effects of Parameter Tuning on Machine Learning Performance in a Software Defect Prediction Context* [Master Thesis, West Virginia University]. Institutional Repository of the West Virginia University. https://core.ac.uk/download/pdf/230455911.pdf.

106. Qiana, J., Nguyena, N., Oya, Y., Kikugawa, G., Okabe, T., Huang, Y. & Ohuchi, F. (2019). Introducing self-organized maps (SOM) as a visualization tool for materials research and education. *Results in Materials*, **4**, 100020.

107. R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

108. Racz, A., Bajusz, D. & Heberger, K. (2021). Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification. *Molecules*, **26**(1111).

109. Raghavan, R. (2020). *Study of the Relationship of Training Set Size to Error Rate in Yet Another Decision Tree and Random Forest Algorithms* [Master Thesis, Texas Tech University]. Institutional Repository of the Texas Tech University. `https://ttu-ir.tdl.org/bitstream/handle/2346/13496/Ratheesh_Raghavan.pdf?sequence=1`.

110. Rotich, T. (2016). Mathematical Modeling of the Spread of HIV/AIDS by Markov Chain Process. *American Journal of Applied Mathematics*, **4**(5), 235–246.

111. Rubio-Terrés, C. & Echeverría, A. (2006). Modelos de Markov: una herramienta útil para el análisis fármaco económico. *Pharmacoeconomics*, **3**(S2), 71–78.

112. Salgado, K. (2015). *Un modelo de Markov probabilístico aplicado en la evaluación económica de datos de rehabilitación cardiaca* [Tesis de Maestría, Universidad Nacional de Colombia]. Repositorio Institucional de la Universidad Nacional de Colombia. `https://repositorio.unal.edu.co/handle/unal/55672`.

113. Sandilands, D. (2013). *Encyclopedia of Quality of Life and Well-Being Research* (1st ed.). A.C. Michalos.

114. Schratz, P., Muenchow, J., Iturritxa, E., Richter, J. & Brenning, A. (2018). Performance evaluation and hyperparameter tuning of statistical and machine-learning models using spatial data. *Ecological Modelling*, **406**, 109–120.

115. Scott, E. & Simon, T. (2011). Poverty, Employment and HIV/AIDS in Trinidad and Tobago. *International Journal of Business and Social Science*, **2**(11).

116. Seger, C. (2018). *An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing* [Bachelor Thesis, Kth Royal Institute of Technology]. Institutional Repository of the Kth Royal Institute of Technology. `https://www.diva-portal.org/smash/get/diva2:1259073/FULLTEXT01.pdf`.

117. Sha, W. (2006). Comment on Design of a Propane Ammoxidation Catalyst Using Artificial Neural Networks and Genetic Algorithms. *Industrial & Engineering Chemistry Research*, **45**(24).

118. Sims, G. (2009). *HIV & AIDS*. Society for General Microbiology. `https://microbiologyonline.org/file/b92698b47294588bc5965c3a7f080389.pdf`.

119. Sonnenberg, F. & Beck, J. (1993). Markov models in medical decision making: a practical guide. *Med Decis Making*, **13**(4), 322–338.

120. Soto, L. E. (2004). Mecanismos patogénicos de la infección por VIH. *Revista de Investigación Clínica*, **56**(2), 143–152.

121. Spedicato, G. A. (2017). Discrete Time Markov Chains with R. *The R Journal*, **9**(2), 84–104.

122. Talukder, A. & Ahammed, B. (2020). Machine learning algorithms for predicting malnutrition among under-five children in Bangladesh. *Nutrition*, **78**, 110861.

123. Tang, D., Zhang, M., Xu, J., Xueliang, Z., Yang, F., Li, H., Feng, L., Wang, K. & Zheng, Y. (2018). Application of Data Mining Technology on Surveillance Report Data of HIV/AIDS High-Risk Group in Urumqi from 2009 to 2015. *Complexity*, **2018**(2), 1–17.

124. Teva, I., Bermudez, M., Ramiro, M. & Buela-Casala, G. (2012). Situación epidemiológica actual del VIH/SIDA en Latinoamérica en la primera década del siglo XXI. Análisis de las diferencias entre países. *Revista médica de Chile*, **140**(1), 50–58.

125. Tian, J., Azarian, M. & Pecht, M. (2014). Anomaly Detection Using Self-Organizing Maps-Based K-Nearest Neighbor Algorithm. *PHM Society European Conference*, **2**(1).

126. Tovar-Cuevas, L. & Arrivillaga-Quintero, M. (2011). VIH/SIDA y determinates sociales estructurales en municipios del Valle del Cauca-Colombia. *Gerencia y Políticas de Salud*, **10**(21).

127. Uçar, M., Nour, M., Sindi, H. & Polat, K. (2020). The Effect of Training and Testing Process on Machine Learning in Biomedical Datasets. *Mathematical Problems in Engineering*, **2020**.

128. UNAIDS (2015). *Country reported data from Global AIDS Monitoring, UNAIDS estimates, WHO Health expenditure database.*. Recuperado de la base de datos en `https://hivfinancial.unaids.org/hivfinancialdashboards.html`.

129. Valova, I., Georgiev, G., Gueorguieva, N. & Olson, J. (2013). Initialization Issues in Self-organizing Maps. *Procedia Computer Science*, **20**, 52–57.

130. Van Rossum, G. & Drake, F. (2009). *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

131. Vesanto, J. & Alhoniemi, E. (2000). Clustering of the Self-Organizing Map. *IEEE Transactions On Neural Networks*, **11**(3).

132. Vesanto, J., Himberg, J., Alhoniemi, E. & Parhankangas, J. (2000). SOM Toolbox for Matlab 5. *SOM Toolbox Team - Helsinki University of Technology*.

133. Wang, L. (1977). On competitive learning. *IEEE Transactions on Neural Networks*, **8**(5), 1214–1217.

134. Welsing, P., Severens, J., van Riel, M. H. P. & Laan, R. (2004). Modeling the 5-year cost effectiveness of treatment strategies including tumor necrosis factor-blocking agents and leflunomide for treating rheumatoid arthritis in the Netherlands. *Arthritis Rheum*, **51**(6), 964–973.

135. Woldemariame, S. (2013). *Factors Determining the Prevalence of HIV/AIDS in Ethiopia* [Master Thesis, University of Stockholm]. Institutional Repository of the University of Stockholm.
`https://www2.math.su.se/matstat/reports/master/2013/rep3/report.pdf`.

136. World Bank (2020). *World Development Indicators: Total of Population - Peru*. Recuperado de la base de datos en
`https://data.worldbank.org/indicator/SP.POP.TOTL?locations=PE`.

137. Xia, X. (2017). Self-Organizing Map for Characterizing Heterogeneous Nucleotide and Amino Acid Sequence Motifs. *Computation*, **5**(4), 43.

138. Xu, Y. & Goodacre, R. (2018a). On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *Journal of Analysis and Testing*, **2**(3).

139. Xu, Y. & Goodacre, R. (2018b). On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *Journal of Analysis and Testing*, **2**(3).

140. Yaesoubi, R. & Cohen, T. (2011). Generalized Markov Models of Infectious Disease Spread: A Novel Framework for Developing Dynamic Health Policies. *European Journal Of Operational Research*, **215**(3), 679–687.

141. Yahaya, H. & Kola, R. (2017). Cluster Analysis of the Incidence of HIV in Nigeria. *International Journal of Mathematics and Statistics Studies*, **5**(1), 29–44.

142. Zapata-Tapasco, A., Pérez-Londoño, S. & Mora-Flórez, J. (2014). Método basado en clasificadores k-NN parametrizados con algoritmos genéticos y la estimación de la reactancia para localización de fallas en sistemas de distribución. *Revista Facultad de Ingeniería Universidad de Antioquia*, **70**, 220–232.

143. Zhang, C., Liua, C., Zhang, X. & Almpanidis, G. (2017). An up-to-date comparison of state-of-the-art classification algorithms. *Expert Systems With Applications*, **82**, 128–150.