UNIVERSIDAD ESAN

FACULTAD DE INGENIERÍA

INGENIERÍA DE SISTEMAS

Modelo de clasificación del estado de vacunación infantil en el primer año de vida en la jurisdicción de una Dirección de Redes Integradas de Salud de Lima Metropolitana utilizando técnicas de Machine Learning

Trabajo de Suficiencia Profesional presentado en satisfacción parcial de los requerimientos para obtener el título profesional de Ingeniero(a) de Sistemas

AUTORES

Alocen Tito, Claudia Vanessa

Diaz Espiritu, Jorge Manuel

Pelaez Flores, Antonio Ruis

ASESOR

Mamani Ticona, Wilfredo

ORCID N° 0000-0003-1489-9056

Noviembre, 2025

RESULTADOS DEL INFORME DE SIMILITUD

ORIGINALITY REPORT

12%	12%	6%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	repositorio.unac.edu.pe Internet Source	5%
2	100freezoosites.site Internet Source	3%
3	www.mesadeconcertacion.org.pe Internet Source	2%
4	cdn.www.gob.pe Internet Source	1%
5	www.gob.pe Internet Source	1%
6	repositorio.esan.edu.pe Internet Source	1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		

RESUMEN

La cobertura de vacunación infantil en el Perú presenta importantes retos para el sistema de salud pública, especialmente en Lima Metropolitana, donde se presentan brechas significativas en el cumplimiento del esquema de inmunización en el primer año de vida. El presente trabajo de investigación tuvo como objetivo desarrollar un modelo de clasificación del estado de vacunación infantil en el primer año de vida en la jurisdicción de una Dirección de Redes Integradas de Salud de Lima Metropolitana utilizando técnicas de *machine learning*. Para ello, se emplearon datos provenientes de los sistemas HISMINSA y del Padrón Nominal, aplicando procesos de preprocesamiento, selección de características y entrenamiento de modelos supervisados tales como *Logistic Regression*, *K-Nearest Neighbors* (KNN), *Decision Tree*, *Random Forest*, *AdaBoost*, *XGBoost* y *Bagging*, incorporando el ajuste de hiperparámetros y validación cruzada para evaluar la estabilidad del modelo. Como resultado, el modelo XGBoost combinado con el conjunto de características obtenidas por LightGBM, alcanzó el mejor desempeño con un *accuracy* de 74%, *precision* de 75%, *recall* de 97%, *F1-score* de 85% y AUC de 66%. Este estudio aporta una base sólida para la integración de herramientas de analítica avanzada en la gestión de programas de inmunización y en la toma de decisiones.

Palabras clave: vacunación infantil, primer año de vida, selección de características, aprendizaje automático, clasificación, *LightGBM*, *XGBoost*.

ABSTRACT

Childhood vaccination coverage in Peru presents significant challenges for the public health system, particularly in Metropolitan Lima, where considerable gaps persist in meeting the immunization schedule during the first year of life. This research aimed to develop a classification model to determine the vaccination status of infants in their first year of life within the jurisdiction of Dirección de Redes Integradas de Salud in Metropolitan Lima, using machine learning techniques. Data were obtained from the HISMINSA and *Padrón Nominal* systems, applying preprocessing, feature selection, and supervised model training processes. The models evaluated included Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Random Forest, AdaBoost, XGBoost, and Bagging, incorporating hyperparameter tuning and cross-validation to evaluate the stability and generalization capability of the model. Among the results, the XGBoost model combined with the feature set obtained through LightGBM achieved the best performance, with an accuracy of 74%, precision of 75%, recall of 97%, F1-score of 85%, and AUC of 66%. This study provides a solid foundation for integrating advanced analytics tools into immunization program management and evidence-based decision-making.

Keywords**:** childhood vaccination, first year of life, feature selection, machine learning, classification, LightGBM, XGBoost.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

La inmunización infantil constituye una de las estrategias más eficaces y costo-efectivas para proteger la salud pública, al prevenir la propagación de enfermedades transmisibles y reducir significativamente la morbilidad y mortalidad infantil. A pesar de los avances logrados a nivel global y nacional en la cobertura de vacunación, aún persisten brechas importantes que afectan la equidad en salud, especialmente en determinados territorios y grupos poblacionales. En el contexto peruano, si bien se ha evidenciado una mejora sostenida en los indicadores de vacunación durante los últimos años, persisten desigualdades entre regiones, con coberturas heterogéneas que reflejan limitaciones en el acceso, seguimiento y continuidad de la atención.

En la jurisdicción de la Dirección de Redes Integradas de Salud (DIRIS) Lima Centro, se observa una baja cobertura acumulada en los controles de Crecimiento y Desarrollo (CRED) en menores de doce meses, alcanzando solo un 16.9%, muy por debajo de la meta establecida del 41%. Esta brecha representa más de 18,000 niños que no han completado oportunamente su esquema de vacunación, lo que pone en riesgo su salud y la de la comunidad. Esta situación evidencia la necesidad de fortalecer los mecanismos de identificación temprana de niños con esquemas incompletos, así como de optimizar las estrategias de intervención y seguimiento para mejorar las coberturas de vacunación.

En este escenario, el uso de herramientas de análisis de datos e inteligencia artificial ofrece nuevas oportunidades para apoyar la toma de decisiones en salud pública. En particular, las técnicas de aprendizaje automático, o *machine learning*, permiten identificar patrones complejos y predecir comportamientos futuros a partir de grandes volúmenes de información, facilitando la clasificación automatizada y precisa de poblaciones en riesgo. Su aplicación en el ámbito de la vacunación infantil permite desarrollar modelos predictivos que apoyan la gestión eficiente de programas de inmunización, priorizando intervenciones en grupos vulnerables.

El presente trabajo tiene como objetivo desarrollar un modelo de clasificación basado en técnicas de *machine learning* para predecir el estado de vacunación infantil en el primer año de vida en la jurisdicción de una Dirección de Redes Integradas de Salud de Lima Metropolitana, utilizando datos provenientes de sistemas institucionales como HIS MINSA y el Padrón Nominal. Para ello, se aplicaron procedimientos rigurosos de preprocesamiento, integración, selección de variables y balanceo de clases, con el fin de construir un conjunto de datos de alta calidad para el entrenamiento de diversos modelos de clasificación supervisada.

El documento se encuentra estructurado en seis capítulos. En el Capítulo 1, denominado "Planteamiento del Problema", se describe la realidad problemática, la justificación y la delimitación del trabajo. A continuación, en el Capítulo 2, denominado "Marco Teórico", se realiza una revisión detallada de conceptos fundamentales de inteligencia artificial y antecedentes relevantes sobre el uso de *machine learning* en la vacunación infantil. Posteriormente, en el Capítulo 3, denominado "Entorno Empresarial", se describe el entorno institucional de la DIRIS Lima Centro. En el Capítulo 4, denominado "Metodología de la Investigación", se presenta la metodología a utilizar, así como la población, la muestra y el diseño de la investigación. Luego, en el Capítulo 5, denominado "Desarrollo de la Solución", se detalla el desarrollo del modelo propuesto, incluyendo las etapas de modelado, evaluación y validación. Finalmente, en el Capítulo 6, denominado "Conclusiones y Recomendaciones", se presentan las conclusiones obtenidas del trabajo realizado y algunas recomendaciones derivadas del estudio para investigaciones futuras.

Capítulo 1: Planteamiento del Problema

1.1. Descripción de la Realidad Problemática

La inmunización infantil constituye un método sencillo, confiable y altamente efectivo para resguardar a la población frente a diversas enfermedades peligrosas, ya que prepara al sistema inmunológico para desarrollar defensas específicas antes de que ocurra la exposición a dichos agentes. (OMS, 2024)

De acuerdo con la Organización Panamericana de la Salud (OPS), la inmunización se refiere al mecanismo mediante el cual un individuo desarrolla inmunidad frente a una enfermedad, generalmente a través de la aplicación de una vacuna, resaltando su importancia tanto para la protección personal como para la salud comunitaria. (PAHO, s.f.).

Por su parte, el Ministerio de Salud del Perú (MINSA) define el esquema de vacunación como la programación organizada y secuencial de la aplicación mínima de vacunas que se administran de manera regular, con el propósito de brindar protección integral a la población infantil. (MINSA, 2005)

A nivel internacional, la cobertura de vacunación en menores de un año ha mostrado progresos significativos en las últimas décadas; sin embargo, aún persisten brechas importantes que afectan la equidad en salud. Según datos conjuntos de la Organización Mundial de Salud (OMS) y el Fondo Internacional de Emergencia de las Naciones Unidas para la Infancia, o UNICEF por sus siglas en inglés, en el año 2023 la cobertura mundial de la tercera dosis de la vacuna contra difteria, tétanos y tos ferina (DTP3), utilizada como indicador clave de desempeño de los programas de inmunización, alcanzó el 85%. A pesar de este avance, alrededor de 14.3 millones de niños en el mundo no recibieron ninguna dosis, concentrándose principalmente en países de ingresos bajos y medios (OMS & UNICEF, 2024).

A nivel nacional, los indicadores de cobertura de vacunación infantil evidencian avances importantes en los últimos años, aunque persisten desigualdades entre departamentos. De acuerdo con la Encuesta Demográfica y de Salud Familiar (ENDES 2023), publicada en marzo de 2024 por el Instituto Nacional de Estadística e Informática (INEI) en coordinación con el MINSA, el 77.5% de los menores de 12 meses recibieron vacunas según su edad, lo que representa un incremento de 7.9 puntos porcentuales respecto al año 2022. Esta tendencia positiva también se observa en los grupos de 15 y 36 meses, con coberturas de 75.2% y 62.1%, respectivamente. Sin embargo, al analizar los datos por departamento, se aprecian brechas

considerables: Lambayeque, Loreto y Pasco mostraron incrementos superiores a 16 puntos porcentuales en comparación con 2022, mientras que Cajamarca y Moquegua registraron retrocesos de 1.2 y 3.6 puntos porcentuales, respectivamente. Lima Metropolitana presentó un aumento moderado de 4.3 puntos porcentuales, ubicándose por debajo de departamentos con mayores mejoras porcentuales. Estas variaciones territoriales reflejan desigualdades en el acceso y la oportunidad de la vacunación, lo que constituye un desafío para el logro de coberturas homogéneas a nivel nacional.

Tabla 1

Vacunación en menores de 12 meses por departamento

Departamento	2018	2019	2020	2021	2022	2023	Diferencia 2023-2022
Total	73.3	76.7	61.1	69.1	69.6	77.5	7.9
Lambayeque	75.0	74.6	58.8	72.0	57.1	74.6	17.5
Loreto	67.1	65.3	43.4	42.4	49.1	66.6	17.5
Pasco	76.4	79.8	63.3	69.4	61.8	78.2	16.4
Departamento Lima	75.6	83.9	57.4	68.5	77.6	91.4	13.8
Callao	67.2	65.2	64.3	73.9	75.1	88.5	13.4
Tumbes	84.0	85.7	74.0	80.4	76.0	89.4	13.4
Puno	58.4	61.5	57.1	52.6	45.6	58.0	12.4
Piura	76.7	77.3	67.8	77.7	68.8	81.1	12.3
Madre de Dios	72.9	75.3	58.4	69.3	60.5	72.8	12.3
Arequipa	73.7	79.3	64.3	72.6	65.4	76.8	11.4
Apurímac	87.3	82.0	68.4	68.9	80.9	89.8	8.9
Ica	78.7	77.5	59.4	74.1	73.4	82.1	8.7
Ancash	89.8	87.8	62.6	79.5	79.5	87.9	8.4
Huánuco	84.6	85.5	66.7	77.5	77.1	85.0	7.9
Ayacucho	75.2	81.4	67.2	71.5	75.1	81.9	6.8
San Martín	82.3	84.0	67.5	67.7	73.1	79.7	6.6
Junín	73.6	83.9	69.2	72.6	75.3	81.9	6.6

Amazonas	69.0	80.1	67.6	74.0	60.2	66.7	6.5
Huancavelica	76.9	81.0	67.6	60.2	75.8	81.7	5.9
Cusco	83.1	86.4	60.2	75.5	75.4	80.4	5.0
Lima Metropolitana	66.5	71.7	57.7	67.8	69.9	74.2	4.3
Ucayali	72.2	74.1	55.6	61.1	68.8	72.9	4.1
Tacna	82.9	77.2	65.0	79.8	77.2	79.8	2.6
La Libertad	75.6	81.0	64.9	66.0	81.0	81.9	0.9
Cajamarca	74.8	77.0	67.7	74.9	71.2	70.0	-1.2
Moquegua	81.2	84.1	61.3	67.6	75.1	71.5	-3.6

Nota. Adaptado de *Balance de los Avances en la Vacunación Nacional 2023 y Desafíos para el Año 2024*, por Mesa de Concertación para la Lucha contra la Pobreza y Ministerio de Salud del Perú, 2024.

A nivel de DIRIS Lima Centro enfrenta una baja cobertura en los controles CRED en menores de 12 meses, con un avance acumulado de apenas 16.9%, muy por debajo de la meta del 41%, lo que representa una brecha de atención de más de 18,000 niños. Esta situación se manifiesta en una tendencia descendente durante el año, con desigualdades entre redes y establecimientos de salud, reflejando limitaciones en la captación, seguimiento y continuidad de la atención infantil. Esta brecha compromete la detección oportuna de problemas de crecimiento y desarrollo, afectando directamente el bienestar y la salud integral de la población infantil en Lima Centro. (DLC, 2024).

El presente estudio tiene como propósito desarrollar un modelo basado en técnicas de *machine learning* para la clasificación del estado de vacunación infantil en el primer año de vida en la jurisdicción de la Dirección de Redes Integradas de Salud (DIRIS) Lima Centro. Para el desarrollo de este trabajo, se han obtenido datos provenientes del sistema HIS MINSA y del Padrón Nominal de la DIRIS Lima Centro, aplicando técnicas de preprocesamiento, integración, selección de variables y balanceo de clases, con el fin de construir un conjunto de datos representativo y de calidad para el entrenamiento de modelos de clasificación.

1.2. Objetivos de la Investigación

Para la identificación del objetivo general y de los objetivos específicos, se ha utilizado la matriz de consistencia (Anexo N° 02). Los objetivos han sido definidos en función de los problemas previamente identificados, con el fin de establecer un marco de trabajo que permita abordar y resolver dichos problemas.

1.2.1. Objetivo general

Desarrollar un modelo de clasificación del estado de vacunación infantil en el primer año de vida en la jurisdicción de una Dirección de Redes Integradas de Salud de Lima Metropolitana Utilizando técnicas de Machine Learning.

1.2.2. Objetivos específicos

OE1: Obtener un *dataset* de vacunación infantil en el primer año de vida sobre estado de vacunación.

OE2: Analizar el impacto del preprocesamiento en el entrenamiento de un modelo para clasificar el estado de vacunación.

OE3: Aplicar técnicas de selección de características para mejorar la clasificación del estado de vacunación infantil en el primer año de vida.

OE4: Utilizar técnicas de *machine learning* para clasificación del estado de vacunación infantil en el primer año de vida.

OE5: Utilizar métricas para evaluar el desempeño de los modelos de *machine learning* utilizados para la clasificación del estado de vacunación infantil en el primer año de vida.

1.3. Justificación de la Investigación

1.3.1. Teórica

La investigación se justifica en el plano teórico porque busca aportar al conocimiento existente sobre la aplicación de técnicas de machine learning en el ámbito de la salud pública, específicamente en la vigilancia y monitoreo de la vacunación infantil. El uso de modelos de clasificación permitirá comprender mejor los patrones y factores asociados al cumplimiento del esquema de vacunación en el primer año de vida. De esta manera, se fortalece la literatura académica y científica sobre la intersección entre inteligencia artificial y salud preventiva, ofreciendo una base conceptual para futuros estudios en el campo.

1.3.2. Práctica

La investigación proporcionará un modelo de *machine learning* capaz de clasificar el estado de vacunación infantil en el primer año de vida dentro en la jurisdicción de una Dirección de Redes Integradas de Salud de Lima Metropolitana. Este aporte facilitará a los profesionales de la salud identificar de manera temprana a los niños con esquemas incompletos de vacunación y, en consecuencia, implementar estrategias de intervención más oportunas y eficaces. Asimismo, la herramienta contribuirá a optimizar la gestión de los programas de inmunización, favoreciendo la toma de decisiones basada en evidencia.

1.3.3. Metodológica

La investigación se centra en el desarrollo de modelos mediante la aplicación de técnicas de *machine learning* a los datos recolectados entre el 26 de setiembre de 2019 y el 23 de setiembre de 2025, con el objetivo de clasificar el estado de vacunación infantil.

La construcción de un modelo de machine learning constituye el núcleo central del estudio y se lleva a cabo siguiendo la ruta metodológica descrita en los artículos revisados, por lo cual se garantiza coherencia con los antecedentes consultados. En este marco, se ejecutan distintas actividades que abarcan desde la obtención del dataset hasta la evaluación de los modelos.

1.4. Delimitación del Estudio

1.4.1. Espacial

El presente trabajo de investigación se llevará a cabo utilizando información sobre vacunación infantil de niños pertenecientes a la jurisdicción en la Dirección de Redes Integradas de Salud de Lima Metropolitana. Para ello, se gestionará con la organización la obtención del dataset para el análisis. Asimismo, se hará uso de papers que contengan resultados de investigaciones previas relacionadas con la clasificación de estado de vacunación infantil para poder hacer una comparación de la efectividad y desempeño con el modelo a desarrollar.

1.4.2. Temporal

La presente investigación analizará un conjunto de datos recolectados desde el 26 de setiembre de 2019 hasta el 23 de setiembre de 2025.

1.4.3. Conceptual

Este estudio se enfoca en los principios del Machine Learning, empleando modelos de clasificación y análisis de datos para identificar patrones asociados al estado de vacunación infantil en el primer año de vida. Asimismo, se incluyen determinantes sociales y condiciones de atención sanitaria con el propósito de clasificar oportunamente el cumplimiento del esquema de vacunación en la jurisdicción de una Dirección de Redes Integradas de Salud de Lima Metropolitana.

Capítulo 2: Marco Teórico

2.1. Antecedentes de la Investigación

A continuación, se presenta algunos antecedentes relevantes para poder entender cómo es que se han desarrollado en otros casos vinculados al tema de investigación elegido, los objetivos, soluciones y los resultados a los cuales se ha llegado.

Asnake, A.A., Seifu, B.L., Gebrehana, A.K. (2025). Prediction of zero-dose children using supervised machine learning algorithm in Tanzania: evidence from the recent 2022 Tanzania Demographic and Health Survey. BMJ Open 2025, 15: e097395. https://doi.org/10.1136/bmjopen-2024-097395

- **Resumen:**

El estudio aborda la problemática de la baja cobertura de inmunización infantil en Tanzania, específicamente concentrándose en identificar a los niños que no han recibido ninguna dosis de vacunas rutinarias. Utilizando un enfoque basado en Machine Learning, se evaluaron 7 algoritmos de clasificación, destacando el clasificador de bosque aleatorio (RF) como el más efectivo por su alta precisión y capacidad predictiva. La investigación empleó datos de la Encuesta Demográfica y de Salud de Tanzania 2022, que abarcó una muestra representativa de 2,120 niños de entre 12 y 23 meses, así como información sociodemográfica, de salud y comportamiento de los hogares. El objetivo de la investigación fue detectar los principales factores asociados con la no vacunación, como la falta de educación materna, desempleo de los padres, tamaño familiar y acceso a servicios de salud, para dirigir intervenciones específicas y mejorar la cobertura de vacunas en ámbitos rurales y urbanos.

El análisis fue exhaustivo, implementando técnicas de preprocesamiento de datos, validación cruzada de 10 pliegues y ajuste de hiperparámetros para optimizar los modelos. Se utilizó la interpretación mediante valores SHAP para comprender la importancia relativa de cada variable en la predicción de niños sin dosis, revelando que el desempleo materno, la ausencia de educación formal y el tamaño reducido de la familia fueron los principales predictores. Los resultados mostraron que aproximadamente el 7.45% de los niños en la muestra no habían recibido ninguna vacuna, una prevalencia que evidencia una cobertura inmunitaria aún por debajo de los objetivos nacionales e internacionales. La alta precisión del modelo RF (con métricas como precisión de 0.94 y AUC de 0.99) demuestra la utilidad de aplicar

algoritmos de machine learning para identificar poblaciones vulnerables, optimizar recursos y diseñar políticas públicas más efectivas para incrementar la vacunación infantil en Tanzania y otros países en desarrollo.

- **Base de datos:**

La investigación utilizó datos de la Encuesta Demográfica y de Salud (TDHS) de Tanzania 2022, que incluye información de 2,120 niños y sus hogares, representando todo el país. Los datos fueron obtenidos de manera pública a través del sitio web del DHS, con permisos adecuados para su uso. Las variables utilizados parte esta investigación fueron los siguiente: Edad materna, Nivel económico del hogar, Educación materna, Empleo materno, Tamaño de la familia, Educación y ocupación del esposo, Edad al primer parto, Participación en atención prenatal, Lugar de parto, Edad y región del niño, Lugar de residencia (urbano/rural), Acceso a servicios de salud. La variable de objetivo es: Niños sin dosis (Recibieron ninguna de las vacunas principales).

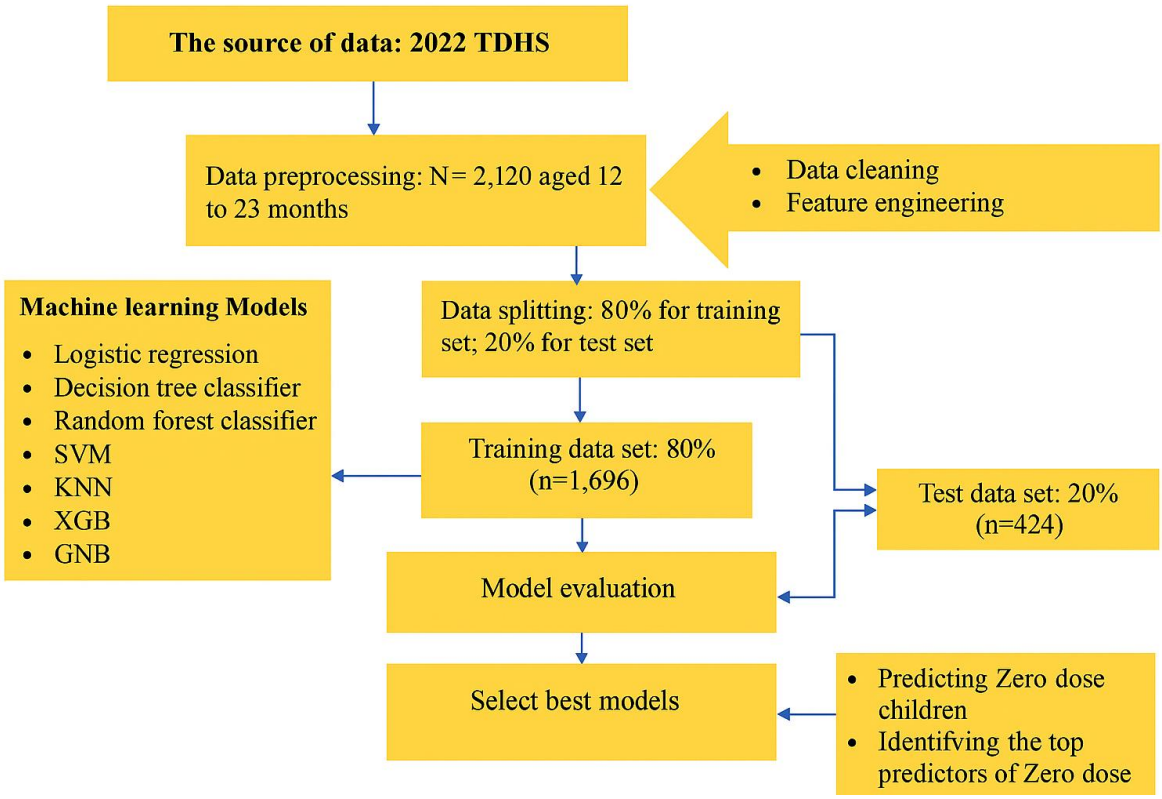
- **Metodología:**

La metodología utilizada de manera exhaustivo con técnicas de preprocesamiento de datos, incluyendo imputación de valores faltantes y codificación de variables categóricas. Se dividió el conjunto de datos en entrenamiento (80%) y prueba (20%) para evaluar diversos algoritmos de aprendizaje automático: regresión logística, árboles de decisión, bosque aleatorio, SVM, KNN, XGBoost y Naive Bayes. Para validar la robustez del modelo, se empleó validación cruzada de 10 pliegues. La interpretabilidad de los resultados se realizó mediante valores SHAP, que explican la contribución de cada variable en las predicciones.

En la Figura 1, se muestra la arquitectura para predecir la probabilidad de niños sin dosis en 12 a 23 meses.

Figura 1

La arquitectura de los algoritmos de machine learning



Nota. Adaptado de *Prediction of zero-dose children using supervised machine learning algorithm in Tanzania: evidence from the recent 2022 Tanzania Demographic and Health Survey* (p. 2), por Asnake, A. A., Seifu, B. L. y Gebrehana, A. K., 2025.

- **Resultados:**

Se encontró que aproximadamente el 7.45% de los niños en la muestra no habían recibido ninguna vacuna. El clasificador de bosque aleatorio superó a los otros algoritmos, logrando métricas como precisión de 0.94, recall de 0.96, puntuación F1 de 0.95 y un AUC de 0.99, siendo el mejor método para identificar niños sin dosis. Los factores más influyentes en la predicción fueron el desempleo materno, la falta de educación materna y el tamaño reducido de la familia, sugiriendo que condiciones socioeconómicas y de acceso a servicios de salud impactan significativamente en la inmunización infantil.

Demsash, A.W., Chereka, A.A., Walle, A.D., Kassie, S.Y., Bekele, F., Bekana, T. (2023). Machine learning algorithms' application to predict childhood vaccination among children aged 12–23 months in Ethiopia: Evidence 2016 Ethiopian Demographic and Health Survey dataset. PLoS ONE 18(10): e0288867. https://doi.org/10.1371/journal.pone.0288867

- **Resumen:**

La investigación muestra la problemática de la baja cobertura de inmunización infantil en países en vías de desarrollo como Etiopía, donde diversos factores sociodemográficos, culturales y de acceso a servicios de salud contribuyen a que muchos niños no completen su esquema de vacunación. A pesar del reconocimiento de la vacunación como una de las intervenciones más rentables para reducir la mortalidad infantil, persisten tasas de inmunización incompletas, retrasadas o abandonadas, lo que compromete la salud pública y el logro de metas nacionales e internacionales. La identificación de los factores asociados y la predicción temprana de niños en riesgo de no vacunarse o retrasar su vacunación tradicionalmente se han abordado mediante análisis estadísticos convencionales, como la regresión logística. Sin embargo, estos métodos presentan limitaciones para captar relaciones complejas y no lineales entre múltiples variables. En respuesta, la aplicación de técnicas avanzadas de minería de datos y algoritmos de *machine learning* ha demostrado ser una estrategia eficaz para analizar grandes volúmenes de datos biomédicos y demográficos, permitiendo la detección de patrones subyacentes y la predicción precisa del estado de vacunación.

El uso de reglas de asociación y modelos predictivos no solo facilita la identificación de atributos clave que influyen en la inmunización infantil, sino que también apoya la formulación de intervenciones dirigidas. Esto resulta esencial para mejorar las políticas de salud pública y optimizar los recursos en contextos donde la cobertura vacunal todavía presenta importantes brechas.

Este estudio tiene como objetivo evaluar diferentes algoritmos de *machine learning* y minería de reglas de asociación para predecir la vacunación infantil en niños de 12 a 23 meses en Etiopía, generando conocimientos que puedan orientar acciones eficaces para incrementar la cobertura de vacunación y reducir las desigualdades en la inmunización infantil.

- **Base de datos:**

La base de datos utilizada en este estudio proviene de la Encuesta Demográfico y de Salud de Etiopía (EDHS) de 2016. Esta fuente proporciona datos detallados sobre diversos aspectos relacionados con la salud, demografía y sociodemográfica de la población, de esta encuesta se extrajeron los registros de 1,617 niños de entre 12 y 23 meses de edad que vivían con sus madres, constituyendo la muestra final para el análisis sobre el estado de vacunación infantil.

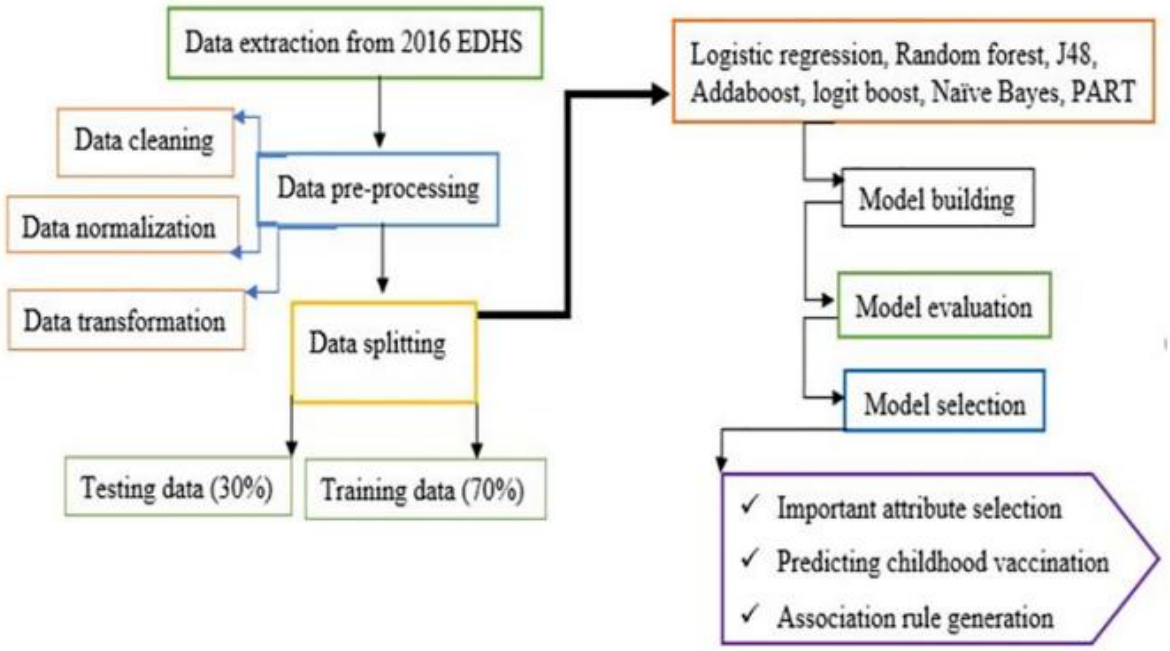
Las variables empleadas para predecir la vacunación infantil incluyen tanto factores sociodemográficos como de acceso y comportamiento, entre ellas: Nivel de educación de las madres, Región de residencia, Residencia rural o urbana, Edad de la madre, Estado de vacunación infantil, Visitas a atención prenatal, Lugar de parto , Nivel socioeconómico, Sexo del niño, Edad del niño, Intervalo entre nacimientos, Orden de nacimiento, Exposición a medios de comunicación, Distancia a instalaciones de salud, Nivel de ingresos del hogar, Migración de cuidadores, Nivel de conocimiento y actitud respecto a vacunas, Temor a efectos secundarios de vacunas, Participación en programas de salud y vacunación. Como variable dependiente es Vacunación infantil en niños de 12 a 23 meses.

- **Metodología:**

El estudio siguió un proceso estructurado dividido en varias fases. Primero se realizó la extracción de datos, utilizando como fuente la Encuesta Demográfica y de Salud de Etiopía (EDHS 2016), que proporcionó información sobre las características sociodemográficas, de salud materna e infantil y el estado de vacunación de niños de 12 a 23 meses. Posteriormente, se llevó a cabo el preprocesamiento de los datos, que incluyó la limpieza de registros incompletos, inconsistentes o duplicados, la normalización de valores numéricos para hacer comparables las variables y la transformación de datos categóricos mediante recodificación y ajustes de formato para su procesamiento en los algoritmos de m*achine learning*. Una vez depurada la base, se realizó la división de los datos, destinando el 70% al entrenamiento de los modelos y el 30% a la prueba, con el fin de evaluar objetivamente su rendimiento predictivo. En la fase de construcción y evaluación de modelos, se aplicaron ocho algoritmos de Machine Learning (Regresión Logística, Random Forest, J48, AdaBoost, Logit Boost, Naïve Bayes y PART), los cuales fueron entrenados con el conjunto de entrenamiento y validados con el de prueba. La selección del modelo óptimo se basó en la comparación de métricas derivadas de la matriz de confusión (precisión, sensibilidad, especificidad y F1-score), así como en el área bajo la curva ROC (AUC) y el estadístico Kappa. Finalmente, en la etapa de análisis, el modelo con mejor desempeño permitió llevar a cabo la selección de atributos más influyentes en la predicción de la vacunación infantil, la clasificación de los niños en vacunados o no vacunados y la generación de reglas de asociación tipo if/then, que describieron los patrones y combinaciones de factores determinantes en la cobertura de vacunación.

Figura 2

Flujo de la metodología



Nota. Adaptado de *Machine learning algorithms'application to predict childhood vaccination among children aged 12–23 months in Ethiopia: Evidence from the 2016 Ethiopian Demographic and Health Survey dataset* (p. 10), por Demsash, A. W., Chereka, A. A., Walle, A. D., Kassie, S. Y., Bekele, F. y Bekana, T., 2023.

- **Resultados:**

En esta investigación se evidenciaron que, de los 1,617 niños de 12 a 23 meses analizados, apenas el 38.9% contaba con el esquema completo de vacunación, mostrando coberturas relativamente altas en la vacuna BCG (68.6%) y en la primera dosis de polio (80.8%), pero descensos notables en las dosis sucesivas como la tercera de DPT (53%) y la vacuna de sarampión (54.1%). En la comparación de algoritmos de aprendizaje automático, el modelo PART alcanzó el mejor desempeño con una exactitud de 95.5 % y un área bajo la curva ROC de 91.89%, seguido de J48 (89.24%), perceptrón multicapa (87.20%) y bosque aleatorio (82.37%), superando a los demás modelos evaluados. Asimismo, el análisis de importancia de atributos identificó como factores más influyentes en la vacunación infantil las visitas prenatales adecuadas, el parto institucional, la asistencia a centros de salud en los últimos 12 meses, el nivel educativo de la madre y el nivel socioeconómico. Finalmente, las reglas de asociación generadas mostraron que los niños de madres con mejor nivel económico, controles

prenatales suficientes y residencia urbana tenían hasta un 86.7% de probabilidad de contar con el esquema de vacunación completo.

Alemayehu, M.A. (2024). Machine learning algorithms for prediction of measles one vaccination dropout among 12-23 months children in Ethiopia. BMJ Open. https://doi.org/10.1136/bmjopen-2024-089764

- **Resumen:**

La investigación busca como objetivo identificar los principales factores que contribuyen al abandono de la vacunación contra el sarampión en niños de 12 a 23 meses en Etiopía, usando técnicas avanzadas de aprendizaje automático (AA). A pesar de la disponibilidad de una vacuna segura y eficaz, Etiopía ha presentado brotes recurrentes de sarampión en los últimos años, en parte debido a altas tasas de abandono en los programas de inmunización. La falta de vacunación contribuye significativamente a estos brotes y a la mortalidad infantil relacionada.

Para abordar la problemática, los investigadores utilizaron datos de la Encuesta Demográfica y de Salud de Etiopía de 2016, que proporciona información detallada sobre características sociodemográficas, salud y acceso a servicios de los hogares. Después del procesamiento y limpieza de datos, que incluyó la eliminación de registros con valores faltantes y la aplicación de técnicas de balanceo de clases (SMOTE), se construyeron modelos predictivos para identificar qué factores estaban relacionados con el abandono de la vacunación.

Se implementaron ocho algoritmos de aprendizaje automático supervisado, incluyendo XGBoost, *Random Forest*, *Gradient Boosting*, SVM, *Decision Tree*, *Naive Bayes*, *K-Nearest Neighbors* y Regresión Logística. Para evaluar la efectividad de los modelos, se emplearon métricas como precisión, recuperación, puntuación F1 y el área bajo la curva ROC (AUC). El algoritmo XGBoost resultó ser el más eficiente, logrando una precisión de 73.9% y una AUC de 0.813, indicando una buena capacidad discriminativa para predecir los casos de abandono de la vacunación.

- **Base de datos:**

Se utilizó la Encuesta Demográfica y de Salud (EDHS) de Etiopía de 2016, que proporciona datos de hogares y madres residentes o visitantes en el país. En total, se analizaron 3,893 observaciones tras el proceso de limpieza y preprocesamiento de los datos, con

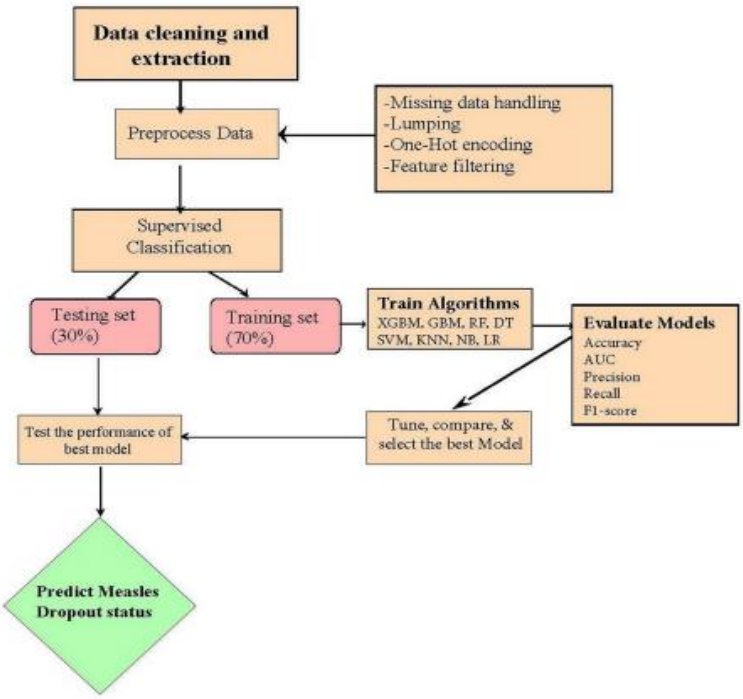
información sobre variables sociodemográficas, de salud, y de acceso a servicios. La variable objetivo fue si el niño abandonó la vacunación contra el sarampión (sí/no).

- **Metodología:**

La investigación realizo un proceso estructurado dividido en varias fases. Primero se realizó la extracción de datos, utilizando como fuente la Encuesta Demográfica y de Salud de Etiopía (EDHS 2016), que proporcionó información sobre las características sociodemográficas, de salud materna e infantil y el estado de vacunación de niños de 12 a 23 meses. Posteriormente, se llevó a cabo el preprocesamiento de los datos, que incluyó la limpieza de registros incompletos, inconsistentes o duplicados, la normalización de valores numéricos para hacer comparables las variables y la transformación de datos categóricos mediante recodificación y ajustes de formato para su procesamiento en los algoritmos de Machine Leaning. Una vez depurada la base, se realizó la división de los datos, destinando el 70% al entrenamiento de los modelos y el 30% a la prueba, con el fin de evaluar objetivamente su rendimiento predictivo. En la fase de construcción y evaluación de modelos, se aplicaron ocho algoritmos de *machine learning*.

Figura 3

Arquitectura del modelo



Nota. Adaptado de *Machine learning algorithms for prediction of measles one vaccination dropout among 12–23 months children in Ethiopia* (p. 4), por Alemayehu, M. A., 2024.

- **Resultados:**

Los resultados de la investigación evidenciaron que, a partir del análisis de datos de la Encuesta Demográfica y de Salud de Etiopía 2016 (n = 3 893), el 39.9% de los niños abandonó la vacunación contra el sarampión, superando ampliamente el umbral recomendado por la OMS. Entre los ocho algoritmos de aprendizaje automático evaluados, el modelo XGBoost presentó el mejor desempeño, alcanzando una precisión del 73.9%, sensibilidad del 93.7%, AUC del 88.1% y una puntuación F1 del 81.3%. Asimismo, se identificaron como principales factores asociados al abandono la edad materna joven, la pertenencia a la religión Jehová/Adventista, el bajo nivel educativo de los padres, la falta de ocupación materna, el tamaño familiar numeroso y la residencia en regiones con menores recursos, como Oromia y Somalí; en contraste, un mayor nivel educativo, el empleo materno y la residencia urbana se relacionaron con un menor riesgo.

Figura 4

Comparativa de los ocho algoritmos de aprendizaje automático utilizados

ML model	Accuracy (95% CI)	SS/recall	Precision/PPV	AUC–ROC	F1-score	Kappa
Xgboost	0.739 (0.722, 0.755)	0.937	0.718	0.881	0.813	0.405
GBM	0.604 (0.586, 0.623)	0.709	0.020	0.646	0.038	0.017
RF	0.603 (0.594, 0.621)	1.000	0.601	0.820	0.750	0.008
LR	0.616 (0.598, 0.635)	0.861	0.632	0.627	0.729	0.122
SVM	0.703 (0.685, 0.720)	0.940	0.684	0.644	0.791	0.317
NB	0.600 (0.581, 0.618)	1.000	0.600	0.593	0.750	0.001
KNN	0.617 (0.599, 0.636)	0.918	0.623	0.620	0.742	0.096
DT	0.630 (0.612, 0.648)	0.880	0.639	0.597	0.741	0.150

Nota. Adaptado de *Machine learning algorithms for prediction of measles one vaccination dropout among 12-23 months children in Ethiopia* (p. 6), por Alemayehu, M. A., 2024.

Chandir, S., Siddiqi, D.A., Hussain, O.A., Niazi, T., Shah, M.T., Dharma, V.K., Habib, A., Khan, A.J. (2018). Using Predictive Analytics to Identify Children at High Risk of Defaulting from a Routine Immunization Program: Feasibility Study. JMIR Public Health Surveill 2018;4(3):e63. https://doi.org/10.2196/publichealth.9681

- **Resumen:**

Esta investigación de factibilidad tiene como objetivo evaluar la viabilidad y precisión de una herramienta de análisis predictivo basada en machine Learning para identificar a los niños en riesgo de no completar sus visitas de inmunización en países de ingresos bajos y medianos. La problemática central en la alta tasa de incumplimiento y abandono en los programas de inmunización infantil en estos países, que limita la cobertura, reduce la protección comunitaria y favorece la persistencia de enfermedades prevenibles por vacunación. A pesar de que las vacunas son gratuitas y están disponibles, muchos niños no completan el esquema de vacunación debido a factores socioeconómicos, logísticos y de desconocimiento, lo que contribuye a tasas elevadas de rechazo, retrasos y pérdidas de seguimiento. La herramienta de análisis predictivo busca solucionar esta problemática al facilitar la identificación temprana y precisa de los niños en riesgo, permitiendo que los vacunadores y las autoridades específicas intervenciones dirigidas y oportunas. Para ello, los autores utilizaron una base de datos de más de 47,000 registros longitudinales de vacunación, que fueron divididos en conjuntos de entrenamiento y validación para desarrollar y evaluar diferentes modelos de *machine learning*, incluyendo *random forest*, *recursive partitioning, support vector machines* y *C-forest*. Estos modelos analizaron variables como género, idioma, residencia, cronología de las vacunas, puntualidad, edad y personal de vacunación. Los resultados mostraron que el modelo de árboles de decisión (*recursive partitioning*) alcanzó la mejor capacidad predictiva, con un área bajo la curva (AUC) de 0.791, una sensibilidad de 94.9% y una precisión general de aproximadamente 79%. Esto indica que la herramienta puede identificar correctamente a la mayoría de los niños en riesgo de incumplimiento con alta fiabilidad. La implementación y validación de este sistema sugieren que el uso de análisis predictivo en entornos con bajos recursos puede ser una estrategia efectiva para superar las dificultades existentes, facilitando la identificación temprana de quienes necesitan intervenciones específicas, optimizando esfuerzos y recursos para incrementar las tasas de inmunización completa, oportuna y equitativa.

- **Metodología:**

Este estudio ha contemplado tres etapas principales para el desarrollo de la metodología:

Datos: Se utilizó un conjunto de registros longitudinales de inmunización, con un total de 47,554 entradas provenientes del Registro Digital de Inmunizaciones de Zindagi Mehfooz, en Sindh y Punjab, Pakistán. Los datos incluyeron variables como genero del niño, idioma hablado en el hogar del niño, lugar de residencia del niño,vacuna de inscripcion,puntualidad de la vacunación, personal de inscripción, fecha nacimiento, grupo de edad del niño. Se filtraron

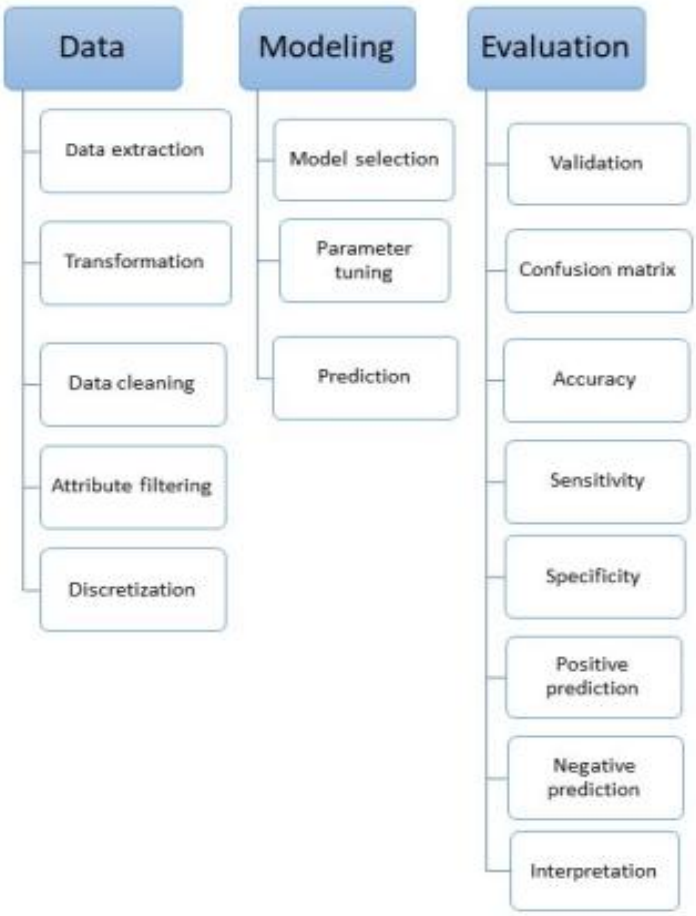
registros inválidos y registros con poca información, dejando una muestra limpia para el análisis. Para la validación del modelo, se utilizó un conjunto de 11,889 casos del conjunto de datos de validación.

Modelado: Se desarrollaron y compararon varios modelos de aprendizaje automático, incluyendo árboles de decisión recursivos, Random Forest, Support Vector Machines (SVMs) y C-forest. Estos modelos fueron entrenados usando las variables predictoras seleccionadas y fueron encapsulados en un motor predictivo que seleccionaba automáticamente el método más adecuado. Cada modelo se configuró para predecir la probabilidad de que un niño incumpliera su próxima visita de inmunización.

Evaluación: Cada modelo se evaluó utilizando métricas como precisión, sensibilidad, especificidad, valor predictivo positivo y negativo, además del área bajo la curva ROC (AUC).

Figura 5

Procedimiento del estudio



Nota. Adaptado de *Using Predictive Analytics to Identify Children at High Risk of Defaulting from a Routine Immunization Program: Feasibility Study* (P. 4), por Chandir, S., Siddiqi, D.A., Hussain, O.A., Niazi, T., Shah, M.T., Dharma, V.K., Habib, A., Khan, A.J. 2018.

- **Resultados:**

Los resultados de la investigación muestran que el modelo de bosques aleatorios (*Random Forest*) obtuvo un alto rendimiento en la predicción del incumplimiento de citas de inmunización. Específicamente, alcanzó una sensibilidad del 94.9%, lo que indica que pudo identificar casi todos los niños que no acudirían a las siguientes citas de vacunación. No obstante, presentó una precisión más baja en comparación con otros modelos en cuanto a la tasa global de aciertos (exactitud), aunque destacó por su superioridad en la detección de casos positivos (niños que faltarían).

En términos generales, el modelo de bosques aleatorios resulta útil para identificar a la mayoría de los niños en riesgo de incumplimiento de sus citas de vacunación, lo que facilita la implementación de intervenciones preventivas y la optimización de recursos en los programas de inmunización.

Figura 6

Métricas de rendimiento de todos los modelos predictivos del estudio

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	Negative predicted value (%)
Recursive partitioning	78.9	74.0	84.2	83.4	75.1
Support vector machines	78.8	88.9	68.0	74.9	85.1
Random forests	75.6	94.9	54.9	69.3	91.0
C-Forest	78.6	90.5	65.8	74.0	86.6

Nota. Adaptado de *Using Predictive Analytics to Identify Children at High Risk of Defaulting from a Routine Immunization Program: Feasibility Study* (P. 8), por Chandir, S., Siddiqi, D.A., Hussain, O.A., Niazi, T., Shah, M.T., Dharma, V.K., Habib, A., Khan, A.J. 2018.

2.2. Bases Teóricas

2.2.1. Inteligencia Artificial

A lo largo de la historia, el ser humano ha intentado explicar los procesos que le permiten pensar, actuar y adaptarse al entorno, cuestionándose cómo el cerebro humano, una simple cantidad de materia, logra percibir, interpretar, anticipar y transformar un mundo mucho

más complejo que él mismo. En este contexto, la inteligencia artificial, conocida como IA, surge no solo como una disciplina orientada a comprender estos mecanismos, sino también a reproducirlos mediante el desarrollo de entidades inteligentes, es decir, máquinas capaces de tomar decisiones de forma eficaz y segura frente a diversas y nuevas situaciones. (Russel & Norvig, 2021)

Según Russel y Norvig (2021), la inteligencia artificial ha sido abordada desde distintas perspectivas. Mientras algunos investigadores lo han definido en función del desempeño humano, otros prefieren una definición más abstracta u formal basada en el concepto de racionalidad. Asimismo, existen diferencias en el enfoque, algunos consideran la inteligencia como una característica de los procesos internos y del razonamiento, mientras que otros la entienden a partir del comportamiento inteligente, una caracterización externa. A partir de estas dos dimensiones, humana vs racional y pensamiento vs comportamiento, se derivan cuatro combinaciones cuatro enfoques principales, que se describen a continuación:

- ***Comportarse como humano (el enfoque del test de Turing)***: propuesto por Alan Turing en 1950, en la que una computadora supera la prueba si el interrogador humano, después de realizar ciertas preguntas escritas, no puede distinguir si las respuestas provienen de una persona o una máquina. (Russel & Norvig, 2021)

- ***Pensar como humano (el enfoque del modelo cognitivo)***: para validar este enfoque, es necesario conocer cómo piensan los humanos mediante la introspección, intentando capturar los propios pensamientos mientras pasan; los experimentos psicológicos, observando a una persona en acción; y las imágenes cerebrales, observando el cerebro en acción. (Russel & Norvig, 2021)

- ***Pensar racionalmente (el enfoque de las leyes del pensamiento)***: el filósofo griego Aristóteles fue uno de los primeros en intentar codificar “el pensamiento correcto”, es decir, los procesos de razonamiento irrefutables. Sus silogismos proveían esquemas de estructuras de argumentación mediante las que siempre se llega a conclusiones correctas a partir de premisas correctas. Se creía que estas leyes del pensamiento debían gobernar el funcionamiento de la mente; su estudio dio inicio al campo de la lógica. Sin embargo, debido a las condiciones de incertidumbre del mundo real, se requiere complementar este enfoque con teorías que no solo aborden el pensamiento racional, sino también la acción racional. (Russel & Norvig, 2021)

- ***Comportarse racionalmente (el enfoque del agente racional)***: un agente racional actúa para alcanzar el mejor resultado o, en contextos de incertidumbre, el resultado esperado. Este enfoque tiene dos ventajas principales sobre los otros enfoques. En primer lugar, es más general que el enfoque basado en las leyes del pensamiento, pues la inferencia lógica es solo uno de los medios para garantizar la racionalidad; y, en segundo lugar, se alinea mejor al avance científico que los enfoques basados en el comportamiento o el pensamiento humano, dado que la racionalidad se define de forma clara es aplicable de forma general. Por otro lado, el comportamiento humano se adapta a un entorno particular y, en parte, es producto de un proceso evolutivo complejo que aún no se comprende por completo y dista de ser perfecto. (Russel & Norvig, 2021)

Desde una perspectiva práctica, Mahajan (2022) define la inteligencia artificial (IA) como un concepto amplio que abarca cualquier tecnología, algoritmo o código diseñado para imitar la inteligencia o el comportamiento humano, y que se compone de diversos subdominios aplicables a distintas capacidades humanas.

La IA puede entenderse como la versión creada por el hombre de todos los procesos humanos, manifestada en máquinas, dispositivos o sistemas programados para mostrar patrones de razonamiento y pensamiento similares a los de los humanos. Estas máquinas cumplen funciones específicas en respuesta a determinadas acciones, basándose en la programación y en los conjuntos de datos utilizados como base para su aprendizaje. (Mahajan, 2022)

2.2.1.1. Machine Learning

Según Mitchell (1997), el *machine learning*, o aprendizaje automático, se centra en resolver la pregunta de cómo desarrollar programas capaces de mejorar su rendimiento de manera autónoma a partir de la experiencia. Una perspectiva útil sobre machine learning es que implica examinar un amplio conjunto de hipótesis posibles con el fin de identificar aquella que se ajusta de forma más adecuadas a los datos disponibles y al conocimiento previo existente. Asimismo, Mitchell (1997), define que un programa de computadora aprende de la experiencia E con respecto a alguna clase de tareas T y medida de desempeño P, si su desempeño en las tareas en TI, medido por P, mejora con la experiencia E" (p. 2).

Según Russel y Norvig (2021), se considera que un agente está aprendiendo cuando su desempeño mejora a partir de las observaciones que realiza sobre su entorno. Cuando el agente es una computadora, se denomina machine learning, en el que este agente observa algunos

datos, construye un modelo a partir de dichos datos y lo utiliza como una hipótesis sobre el mundo como una pieza de software que puede resolver problemas.

Existen tres tipos de retroalimentación que pueden acompañar a los datos de entrada, las cuales dan lugar a tres tipos de aprendizaje: supervisado, no supervisado y por refuerzo. (Russel & Norvig, 2021)

2.2.1.1.1. Aprendizaje Supervisado

Para Harrington (2012), el aprendizaje supervisado implica especificar al algoritmo qué predecir. La clasificación y regresión son ejemplos de este tipo de aprendizaje; la primera, busca determinan la clase a la que pertenece una determinada instancia de datos, mientras que la segunda consiste en predecir un valor numérico.

Según Russel y Norvig (2021), en este tipo de aprendizaje el agente analiza pares de entrada-salida y construye una función que relaciona las entradas con las salidas. Por ejemplo, las entradas pueden ser imágenes acompañadas de una salida que indique se corresponden a un "bus", "peatón" u otra clase, lo que se conoce como etiqueta. A partir de estos datos, el agente aprende una función capaz de predecir la etiqueta correspondiente cuando se le proporciona una nueva imagen.

A continuación, se presentan algunos de los principales algoritmos de aprendizaje supervisado:

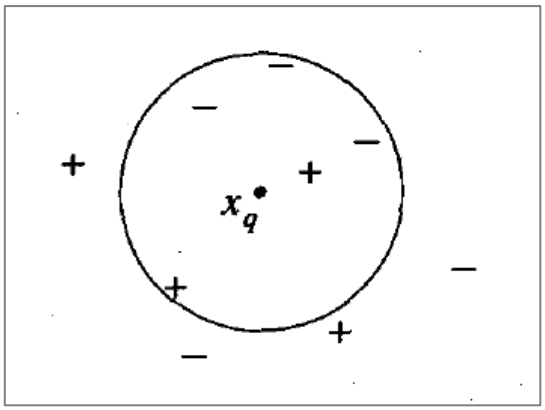
2.2.1.1.1.1. K-Nearest Neighbor (KNN)

Según Mitchell (1997), este algoritmo considera que cada instancia puede representarse como un punto dentro de un espacio de n dimensiones. La determinación de los vecinos más cercanos se realiza utilizando la distancia euclidiana estándar.

La Figura 7 muestra el funcionamiento del algoritmo KNN en un escenario donde las instancias se representan como puntos dentro de un espacio bidimensional y la función objetivo tiene valores booleanos, es decir "+" y "-". También se muestra un punto de consulta x. un algoritmo 1-NN clasificaría como un ejemplo positivo, mientras que un algoritmo 5-NN lo clasificaría como un ejemplo negativo.

Figura 7

KNN. Conjunto de ejemplos de entrenamiento positivos y negativos



Nota. Adaptado de *Machine Learning (*p. 233), por Mitchell, T., 1997, McGraw-Hill Science

2.2.1.1.1.2. Árboles de Decisión

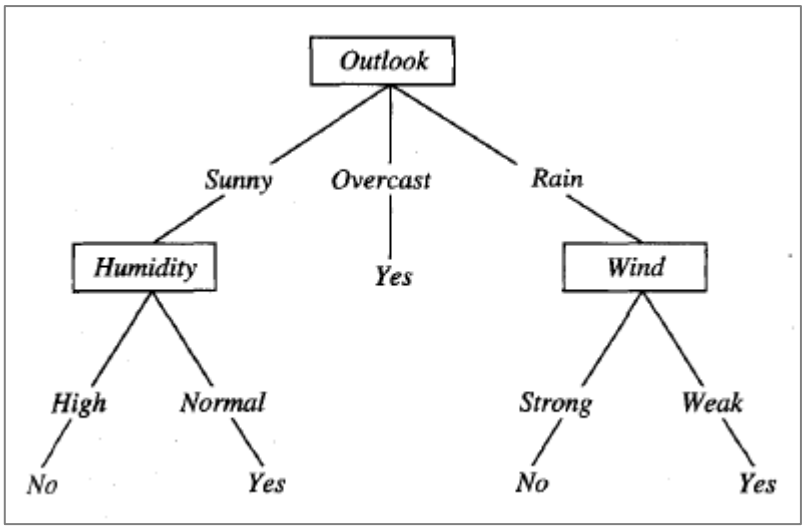
Para Russel y Norvig (2021), un árbol de decisión es una representación de una función que mapea un vector de valores de atributos a un único valor de salida, es decir, a una "decisión". Este tipo de modelo llega a su decisión mediante la ejecución de una secuencia de pruebas, que inicia en la raíz del árbol y sigue la rama apropiada hasta llegar a una hoja. Cada nodo interno en el árbol corresponde una prueba de valor de uno de los atributos de entrada, las ramas desde el nodo están etiquetadas con los posibles valores del atributo, y los nodos hoja indican el valor que la función debe devolver como resultado final.

Para Mitchell (1997), el aprendizaje de árboles de decisión es un método utilizado para aproximar funciones objetivo de valores discretos, en el cual la función resultante se representa mediante un árbol de decisión. Los árboles aprendidos pueden expresarse como un conjunto de reglas *if-then* para facilitar su comprensión. Este método de aprendizaje es ampliamente utilizado dentro de los algoritmos de inferencia inductiva y ha demostrado ser efectivo en diversas aplicaciones, que van desde diagnósticos médicos hasta la evaluación de riesgo crediticio de solicitantes de préstamos.

La Figura 8 muestra un árbol de decisión típico aprendido, el cual clasifica las mañanas de los sábados según si son adecuadas o no para jugar tenis. (Mitchell, 1997)

Figura 8

Árbol de decisión para el concepto "jugar tenis"



Nota. Adaptado de *Machine Learning* (p. 53), por Mitchell, T., 1997, McGraw-Hill Science.

2.2.1.1.1.3. Logistic Regression

La regresión logística es un modelo de aprendizaje supervisado ampliamente utilizado en problemas de clasificación binaria. A diferencia de los clasificadores lineales tradicionales, que aplican una función umbral rígida para decidir entre dos clases, por ejemplo 0 o 1, la regresión logística reemplaza dicha función por una función continua y diferenciable, lo que mejora el proceso de aprendizaje.

La regresión logística introduce la función logística o sigmoide, la cual transforma la salida lineal del modelo en un valor continuo comprendido entre 0 y 1. Esta función se expresa de la siguiente manera:

$$Logistic(z) = \frac{1}{1 + e^{-z}} \quad (1)$$

Aplicando esta función, la hipótesis del modelo se define como:

$$h_w(x) = Logistic(w \cdot x) = \frac{1}{1 + e^{-w \cdot x}} \quad (2)$$

Donde w representa el vector de parámetros o pesos, y x el vector de características de entrada. El resultado $h_w(x)$ puede interpretarse como la probabilidad de que el ejemplo pertenezca a la clase positiva (1).

El proceso de ajustar los pesos de este modelo para minimizar la pérdida en un conjunto de datos se denomina regresión logística. No existe una solución analítica sencilla (forma

cerrada) para encontrar el valor óptimo w con este modelo, pero el cálculo mediante descenso por gradiente es directo. (Russel & Norvig, 2021)

2.2.1.1.1.4. Random Forest

El modelo de bosques aleatorios, o *random forest*, es una variante del *bagging* de árboles de decisión en la que se implementan pasos adicionales para hacer que el conjunto de K árboles sea más diverso, reduciendo así la varianza. El modelo puede aplicarse tanto para problemas de clasificación como de regresión.

La idea principal consiste en introducir aleatoriedad en la selección de atributos, en lugar de únicamente variar los ejemplos de entrenamiento. En cada punto de división durante la construcción del árbol, se selecciona una muestra aleatoria de atributos y, entre ellos, se elige el que proporciona la mayor ganancia de información. Por lo general, si hay n atributos, se considera $\sqrt{n}$ atributos en cada división en problemas de clasificación, o $n/3$ en el caso de problemas de regresión.

Una mejora adicional consiste en introducir aleatoriedad en la selección del valor de división; para cada atributo seleccionado, se toman varias muestras de valores candidatos de una distribución uniforme dentro del rango del atributo. Luego, se elige el valor que produce la mayor ganancia de información. Esto aumenta la probabilidad de que cada árbol del bosque sea diferente. Los árboles construidos de esta forma se denominan árboles extremadamente aleatorizados, o *ExtraTrees*.

El modelo de *random forest* ha demostrado ser altamente exitoso en una amplia variedad de aplicaciones. Por ejemplo, en las competencias de ciencia de datos de *Kaggle*, fueron el enfoque más utilizado por los equipos ganadores entre 2011 y 2014, y continúan siendo una técnica popular en la actualidad, aunque el *deep learning* y el *gradient boosting* se han vuelto aún más comunes entre los ganadores recientes.

En el ámbito financiero, este modelo se ha utilizado en tareas como la predicción de incumplimiento de tarjetas de crédito, la estimación de ingresos familiares y la valoración de opciones. En el sector industrial, se ha aplicado en el diagnóstico de fallas en maquinaria y la teledetección. En bioinformática y medicina, han sido aplicados en la detección de retinopatía diabética, el análisis de expresión génica mediante *microarrays*, el análisis de proteínas por espectrometría de masas, el descubrimiento de biomarcadores y la predicción de interacciones proteína-proteína. (Russel & Norvig, 2021)

2.2.1.1.1.5. Adaptive Boosting (AdaBoost)

Según Harrington (2012), *AdaBoost* basa su funcionamiento en el principio de que a cada ejemplo del conjunto de entrenamiento se le asigna un peso, representado por un vector denominado D. Inicialmente, todos estos pesos son iguales.

Primero, se entrena un clasificador débil utilizando los datos de entrenamiento. A continuación, se calculan los errores producidos por este clasificador y se entrena nuevamente con el mismo conjunto de datos. En esta segunda iteración, los pesos del conjunto de entrenamiento se ajustan: los ejemplos que fueron clasificados correctamente en la primera pasada reciben un peso menor, mientras que aquellos que fueron clasificados de manera incorrecta reciben un peso mayor. Para obtener una única predicción a partir de todos estos clasificadores débiles, *AdaBoost* les asigna valores α a cada uno de los clasificadores. Los valores α están basados en los errores de cada clasificador. El error ε se define como:

$$\varepsilon = \frac{número\ de\ ejemplos\ clasificados\ incorrectamente}{número\ total\ de\ ejemplos} \tag{3}$$

Y el valor α se define como:

$$\alpha = \frac{1}{2}\ln\left(\frac{1-\varepsilon}{\varepsilon}\right) \tag{4}$$

Una vez calculado α, es posible actualizar el vector de pesos D de modo que los ejemplos correctamente clasificados disminuyan su peso, mientras que aquellos que fueron clasificados de forma incorrecta incrementen su peso. El vector D se expresa de la siguiente manera si el ejemplo fue clasificado correctamente:

$$D_i^{(t+1)} = \frac{D_i^{(t)} e^{-\alpha}}{Sum(D)} \tag{5}$$

Y si el ejemplo fue clasificado incorrectamente:

$$D_i^{(t+1)} = \frac{D_i^{(t)} e^{\alpha}}{Sum(D)} \tag{6}$$

Luego de calcular el nuevo vector de pesos D, *AdaBoost* inicia una nueva iteración. Este proceso de entrenamiento y ajuste de pesos se repite hasta que el error de entrenamiento sea igual a 0 o hasta que el número de clasificadores débiles alcance un valor definido previamente por el usuario.

2.2.1.1.1.6. Gradient Boosting

El *gradient boosting*, también conocido como *gradient boosting machines* (GBM) o *gradient boosted regression trees* (GBRT), es un método ampliamente utilizado en la actualidad. Como su nombre lo indica, se trata de una técnica de *boosting* que se basa en el uso del descenso por gradiente para optimizar el rendimiento del modelo.

A diferencia de *AdaBoost*, donde se parte de una hipótesis inicial h_1 y se añaden sucesivamente nuevas hipótesis que prestan especial atención a los ejemplos que fueron clasificados incorrectamente, en *gradient boosting* también se agregan nuevas hipótesis en cada iteración, pero estas no se enfocan en ejemplos específicos, sino en el gradiente entre las respuestas correctas y las respuestas dadas por las hipótesis anteriores.

Al igual que en otros algoritmos que emplean descenso por gradiente, se parte de una función de pérdida diferenciable; por ejemplo, el error cuadrático medio en problemas de regresión o la pérdida logarítmica en problemas de clasificación. De manera similar a *AdaBoost*, se construye un árbol de decisión. En el *gradient boosting*, en lugar de actualizar los parámetros del modelo existente, se actualizan los parámetros del siguiente árbol, pero este proceso debe hacerse de modo que se reduzca la pérdida avanzando en la dirección del gradiente.

2.2.1.1.1.7. Bootstrap Aggregating (Bagging)

Con la técnica *bagging*, se generan K conjuntos de entrenamiento distintos a partir del conjunto de entrenamiento original utilizando muestreo con reemplazo. Es decir, se seleccionan aleatoriamente N ejemplos del conjunto original, pero cada selección puede incluir ejemplos que ya hayan sido elegidos previamente. Luego, se ejecuta el algoritmo de aprendizaje automático sobre esos N ejemplos para obtener una hipótesis. Este proceso se repite K veces, obteniendo así K hipótesis diferentes.

Cuando se necesita predecir el valor de una nueva entrada, se agregan las predicciones de todas las hipótesis. En problemas de clasificación, esto significa tomar el voto plural, o el voto mayoritario en el caso de clasificación binaria, mientras que, en problemas de regresión, la salida final se calcula como el promedio de todas las predicciones individuales.

El *bagging* tiende a reducir la varianza y es un enfoque común cuando se dispone de datos limitados o cuando el modelo base muestra signos de sobreajuste (*overfitting*). Este método puede aplicarse a cualquier tipo de modelo, aunque se utiliza con mayor frecuencia en árboles de decisión, debido a que estos suelen ser inestables, ya que un conjunto de ejemplos

ligeramente distinto puede producir un árbol muy diferente. El uso de *bagging* suaviza esta varianza. Además, su implementación es eficiente en entornos con múltiples recursos de cómputos, ya que el entrenamiento de cada hipótesis puede llevarse a cabo de forma paralela. (Russel & Norvig, 2021)

2.2.1.1.1.8. Support Vector Machines (SVM)

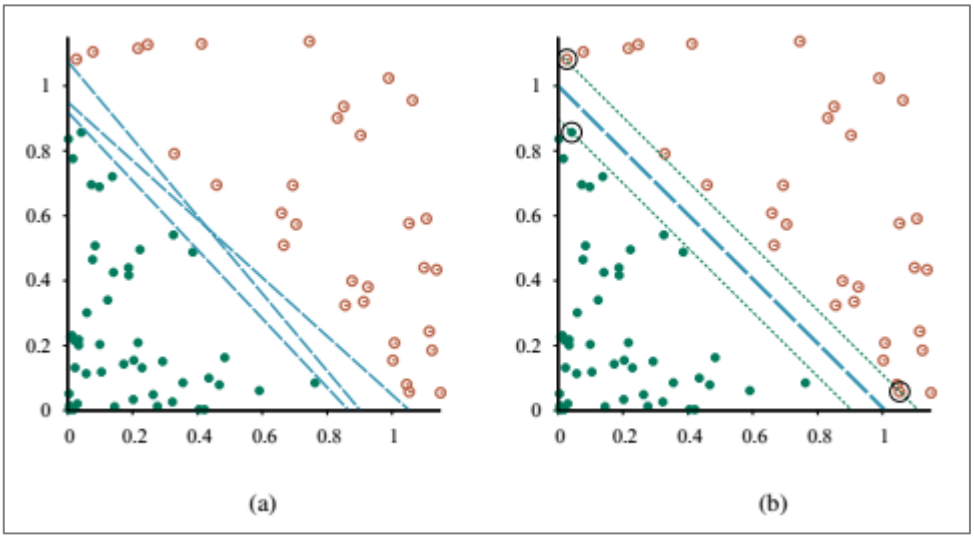
Según Russel y Norvig (2021), a comienzos de la década del 2000, el modelo de *support vector machine* (SVM) fue uno de los métodos más utilizados para el ámbito del aprendizaje supervisado, especialmente en escenarios donde no se contaba con conocimiento previo sobre un dominio. Actualmente ese lugar ha sido ocupado por las redes de *deep learning* y *random forests*, pero las SVM conservan tres propiedades atractivas:

- Construyen un separador de margen máximo, es decir, un límite de decisión con la mayor distancia posible a los puntos de ejemplo, lo que favorece su capacidad de generalización ante datos nuevos.
- Crean un hiperplano de separación lineal, pero tienen la capacidad de incrustar los datos en un espacio de mayor dimensión. A menudo, los datos que no son linealmente separables en el espacio de entrada original son fácilmente separables un espacio de mayor dimensión.
- Son no paramétricas, es decir, el hiperplano de separación está definido por un conjunto de datos de ejemplo, no por una colección de valores de parámetros. Sin embargo, mientras que los modelos de vecinos más cercanos necesitan retener todos los ejemplos, un modelo SVM conserva solo los ejemplos que están más cerca del plano de separación. Es por ello que las SVM combinan las ventajas de modelos no paramétricos y paramétricos, tienen la flexibilidad para representar funciones complejas, pero son resistentes al sobreajuste.

En la Figura 9, se muestra un problema de clasificación binaria con tres posibles límites de decisión, cada uno de los cuales es un separador lineal. Aunque cada uno clasifica correctamente todos los ejemplos, es importante destacar que algunos ejemplos pueden ser más significativos que otros para una generalización exitosa. Por ejemplo, un límite de separación que se acerque demasiado a ciertos ejemplos podría indicar un riesgo de sobreajuste si otros ejemplos no clasificados correctamente terminan en el lado incorrecto de la línea.

Figura 9

SVM. (a) Dos clases de puntos y tres posibles separadores lineales. (b) El separador del margen máximo está en el punto medio del margen



Nota. Adaptado de Artificial Intelligence: A Modern Approach (4th ed.) (p. 693), por Russel, S. J., Norvig P., 2021, Pearson.

2.2.1.1.2. Aprendizaje No Supervisado

En este tipo de aprendizaje, no se dispone de una etiqueta o valor objetivo para los datos. Una de las tareas más comunes consiste en agrupar elementos con características similares, proceso conocido como *clustering*. Otra aplicación importante es la estimación de densidad, que permite calcular valores estadísticos representativos de los datos. Asimismo, el aprendizaje no supervisado puede emplearse para reducir la dimensionalidad de un conjunto de datos con muchas características, facilitando así su visualización en dos o tres dimensiones. (Harrington, 2012)

Según Russel y Norvig (2021), en el aprendizaje no supervisado el agente es capaz de identificar patrones en los datos de entrada sin recibir retroalimentación explícita. La tarea más común es el *clustering*, que consiste en detectar grupos potencialmente útiles de ejemplos de entrada. Por ejemplo, al mostrarle millones de imágenes obtenidas de internet, un sistema de visión computacional puede identificar un gran grupo de imágenes con características similares que un hablante de inglés llamaría “cats”.

2.2.1.1.2.1. K-Means Clustering

Según Harrington (2012), el *clustering* es un tipo de aprendizaje no supervisado que permite agrupar automáticamente elementos con características similares, es decir, realiza una clasificación automática. Casi todo se puede agrupar, y cuanto más similares sean los elementos en el grupo, se realizarán mejor las agrupaciones. La diferencia con la clasificación es que en esta última se conoce lo que se está buscando. Ese no es el caso en el *clustering* que a veces se le denomina clasificación no supervisada, ya que genera un resultado similar al de la clasificación, pero sin requerir etiquetas previas.

K-Means es una técnica de agrupamiento basada en la distancia cuyo mecanismo central implica determinar un número preestablecido de grupos (el valor *K*) y asignar iterativamente puntos de datos a estos grupos, recalculando los centroides en cada iteración. En este método, los centroides iniciales de los clústeres se seleccionan de forma aleatoria, y cada punto se asigna al clúster correspondiente según su distancia al centroide. Posteriormente, el centroide de cada clúster se actualiza para ser la media de todos los puntos asignados a ese clúster. Este proceso se repite de manera iterativa hasta cumplir con los criterios de convergencia o finalización establecidos. (Ao et al., 2024)

Según Ao et al. (2024), en la aplicación del algoritmo *K-Means* al análisis de color de imágenes, el paso inicial es establecer el número deseado de clústeres (el valor *K*), que corresponde al número de centroides. Posteriormente, para cada píxel, se calcula la distancia del valor de color a cada centroide como *D*, y el píxel se asigna al clúster más cercano. Después, los valores de color de todos los píxeles dentro de cada clúster se promedian para actualizar los centroides. Este proceso iterativo se repite hasta que se cumplen los criterios predefinidos de convergencia, típicamente evaluados mediante el error cuadrático medio, a. El algoritmo concluye cuando las diferencias entre los resultados de las iteraciones consecutivas son menores que 1, momento en el cual los valores de los centros de clúster de salida representan los valores de color extraídos. Este método optimiza los resultados del agrupamiento mediante actualizaciones sucesivas de los centroides durante cada iteración, alcanzando en última instancia los criterios de convergencia establecidos.

$$D = \sqrt{(r - r_i) - (g - g_i) - (b - b_i)^2} \tag{7}$$

En esta ecuación, *D* es la distancia desde el punto del píxel hasta el centroide; y r_i, g_i, b_i son los valores de color de los otros píxeles.

$$a = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \quad (8)$$

En esta segunda ecuación, a es el error cuadrático medio; n es el número de puntos de píxeles en el clúster; x_i son los puntos de píxeles en el clúster; y μ es la media aritmética de todos los puntos de píxeles en el clúster.

2.2.1.1.3. Aprendizaje por Refuerzo

Según Russel y Norvig (2021), en este tipo de aprendizaje, el agente aprende de una serie de refuerzos (recompensas y castigos). Por ejemplo, al finalizar una partida de ajedrez, el agente recibe la notificación de si ha ganado (una recompensa) o perdido (un castigo). Es responsabilidad del agente determinar cuáles de sus acciones anteriores contribuyeron más al resultado obtenido y ajustar su comportamiento para maximizar las recompensas en el futuro.

Según Russel y Norvig (2021), se categorizan los enfoques de la siguiente manera:

2.2.1.1.3.1. Aprendizaje por Refuerzo basado en Modelos

En estos enfoques, el agente utiliza un modelo del entorno para ayudar a interpretar las señales de recompensa y tomar decisiones sobre cómo actuar. El modelo puede ser inicialmente desconocido, en cuyo caso el agente aprende el modelo observando los efectos de sus acciones, o puede ser conocido de antemano. Por ejemplo, un programa de ajedrez puede conocer las reglas del juego, aunque no sepa cómo elegir las mejores jugadas. En entornos parcialmente observables, el modelo de transición también es útil para la estimación del estado. Los sistemas de aprendizaje por refuerzo basados en modelos a menudo aprenden una función de utilidad U(s), definida en términos de la suma de recompensas desde el estado s en adelante. (Russel & Norvig, 2021)

2.2.1.1.3.2. Aprendizaje por Refuerzo sin Modelo

En estos enfoques, el agente no conoce ni aprende un modelo de transición para el entorno. En cambio, adquiere una representación más directa de cómo comportarse. Esto se presenta en una de dos variedades:

- ***Aprendizaje de utilidad en acción***: La forma más común de este tipo de aprendizaje es el *Q-learning*, donde el agente aprende una función *Q*, o función de calidad, *Q(s,a)*, que representa la suma de recompensas desde el estado s en adelante si se

toma la acción a. Dada una función *Q*, el agente puede determinar qué hacer en s eligiendo la acción con el valor *Q* más alto.

- ***Búsqueda de política***: El agente, denominado agente reflejo, aprende una política π(s) que mapea directamente de estados a acciones.

2.2.1.2. Deep Learning

Según Goodfellow et al. (2016), los algoritmos básicos de *machine learning* tienen un buen desempeño en una amplia variedad de problemas importantes. Sin embargo, no han logrado resolver los principales desafíos en inteligencia artificial, como el reconocimiento de voz o de objetos. La necesidad de desarrollar el aprendizaje profundo, o *deep learning*, surgió en parte debido a la limitación de los algoritmos tradicionales para generalizar eficazmente en estas tareas de IA. La generalización hacia nuevos ejemplos se complejiza enormemente al trabajar con datos de alta dimensionalidad, y los mecanismos utilizados en *machine learning* convencional son insuficientes para aprender funciones complejas en espacios de alta dimensión, que además suelen requerir elevados recursos computacionales. El aprendizaje profundo fue diseñado específicamente para superar estos obstáculos y otros similares.

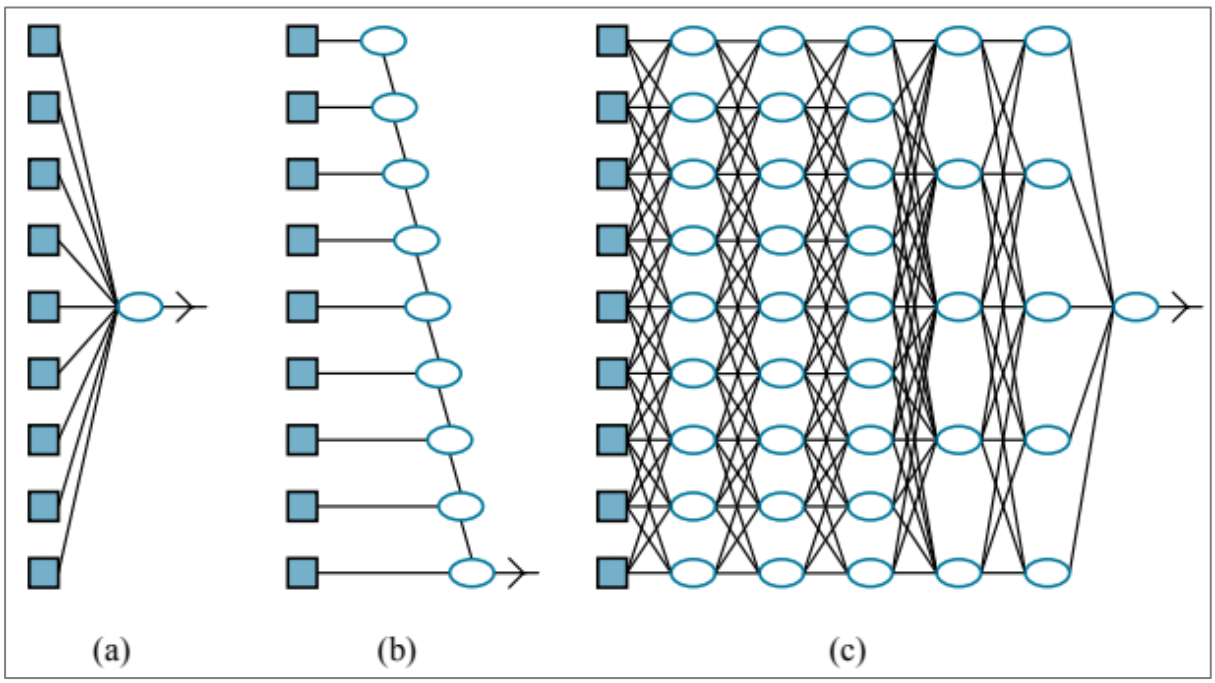
Según Russel y Norvig (2021), *deep learning* es una amplia familia de técnicas de *machine learning* en la que las hipótesis adoptan la forma de circuitos algebraicos complejos con conexiones ajustables. La referencia a "*deep*" se debe al hecho que los circuitos suelen estar organizados en múltiples capas, lo que implica que los caminos de computación desde las entradas hasta las salidas tienen muchos pasos. *Deep learning* es el enfoque más ampliamente utilizado en aplicaciones como reconocimiento visual de objetos, la traducción automática, entre otros. Además, desempeña un papel importante en aplicaciones de aprendizaje por refuerzo.

Deep learning tiene sus orígenes en investigaciones anteriores que buscaban modelar redes de neuronas en el cerebro con circuitos computacionales. Por ello, las redes entrenadas por métodos de *deep learning* son comúnmente conocidas como redes neuronales. Por ejemplo, mientras que métodos como la regresión lineal y logística pueden manejar múltiples variables de entrada, el proceso computacional desde cada entrada hasta la salida es muy corto: se multiplica por un único peso y luego se suma al resultado agregado. Además, cada variable de entrada contribuye independientemente a la salida, sin interactuar entre sí (Figura 10(a)). Esto limita considerablemente la capacidad expresiva de estos modelos, ya que solo pueden representar funciones y límites lineales en el espacio de entrada, mientras que los conceptos del

mundo real suelen ser mucho más complejos. Por otro lado, las listas de decisiones y los árboles de decisiones permiten caminos de computación largos que pueden depender de múltiples variables de entrada, pero solo para una fracción relativamente pequeña de los vectores de entrada posibles (Figura 10(b)). Si un árbol de decisiones tiene caminos de computación largas para una fracción significativa de las entradas posibles, debe ser exponencialmente grande en el número de variables de entrada. La idea de d*eep learning* es entrenar circuitos de tal manera que los caminos de computación sean largos, permitiendo que todas las variables de entrada interactúen de formas complejas (Figura 10(c)). Estos modelos de circuitos resultan ser lo suficientemente expresivos como para capturar la complejidad de los datos del mundo real en muchos tipos importantes de problemas de aprendizaje. (Russel & Norvig, 2021)

Figura 10

(a) Modelo de regresión lineal. (b) Red de lista de decisiones. (c) Red de Deep learning



Nota. Adaptado de Artificial Intelligence: A Modern Approach (4th ed.) (p. 751), por Russel, S. J., Norvig P., 2021, Pearson.

2.2.1.3. Ajuste de hiperparámetros

Los hiperparámetros son variables de configuración externas al modelo que controlan su estructura y comportamiento durante el entrenamiento, podemos tomar en cuenta al número de árboles, la tasa de aprendizaje o la profundidad máxima (Géron, 2022). El ajuste de dichos

hiperparámetros tiene como finalidad encontrar la mejor combinación que maximice el rendimiento del modelo y reduzca el sobreajuste.

Existen diferentes métodos de optimización, como la búsqueda en cuadrícula (Grid Search), la búsqueda aleatoria (Random Search) y los enfoques bayesianos.

2.2.1.4. Validación cruzada

Según Asnake, Seifu y Gebrehana (2025), la validación cruzada es una técnica utilizada para estimar el rendimiento real de un modelo de aprendizaje automático, reduciendo el riesgo de sobreajuste y garantizando una evaluación más estable.

El procedimiento consiste en dividir el conjunto de datos en k subconjuntos o pliegues, utilizando cada uno de ellos de forma alternada como conjunto de prueba y los restantes para entrenamiento.

A partir de los resultados obtenidos en cada iteración, se calculan los promedios y desviaciones estándar de las métricas de Accuracy, Precision, Recall, F1-score y ROC-AUC, con el propósito de analizar la variabilidad entre los pliegues.

De esta manera, se confirma o se descarta la consistencia y robustez del modelo frente a diferentes subconjuntos de datos.

2.2.2. Vacunación Infantil en el Perú

2.2.2.1. Definición de Vacunación

La vacunación es una forma simple, segura y eficaz para proteger al organismo frente a diversas enfermedades infecciosas antes de que ocurra la exposición directa a ellas. Este proceso utiliza las defensas naturales del organismo, estimulando su capacidad de generar resistencia frente a infecciones específicas y fortalecer el sistema inmunológico.

Las vacunas entrenan al sistema inmunológico para que produzca anticuerpos, de la misma manera que lo haría si estuviera expuesto a una enfermedad. Sin embargo, dado que las vacunas contienen únicamente formas muertas o debilitadas de los gérmenes, como virus o bacterias, no causan la enfermedad ni representan un riesgo de sus complicaciones. (OMS, 2024)

2.2.2.2. Importancia y Beneficios

La vacunación infantil representa una de las estrategias más eficaces y costo-efectivas en el ámbito de la salud pública, debido a que contribuye de manera significativa a la prevención de enfermedades infecciosas graves y a la reducción de la mortalidad infantil a nivel mundial. A través de la inmunización, es posible evitar la aparición de patologías que pueden generar complicaciones severas, discapacidades permanentes e incluso la muerte, como el sarampión, la meningitis, la neumonía, el tétanos o la poliomielitis. Según datos de la Organización Mundial de la Salud (OMS), las vacunas aplicadas durante la infancia permiten salvar más de cuatro millones de vidas cada año, lo que evidencia su impacto trascendental en la salud global.

A pesar de que muchas de estas enfermedades han disminuido notablemente en frecuencia, los agentes patógenos que las provocan continúan circulando en distintas regiones del mundo. En el contexto actual, caracterizado por la alta movilidad humana, las enfermedades infecciosas pueden traspasar fronteras con facilidad, exponiendo a las personas no inmunizadas a un riesgo considerable de contagio.

La vacunación no solo tiene beneficios a nivel individual, al ofrecer protección directa al organismo, sino también a nivel colectivo, al contribuir a la creación de una inmunidad comunitaria. Este efecto es especialmente relevante para grupos vulnerables que no pueden recibir vacunas, como los bebés menores de edad, las personas con enfermedades crónicas o aquellas con alergias severas. Estas poblaciones dependen de la inmunización del resto de la comunidad para mantenerse protegidas frente a enfermedades prevenibles mediante vacunación.

Por último, la vacunación infantil constituye un pilar fundamental en la salud pública, no solo por su papel en la prevención de brotes epidémicos, sino también por su impacto positivo en la calidad y esperanza de vida, así como en la reducción de los costos sanitarios asociados al tratamiento de enfermedades. (OMS, 2024)

2.2.2.3. Marco Normativo en el Perú

El marco normativo se sustenta en la Ley N.º 26842 – Ley General de Salud, que reconoce la vacunación como un derecho fundamental, así como en el Decreto Supremo N.º 017-2018-SA, que aprueba el Reglamento de Inmunizaciones, y en diversas resoluciones ministeriales que actualizan el Esquema Nacional de Vacunación (MINSA, 2022).

Capítulo 3: Entorno Empresarial

3.1. Descripción de la empresa

La Dirección de Redes Integradas de Salud Lima Centro (DIRIS Lima Centro) es una institución pública perteneciente al Ministerio de Salud del Perú, encargada de planificar, coordinar y supervisar las acciones de prevención, promoción y atención en salud dentro de su jurisdicción. Su ámbito de acción comprende diversos distritos de la ciudad de Lima, donde articula esfuerzos con establecimientos de salud, municipalidades y actores comunitarios, con el fin de garantizar la cobertura y calidad de los servicios sanitarios.

En el marco de la presente investigación, la DIRIS Lima Centro constituye el entorno organizacional donde se gestionan los programas de inmunización infantil, siendo la entidad que proporciona la información y lineamientos necesarios para el análisis y la implementación de herramientas tecnológicas orientadas a mejorar la cobertura de vacunación.

3.1.1. Reseña histórica y actividad económica

La Dirección de Redes Integradas de Salud (DIRIS) Lima Centro fue creada el 14 de junio de 2017 mediante Resolución Ministerial N° 467-2017/MINSA, como órgano desconcentrado del Ministerio de Salud (MINSA). Su conformación respondió al proceso de modernización y descentralización del sistema sanitario peruano, buscando fortalecer el trabajo del MINSA y garantizar un acceso más equitativo a los servicios de salud. (MINSA, 2017).

La jurisdicción de la DIRIS Lima Centro abarca catorce (14) distritos de la capital: Breña, Jesús María, La Victoria, Lima, Lince, Magdalena del Mar, Miraflores, Pueblo Libre, San Borja, San Isidro, San Juan de Lurigancho, San Luis, San Miguel y Surquillo. Para cumplir con sus funciones, articula esfuerzos en el marco de la Red Integrada de Salud, conformada por 8 hospitales, 64 establecimientos de Salud del Primer Nivel de Atención y 6 centros especializados (Targa-VIH, Zoonosis y Salud Mental). (DIRIS Lima Centro, s.f.).

En cuanto a su actividad económica, la DIRIS Lima Centro, como organismo público, no persigue fines de lucro, sino que se orienta a la gestión de servicios de salud pública financiados con recursos. Entre sus principales actividades destacan la vigilancia epidemiológica, la prevención y control de enfermedades, la promoción de la salud y la gestión de campañas de inmunización infantil.

De manera particular, el área de Inmunizaciones desempeña un rol estratégico en la cobertura vacunal de niños menores de un año. Este proceso se complementa con el trabajo del área de Estadística, perteneciente a la Oficina de Epidemiología, Inteligencia Sanitaria y Docencia e Investigación, que se encarga de procesar y digitalizar los datos de vacunación, generando herramientas de gestión que permiten mejorar la eficiencia y la toma de decisiones.

De esta forma, la DIRIS Lima Centro desempeña un papel importante al garantizar que la población de su jurisdicción tenga acceso a servicios de vacunación gratuitos, seguros y de calidad. Para el Ministerio de Salud, donde está incluida la DIRIS Lima Centro, la vacunación es reconocida por la Organización Mundial de la Salud (OMS) y la Organización Panamericana de la Salud (OPS) como una de las intervenciones más costo-efectivas para prevenir enfermedades, evita millones de muertes cada año a nivel mundial. En el Perú, el Esquema Nacional de Vacunación incluye 18 vacunas que protegen contra 28 enfermedades, y su cumplimiento protege especialmente a los niños menores de un año, que constituyen uno de los grupos más vulnerables (MINSA 2024).

3.1.2. Descripción de la organización

La DIRIS Lima Centro organiza sus actividades a través de una estructura que combina la gestión administrativa con la coordinación operativa de los servicios de salud. Su funcionamiento busca que las acciones de prevención, promoción y atención lleguen de manera oportuna y eficiente a los catorce distritos que conforman su jurisdicción.

3.1.2.1. Organigrama

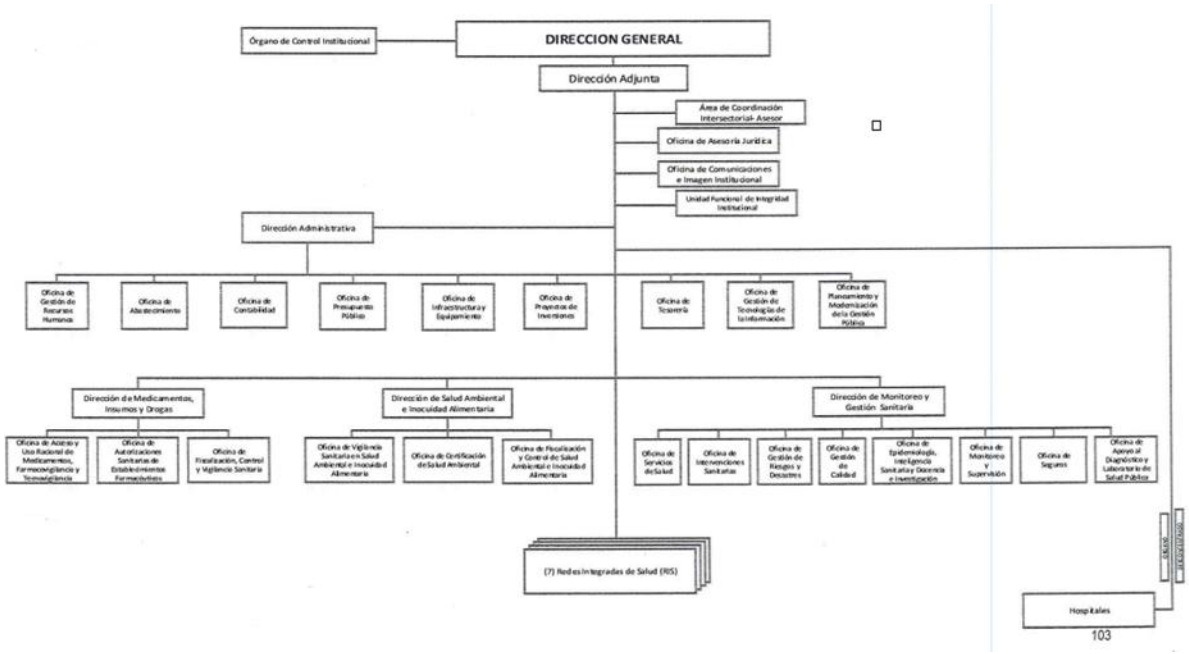
La DIRIS Lima Centro se estructura en torno a la Dirección General, que constituye el órgano de mayor jerarquía y del cual dependen directamente las áreas de Asesoría Jurídica y Comunicaciones. A su vez, cuenta con un Órgano de Control Institucional, encargado de supervisar la correcta gestión de los recursos y velar por la transparencia de los procesos internos. La organización se apoya en la Dirección Administrativa, responsable de la gestión de recursos humanos, financieros, logísticos y tecnológicos, así como de la mejora continua de los procesos administrativos.

En cuanto a los órganos de línea, se tienen: la Dirección de Monitoreo y Gestión Sanitaria, la Dirección de Medicamentos, Insumos y Drogas, la Dirección de Salud Ambiental e Inocuidad Alimentaria, las Redes Integradas de Salud (RIS), y los Establecimientos de Salud. (DIRIS Lima Centro, s.f.)

Esta estructura busca asegurar un funcionamiento articulado, donde cada área cumple un rol definido en la provisión de servicios de salud, con especial atención en la cobertura de inmunizaciones y vigilancia epidemiológica.

Figura 11

Organigrama general de la DIRIS Lima Centro



Nota. Adaptado de *Organigrama Institucional*, por Dirección de Redes Integradas de Salud Lima Centro, s. f.

3.1.2.2. Cadena de suministros

La cadena de suministro se caracteriza por ser un proceso estructurado, que abarca desde la planificación y adquisición de bienes y servicios hasta su distribución y disponibilidad en los establecimientos de salud de su jurisdicción. Este sistema logístico busca garantizar la continuidad operativa de los servicios de salud, asegurando el abastecimiento oportuno de medicamentos, insumos médicos, vacunas, materiales sanitarios y bienes administrativos esenciales para la atención a la población.

3.1.2.2.1. Planificación y programación de requerimientos

La identificación y consolidación de necesidades de bienes e insumos médicos mediante la coordinación con las redes integradas de salud y los establecimientos asistenciales.

Este proceso permite elaborar el Plan Anual de Adquisiciones y Contrataciones (PAAC) de la entidad, priorizando los productos estratégicos de salud pública, como medicamentos esenciales, vacunas, equipos de protección personal y material de laboratorio. La planificación se basa en el análisis de consumo histórico, la demanda proyectada y los lineamientos del MINSA. (DIRIS Lima Centro, 2025)

3.1.2.2.2. Adquisiciones y abastecimiento institucional

El proceso de adquisiciones se ejecuta conforme a la Ley N° 30225 – Ley de Contrataciones del estado y sus reglamentos (MEF, 2019; MEF, 2018), garantizando la transparencia, la eficiencia y la calidad del suministro. La DIRIS Lima Centro establece relaciones de coordinación con proveedores nacionales y con el CENARES (Centro Nacional de Abastecimiento de Recursos Estratégicos en Salud) para la provisión de medicamentos y vacunas. El objetivo principal es asegurar el acceso a productos certificados, en los plazos establecidos y con las condiciones técnicas requeridas para el sector salud.

3.1.2.2.3. Recepción y control de productos

Una vez entregados los bienes por los proveedores, estos son recepcionados en los almacenes institucionales, donde se verifica el cumplimiento de las especificaciones técnicas, cantidad, calidad, fechas de vencimiento y condiciones de embalaje. Este proceso se realiza bajo procedimientos estandarizados que garantizan la trazabilidad y la transparencia.

Los productos son registrados en el sistema SIGA (MEF, 2024) para su control, y los medicamentos o biológicos son validados por la Unidad de Farmacia y la Unidad de Inmunizaciones, según corresponda.

3.1.2.2.4. Almacenamiento y control de inventarios

Se cuenta con almacenes institucionales equipados para conservar medicamentos, vacunas y material médico bajo condiciones controladas. En el caso de productos biológicos, se mantiene la cadena de frío mediante cámaras refrigeradas con monitoreo continuo de temperatura, asegurando la calidad y eficacia de las vacunas.

El control de inventarios se realiza en tiempo real a través del sistema SIGA, lo que permite conocer los niveles de stock, las fechas de caducidad y la rotación de productos.

Este sistema garantiza el abastecimiento continuo y minimiza pérdidas por vencimiento o sobre almacenamiento.

3.1.2.2.5. Distribución y transporte

Los bienes e insumos almacenados son distribuidos periódicamente a las redes integradas y establecimientos de salud, de acuerdo con cronogramas de entrega planificados. La Unidad de Logística y el Área de Almacén coordinan la distribución utilizando vehículos institucionales y operadores logísticos, garantizando la entrega en condiciones óptimas.

Para vacunas e insumos refrigerados, se emplean equipos de transporte con cadena de frío que aseguran la conservación adecuada durante el traslado. Este proceso tiene como objetivo asegurar la continuidad operativa de los servicios de salud en toda la jurisdicción de Lima Centro.

3.1.2.2.6. Supervisión y monitoreo del suministro

En esta etapa se realiza el seguimiento del proceso logístico a través de indicadores de cumplimiento, cobertura de distribución y disponibilidad de productos.

El monitoreo permite identificar quiebres de stock, demoras en la entrega o desviaciones en los registros, implementando acciones correctivas oportunas. Asimismo, se coordina con los responsables de cada establecimiento para verificar la recepción conforme de los productos y la adecuada gestión de los insumos recibidos.

3.1.2.2.7. Servicio post distribución y retroalimentación

Con el fin de mejorar continuamente el proceso logístico, se establecen mecanismos de retroalimentación entre las redes, los establecimientos de salud y la Unidad de Logística. Este servicio incluye la gestión de devoluciones por productos vencidos o dañados, el control de inventarios en campo y la actualización de requerimientos. La información recogida en esta etapa permite ajustar la planificación y optimizar la cadena de suministro, contribuyendo a la eficiencia y transparencia en la gestión de los recursos públicos.

3.1.3. Datos generales estratégicos de la empresa

La Dirección de Redes Integradas de Salud Lima Centro (DIRIS Lima Centro) es una entidad descentralizada del Ministerio de Salud del Perú (MINSA), con más de una década de experiencia en la gestión, supervisión y articulación de los servicios de salud en su jurisdicción.

Su ámbito de acción abarca los distritos del Cercado de Lima. La propuesta de valor institucional se centra en garantizar el acceso oportuno, continuo y equitativo a los servicios de salud de la población, fortaleciendo la gestión sanitaria, la prevención de enfermedades y la

promoción de la salud. Además, busca consolidar un modelo de atención integral y territorial, basado en las Redes Integradas de Salud (RIS), que permita brindar atención de calidad en los diferentes niveles asistenciales.

Como parte de su estrategia institucional, promueve la modernización de la gestión pública en salud mediante el uso de herramientas digitales, la optimización de procesos logísticos, como el abastecimiento de medicamentos y vacunas, y la mejora del desempeño del recurso humano. Asimismo, fomenta la articulación intersectorial y la participación comunitaria como pilares para fortalecer la respuesta del sistema de salud frente a emergencias y necesidades locales.

Para evaluar el cumplimiento de sus objetivos estratégicos, monitorea indicadores de desempeño como la cobertura de inmunización, la disponibilidad de medicamentos esenciales, la oportunidad de atención, la satisfacción del usuario, y el cumplimiento de metas sanitarias. Estos indicadores permiten realizar un seguimiento continuo de la gestión institucional, orientar la toma de decisiones y asegurar la mejora continua en la prestación de los servicios de salud.

3.1.3.1. Visión, misión y valores o principios

Visión: Sector líder que establece políticas públicas en salud centrado en las personas que gozan de una vida más saludable, con acceso universal a los servicios de salud con calidad, integrales, oportunos y eficientes; basados en enfoques de derechos en salud e interculturalidad. (DIRIS Lima Centro, s.f.)

Misión: Ejercer la rectoría del sector y conducir con eficiencia el sistema de salud en concertación con el sector público, privado y actores sociales, centrado en las personas, en la prevención de enfermedades; fortaleciendo el primer nivel de atención, asegurando el acceso y calidad a servicios de salud con infraestructura moderna e interconectada, revalorizando al personal de salud y fortaleciendo una gestión, transparente, oportuna y resolutiva. (DIRIS Lima Centro, s.f.)

Valores: La DIRIS Lima Centro promueve valores institucionales que orientan la conducta de sus trabajadores y consolidan una cultura organizacional basada en la ética y el servicio público. Entre ellos destacan la integridad, la independencia y objetividad, la competencia y el comportamiento profesional, así como la confidencialidad, la transparencia, la inclusión y la no discriminación. Estos valores reflejan el compromiso de la institución con una atención de salud de calidad, equitativa y centrada en las personas, fomentando un entorno

de confianza tanto dentro de la organización como hacia la ciudadanía. (DIRIS Lima Centro, 2024)

Principios: La institución se rige por principios que guían el desempeño diario de sus funcionarios y trabajadores. Entre los más relevantes se encuentran el respeto, la probidad, la eficiencia, la idoneidad y la veracidad, además de la lealtad y obediencia, la justicia y equidad, y la lealtad al Estado de Derecho. Estos principios actúan como lineamientos éticos y normativos que fortalecen la transparencia y la rendición de cuentas, asegurando que la gestión pública responda a estándares de responsabilidad social y legalidad en beneficio de la población. (DIRIS Lima Centro, 2024)

3.1.3.2. Objetivos estratégicos

Según DIRIS Lima Centro (2024), la DIRIS Lima Centro en su Plan Operativo Institucional Anual 2025, aprobado mediante Resolución Directoral N°1378-2024-DG-DIRIS-LC, orienta a la institución a cumplir una serie de objetivos estratégicos que reflejan su compromiso con la salud pública y la mejora continua de los servicios sanitarios, los cuales se mencionan a continuación:

- Prevenir, vigilar, controlar y reducir el impacto de las enfermedades, daños y condiciones que afectan la salud de la población, con énfasis en las prioridades nacionales.
- Garantizar el acceso a cuidados y servicios de salud de calidad organizados en redes integradas, centrados en la persona, familia y comunidad, con énfasis en la promoción de la salud y la prevención de la enfermedad.
- Velar por la eficacia, seguridad y calidad de los productos farmacéuticos, dispositivos médicos y productos sanitarios, así como la inocuidad de los alimentos y la calidad del agua.
- Fortalecer la rectoría y gobernanza del sistema de salud, y la gestión institucional, para un desempeño eficiente, ético e íntegro, en el marco de la modernización de la gestión pública.
- Mejorar la gestión y el desarrollo de los recursos humanos en salud, con énfasis en competencias y en la disponibilidad equitativa en el país.

- Mejorar la toma de decisiones, la prestación de servicios públicos, el empoderamiento y la satisfacción de la población a través del Gobierno Digital en Salud.
- Fortalecer las capacidades y la gestión de la generación, análisis, uso y transferencia del conocimiento en salud.
- Fortalecer la gestión del riesgo y la respuesta ante emergencias y desastres.

3.1.3.3. Evaluación interna y externa. FODA cuantitativo

3.1.3.3.1. Evaluación Interna

Fortalezas:

- Amplia cobertura territorial en 14 distritos de Lima Metropolitana.
- Red integrada con 8 hospitales, 5 institutos especializados, más de 60 establecimientos de primer nivel y centros de salud mental.
- Experiencia consolidada en vigilancia epidemiológica, control de brotes y campañas sanitarias.
- Personal de salud con experiencia en programas de inmunización y salud comunitaria.
- Articulación con el MINSA y alineación a los lineamientos nacionales.

Debilidades:

- Limitaciones en infraestructura en algunos establecimientos de primer nivel.
- Brechas en digitalización y registro oportuno de información en salud.
- Alta demanda poblacional que supera la capacidad instalada en algunos servicios.
- Procesos administrativos burocráticos que ralentizan la gestión.
- Rotación y déficit de personal especializado en áreas críticas.

3.1.3.3.2. Evaluación Externa:

Oportunidades:

- Prioridad nacional e internacional en fortalecer los sistemas de salud (OPS, OMS, CEPLAN).
- Acceso a financiamiento estatal y cooperación internacional en proyectos de salud pública.
- Desarrollo de plataformas digitales y gobierno digital para mejorar procesos.
- Mayor conciencia ciudadana sobre la importancia de la salud preventiva tras la pandemia.
- Alineación con políticas multisectoriales (salud, educación, inclusión social).

Amenazas:

- Limitaciones presupuestales y dependencia de transferencias del gobierno central.
- Brechas sociales y económicas en la población que afectan el acceso a los servicios.
- Resistencia o desinformación en temas de salud (ej. vacunación, alimentación saludable).
- Riesgo de emergencias sanitarias (brotes epidémicos, desastres naturales).
- Competencia con otros sectores por recursos limitados en el Estado.

3.1.3.3.3. FODA cuantitativo:

En la siguiente matriz se muestra el análisis FODA cuantitativo, donde se presentan los factores estratégicos de la DIRIS Lima Centro clasificados en fortalezas, debilidades, oportunidades y amenazas. Cada factor ha sido valorado según su peso relativo e impacto, obteniéndose un puntaje ponderado que permite establecer un panorama cuantitativo de la situación institucional.

Tabla 2

FODA Cuantitativo

Categoría	Factor	Peso	Calificación	Valor ponderado	Valor por Categoría

Fortaleza	Amplia cobertura territorial en 14 distritos de Lima Metropolitana	0.08	4	0.32	1.44
	Red integrada con hospitales, institutos especializados y centros de salud	0.08	4	0.32	
	Experiencia en vigilancia epidemiológica y campañas sanitarias	0.08	3	0.24	
	Personal de salud con experiencia en salud comunitaria	0.08	3	0.24	
	Articulación con el MINSA y alineación a lineamientos nacionales	0.08	4	0.32	
Debilidad	Limitaciones en infraestructura en establecimientos de primer nivel	0.05	2	0.1	0.5
	Brechas en digitalización y registro oportuno de información	0.05	2	0.1	
	Alta demanda poblacional que supera la capacidad instalada	0.05	2	0.1	
	Procesos administrativos burocráticos	0.05	2	0.1	
	Rotación y déficit de personal especializado	0.05	2	0.1	
Oportunidad	Prioridad nacional e internacional en fortalecer sistemas de salud	0.04	4	0.16	0.64
	Acceso a financiamiento estatal y cooperación internacional	0.04	3	0.12	
	Desarrollo de plataformas y gobierno digitales en salud	0.04	3	0.12	
	Mayor conciencia ciudadana sobre salud preventiva	0.04	3	0.12	
	Alineación con polí-ticas multisectoriales de salud pública	0.04	3	0.12	

Amenaza	Limitaciones presupuestales y dependencia del gobierno central	0.03	2	0.06	0.33
	Brechas sociales y económicas en la población	0.03	2	0.06	
	Resistencia o desinformación en temas de salud	0.03	2	0.06	
	Riesgo de emergencias sanitarias y desastres naturales	0.03	3	0.09	
	Competencia con otros sectores por recursos estatales	0.03	2	0.06	

Nota. Elaboración Propia.

Los resultados evidencian que las fortalezas (1.44) y oportunidades (0.64) superan ampliamente a las debilidades (0.50) y amenazas (0.33), lo que refleja un perfil estratégico favorable. Esto significa que la DIRIS Lima Centro cuenta con capacidades y condiciones externas que le permiten sostener y ampliar su rol en la gestión de la salud pública, aunque se requiere atender con prioridad las brechas en infraestructura, digitalización y recursos humanos para garantizar un desempeño más eficiente y equitativo.

3.2. Modelo de negocio actual (CANVAS)

En el presente análisis se empleó el modelo CANVAS para representar la forma en que la DIRIS Lima Centro organiza su gestión y cumple con su misión institucional. A diferencia de una empresa privada, donde el enfoque suele estar centrado en la rentabilidad, en este caso el modelo refleja cómo la institución crea, entrega y mantiene valor viéndolo desde una perspectiva social y sanitaria. Su propuesta de valor se centra en la búsqueda de garantizar servicios de salud integrales, gratuitos y de calidad para la población de su jurisdicción, priorizando la prevención, la cobertura de vacunación infantil y la atención a grupos vulnerables.

Propuesta de valor

- Brindar servicios de salud integrales y oportunos a la población de 14 distritos de Lima.
- Garantizar la cobertura de vacunación infantil y programas de prevención.

- Asegurar acceso gratuito y equitativo a servicios de salud públicos.
- Promover la salud comunitaria mediante campañas, vigilancia epidemiológica y atención preventiva.

Segmentos de clientes

- Población general de la jurisdicción compuesta por 14 distritos, Breña, Jesús María, La Victoria, Lima, Lince, Magdalena del Mar, Miraflores, Pueblo Libre, San Borja, San Isidro, San Juan de Lurigancho, San Luis, San Miguel y Surquillo, en todas sus etapas de curso de vida: niños, adolescentes, adultos y adultos mayores.
- Grupos vulnerables (niños menores de 1 año, gestantes, adultos mayores, personas con enfermedades crónicas).
- Establecimientos de salud que forman parte de la red.

Canales

- Establecimientos de salud (hospitales, centros de salud, puestos de salud, centros de salud mental).
- Campañas itinerantes en comunidades y espacios públicos.
- Plataformas digitales (SIS, aplicativos internos de registro, portales web del MINSA y DIRIS).
- Medios de comunicación y redes sociales institucionales.

Relación con los clientes

- Atención directa en establecimientos de salud.
- Promoción de la salud comunitaria (charlas, ferias, visitas domiciliarias).
- Comunicación institucional para informar sobre campañas y servicios.
- Enfoque de cercanía, equidad e inclusión.

Actividades clave

- Prestación de servicios médicos preventivos, promocionales y asistenciales.

- Gestión de programas de inmunización infantil y salud pública.
- Vigilancia epidemiológica y control de enfermedades.
- Gestión de emergencias y desastres en salud.
- Digitalización y análisis de datos estadísticos de salud.

Recursos clave

- Personal de salud (médicos, enfermeras, técnicos, promotores de salud).
- Infraestructura sanitaria (hospitales, centros y puestos de salud).
- Equipamiento médico y cadena de frío para vacunas.
- Sistemas de información en salud.
- Financiamiento estatal (recursos del MINSA y cooperación internacional).

Socios clave

- Ministerio de Salud (MINSA).
- Gobiernos locales y regionales.
- Organismos internacionales (OPS, OMS).
- Sociedad civil y organizaciones comunitarias.
- Universidades e institutos de investigación.

Estructura de costos

- Operación de establecimientos de salud.
- Adquisición y distribución de medicamentos e insumos (especialmente vacunas).
- Remuneraciones del personal de salud y administrativo.
- Inversión en infraestructura, equipamiento y tecnología.
- Campañas de comunicación y promoción de la salud.

Fuentes de ingreso / financiamiento

- Presupuesto público asignado por el Estado peruano vía MINSA.
- Recursos ordinarios y transferencias presupuestales.
- Cooperación internacional y donaciones en programas específicos.

Figura 12

Modelo CANVAS de DIRIS Lima Centro

Nota. Elaboración propia.

El modelo CANVAS de la DIRIS Lima Centro evidencia que la institución cuenta con una estructura sólida de socios estratégicos, recursos clave y actividades que le permiten sostener su propuesta de valor en beneficio de la población. La articulación con el MINSA, los gobiernos locales y los organismos internacionales, sumada a la red de establecimientos de salud y al personal especializado, garantizan la capacidad de respuesta frente a las necesidades sanitarias. Aunque su gestión no persigue fines de lucro, el análisis muestra que la DIRIS Lima Centro genera un alto valor social, centrado en la prevención, la equidad y la protección de la salud de los ciudadanos.

3.3. Mapa de procesos actual

El Mapa de Procesos de la DIRIS Lima Centro, aprobado mediante la Resolución Directoral N.° 141-2020-DG-DIRIS-LC, representa de manera estructurada las actividades clave que aseguran una gestión sanitaria eficiente, orientada a la satisfacción de los usuarios y al logro de los objetivos institucionales (DIRIS Lima Centro, 2020). En concordancia con la gestión por procesos en el sector público, el mapa se organiza en tres categorías: procesos estratégicos, misionales u operativos y de soporte (PCM, 2025; Poder Judicial del Perú, s. f.).

3.3.1. Procesos Estratégicos

Los procesos estratégicos establecen la dirección y las políticas institucionales que orientan el funcionamiento. Estos procesos aseguran la planificación, control, calidad y gestión del conocimiento, permitiendo el cumplimiento de los objetivos a mediano y largo plazo.

- **PE.0.1 Gestión de la Planificación y Desarrollo Institucional:** este proceso se encarga de formular, implementar y evaluar los planes estratégicos y operativos de la institución, asegurando la alineación con las políticas nacionales de salud y la mejora de la gestión organizacional.
- **PE.0.2 Gestión de la Inteligencia Sanitaria:** se orienta a la recolección, análisis y uso de la información epidemiológica y sanitaria para la toma de decisiones basadas en evidencia. Facilita la identificación de problemas prioritarios de salud y la formulación de intervenciones efectivas.
- **PE.0.3 Gestión del Control Interno:** su propósito es garantizar el uso adecuado de los recursos institucionales, promoviendo la transparencia, el cumplimiento normativo y la evaluación de riesgos operacionales y administrativos.

- **PE.0.4 Gestión de la Calidad:** este proceso busca la mejora continua de los servicios y procesos institucionales, estableciendo estándares de calidad, auditorías internas y acciones correctivas que aseguren la satisfacción del usuario.

3.3.2. Procesos Misionales

Los procesos misionales constituyen el núcleo operativo. Son aquellos directamente vinculados con la gestión sanitaria, la vigilancia, la prevención y la atención integral de la salud de la población.

- **PM.0.1 Supervisión y Monitoreo a la Gestión Sanitaria:** encargado de supervisar y evaluar el desempeño de las redes y establecimientos de salud, asegurando el cumplimiento de los estándares de calidad y la eficiencia en la prestación de los servicios.
- **PM.0.2 Administración del Otorgamiento de Derechos en Salud:** gestiona el acceso y reconocimiento de los derechos en salud de los ciudadanos, garantizando que los servicios sean brindados con equidad, oportunidad y sin discriminación.
- **PM.0.3 Vigilancia y Prevención de Riesgos en Salud:** comprende la vigilancia epidemiológica, la prevención y control de enfermedades, y la gestión de emergencias sanitarias, contribuyendo a la reducción de riesgos y daños a la salud pública.
- **PM.0.4 Gestión Territorial en Salud:** se enfoca en la articulación de acciones intersectoriales e interinstitucionales a nivel territorial, promoviendo la planificación local de la salud según las características y necesidades de cada comunidad.
- **PM.0.5 Atención Integral de Salud:** dirigido a garantizar una atención continua y centrada en la persona, familia y comunidad, integrando la promoción, prevención, recuperación y rehabilitación de la salud en los diferentes niveles de atención.

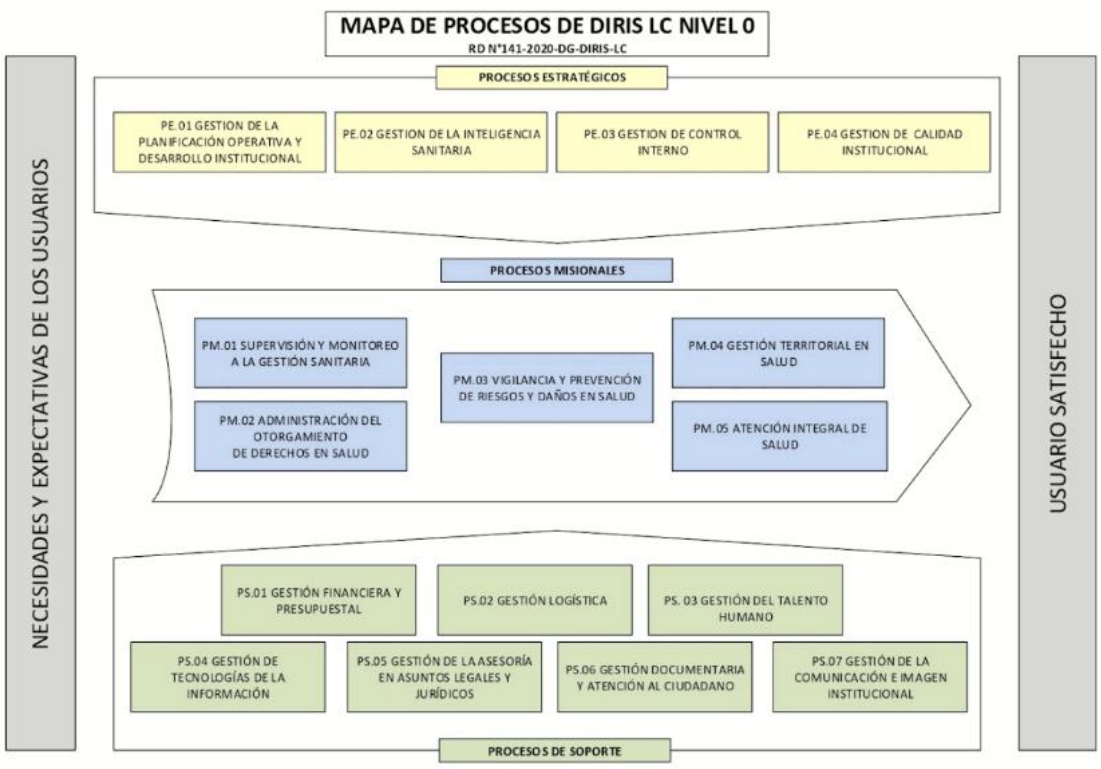
3.3.3. Procesos de Soporte

Los procesos de soporte proporcionan el respaldo administrativo, técnico y logístico necesario para el adecuado funcionamiento de los procesos estratégicos y misionales. cuenta con siete procesos de soporte:

- **PS.0.1 Gestión Financiera y Presupuestal:** administra los recursos financieros y presupuestales asegurando su uso eficiente, transparente y alineado a los objetivos institucionales.
- **PS.0.2 Gestión Logística:** responsable de la adquisición, almacenamiento y distribución de bienes y servicios requeridos por las áreas operativas y administrativas, garantizando la disponibilidad oportuna de recursos.
- **PS.0.3 Gestión del Talento Humano:** comprende la planificación, selección, capacitación y bienestar del personal, promoviendo un clima laboral favorable y el desarrollo de capacidades institucionales.
- **PS.0.4 Gestión de Tecnologías de la Información:** se encarga de administrar los sistemas informáticos, redes y plataformas digitales institucionales, asegurando la seguridad, disponibilidad y eficiencia en el uso de la información.
- **PS.0.5 Gestión de la Asesoría en Asuntos Legales y Jurídicos:** brinda soporte legal a todas las unidades de la institución, garantizando que las acciones administrativas y contractuales se realicen conforme a la normativa vigente.
- **PS.0.6 Gestión Documentaria y Atención al Ciudadano:** asegura la adecuada tramite de sus documentos institucionales y la atención oportuna a los requerimientos de los usuarios internos y externos.
- **PS.0.7 Gestión de la Comunicación e Imagen Institucional:** promueve la difusión de información institucional, la gestión de la imagen pública y la comunicación efectiva con la ciudadanía, fortaleciendo la transparencia y confianza en la institución.

Figura 13

Mapa de Procesos de la DIRIS Lima Centro



Nota. Adaptado de *Resolución Directoral N.° 141-2020-DG-DIRIS-LC: Aprueba el Mapa de Procesos y las Fichas Técnicas Nivel 0 de la DIRIS Lima Centro* [Norma legal], por Dirección de Redes Integradas de Salud Lima Centro, Gob.pe, 2020.

Capítulo 4: Metodología de la Investigación

4.1. Diseño de la Investigación

4.1.1. Diseño

Según Hernández Sampieri et al. (2010), un diseño experimental se caracteriza por la manipulación intencionada de una o más variables independientes con el propósito de observar y analizar los efectos que estas producen sobre las variables dependientes dentro de un contexto controlado. Este tipo de diseño resulta pertinente cuando se busca determinar el impacto de una causa tras haber sido objeto de intervención.

Este estudio adoptará un diseño experimental, dado que posibilita un control preciso de las variables, tanto independientes como dependientes. Mediante la realización de diversos experimentos, se obtendrá un conjunto amplio de resultados que serán evaluados y contrastados, con el propósito de determinar la alternativa más eficiente y, en consecuencia, seleccionar el modelo de clasificación más apropiado.

4.1.2. Tipo

Según Hernández Sampieri et al. (2010), una investigación aplicada tiene como finalidad central atender un problema específico y dar respuesta a interrogantes puntuales. Este tipo de enfoque se caracteriza por priorizar la búsqueda de soluciones prácticas frente a las problemáticas planteadas.

De esta manera, el presente estudio es una investigación aplicada, pues persigue la construcción de un modelo de clasificación del estado de vacunación infantil en el primer año de vida en la jurisdicción de una Dirección de Redes Integradas de Salud de Lima Metropolitana utilizando técnica de Machine Learning. Al abordar un problema social específico, como es la cobertura de vacunación infantil, y plantear una alternativa de solución basada en herramientas tecnológicas modernas, este estudio no se limita únicamente a describir el fenómeno, sino que pretende aportar un recurso práctico que facilite su análisis y gestión.

4.1.3. Enfoque

El enfoque cuantitativo se distingue por el uso de información numérica, la aplicación de pruebas de hipótesis y un desarrollo de carácter secuencial, lo que significa que cada fase

debe cumplirse rigurosamente y en el orden previsto, sin omitir ninguna etapa (Hernández & Mendoza, 2018).

El presente estudio adopta un enfoque cuantitativo, al basarse en el análisis de datos numéricos. Para ello, las variables categóricas serán convertidas en valores numéricos a través de técnicas de codificación, lo que posibilitará su correcto tratamiento. Posteriormente, se llevará a cabo el entrenamiento del modelo mediante métodos de Machine Learning, asegurando un proceso analítico riguroso y estructurado en cada una de sus fases.

4.1.4. Población y Muestra

Según Hernández Sampieri et al. (2010), la población se entiende como el total de elementos o casos que poseen determinadas características o cumplen con criterios específicos establecidos para una investigación.

La población está compuesta por 185,220 registros de estado de vacunación, los cuales fueron recopilados a nivel de Dirección de Redes Integrados de Salud de Lima Metropolitana a través del sistema Padrón nominal y HIS MINSA.

Según Hernández Sampieri et al. (2010) el muestreo probabilístico consiste en elegir de forma aleatoria a los elementos de una población, garantizando que cada uno tenga una posibilidad conocida de ser incluido. Este método se aplica con el fin de obtener una muestra que refleja fielmente las características de la población.

Se selecciono una muestra 122,142 registros mediante muestreo aleatorio, tras un proceso de limpieza de datos que redujo de la base original, garantizando la calidad y representabilidad de la información para el entrenamiento del modelo de clasificación.

4.1.5. Operacionalización de Variables

Tabla 3

Operacionalización de variables

Variable	Indicador	Fórmula
Técnicas de Machine Learning Según (Mitchell,1997) se analiza el desarrollo de sistemas informáticos capaces de perfeccionarse de forma	Accuracy	$\frac{TP + TN}{TP + FP + FN + TN}$
	Recall	$\frac{TP}{TP + FN}$

autónoma mediante procesos de aprendizaje.	Precision	$\frac{TP}{TP + FP}$
	F1-score	$2 \cdot \left(\frac{Precision \cdot Recall}{precision + Recall}\right)$
Estado de Vacunación Situación en la que se encuentra un niño en relación con el cumplimiento de las dosis establecidas en el calendario nacional de inmunización para su edad, diferenciando entre esquema completo, incompleto o no iniciado. Este estado refleja si las vacunas han sido aplicadas en la cantidad y tiempo recomendados por el programa de salud. Según la Organización Mundial de la Salud (OMS, 2023)	Porcentaje de nivel de estado de vacunación infantil	$\left(\frac{Registro\ EsquemaCompleto}{Total\ de\ registros}\right) \cdot 100$ $\left(\frac{Registro\ Esquema\ No\ Completo}{Total\ de\ registro}\right) \cdot 100$

Nota. Elaboración propia.

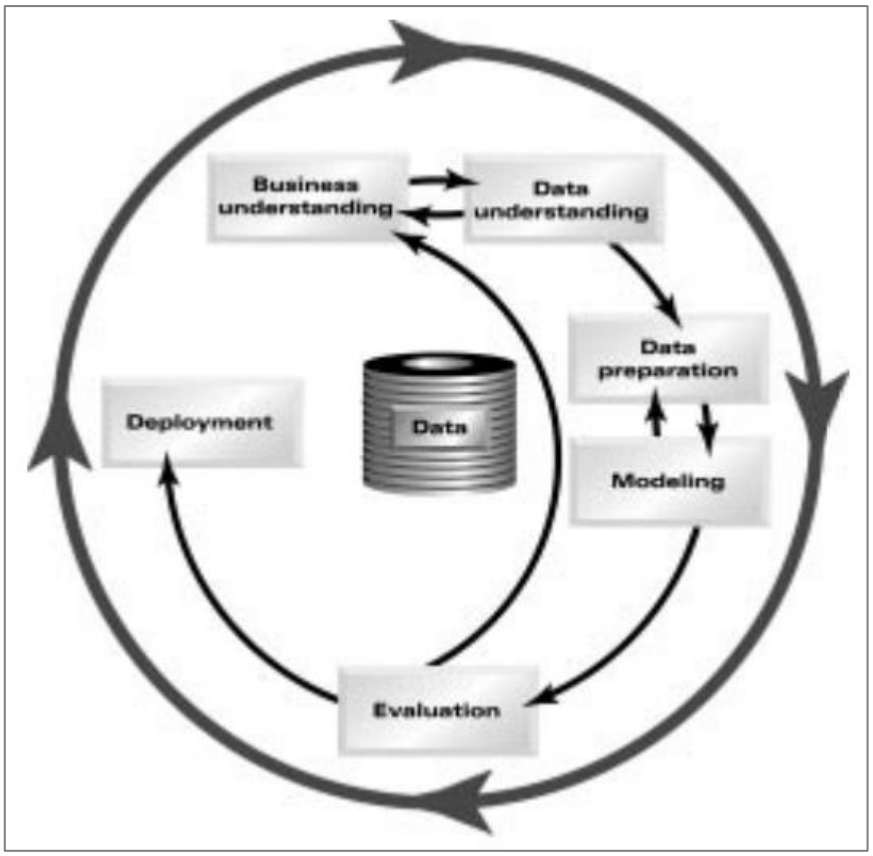
4.2. Metodología de Implementación de la Solución

La metodología que se utilizará en el presente trabajo se basa en la metodología CRISP-DM (*Cross-Industry Standard Process for Data Mining*), la cual consiste en un estándar industrial y académico ampliamente adoptado, diseñado para guiar proyectos de minería de datos y *machine learning* de forma estructurada e iterativa.

Según Chapman et al. (2000), la metodología CRISP-DM comprende seis etapas: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación y despliegue, tal como se muestra en la Figura 14.

Figura 14

Etapas del modelo de referencia CRISP-DM



Nota. Adaptado de CRISP-DM 1.0: Step-by-step data mining guide (p. 13), por Chapman, P. et al, 2000, SPSS.

A continuación, se describe brevemente, según Chapman et al. (2020), cada una de las etapas que conforman la metodología CRISP-DM y cómo se vincula con las fases de la metodología que se utilizará en el presente trabajo de investigación:

- **Entendimiento del negocio**: esta etapa inicial se centra en comprender los objetivos y los requisitos del proyecto desde una perspectiva de negocio, para luego convertir este conocimiento en una definición del problema de minería de datos y en un plan preliminar diseñado para alcanzar dichos objetivos. En el presente trabajo de investigación, esta etapa se ha abordado en el Capítulo 1, denominado “Planteamiento del Problema”, donde se detalló la problemática asociada al estado de vacunación infantil en el primer año de vida; y en el Capítulo 3, denominado “Entorno Empresarial”, en el cual se profundizó en el contexto institucional y social en el cual se enmarca el presente estudio. Por tal motivo, esta etapa se omite en este Capítulo 4, denominado “Metodología de la Investigación” y en el Capítulo 5, denominado “Desarrollo de la Solución”.

- **Entendimiento de los datos**: comienza con la recolección de los datos y continúa con una serie de actividades orientadas a familiarizarse con los datos, identificar posibles problemas de calidad, descubrir primeros hallazgos o patrones relevantes, así como detectar subconjuntos de interés que permitan formular hipótesis sobre información oculta. En el presente trabajo de investigación, esta etapa se abordará en la Fase 1, denominada "Obtención de los Datos".

- **Preparación de los datos**: esta etapa comprende todas las actividades necesarias para construir el *dataset* final, que será utilizado en la etapa de modelado, a partir de los datos originales. Las tareas de preparación de datos suelen realizarse en múltiples iteraciones y no siguen un orden estrictamente definido. Estas tareas incluyen la selección de tablas, registros y atributos, así como la transformación y limpieza de los datos para su uso en las herramientas de modelado. En el presente trabajo de investigación, esta etapa se abordará en la Fase 2, denominada "Preprocesamiento", y en la Fase 3, denominada "Selección de Características".

- **Modelado**: esta etapa consiste en la selección y aplicación de diversas técnicas de modelado, ajustando sus parámetros con el fin de optimizar su rendimiento. Por lo general, existen múltiples técnicas disponibles para abordar un mismo tipo de problema de machine learning, cada una con requisitos específicos en cuanto al formato o la estructura de los datos. Por ello, es común que con frecuencia sea necesario regresar a la etapa de preparación de datos. En el presente trabajo de investigación, esta etapa se abordará en la Fase 4, denominada "Machine Learning".

- **Evaluación**: en esta etapa se dispone de uno o varios modelos que muestran un alto nivel de calidad desde la perspectiva del análisis de datos. En este punto, es fundamental realizar una evaluación más exhaustiva y revisar los pasos ejecutados durante su construcción, con el fin de asegurarse de que cumpla adecuadamente con los objetivos del negocio. En el presente trabajo de investigación, esta etapa se abordará en la Fase 5, denominada "Evaluación".

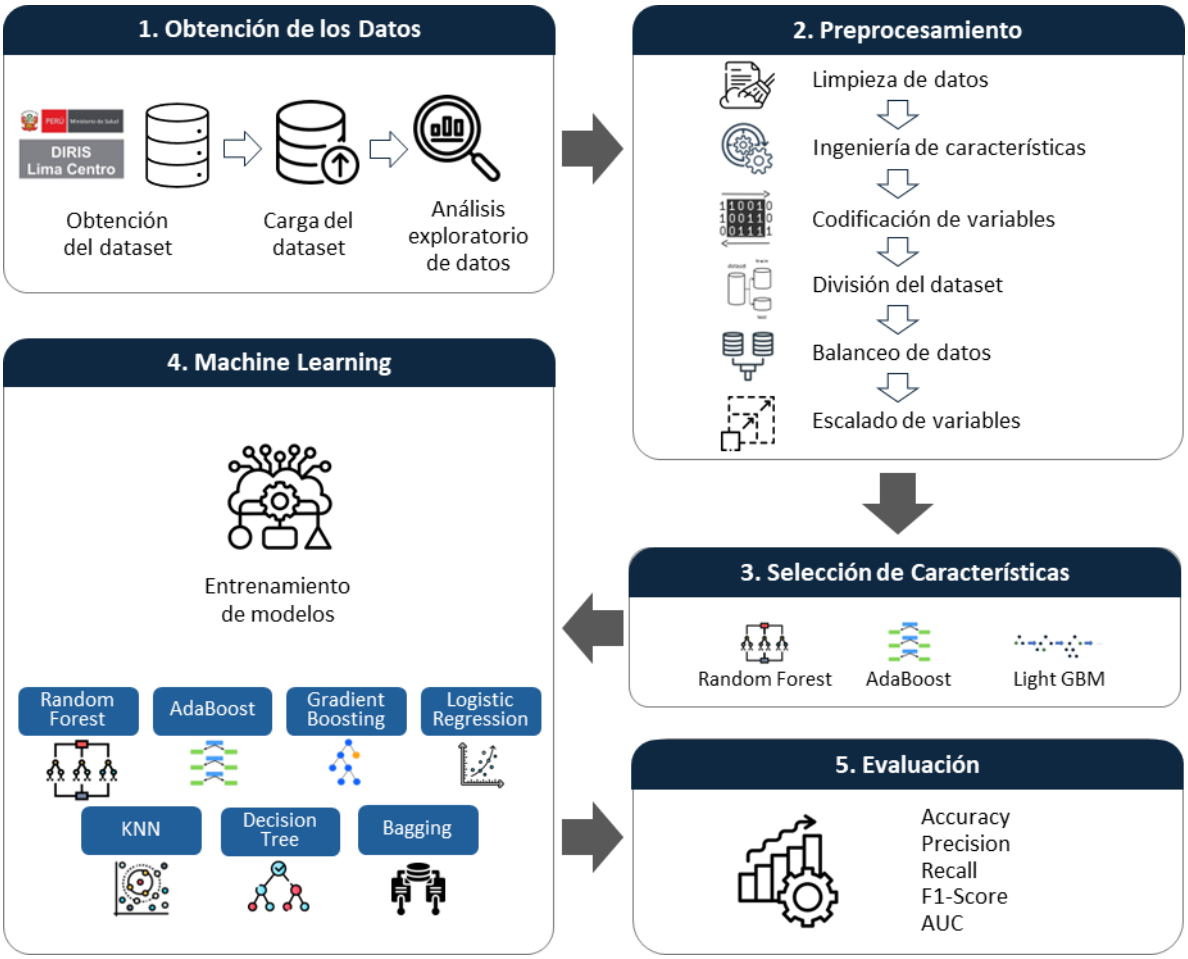
- **Despliegue**: esta etapa implica la aplicación de modelos "en vivo" dentro de los procesos de toma de decisiones de una organización. Sin embargo, dado que el

alcance del presente trabajo de investigación se limita al diseño, construcción y validación de modelos de clasificación. Esta etapa se considera fuera del alcance del estudio.

En consecuencia, la metodología que se utilizará para la presente investigación se muestra en la Figura 15.

Figura 15

Metodología de la investigación



Nota. Elaboración propia.

4.2.1. Obtención de los datos

Tabla 4

Actividades de la fase de obtención de los datos

Actividades	Descripción	Tareas

1. Obtención de dataset	Obtención de los datos que servirán como base del análisis	Obtención de dataset desde fuente oficial
2. Carga de dataset	Importación y verificación inicial del dataset para garantizar su correcta lectura y estructura.	Importar dataset
		Verificar cantidad de registros y columnas
3. Análisis exploratorio de datos (EDA)	Exploración del dataset orientada a comprender su estructura y patrones generales antes del preprocesamiento.	Analizar estadísticas descriptivas

Nota. Elaboración propia.

4.2.1.1. Actividad 1: Obtención del dataset

Esta actividad consistió en la recopilación de los datos que servirán como base para el análisis y posterior modelado del estado de vacunación infantil en menores de 1 año.

El propósito principal fue garantizar que las fuentes de información fueran oficiales, actualizadas y representativas de la población objetivo. Para ello, se gestionó formalmente el acceso a las bases institucionales administradas por la Oficina de Estadística e Informática de la DIRIS Lima Centro, las cuales integran información proveniente de los sistemas HIS-MINSA y Padrón Nominal (PN).

La solicitud se realizó mediante un oficio institucional (Anexo N.º 01), autorizando el uso académico de los datos. Los archivos fueron entregados en formato CSV, totalmente anonimizados conforme a la Ley N.º 29733 – Ley de Protección de Datos Personales y su reglamento (D.S. N.º 003-2013-JUS).

El dataset consolidado comprendió 185,220 registros de niños menores de cinco años atendidos entre septiembre de 2019 y setiembre de 2025, conteniendo información relevante y necesaria para el presente estudio.

4.2.1.2. Actividad 2: Carga del dataset

Una vez obtenidos los datos, se procederá con su importación en el entorno de análisis y desarrollo de Google Colab (Python). Esta actividad tiene como finalidad asegurar que los

datos sean correctamente leídos, estructurados y compatibles con las herramientas de procesamiento.

Luego de la carga de datos, se verificará la cantidad de registros y columnas en relación con el formato esperado.

4.2.1.3. Actividad 3: Análisis Exploratorio de Datos (EDA)

En esta actividad se realizará un análisis exploratorio de los datos (EDA) con el objetivo de comprender la estructura del dataset, identificar patrones generales y reconocer posibles anomalías antes del preprocesamiento.

4.2.2. Preprocesamiento

Tabla 5

Actividades de la fase de preprocesamiento

Actividades	Descripción	Tareas
1. Limpieza de datos	Depuración del dataset para eliminar errores, valores faltantes y columnas irrelevantes, asegurando que la información esté lista para el análisis posterior	Eliminar columnas innecesarias
		Imputar datos faltantes
		Eliminar valores nulos y duplicados
		Limpiar y normalizar texto
2. Ingeniería de características	Creación y/o modificación de variables relevantes que mejoren la capacidad predictiva del modelo	Eliminar registros fuera del alcance del estudio
		Generar variable objetivo “target”
3. Codificación de variables	Transformación de variables categóricas en valores numéricos para que puedan ser interpretadas correctamente por los algoritmos de aprendizaje automático	Aplicar codificación a variables numéricas ordinales
		Aplicar codificación a variables categóricas
4. División del dataset	Separación del conjunto de datos en subconjuntos de *train* y *test* con el fin de evaluar el desempeño de los modelos a utilizar	Dividir el dataset con train_test_split con una proporción 80-20

5. Balanceo de datos	Manejar el desbalance en las clases de la variable objetivo 'target' mediante técnicas de sobremuestreo o submuestreo, para mejorar la capacidad general del modelo	Implementar el balanceo de datos
		Validar que el balance de clases sea adecuado
6. Escalado de variables	Ajuste de variables numéricas a una escala uniforme con el fin de optimizar el rendimiento de los modelos de machine learning	Normalizar las variables numéricas
		Validar la distribución de valores escalados

Nota. Elaboración propia.

4.2.2.1. Actividad 1: Limpieza de datos

En esta actividad se realizará la depuración del dataset con el propósito de garantizar la calidad, consistencia e integridad de los datos.

Durante esta fase se llevarán a cabo las siguientes acciones:

- Eliminar columnas irrelevantes que no aporten información significativa al modelo.
- Imputar datos faltantes mediante estrategias racionales asociadas a la vacunación infantil.
- Eliminar valores nulos y duplicados que puedan distorsionar las métricas o el entrenamiento.

Limpiar y normalizar texto, corrigiendo inconsistencias en espacios en blanco innecesarios, mayúsculas, tildes y caracteres especiales.

4.2.2.2. Actividad 2: Ingeniería de características

Durante esta etapa se realizarán las siguientes tareas:

- Generar la variable objetivo "target", que representará el estado del esquema de vacunación infantil (1: incompleto o 0: completo), a partir del estado de las variables asociadas a las vacunas de acuerdo con el esquema de vacunación infantil al primer año de vida.

4.2.2.3. Actividad 3: Codificación de variables

En esta actividad, se realizará la conversión de variables categóricas a valores numéricos, con el objetivo de que puedan ser interpretados de manera correcta por los algoritmos de *machine learning*.

Se aplicarán técnicas de codificación acordes al tipo de variable:

- Aplicar *OrdinalEncoder* a variables categóricas con un orden lógico (por ejemplo, el nivel educativo de la madre).
- Aplicar *LabelEncoder* a variables categóricas sin jerarquía (por ejemplo, distrito de residencia o tipo de establecimiento).
- Verificar la coherencia de los valores codificados.

4.2.2.4. Actividad 4: División del dataset

En esta actividad se realizará la división del dataset en dos subconjuntos: uno de entrenamiento o *train* (80%) y otro de prueba o *test* (20%), con la finalidad de evaluar de forma objetiva el desempeño de los modelos seleccionados.

La separación se efectuará utilizando la función *train_test_split* de la librería *scikit-learn*.

4.2.2.5. Actividad 5: Balanceo de datos

En esta actividad se abordará el desbalance de clases presente en la variable objetivo "target". Para corregir este problema se aplicará una técnica de sobremuestreo, específicamente el SMOTE (*Synthetic Minority Over-sampling Technique*), a fin de generar nuevas instancias sintéticas de la clase minoritaria hasta alcanzar una distribución equilibrada. Las tareas específicas serán:

- Implementar SMOTE sobre el dataset.
- Verificar que el balance de clases sea el adecuado.

4.2.2.6. Actividad 5: Escalado de variables

Finalmente, se realizará el escalado de las variables numéricas para que todas las características se encuentren en una misma escala de magnitud. Para ello se utilizará el método

StandardScaler, el cual transforma los valores restando la media y dividiendo entre la desviación estándar.

Las acciones específicas incluirán:

- Aplicar el escalado a las variables numéricas seleccionadas.
- Verificar la media y desviación estándar de las variables transformadas para confirmar la normalización.

Este ajuste mejorará la estabilidad numérica y la eficiencia de los algoritmos de clasificación que se aplicarán en la fase 4 de Machine Learning.

4.2.3. Selección de Características

Tabla 6

Actividades de la fase de selección de características

Actividades	Descripción	Tareas
1. Aplicación de técnicas de selección de características	Identificación de las variables más relevantes del dataset que aportan mayor valor predictivo al modelo, reduciendo la dimensionalidad y mejorando la eficiencia del entrenamiento	Aplicar técnica RandomForest
		Aplicar técnica AdaBoost
		Aplicar técnica LightGBM

Nota. Elaboración propia.

4.2.3.1. Actividad 1: Aplicación de técnicas de selección de características

Durante esta actividad se implementarán técnicas de selección de características con el fin de identificar las variables más relevantes que aporten mayor valor predictivo al modelo. Este proceso permitirá reducir la dimensionalidad del dataset, optimizar el tiempo de entrenamiento, disminuir el riesgo de sobreajuste (*overfitting*) y mejorar la interpretabilidad del modelo.

Para ello, se aplicarán tres técnicas basadas en modelos de *ensemble learning* que evalúan la importancia de las variables mediante el peso asignado a cada una durante el entrenamiento:

- *Random Forest*: se utilizará para determinar la relevancia de las variables en función de la reducción de la impureza y la ganancia en precisión obtenida por cada árbol del conjunto.
- *AdaBoost*: se empleará para evaluar la influencia de cada característica a partir de los pesos ajustados en cada iteración del algoritmo.
- *LightGBM*: se aplicará para calcular la importancia de las variables según la ganancia promedio generada por cada división en los árboles.

Una vez obtenidos los resultados de cada técnica, se determinará las variables que aporta mayor valor significativo a cada uno de acuerdo con los rankings de importancia de característica. Esto servirá como base para la fase siguiente de entrenamiento y evaluación de los modelos de clasificación.

4.2.4. Machine Learning

Tabla 7

Actividades de la fase de machine learning

Actividades	Descripción	Tareas
1. Entrenamiento de modelos sin ajuste de hiperparámetros	Entrenamiento de los modelos de machine learning seleccionados utilizando los distintos conjuntos de datos generados a partir de las técnicas de selección de características	Entrenar los modelos LogisticRegression, RandomForest, XGBoost, KNN, AdaBoost, DecisionTree y Bagging
2. Entrenamiento con ajuste de hiperparámetros	Entrenamiento y optimización de los siete modelos, utilizando la búsqueda automática de parámetros óptimos.	Entrenar los modelos LogisticRegression, RandomForest, XGBoost, KNN, AdaBoost, DecisionTree y Bagging con ajuste de hiperparámetros.

Nota. Elaboración propia.

4.2.4.1. Actividad 1: Entrenamiento de modelos sin ajuste de hiperparámetros

En esta actividad se realizará el entrenamiento de los modelos de machine learning seleccionados, empleando los distintos conjuntos de datos resultantes de las técnicas de selección de características aplicadas en la fase previa.

Durante esta etapa, se entrenarán los siguientes modelos de clasificación: *Logistic Regression, Random Forest, XGBoost, K-Nearest Neighbors (KNN), AdaBoost, Decision Tree y Bagging Classifier en combinación* con las tres técnicas de selección de características (*LightGBM_FS, RandomForest_FS y AdaBoost_FS*).

Cada modelo se ajustará utilizando el *dataset* de *train* y se evaluará con el *dataset* de *test* (80-20), asegurando que las métricas obtenidas reflejen el comportamiento general del modelo ante nuevos datos.

4.2.4.2. Actividad 2: Entrenamiento con ajuste de hiperparámetros

En esta actividad se llevará a cabo el entrenamiento de los modelos de machine learning aplicando un proceso de ajuste de hiperparámetros con el propósito de optimizar su rendimiento. Este procedimiento consiste en explorar de manera automática distintas combinaciones de parámetros que permitan mejorar la capacidad predictiva de cada modelo.

Durante esta etapa, se emplearán los mismos siete modelos de clasificación utilizados previamente (Logistic Regression, Random Forest, XGBoost, K-Nearest Neighbors (KNN), AdaBoost, Decision Tree y Bagging Classifier), combinados con las tres técnicas de selección de características anteriormente descritas.

El ajuste se realizará mediante una técnica de optimización bayesiana, la cual evalúa de forma iterativa diferentes configuraciones de parámetros para maximizar las métricas de desempeño (*Accuracy, Precision, Recall, F1-score y AUC*). Una vez completado este proceso, los resultados obtenidos permitirán identificar las combinaciones de modelos y parámetros con mejor comportamiento general.

4.2.5. Evaluación

Tabla 8

Actividades de la fase de evaluación

Actividades	Descripción	Tareas

1. Evaluación de resultados	Análisis del desempeño de los modelos entrenados, comparando sus métricas y visualizando los resultados obtenidos para seleccionar el modelo óptimo	Analizar las métricas de desempeño accuracy, precisión, recall, f1-score, ROC-AUC obtenidas por cada modelo
		Graficar la matriz de confusión y la curva ROC del modelo con el mejor desempeño
2. Validación cruzada del modelo final	Aplicación de la validación cruzada estratificada al modelo con mejor desempeño para evaluar su estabilidad y capacidad de generalización.	Calcular métricas promedio y desviaciones estándar de Accuracy, Precision, Recall, F1-score y ROC-AUC mediante 5 pliegues (k=5).

Nota. Elaboración propia.

4.2.5.1. Actividad 1: Evaluación de resultados

En esta actividad se evaluará el desempeño de los modelos de *machine learning* entrenados en la fase anterior, con el objetivo de comparar su rendimiento y seleccionar el modelo más adecuado para el problema de clasificación asociado al presente estudio.

La evaluación se realizará utilizando el *dataset* de *test* (20%), que no habrá sido empleado durante el entrenamiento, garantizando así una medición imparcial.

Durante esta fase se analizarán las métricas de desempeño más representativas, entre ellas:

- *Accuracy*: proporción de predicciones correctas sobre el total de casos.
- *Precision*: porcentaje de verdaderos positivos entre todas las predicciones positivas, indicando la exactitud del modelo.
- *Recall*: proporción de verdaderos positivos detectados entre todos los casos reales positivos.
- *F1-score*: media armónica entre precisión y *recall*, que permite equilibrar ambas métricas y reflejar el rendimiento global.
- ROC-AUC: área bajo la curva ROC, que mide la capacidad del modelo para distinguir entre clases.

Asimismo, se generarán gráficas que facilitarán la interpretación de los resultados, tales como: matriz de confusión, para observar la distribución de aciertos y errores del modelo en cada clase, así como curva ROC para analizar gráficamente la sensibilidad y especificidad del modelo y comparar su rendimiento con los demás algoritmos evaluados.

Finalmente, se compararán los valores de métricas obtenidos entre todos los modelos y se seleccionará el modelo óptimo considerando tanto su desempeño cuantitativo (mayor F1-score y ROC-AUC) como su estabilidad y capacidad de generalización.

4.2.5.2. Actividad 2: Validación cruzada del modelo final

Con el propósito de evaluar la estabilidad y capacidad de generalización del modelo con mejor desempeño, se aplicará la técnica de validación cruzada estratificada (Stratified K-Fold). Este procedimiento divide el conjunto de datos en k particiones o pliegues, de modo que el modelo se entrena en k−1 subconjuntos y se valida en el restante, repitiendo el proceso k veces hasta que todos los pliegues (*folds*) hayan sido utilizados para validación.

En este estudio se utilizará una configuración de 5 pliegues (k=5), lo que permitirá obtener una estimación más robusta del rendimiento real del modelo y medir la variabilidad entre las ejecuciones. Asimismo, se calcularán las métricas promedio y desviaciones estándar de *Accuracy, Precision, Recall, F1-score y ROC-AUC*, para demostrar que el modelo presenta un desempeño consistente y estable en diferentes subconjuntos de datos, lo que valida su capacidad predictiva para la clasificación del estado de vacunación infantil.

4.3. Metodología para la Medición de Resultados de la Implementación

4.3.1. Matriz de Confusión

Según Harrington (2012), la medición de resultados es un componente fundamental en la evaluación de modelos de clasificación, ya que permite determinar el grado de acierto y la capacidad predictiva de los algoritmos utilizados. Una de las herramientas más empleadas para este propósito es la matriz de confusión, la cual proporciona una visión estructurada de las predicciones del modelo frente a los valores reales de las clases.

Figura 16

Matriz de confusión de dos clases

		redicted	
		+1	-1
Actual	+1	True Positive (TP)	False Negative (FN)
	-1	False Positive (FP)	True Negative (TN)

Nota. Adaptado de *Machine Learning in Action* (p. 144), por Harrington, 2012, Manning.

En el contexto de los modelos de clasificación, un Verdadero Positivo (TP) ocurre cuando el sistema identifica correctamente una instancia perteneciente a la clase positiva. De manera similar, un Verdadero Negativo (TN) se presenta cuando una instancia negativa es clasificada adecuadamente como tal. Por el contrario, un Falso Positivo (FP) sucede cuando una instancia negativa es erróneamente clasificada como positiva, mientras que un Falso Negativo (FN) corresponde al caso en que una instancia positiva es clasificada incorrectamente como negativa.

4.3.2. Accuracy

Se define como la proporción de aciertos obtenidos por el modelo respecto al total de predicciones efectuadas. En otras palabras, indica con qué frecuencia el modelo clasifica correctamente las instancias, proporcionando una medida general de su desempeño.

$$\frac{TP + TP}{TP + FP + FN + TN} \tag{9}$$

En esta fórmula, TP corresponde a la cantidad de verdaderos positivos, TN representa el número de verdaderos negativos, FP indica la cantidad de falsos positivos, y FN hace referencia al número de falsos negativos.

4.3.3. Precision

Hace referencia a la proporción de casos que el modelo identificó correctamente como positivos respecto al total de instancias que predijo como positivas, indicando así el grado de exactitud en las predicciones positivas.

$$\frac{TP}{TP + FP} \tag{10}$$

4.3.4. Recall

También denominado sensibilidad o tasa de verdaderos positivos, representa la proporción de casos positivos que el modelo logra identificar correctamente, mostrando así su capacidad para detectar instancias de la clase positiva.

$$\frac{TP}{TP + FN} \tag{11}$$

4.3.5. F1-score

Corresponde a la media armónica entre la precisión y el *recall*, lo que permite obtener un único indicador que equilibra el rendimiento entre ambas métricas y ofrece una evaluación más completa del modelo, especialmente en contextos con clases desbalanceadas.

$$2 \cdot \left(\frac{\mathrm{Pr}\,ecision \cdot recall}{\mathrm{Pr}\,ecision + Recall}\right) \tag{12}$$

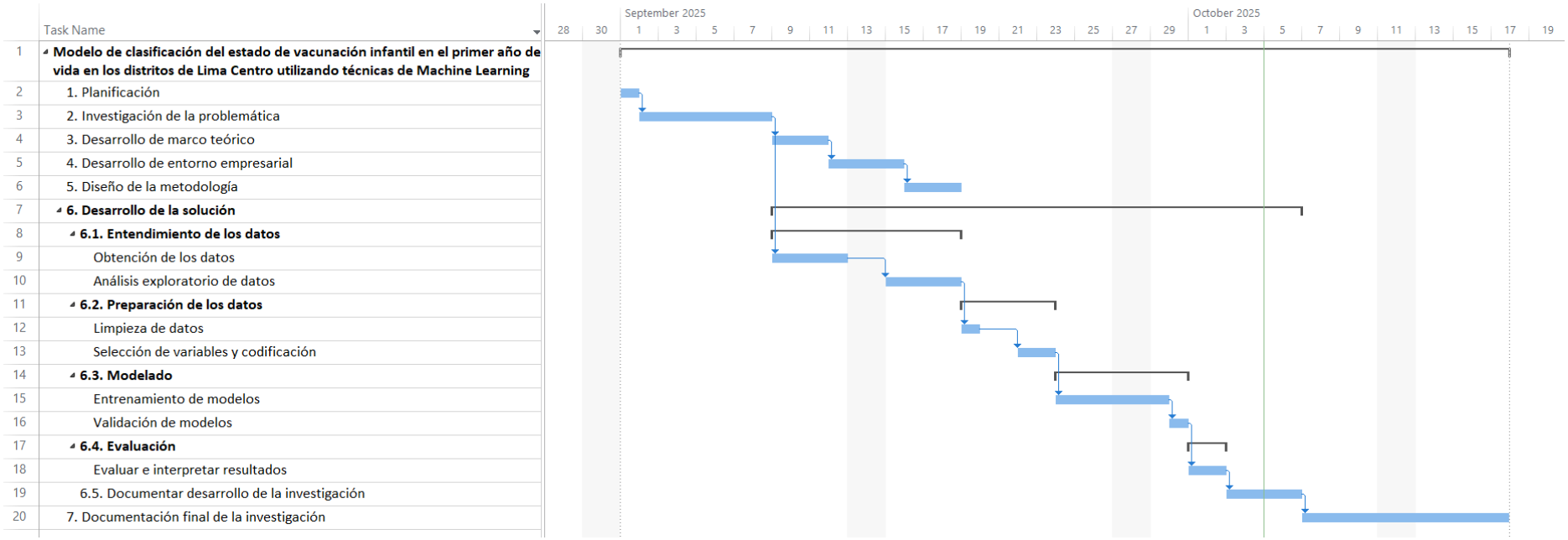
4.3.6. Validación cruzada del modelo final

La validación cruzada se aplica al modelo entrenado con machine learnig con el objetivo de evaluar su estabilidad y capacidad de generalización. Se utiliza una configuración de pliegues estratificados (k), calculando los promedios y desviaciones estándar de las principales métricas de desempeño. Este procedimiento permite confirmar la consistencia del modelo ante distintas particiones de los datos y reforzar la confiabilidad de los resultados obtenidos.

4.4. Cronograma de Actividades

Figura 17

Cronograma de actividades



Nota. Elaboración propia.

4.5. Presupuesto

Tabla 9

Tabla de gastos

Recurso	Descripción de uso / gasto	Costo soles (S/)	Costo dólares ($)
Libro Artificial Intelligence: A Modern Approach	Material bibliográfico utilizado para el desarrollo del marco teórico de la presente investigación		$80.00
Google Colab Pro	Ejecución de código Python en la nube de Google con recursos de cómputo y tiempos de ejecución adecuados para el presente trabajo. (200 unidades de procesamiento - 2 meses de uso).		$20.00
TOTAL		S/	$100.00

Nota. Elaboración propia.

Capítulo 5: Desarrollo de la Solución

En el presente capítulo se detalla el proceso de implementación de la solución propuesta para la clasificación del estado de vacunación infantil en el primer año de vida en la jurisdicción de DIRIS Lima Centro utilizando técnicas de Machine Learning. Se detalla el entendimiento y la preparación de los datos, el modelado y la evaluación de los modelos. Además, se presentan los criterios utilizados para seleccionar la mejor arquitectura de modelo en función de su desempeño.

5.1. Propuesta Solución

5.1.1. Obtención de los Datos

En esta etapa se tiene como objetivo la recopilación, organización y estructuración de los datos necesarios para la presente investigación.

Actividad 1: Obtención de datase

Tarea 1: Obtención de dataset desde fuente oficial

En esta actividad se gestionó la recopilación y consolidación de la base de datos de vacunación infantil proveniente de los sistemas institucionales de la DIRIS Lima Centro, específicamente del Sistema de Historia Clínica HIS MINSA y del Padrón Nominal, los cuales concentran los registros de atención de los menores de cinco años en los establecimientos de la jurisdicción.

La solicitud de acceso se realizó formalmente mediante mesa de partes institucional (ver Anexo 01) y el dataset fue entregado en formato CSV anonimizado, en cumplimiento de la Ley N.º 29733 – Ley de Protección de Datos Personales y su reglamento (D.S. N.º 003-2013-JUS). De esta forma se garantizó la confidencialidad y la utilización exclusiva de la información con fines de investigación académica.

El archivo recibido contenía 185,220 registros correspondientes a los niños menores de cinco años atendidos entre setiembre de 2019 y setiembre de 2025, junto con variables demográficas y clínicas relacionadas con el nacimiento, la madre y el estado de vacunación. Cada registro representa la información individual del niño, incluyendo fechas de aplicación de dosis, establecimiento de salud de atención, tipo de parto, peso al nacer, nivel educativo de la madre, distrito de residencia, entre otros atributos, detallados en laTabla 10.

Tabla 10

Detalle de dataset de vacunación infantil

Datos	Descripción
RIS_RN	RIS donde nació el menor
EESS_PN	EESS donde nació el menor
EESS_HIS_O_EESS_CPRO1	EESS donde se atiende el menor
FECHA_NACIMIENTO_PN	Fecha de Nacimiento del menor
GENERO	Género
GRUPO_DE_EDAD	Grupo de Edad
IPV_1¬ dosis	Fecha IPV 1ra dosis
IPV_2¬ dosis	Fecha IPV 2da dosis
IPV_3¬ dosis	Fecha IPV 3ra dosis
PENTA_1¬ dosis	Fecha Pentavalente 1ra dosis
PENTA_2¬ dosis	Fecha Pentavalente 2da dosis
PENTA_3¬ dosis	Fecha Pentavalente 3ra dosis
NEUMOCOCO_1¬ dosis	Fecha Neumococo 1ra dosis
NEUMOCOCO_2¬ dosis	Fecha Neumococo 2da dosis
ROTAVIRUS_1¬ dosis	Fecha Rotavirus 1ra dosis
ROTAVIRUS_2¬ dosis	Fecha Rotavirus 2da dosis
INFLUENZA_3-_1¬ dosis	Fecha Influenza 1ra dosis
INFLUENZA_3-_2¬ dosis	Fecha Influenza 2da dosis
SPR_1¬ dosis	Fecha Sarampión 1ra dosis
SPR_2¬ dosis	Fecha Sarampión 2da dosis
DPT_1¬ Dosis refuerzo	Fecha DPT 1ra dosis refuerzo
IPV_1¬ Dosis refuerzo	Fecha IPV dosis refuerzo
INFLUENZA_3+_Dosis ·nica	Fecha Influenza 3+ dosis única
HAV_Dosis ·nica	Fecha HAV dosis única

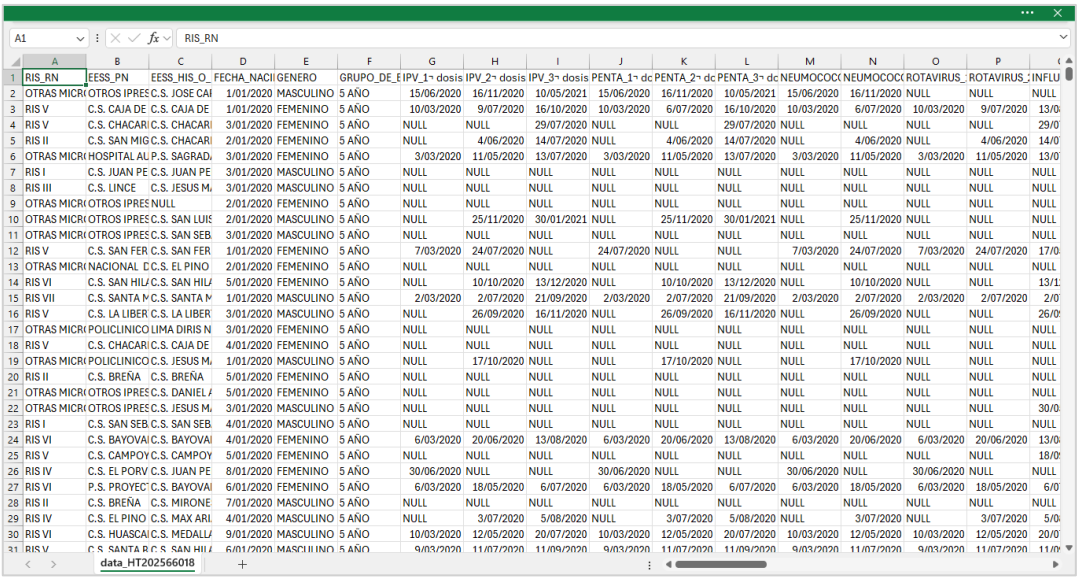
VARICELA_Dosis ·nica	Fecha Varicela dosis única
DPT_2¬ Dosis refuerzo	Fecha DPT 2da dosis refuerzo
IPV_2¬ Dosis refuerzo	IPV 2da dosis refuerzo
esquema_completo	Cumple Esquema Completo
Ubigeo_LugarNacido	Nacimiento Ubigeo
CNVLC_DPTO_EESS	Nacimiento Departamento
CNVLC_PROV_EESS	Nacimiento Provincia
CNVLC_DIST_EESS	Nacimiento Distrito
CNVLC_Ipress	Nacimiento IPRESS
CNVLC_CO_LOCAL	Nacimiento Código Local
CNVLC_Nombre_EESS	Nacimiento Nombre EESS
CNVLC_Diresa_Diris	Nacimiento DIRIS
CNVLC_Institucion	Nacimiento Institución
CNVLC_Categoria	Nacimiento Categoría
CNVLC_PERIODO	Nacimiento Periodo (año mes)
CNVLC_FE_NACIDO	Nacimiento Periodo (año mes día)
CNVLC_PESO_NACIDO	Nacimiento Peso al nacer
CNVLC_TALLA_NACIDO	Nacimiento Talla al nacer
CNVLC_APGAR_5_NACIDO	Nacimiento Apgar
CNVLC_DUR_EMB_PARTO	Nacimiento Duración embarazo en semanas
CNVLC_Condicion_Parto	Nacimiento Condición Parto
CNVLC_sexo_nacido	Nacimiento Sexo
CNVLC_Tipo_Parto	Nacimiento Tipo de Parto
CNVLC_Financiador_Parto	Nacimiento Financiador Parto
CNVLC_Ligadura_corte	Nacimiento Ligadura-corte
CNVLC_Malformacion_Congenita	Nacimiento Malformación congénita
CNVLC_Lactancia_Precoz	Nacimiento Lactancia Precoz
CNVLC_PERCEF	Nacimiento Perímetro Cefálico

CNVLC_PERTOR	Nacimiento Perímetro Torácico
CNVLC_Lugar_Nacido	Nacimiento Lugar
CNVLC_Estado_Civil	Madre Estado Civil
CNVLC_Nivel_Intruccion_Madre	Madre Nivel de Instrucción
CNVLC_Num_embar_madre	Madre Numero embarazos
CNVLC_Hijos_vivo_madre	Madre Hijos Vivos
CNVLC_Dpto_Madre	Madre Dpto. Residencia
CNVLC_Prov_Madre	Madre Prov. Residencia
CNVLC_Dist_Madre	Madre Distrito Residencia
CNVLC_Ubigeo_DOM_Madre	Madre Ubigeo Domicilio
CNVLC_Edad_Madre	Madre Edad

Nota. Elaboración propia.

Figura 18

Dataset data_HT202566018.csv enviado por DIRIS Lima Centro



Nota. Elaboración propia.

Entregable:

- Dataset original anonimizado en formato CSV, proveniente de los sistemas HIS-MINSA y Padrón Nominal, data_HT202566018.csv con un tamaño de 90.6MB, con 63 columnas de datos y 185,220 registros.

Actividad 2: Carga de dataset

Tarea 1: Importar dataset

Una vez obtenido el dataset, se procedió a la carga del dataset en formato csv en Google Colab.

Tarea 2: Verificar cantidad de registros y columnas

La Figura 19 muestra un ejemplo de la salida del comando data.shape() y data columns(), que evidencia el número de registros (185,220) y columnas (63), validando la integridad general del dataset.

Figura 19

Verificación de estructura y limpieza inicial del dataset de vacunación infantil implementada en Google Colab

```
print(data.shape)
print(data.columns)
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
(185220, 63)
Index(['RIS_RN', 'EESS_PN', 'EESS_HIS_O_EESS_CPRO1', 'FECHA_NACIMIENTO_PN',
       'GENERO', 'GRUPO_DE_EDAD', 'IPV_1¬ dosis', 'IPV_2¬ dosis',
       'IPV_3¬ dosis', 'PENTA_1¬ dosis', 'PENTA_2¬ dosis', 'PENTA_3¬ dosis',
       'NEUMOCOCO_1¬ dosis', 'NEUMOCOCO_2¬ dosis', 'ROTAVIRUS_1¬ dosis',
       'ROTAVIRUS_2¬ dosis', 'INFLUENZA_3-_1¬ dosis', 'INFLUENZA_3-_2¬ dosis',
       'SPR_1¬ dosis', 'SPR_2¬ dosis', 'DPT_1¬ Dosis refuerzo',
       'IPV_1¬ Dosis refuerzo', 'INFLUENZA_3+_Dosis ·nica', 'HAV_Dosis ·nica',
       'VARICELA_Dosis ·nica', 'DPT_2¬ Dosis refuerzo',
       'IPV_2¬ Dosis refuerzo', 'esquema_completo', 'CNVLC_Ubigeo_LugarNacido',
       'CNVLC_DPTO_EESS', 'CNVLC_PROV_EESS', 'CNVLC_DIST_EESS', 'CNVLC_Ipress',
       'CNVLC_CO_LOCAL', 'CNVLC_Nombre_EESS', 'CNVLC_Diresa_Diris',
       'CNVLC_Institucion', 'CNVLC_Categoria', 'CNVLC_PERIODO',
       'CNVLC_FE_NACIDO', 'CNVLC_PESO_NACIDO', 'CNVLC_TALLA_NACIDO',
       'CNVLC_APGAR_5_NACIDO', 'CNVLC_DUR_EMB_PARTO', 'CNVLC_Condicion_Parto',
       'CNVLC_sexo_nacido', 'CNVLC_Tipo_Parto', 'CNVLC_Financiador_Parto',
       'CNVLC_Ligadura_corte', 'CNVLC_Malformacion_Congenita',
       'CNVLC_Lactancia_Precoz', 'CNVLC_PERCEF', 'CNVLC_PERTOR',
       'CNVLC_Lugar_Nacido', 'CNVLC_Estado_Civil',
       'CNVLC_Nivel_Intruccion_Madre', 'CNVLC_Num_embar_madre',
       'CNVLC_Hijos_vivo_madre', 'CNVLC_Dpto_Madre', 'CNVLC_Prov_Madre',
       'CNVLC_Dist_Madre', 'CNVLC_Ubigeo_DOM_Madre', 'CNVLC_Edad_Madre'],
      dtype='object')
```

Nota. Elaboración propia.

Entregable:

- Dataset cargado y verificado en google colab.

Actividad 3: Análisis exploratorio – EDA

Tarea 1: Analizar estadísticas descriptivas

El análisis exploratorio de datos constituye la fase final del entendimiento de los datos, dentro de la metodología CRISP-DM. Su propósito fue comprender la estructura, distribución y comportamiento de las variables del dataset de vacunación infantil, así como detectar patrones o relaciones que pudieran influir en el estado de vacunación de los menores de un año.

Para esta actividad se emplearon técnicas de estadística descriptiva y visualización, utilizando el entorno Google Colab con las librerías pandas, matplotlib y seaborn. Se calculó un resumen estadístico de las variables numéricas mediante el comando data.describe(), el cual permitió conocer medidas de tendencia central (media, mediana), dispersión (desviación estándar) y valores extremos. Este análisis mostró una variabilidad significativa en las variables relacionadas con las dosis aplicadas y la edad de los menores, evidenciando diferencias en los intervalos de aplicación de vacunas y en los pesos al nacer.

Asimismo, se calculó el porcentaje de valores faltantes por variable, identificándose campos incompletos en algunas variables clínicas del parto. Estos hallazgos sirvieron de insumo para las decisiones de limpieza e imputación en la fase de preprocesamiento.

Figura 20

Resumen estadístico de variables numéricas

Resumen estadístico de variables numéricas:

	count	mean	min	25%	50%	75%	max	std
FECHA_NACIMIENTO_PN	185220	2022-07-03 00:55:41.341107968	2019-09-26 00:00:00	2020-12-21 00:00:00	2022-06-04 00:00:00	2023-12-05 00:00:00	2025-09-23 00:00:00	NaN
CNVLC_Ubigeo_LugarNacido	147055.0	150115.362769	150101.0	150101.0	150115.0	150131.0	150141.0	13.616965
CNVLC_Ipress	147055.0	8184.023012	5617.0	6208.0	6215.0	9123.0	30308.0	3039.559241
CNVLC_CO_LOCAL	147055.0	8184.023012	5617.0	6208.0	6215.0	9123.0	30308.0	3039.559241
CNVLC_PERIODO	147055.0	202230.240957	202001.0	202104.0	202208.0	202402.0	202509.0	165.452056
CNVLC_FE_NACIDO	147055.0	20223039.685118	20200101.0	20210416.0	20220820.0	20240215.0	20250922.0	16545.159084
CNVLC_PESO_NACIDO	147044.0	3319.099882	300.0	3040.0	3350.0	3650.0	9999.0	528.791824
CNVLC_TALLA_NACIDO	147002.0	50.118224	25.0	48.0	50.0	51.0	9999.0	89.925426
CNVLC_APGAR_5_NACIDO	147055.0	9.133277	0.0	9.0	9.0	9.0	99.0	4.043463
CNVLC_DUR_EMB_PARTO	147055.0	38.56264	20.0	38.0	39.0	40.0	42.0	1.681949
CNVLC_PERCEF	147055.0	34.337473	1.0	33.5	34.5	35.5	66.0	1.794268
CNVLC_PERTOR	147055.0	33.686205	3.0	33.0	34.0	35.0	95.5	2.332496
CNVLC_Ubigeo_DOM_Madre	146886.0	148314.317273	10101.0	150115.0	150132.0	150132.0	250305.0	15849.314286
CNVLC_Edad_Madre	147055.0	30.804291	12.0	26.0	31.0	36.0	57.0	6.516296

Porcentaje de valores nulos por columna:

	0

Nota. Elaboración Propia.

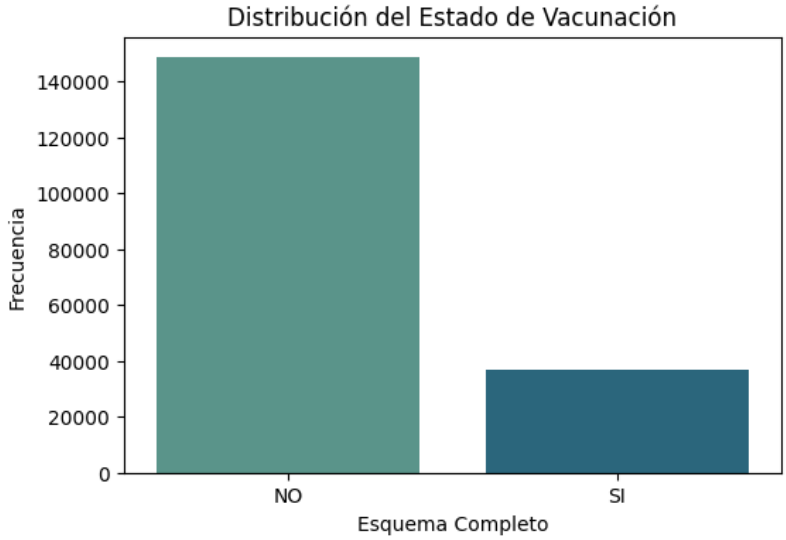
El proceso permitió identificar que el *dataset* final contiene 185,220 registros, con predominancia de variables categóricas y un número reducido de variables numéricas

continuas. Se evidenció también la presencia de ciertos campos con valores nulos, principalmente asociados a variables clínicas del parto, los cuales serían tratados posteriormente durante la etapa de preprocesamiento y limpieza de datos.

Con el fin de caracterizar el comportamiento general de los datos, se elaboraron histogramas y gráficos de barras que muestran la distribución de variables clave como el estado de vacunación completo, el tipo de parto, el financiador de la atención y la condición del parto, así como un mapa de calor de correlaciones basado en el coeficiente de Spearman para las variables numéricas.

Figura 21

Distribución del estado de vacunación infantil según variable esquema_completo, realizado en Google Colab

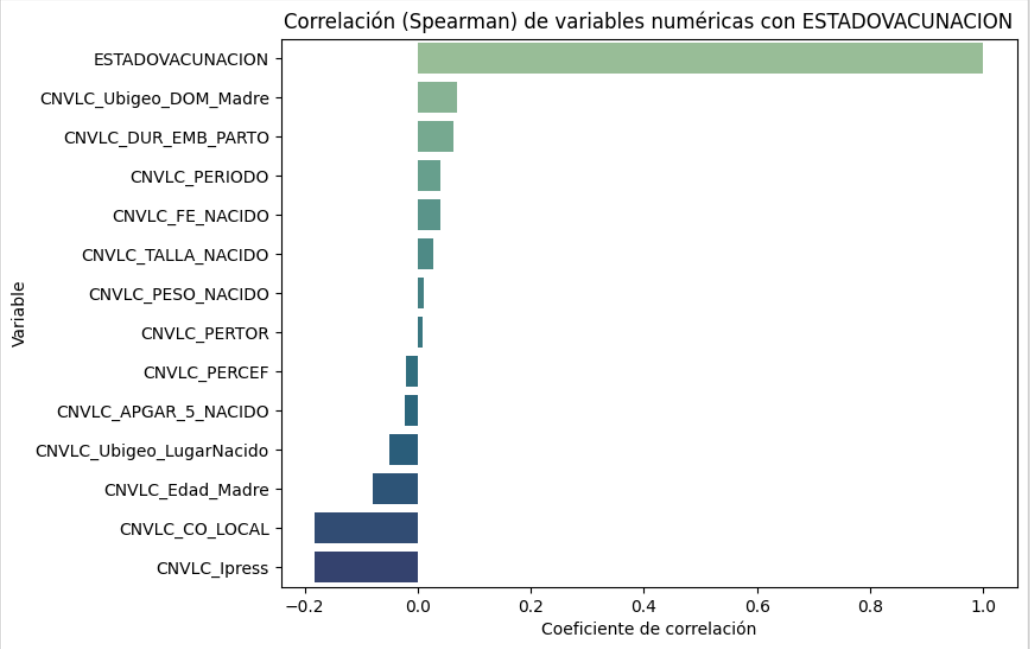


Nota. Elaboración propia.

Adicionalmente, se evaluó la asociación entre las variables independientes y la variable objetivo esquema_completo. Para las variables numéricas se utilizó el coeficiente de correlación de Spearman, mientras que para las variables categóricas se aplicó la medida de V de Cramer, con el fin de identificar aquellas con mayor influencia potencial en el estado de vacunación infantil. Las variables con mayor grado de relación fueron dosis recibidas, dosis faltantes, edad de la madre y nivel educativo de la madre en el caso de las numéricas, y tipo de parto y financiador del parto entre las categóricas.

Figura 22

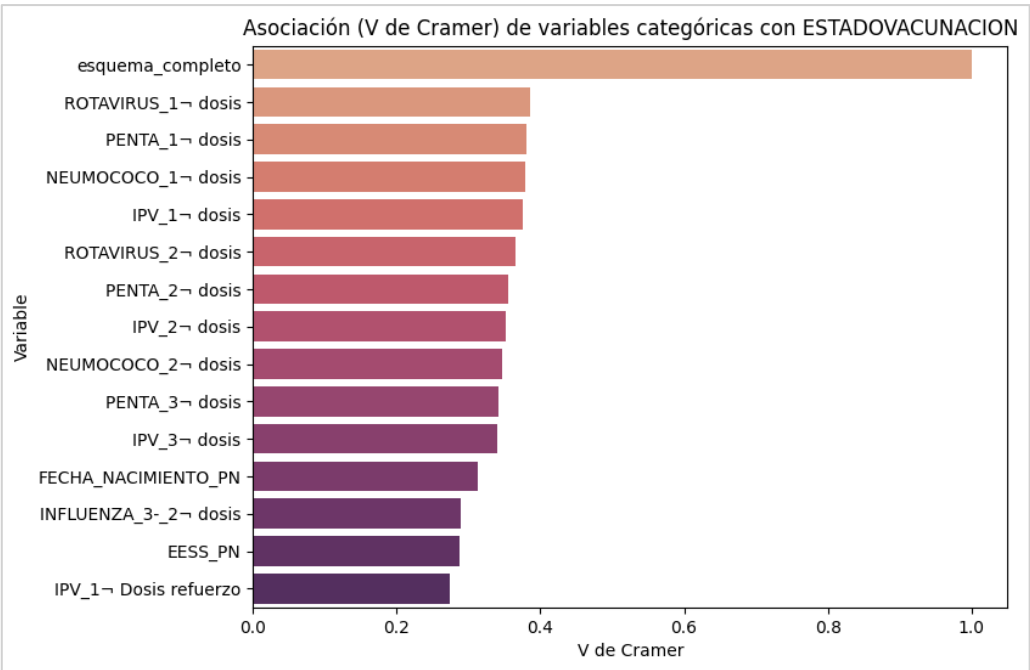
Correlación de Spearman entre variables numéricas y la variable objetivo.



Nota. Elaboración propia.

Figura 23

Asociación de variables categóricas con la variable objetivo mediante el coeficiente V de Cramer



Nota. Elaboración propia.

El análisis permitió confirmar la adecuada consistencia del conjunto de datos y orientar las decisiones de selección y transformación de variables para el modelado posterior. Se observaron correlaciones moderadas entre las variables relacionadas con dosis recibidas, edad de la madre y peso al nacer, las cuales resultaron relevantes para el modelo de clasificación.

Entregable:

- Reporte descriptivo de distribución de variables (estadísticos y gráficos). Gráficos de correlación y relación de variables categóricas con el estado de vacunación.

5.1.2. Preprocesamiento

El objetivo de esta fase fue transformar el conjunto de datos bruto obtenido en la fase anterior en una versión estructurada, consistente y apta para el entrenamiento de los algoritmos de clasificación.

Se aplicaron técnicas de imputación de valores faltantes, codificación de variables, balanceo de clases y escalado, garantizando la integridad y homogeneidad del dataset.

Actividad 1: Limpieza de datos

Tarea 1: Eliminar columnas innecesarias

En esta tarea se depuraron columnas innecesarias, con datos poco relevantes y redundantes ya que se hallaron columnas con datos como Departamento, Provincia y Distrito que se encuentran consolidadas en la columna con datos de ubigeo, así como columnas con datos derivados de la fecha de nacimiento del niño que estaban en otro formato como período, asimismo se encontró que la columna género se encontraba duplicada.

Tarea 2: Imputar datos faltantes

En esta tarea se imputaron valores faltantes, en el caso específico de las dosis de vacunas, se validó en que muchos casos en las dosis del mismo tipo de vacuna se halló que tenía dato la segunda dosis y la primera era NULL, en el levantamiento de información a juicio experto del área de estadística de DIRIS Lima Centro, nos indicaron que si la madre llevaba a su menor hijo al establecimiento de salud y el carnet de vacunas tenía la primera dosis colocada en otro establecimiento que no pertenezca a la red de MINSA (incluyendo a los establecimiento

de salud privados con los que tienen convenio), se consideraba válido mas no se podía colocar como primera dosis ya que el menor ya contaba con dicha vacuna, por lo que se procedió a imputar los datos de aplicación cierta de dichas vacunas que cumplan ese caso específico.

Tarea 3: Eliminar valores nulos y duplicados

En esta tarea se procedió a la eliminación de valores nulos y registros duplicados buscando garantizar la calidad y consistencia del conjunto de datos, por lo que se procedió a reemplazar los valores vacíos como NULL, null, NaN o espacios en blanco por valores nulos con el formato reconocible por el sistema. Asimismo, se eliminaron las filas con al menos un valor faltante y también los registros donde están todas las columnas duplicadas, esto busca reducir los errores en la etapa de análisis.

Tarea 4: Limpiar y normalizar texto

En esta tarea, se realizó la limpieza y normalización de los campos de texto, buscando estandarizar el formato de variables categóricas. Para ello se aplicó la función que elimina espacios en blanco innecesarios y volver los textos a letra mayúscula con el fin de tener una estructura uniforme en las cadenas de caracteres, lo que permite mejorar la calidad de los datos.

Entregable:

- Dataset depurado y estandarizado, libre de valores nulos, duplicados y errores de formato, columnas irrelevantes y con la correcta asignación de vacunas colocadas en establecimientos no MINSA.

Actividad 2: Ingeniería de características

Tarea 1: Eliminar registros innecesarios fuera del alcance del estudio

En esta tarea se valida la cantidad de registros de niños que no cumplen con la edad para el estudio, es decir, que no cuentan con la edad para tener el esquema de vacunación en el primer año de vida.

Tarea 2: Generar variable objetivo “target”

Con el fin de asegurar la calidad de la variable objetivo, se genera la variable dicotómica “target” en base a la cantidad de vacunas correspondientes al primer año de vida, si la cantidad de vacunas es menor que 12, el target es 1 (esquema de vacunación incompleta o nula), caso

contrario es 0 (esquema de vacunación completa). Así nos aseguramos de que la variable contenga exactamente los valores que requerimos para la aplicación en los modelos.

Figura 24

Eliminación de registros con menos de un año de vida

```
Distribución original: Counter({1: 91197, 0: 30945})
```

Nota: Elaboración propia

Entregable:

- Dataset depurado con registros de niños vacunados con la edad adecuada para el presente estudio, asimismo la generación de la variable dicotómica "target" que será la variable dependiente en este estudio.

Actividad 3: Codificación de variables

Tarea 1: Aplicar codificación a variables numéricas ordinales

Para la variable Nivel de Instrucción de la Madre, que tiene un orden jerárquico, se aplicó un proceso de codificación ordinal (*Ordinal Encoding*), asignando valores numéricos de acuerdo con el nivel educativo alcanzado. Se estableció un orden lógico que va desde ningún nivel hasta superior universitario completo, lo que permite reconocer el avance educativo como variable ordinal.

Tarea 2: Aplicar codificación a variables categóricas

El dataset cuenta con variables categóricas nominales como condición del parto, sexo del nacido, tipo de parto, financiador, malformación congénita, lactancia precoz y estado civil, para estas se aplicó la técnica de codificación por etiquetas (*Label Encoding*), asignando un valor numérico entero a cada categoría distinta, esto permitió transformar valores textuales en datos numéricos para que sean usados por los algoritmos de machine learning, sin alterar su significado original.

Entregable:

- Dataset depurado con variables categóricas transformadas en valores numéricos.

Actividad 4: División de Dataset

Tarea 1: Dividir el dataset con train_test_split con una proporción 80-20

El propósito de esta tarea fue separar el *dataset* preprocesado en subconjuntos de entrenamiento o *train* (80%) y prueba o *test* (20%), con la finalidad de evaluar el desempeño de los modelos de forma objetiva y prevenir el sobreajuste.

En primer lugar, se definieron las variables predictoras (X), que incluyó todas las columnas excepto 'target', y la variable objetivo (y), que correspondió a la columna ‘target’. Luego, se realizó la división del dataset utilizando la función *train_test_split* de *scikit-learn*, como se muestra en la Figura 27, con una proporción 80-20, manteniendo un *random_state*=42 para garantizar la reproducibilidad y aplicando estratificación según la variable objetivo para conservar la proporción de clases.

Figura 25

División del dataset en train y test

```
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)
```

Nota. Elaboración propia.

Finalmente, se verificó la distribución de variable objetivo “target” en los subconjuntos de *train* y test, confirmando que la estratificación preservó la proporción original de las clases, como se muestra en la Figura 28.

Figura 26

División del dataset en train y test

```
Distribución del target (train):
target
0    0.5
1    0.5
Name: proportion, dtype: float64
Distribución del target (test):
target
0    0.5
1    0.5
Name: proportion, dtype: float64
```

Nota. Elaboración propia.

Actividad 5: Balanceo de datos

Tarea 1: Implementar el balanceo de datos

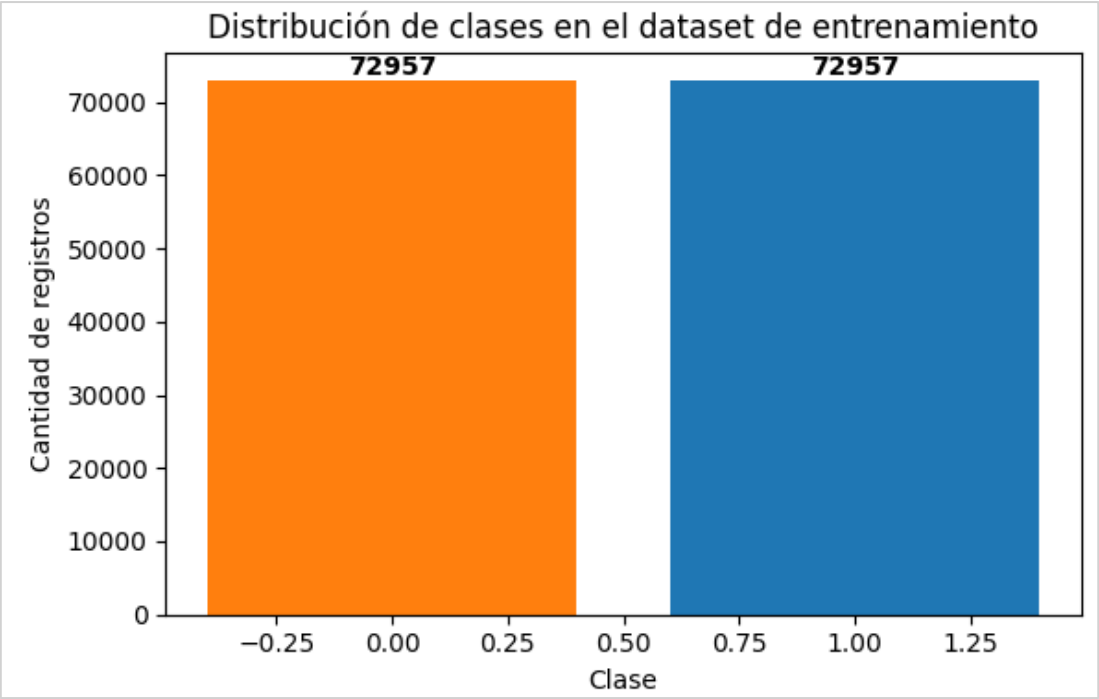
El conjunto de datos presentaba un desbalance entre las clases de la variable objetivo "target", por lo que se aplicó la técnica de sobremuestreo sintético SMOTE, con la finalidad de equilibrar la cantidad de registros de cada categoría, esto hace que se generen nuevas instancias sintéticas de la clase minoritaria y tener una data balanceada.

Tarea 2: Validar que el balanceo de clases sea adecuado

Luego de la aplicación del proceso de SMOTE, se validó la distribución final de clases para asegurar el balanceo adecuado. La clase minoritaria se incrementó a 72,957 registros, igualando a la clase mayoritaria, lo que hace que haya una representación proporcional de ambas clases.

Figura 27

Validar el balanceo de clases



Nota. Elaboración propia

Entregable:

- Dataset validado con un balanceo adecuado.

Actividad 6: Escalado de Variables

Tarea 1: Normalizar las variables numéricas

Con el dataset balanceado, se procedió al escalado de variables mediante la técnica *StandardScaler*, normalizando la media y desviación estándar de cada atributo. Esta transformación garantizó que las variables estuvieran en la misma escala numérica, evitando que las de mayor rango influyeran desproporcionadamente en el modelo.

Tarea 2: Validar la distribución de valores escalados

Se realizó la validación del escalado de variables numéricas, lo que permite una correcta homogeneización de las magnitudes y evita que las diferencias de escala afecten el desempeño de los algoritmos de clasificación.

Figura 28

Validación de valores numéricos escalados

```
Escalado aplicado correctamente a columnas numéricas.

Medias después del escalado:
CNVLC_Edad_Madre         -0.0
CNVLC_Num_embar_madre    -0.0
CNVLC_Hijos_vivo_madre    0.0
CNVLC_PESO_NACIDO        -0.0
CNVLC_TALLA_NACIDO       -0.0
CNVLC_DUR_EMB_PARTO      -0.0
CNVLC_PERCEF              0.0
CNVLC_PERTOR              0.0
dtype: float64

Desviaciones estándar después del escalado:
CNVLC_Edad_Madre          1.0
CNVLC_Num_embar_madre     1.0
CNVLC_Hijos_vivo_madre    1.0
CNVLC_PESO_NACIDO         1.0
CNVLC_TALLA_NACIDO        1.0
CNVLC_DUR_EMB_PARTO       1.0
CNVLC_PERCEF              1.0
CNVLC_PERTOR              1.0
dtype: float64
```

Nota. Elaboración propia

Entregable:

- Dataset con variables numéricas correctamente escaladas.

5.1.3. Selección de Características

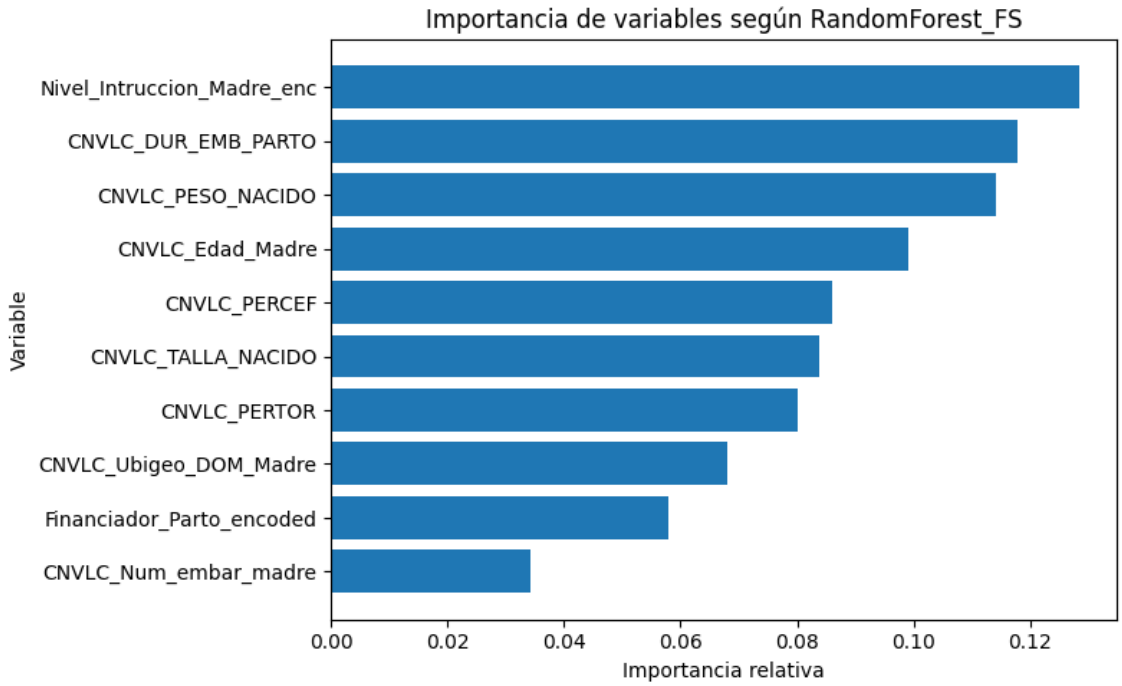
Actividad 1: Aplicación de técnicas de selección de características

Tarea 1: Aplicar técnica *Random Forest*

Se aplicó la técnica *Random Forest* mediante el modelo RandomForestClassifier (random_state=42) entrenado sobre los subconjuntos X_train y y_train. A continuación, se obtuvieron las importancias de las variables mediante el atributo feature_importances_, y se seleccionaron las diez más significativas utilizando la función get_*top*_features, tal como se muestra en la Figura 29.

Figura 29

Top 10 de variables significativas según Random Forest



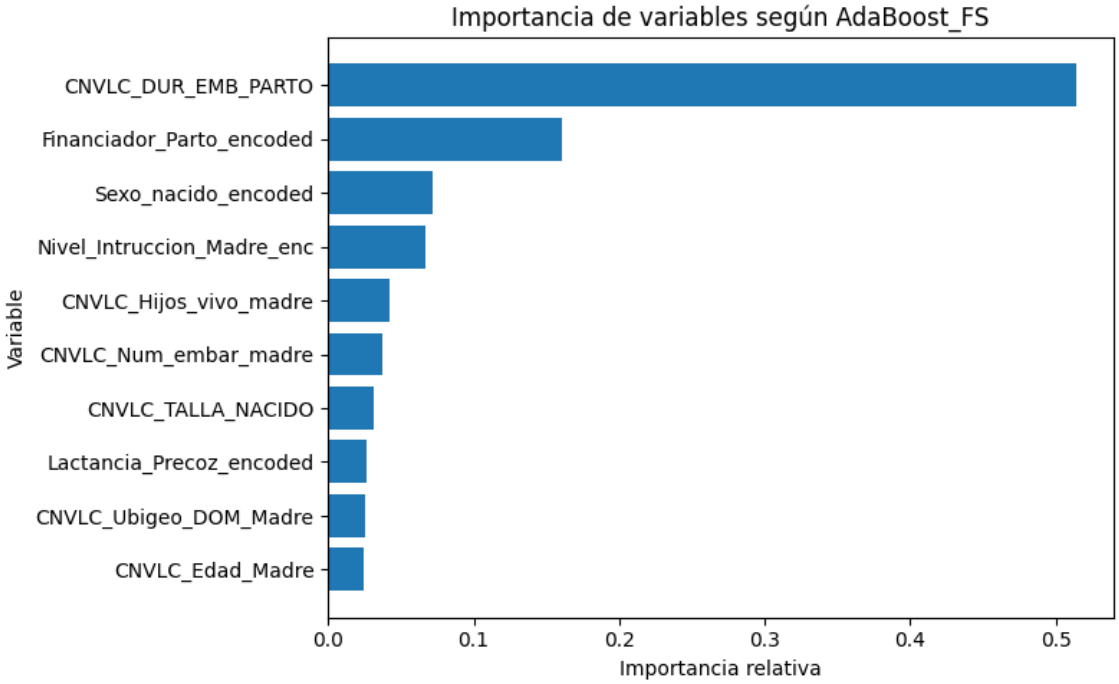
Nota. Elaboración propia.

Tarea 2: Aplicar técnica *AdaBoost*

Se entrenó el modelo AdaBoostClassifier (random_state=42) sobre los datos de *train* (X_train, y_train). Luego, se extrajeron las importancias de las variables desde el atributo feature_importances_, y se ordenaron para obtener las diez más representativas mediante la misma función get_top_features, como se muestra en la Figura 30.

Figura 30

Top 10 de variables significativas según AdaBoost



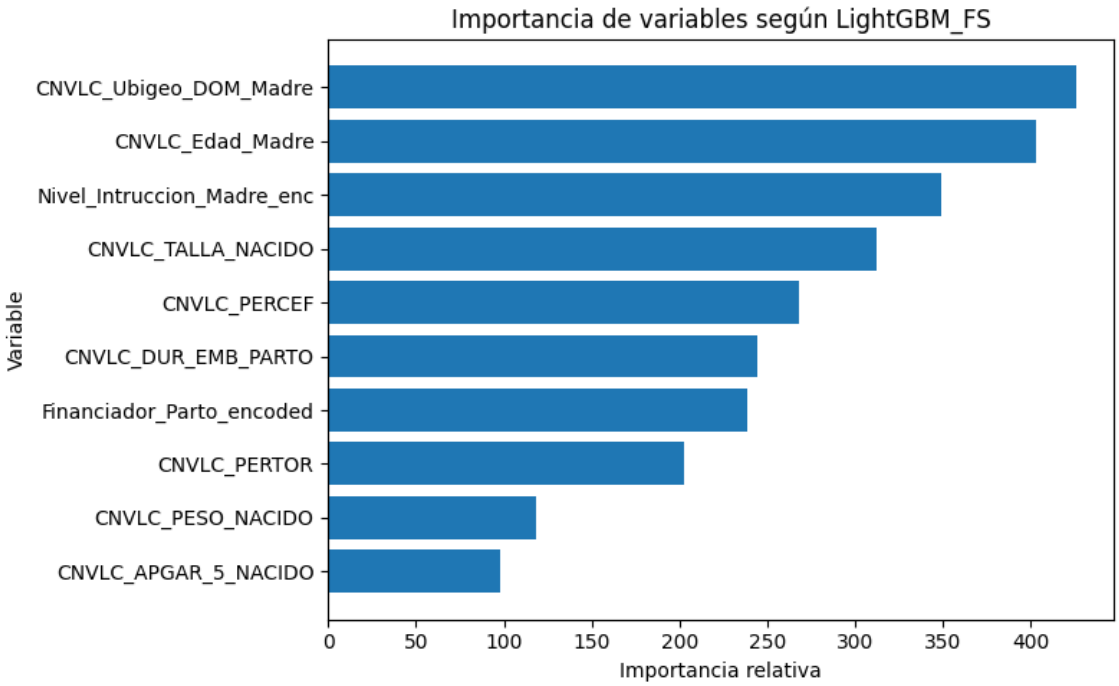
Nota. Elaboración propia.

Tarea 3: Aplicar técnica *LightGBM*

Por último, se aplicó el modelo LightGBM mediante el clasificador *LGBMClassifier* (random_state=42, verbose=-1), entrenado sobre X_train y y_train. Luego, se calcularon las importancias de las variables y se seleccionaron las diez más relevantes mediante la función get_top_features.

Figura 31

Top 10 de variables significativas según LightGBM



Nota. Elaboración propia.

5.1.4. Machine Learning

Actividad 1: Entrenamiento de modelos sin ajuste de hiperparámetros

Tarea 1: Entrenar los modelos *LogisticRegression, RandomForest, XGBoost, KNN, AdaBoost, DecisionTree, Bagging*

El objetivo de esta tarea fue entrenar y evaluar los siete algoritmos de clasificación, como se muestra en la Tabla 11, sobre los distintos conjuntos de variables seleccionadas en la fase previa de selección de características, con el fin de comparar su desempeño y determinar las combinaciones más efectivas. Para ello, se definió un diccionario *models* con los siete algoritmos a evaluar, manteniendo parámetros base y fijando *random_state*=42

- LogisticRegression (max_iter=1000, random_state=42)
- RandomForestClassifier (random_state=42)
- XGBClassifier (random_state=42, eval_metric='logloss', use_label_encoder=False)
- KNeighborsClassifier (parámetros por defecto)

- AdaBoostClassifier (random_state=42)
- DecisionTreeClassifier (random_state=42)
- BaggingClassifier (random_state=42)

Luego, se recorrió el diccionario *feature*_sets generado en la fase previa. Para cada *feature set*, se filtraron las matrices de entrenamiento y prueba con las columnas seleccionadas: X_train_fs = X_train[selected_features] y X_test_fs = X_test[selected_features]

Posteriormente para cada algoritmo en *models* se realizó ajuste, con model.fit(X_train_fs, y_train) y predicción con y_pred = model.predict(X_test_fs), como se muestra en la Figura 32.

Figura 32

Código en python para entrenamiento de modelos

```python
# ============================================================
# 4) Definición de modelos a evaluar (7 modelos)
# ============================================================
models = {
    "LogisticRegression": LogisticRegression(max_iter=1000, random_state=42),
    "RandomForest": RandomForestClassifier(random_state=42),
    "XGBoost": XGBClassifier(random_state=42, eval_metric='logloss', use_label_encoder=
    "KNN": KNeighborsClassifier(),
    "AdaBoost": AdaBoostClassifier(random_state=42),
    "DecisionTree": DecisionTreeClassifier(random_state=42),
    "Bagging": BaggingClassifier(random_state=42)
}

# ============================================================
# 5) Entrenamiento y evaluación (21 combinaciones)
# ============================================================
results = []

for fs_name, selected_features in feature_sets.items():
    print(f"\n===== Evaluando modelos usando {fs_name} =====")
    X_train_fs = X_train[selected_features]
    X_test_fs = X_test[selected_features]

    for model_name, model in tqdm(models.items(), desc=f"Entrenando con {fs_name}"):
        model.fit(X_train_fs, y_train)
        y_pred = model.predict(X_test_fs)

        # Algunos modelos (SVM lineal) no tienen predict_proba
        try:
            y_prob = model.predict_proba(X_test_fs)[:, 1]
        except:
            y_prob = np.zeros(len(y_pred))

        metrics = {
            "FeatureSet": fs_name,
            "Modelo": model_name,
            "Accuracy": accuracy_score(y_test, y_pred),
            "Precision": precision_score(y_test, y_pred),
            "Recall": recall_score(y_test, y_pred),
            "F1-score": f1_score(y_test, y_pred),
            "ROC_AUC": roc_auc_score(y_test, y_prob) if np.any(y_prob) else np.nan
        }
        results.append(metrics)
```

Nota. Elaboración propia.

Tabla 11

Modelos seleccionados para entrenamiento

Modelo	Descripción técnica
Regresión Logística	Modelo lineal probabilístico que estima la probabilidad de pertenencia a una clase mediante la función logística (*sigmoid*). Utilizado como línea base del experimento.
Árbol de Decisión	Clasificador basado en la segmentación iterativa del espacio de atributos, útil para detectar relaciones no lineales y manejar variables mixtas.

Random Forest	Ensamble de árboles de decisión que reduce el sobreajuste y mejora la estabilidad del modelo mediante el método de *bagging*.
Gradient Boosting	Modelo secuencial de árboles que ajusta los errores de modelos previos, optimizando la función de pérdida con gradientes descendentes.
K-Nearest Neighbors (KNN)	Clasificador no paramétrico que asigna etiquetas según la mayoría de clases entre los k vecinos más cercanos en el espacio de características.
Bagging	Método de agregación de múltiples clasificadores entrenados sobre subconjuntos del conjunto de datos, mejorando la generalización.
AdaBoost	Técnica de *boosting* adaptativo que pondera las instancias mal clasificadas, permitiendo una mejora iterativa del rendimiento.

Nota. Elaboración propia.

Entregables:

- Implementación de los 7 modelos de Machine Learning.

***Actividad 2*: Entrenamiento de modelos con ajuste de hiperparámetros**

Tarea 1: Entrenar los modelos *LogisticRegression, RandomForest, XGBoost, KNN, AdaBoost, DecisionTree y Bagging* con ajuste de hiperparámetros.

En esta tarea se llevó a cabo el entrenamiento y optimización de los modelos de *machine learning* mediante el ajuste automático de hiperparámetros, con el propósito de maximizar su desempeño y capacidad de generalización.

Para ello se empleó la biblioteca Optuna, que permite realizar una búsqueda iterativa y adaptativa de combinaciones óptimas de parámetros utilizando técnicas de optimización bayesiana.

El proceso se aplicó a las 21 combinaciones de los siete modelos, mencionados en la Tabla 11 y las tres técnicas de selección de características (RandomForest_FS, AdaBoost_FS y LightGBM_FS).

Durante la optimización, cada modelo fue entrenado y evaluado iterativamente, ajustando de manera automática los principales hiperparámetros que influyen en su rendimiento, tales como el número de estimadores, la profundidad máxima, la tasa de aprendizaje, el tamaño de muestra, el número de vecinos y los criterios de división, entre otros.

El desempeño de cada combinación fue medido mediante las métricas de *Accuracy, Precision, Recall, F1-score y ROC-AUC*, seleccionándose al final aquellas configuraciones que alcanzaron los mejores valores de desempeño general.

Entregables:

- Implementación de los 7 modelos de Machine Learning con la mejor combinación de hiperparámetros.

5.1.5. Evaluación

Actividad 1: Evaluación de resultados

En esta actividad se realizó la evaluación de los modelos desarrollados durante la fase de Machine Learning, considerando los indicadores de desempeño más relevantes para un problema de clasificación binaria: a*ccuracy*, p*recision*, *recall*, *f1-score, ROC-AUC*. En la Figura 33 se muestran los resultados por cada técnica de selección de características y por cada uno de los siete modelos entrenados sin ajuste de *hiperparámetros*, es decir, en total 21 combinaciones.

Figura 33

Resultados comparativos de los modelos entrenados sin ajuste de hiperparámetros

FeatureSet	Modelo	Accuracy	Precision	Recall	F1-score	ROC_AUC
AdaBoost_FS	XGBoost	0.736297	0.755191	0.957072	0.844231	0.673464
AdaBoost_FS	RandomForest	0.698309	0.768104	0.853673	0.808631	0.621513
AdaBoost_FS	Bagging	0.666339	0.773281	0.782566	0.777896	0.601957
AdaBoost_FS	KNN	0.644807	0.780687	0.729112	0.754018	0.608276
AdaBoost_FS	AdaBoost	0.648942	0.790734	0.720504	0.753987	0.635179
AdaBoost_FS	DecisionTree	0.627574	0.768031	0.718092	0.742222	0.544447
AdaBoost_FS	LogisticRegression	0.595399	0.784760	0.631250	0.699684	0.586130
LightGBM_FS	XGBoost	0.738180	0.752107	0.968586	0.846729	0.656478
LightGBM_FS	RandomForest	0.719432	0.758985	0.914693	0.829596	0.623564
LightGBM_FS	Bagging	0.665766	0.766831	0.793695	0.780032	0.592769
LightGBM_FS	DecisionTree	0.637562	0.764693	0.743311	0.753850	0.534582
LightGBM_FS	KNN	0.622089	0.784955	0.680208	0.728837	0.598960
LightGBM_FS	AdaBoost	0.626878	0.810184	0.653344	0.723360	0.645827
LightGBM_FS	LogisticRegression	0.568873	0.807877	0.554441	0.657585	0.605782
RandomForest_FS	XGBoost	0.736461	0.754025	0.960307	0.844755	0.665746
RandomForest_FS	RandomForest	0.721479	0.761741	0.912336	0.830265	0.638024
RandomForest_FS	Bagging	0.666380	0.768494	0.791667	0.779908	0.596029
RandomForest_FS	DecisionTree	0.635188	0.766362	0.735691	0.750713	0.537340
RandomForest_FS	AdaBoost	0.649515	0.802324	0.704057	0.749985	0.653315
RandomForest_FS	KNN	0.624094	0.791353	0.674342	0.728177	0.609096
RandomForest_FS	LogisticRegression	0.592165	0.800916	0.603893	0.688588	0.607169

Nota. Elaboración propia.

Asimismo, también se analizaron los resultados obtenidos tras el proceso de entrenamiento de los modelos con ajuste de hiperparámetros.

La optimización permitió evaluar el efecto de la búsqueda automática de parámetros sobre el rendimiento de cada modelo, considerando las tres técnicas de selección de características (*RandomForest_FS, AdaBoost_FS y LightGBM_FS*) en los 7 modelos en esta investigación como se detalla en la tabla a continuación.

Tabla 12

Modelos seleccionados para entrenamiento

FeatureSet	Modelo	Accuracy	Precision	Recall	F1-score	ROC_AUC	Best_Params
AdaBoost_FS	Bagging	0.696	0.769	0.846	0.806	0.612	n_estimators: 41, max_samples: 0.8919
AdaBoost_FS	Random Forest	0.675	0.788	0.773	0.781	0.657	n_estimators: 167, max_depth: 10, min_samples_split: 3, min_samples_leaf: 1
AdaBoost_FS	XGBoost	0.667	0.786	0.761	0.773	0.646	n_estimators: 104, learning_rate: 0.0290, max_depth: 3, subsample: 0.7013, colsample_bytree: 0.9601
AdaBoost_FS	Decision Tree	0.649	0.786	0.728	0.756	0.638	max_depth: 10, min_samples_split: 8, min_samples_leaf: 2
AdaBoost_FS	AdaBoost	0.655	0.795	0.725	0.758	0.643	n_estimators: 295, learning_rate: 0.2971
AdaBoost_FS	KNN	0.650	0.792	0.720	0.754	0.638	n_neighbors: 15
AdaBoost_FS	Logistic Regression	0.595	0.785	0.630	0.699	0.583	C: 0.4844, solver: lbfgs
LightGBM_FS	XGBoost	0.690	0.781	0.813	0.797	0.648	n_estimators: 139, learning_rate: 0.0130, max_depth: 5, subsample: 0.7050, colsample_bytree: 0.9118
LightGBM_FS	Bagging	0.662	0.764	0.792	0.778	0.588	n_estimators: 10, max_samples: 0.9763
LightGBM_FS	Decision Tree	0.644	0.776	0.735	0.755	0.628	max_depth: 10, min_samples_split: 10, min_samples_leaf: 2
LightGBM_FS	Random Forest	0.662	0.800	0.729	0.763	0.654	n_estimators: 288, max_depth: 8, min_samples_split: 2, min_samples_leaf: 4
LightGBM_FS	AdaBoost	0.642	0.805	0.687	0.741	0.645	n_estimators: 293, learning_rate: 0.2952
LightGBM_FS	KNN	0.615	0.793	0.655	0.718	0.624	n_neighbors: 15
LightGBM_FS	Logistic Regression	0.571	0.808	0.558	0.660	0.604	C: 3.0043, solver: lbfgs
Random Forest_FS	**XGBoost**	**0.711**	**0.778**	**0.859**	**0.816**	**0.665**	**n_estimators: 101, learning_rate: 0.0291, max_depth: 5, subsample: 0.8037, colsample_bytree: 0.7019**
Random Forest_FS	Bagging	0.671	0.770	0.798	0.784	0.601	n_estimators: 10, max_samples: 0.6932
Random Forest_FS	Random Forest	0.662	0.803	0.725	0.762	0.664	n_estimators: 150, max_depth: 9, min_samples_split: 8, min_samples_leaf: 4
Random Forest_FS	AdaBoost	0.629	0.813	0.653	0.724	0.650	n_estimators: 216, learning_rate: 0.2990
Random Forest_FS	KNN	0.620	0.804	0.650	0.719	0.638	n_neighbors: 15
Random Forest_FS	Decision Tree	0.608	0.797	0.637	0.708	0.628	max_depth: 8, min_samples_split: 7, min_samples_leaf: 2
Random Forest_FS	Logistic Regression	0.594	0.800	0.608	0.691	0.605	C: 0.0164, solver: liblinear

Nota. Elaboración propia.

Los resultados mostraron que el modelo con mejor desempeño fue *XGBoost* utilizando las características seleccionadas mediante *RandomForest_FS*, el cual alcanzó un F1-score de 0.816, Accuracy de 0.711 y un Recall de 0.858, evidenciando un equilibrio adecuado entre sensibilidad y precisión.

Sin embargo, a diferencia de la evaluación sin ajuste de hiperparámetros, el proceso de optimización no produjo un incremento sustancial en las métricas.

Por lo que, luego de realizar el análisis de las métricas, podemos indicar que el modelo *XGBoost* con selección de variables *LightGBM* sin ajsute de hiperparámetros, tuvo el mejor desempeño en la clasificación del estado de vacunación infantil. El *accuracy* obtenido indica que el modelo clasificó correctamente 74 % de los registros evaluados, como se observa en la Figura 34.

Figura 34

Modelo óptimo identificado

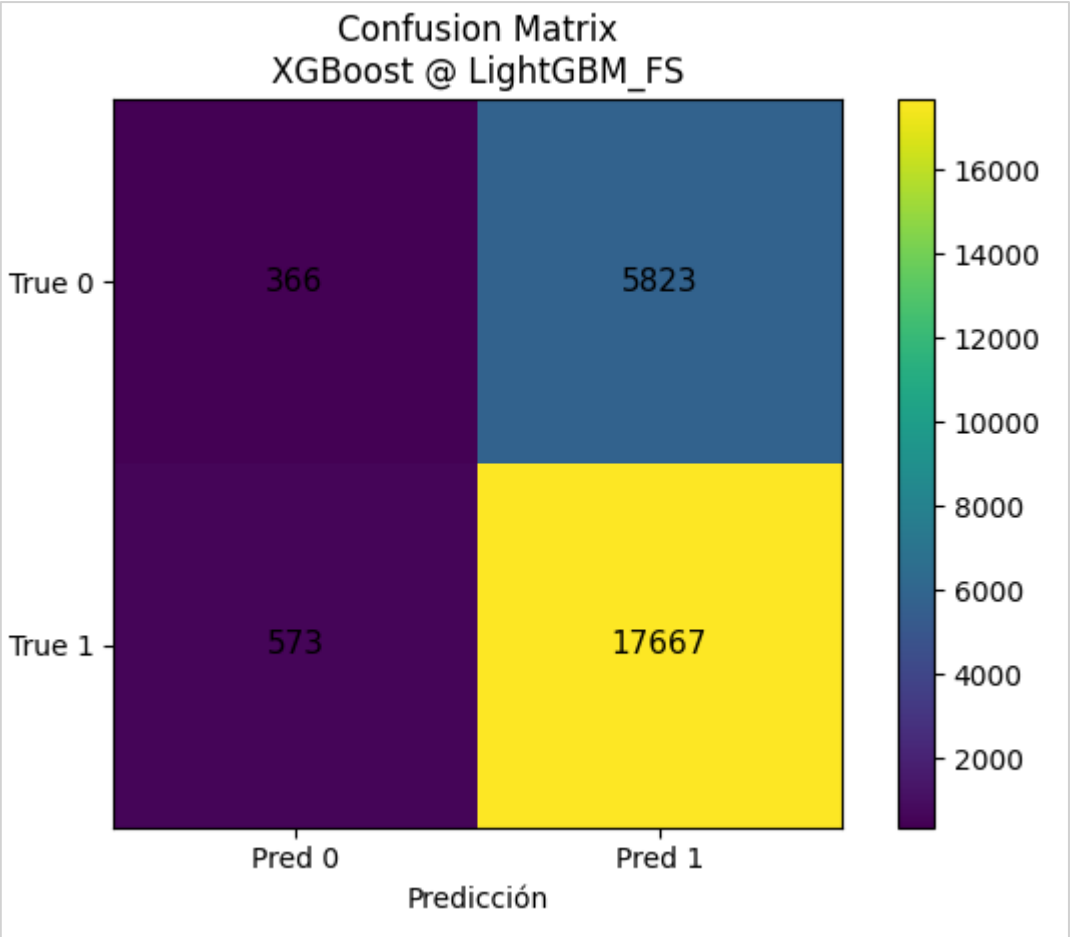
FeatureSet	Modelo	Accuracy	Precision	Recall	F1-score	AUC
LightGBM_FS	XGBoost	0.738180	0.752107	0.968586	0.846729	0.656418

Nota. Elaboración propia.

El *recall* de 97% muestra una alta capacidad del modelo para identificar de manera correcta a los niños con esquema de vacunación completo, maximizando la reducción los falsos negativos. Esto es importante en el contexto de salud pública, ya que permite una detección casi total de los casos positivos. Asimismo, el valor de *precision* de 75% refleja que, aunque el modelo predice correctamente la mayoría de los casos positivos, aún se dan algunos falsos positivos. El *F1-score* obtenido fue de 85%, lo cual evidencia un equilibrio adecuado entre ambos indicadores.

Figura 35

Matriz de confusión de modelo óptimo

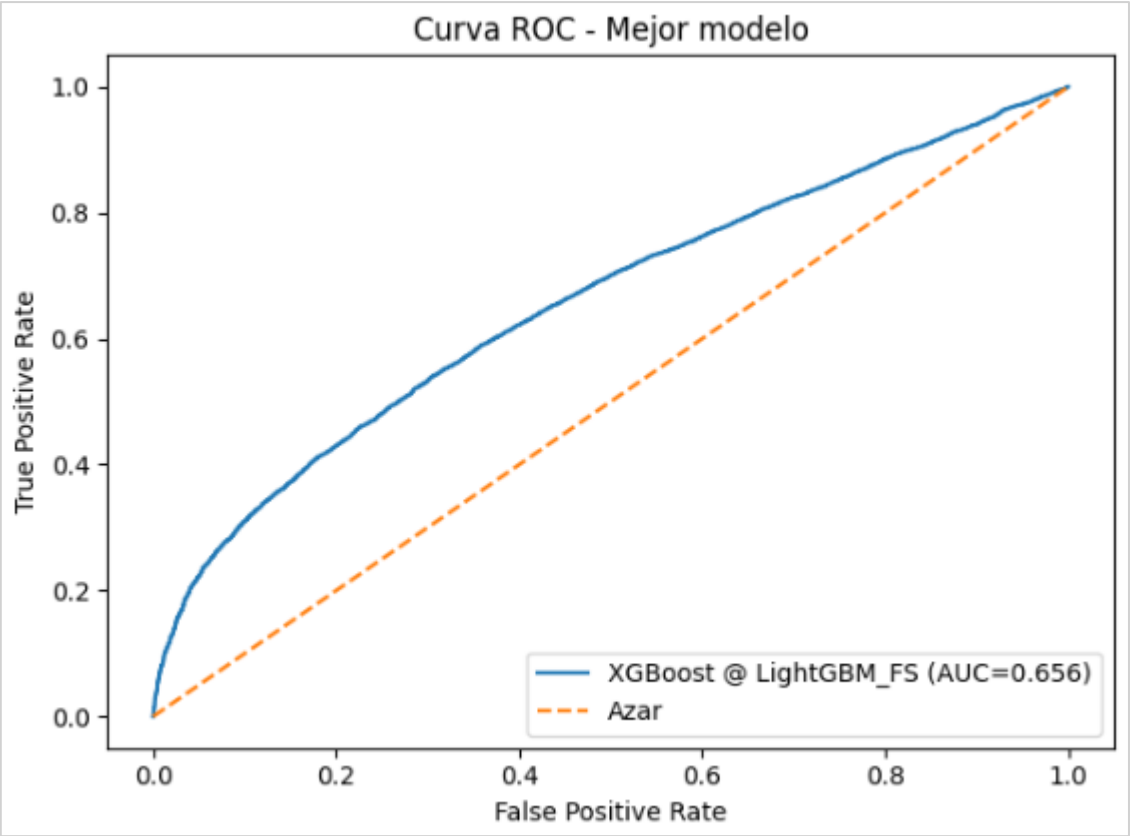


Nota. Elaboración propia.

En la Figura 35 se muestra la matriz de confusión del modelo que obtuvo los mejores resultados a nivel de métricas de desempeño, la cual confirma la interpretación de que el modelo identificó correctamente 17,667 casos positivos (esquema de vacunación incompleta).

Figura 36

Curva ROC de modelo óptimo



Nota. Elaboración propia.

La Figura 36 muestra la curva ROC del modelo XGBoost con la técnica de selección de características LightGBM_FS, donde se observa un área bajo la curva (AUC = 66%).

Este valor indica una capacidad de discriminación moderada del modelo para diferenciar entre los casos con esquema de vacunación completo e incompleto.

Aunque el desempeño no es perfecto, la curva se mantiene consistentemente por encima de la línea diagonal de azar, lo que denota que el modelo tiene una capacidad predictiva superior a la clasificación aleatoria y puede ser utilizado como una herramienta de apoyo en la detección de posibles incumplimientos del esquema de vacunación infantil.

Actividad 2: Validación cruzada del modelo final

Al aplicar la validación cruzada estratificada al modelo con las mejores métricas, *XGBoost* con selección de características *LightGBM,* permitió evaluar su estabilidad y capacidad de generalización. Se utilizó una configuración de 5 pliegues (*folds*) (k=5), con el fin de obtener una estimación más robusta del rendimiento promedio del clasificador.

Los resultados obtenidos evidenciaron un desempeño consistente, con valores promedio de 78% de *Accuracy*, 75% de *Precision*, 83% de *Recall*, 79% de *F1-score* y 86% de *ROC-AUC*.

Las desviaciones estándar fueron muy bajas (inferiores al 0.003 en todas las métricas), lo que demuestra una baja variabilidad entre pliegues y confirma la estabilidad del modelo ante diferentes subconjuntos de datos.

Los resultados de la validación cruzada permitieron confirmar la fiabilidad de las métricas del modelo *XGBoost* con *LightGBM_FS*, evidenciando un comportamiento estable frente a diferentes divisiones de los datos.

Si bien el *Recall* se redujo ligeramente en comparación con la prueba inicial, el aumento del *ROC_AUC* (de 0.65 a 0.86) y la consistencia de las demás métricas demuestran que el modelo mantiene un rendimiento estable y generalizable, lo que valida su capacidad para clasificar adecuadamente el estado de vacunación infantil y apoyar la identificación de posibles casos de incumplimiento del esquema durante el primer año de vida.

En conjunto, los resultados confirman que el modelo *XGBoost* indica que el modelo presenta un rendimiento consistente y una capacidad de generalización confiable, por lo que se considera adecuado para apoyar la identificación de niños con riesgo de no completar su esquema de vacunación durante el primer año de vida.

Capítulo 6: Conclusiones y Recomendaciones

Conclusiones

La investigación permitió desarrollar un modelo de clasificación del estado de vacunación infantil en el primer año de vida en la jurisdicción de una Dirección de Redes Integradas de Salud de Lima Metropolitana, utilizando técnicas de Machine Learning. Para ello, se integraron datos provenientes de los sistemas HIS-MINSA y el Padrón Nominal, complementados con análisis exploratorios que facilitaron la construcción de una base representativa y confiable. Esta base de datos sirvió como insumo fundamental para el análisis y entrenamiento de modelos de clasificación con información real y completa.

El análisis del impacto del preprocesamiento evidenció que las etapas de limpieza, imputación de valores faltantes, codificación de variables, balanceo de clases y escalado fueron determinantes para mejorar la calidad de los datos y optimizar el rendimiento de los algoritmos. Estas acciones contribuyeron a reducir sesgos y a fortalecer la capacidad predictiva de los modelos, asegurando resultados más robustos y generalizables.

Asimismo, la aplicación de técnicas de selección de características como *Random Forest*, *AdaBoost* y *LightGBM* demostró que la reducción de variables irrelevantes y la priorización de aquellas con mayor poder explicativo incrementan significativamente la eficiencia y precisión de los modelos. Destacó particularmente el uso de *LightGBM* como estrategia para identificar un subconjunto óptimo de variables, mejorando el desempeño general de los clasificadores.

Se utilizaron siete modelos de *machine learning*: regresión logística, KNN, árboles de decisión, Random Forest, *AdaBoost*, *XGBoost* y *Bagging*, evaluados con y sin ajuste de hiperparámetros. Por otro lado, el modelo *XGBoost* combinado con la técnica de selección de características mediante *LightGBM* obtuvo el mejor desempeño. Este modelo alcanzó métricas sólidas, con un accuracy de 74%, *precision* de 75%, *recall* de 97%, *F1-score* de 85% y un AUC de 66%, lo que evidencia su alta capacidad para clasificar correctamente el estado de vacunación infantil en el primer año de vida, mostrando un equilibrio adecuado entre exactitud, sensibilidad y capacidad discriminativa. Además, el modelo fue validado con un enfoque de validación cruzada, garantizando la estabilidad y consistencia de su desempeño.

Recomendaciones

Implementar el modelo predictivo dentro del sistema de información institucional, de manera que permita identificar de forma temprana a los niños con riesgo de incumplimiento del esquema de vacunación y generar alertas automáticas para su seguimiento.

Actualizar periódicamente el modelo con nuevos datos provenientes del Sistema HIS MINSA y del Padrón Nominal, asegurando la adaptación del clasificador a posibles cambios en las condiciones sociodemográficas o en los patrones de atención.

Ampliar la base de datos a otras DIRIS o DIRESAS para evaluar la capacidad del modelo de generalizar en diferentes contextos poblacionales y validar su escalabilidad a nivel Lima Metropolitana o a nivel nacional.

Incorporar nuevas variables complementarias, como indicadores de acceso geográfico, factores socioeconómicos o continuidad de la atención, entre otros.

REFERENCIAS

Ao, J., Ye, Z., Li, W. et al. (2024). *Impressions of Guangzhou city in Qing dynasty export paintings in the context of trade economy: a color analysis of paintings based on k-means clustering algorithm*. Herit Sci 12, 77. https://doi.org/10.1186/s40494-024-01195-4

Dirección de Redes Integradas de Salud Lima Centro [DIRIS Lima Centro]. (12 de mayo de 2020). *Resolución Directoral N.° 141-2020-DG-DIRIS-LC: Aprueba el Mapa de Procesos y las Fichas Técnicas Nivel 0 de la DIRIS Lima Centro [Norma legal]. Plataforma digital única del Estado peruano — Gob.pe*. https://www.gob.pe/institucion/dirislimacentro/normas-legales/5863851-141-2020-dg-diris-lc Gobierno del Perú

Dirección de Redes Integradas de Salud Lima Centro [DIRIS Lima Centro]. (2024, 13 de septiembre). *Resolución Directoral N.° 894-2024-DG-DIRIS-LC: Código de conducta de la Dirección de Redes Integradas de Salud Lima Centro [Norma legal]. Plataforma digital única del Estado peruano — Gob.pe*. https://www.gob.pe/institucion/dirislimacentro/normas-legales/6002462-894-2024-dg-diris-lc

Dirección de Redes Integradas de Salud Lima Centro [DIRIS Lima Centro]. (2025, 4 de abril). *Resolución Administrativa N.° 051-2025-DA-DIRIS-LC, que aprueba la décima modificación del Plan Anual de Contrataciones 2025. Plataforma digital única del Estado peruano — Gob.pe*. https://www.gob.pe/institucion/dirislimacentro/normas-legales/6666517-051-2025-da-diris-lc

Dirección de Redes Integradas de Salud Lima Centro [DIRIS Lima Centro]. (30 de diciembre de 2024). *Resolución Directoral N.° 1378-2024-DG-DIRIS-LC: Plan Operativo Institucional (POI) Anual 2025 [Norma legal]. Plataforma digital única del Estado peruano — Gob.pe*. https://www.gob.pe/institucion/dirislimacentro/normas-legales/6357918-1378-2024-dg-diris-lc

Dirección de Redes Integradas de Salud Lima Centro [DIRIS Lima Centro]. (s. f.). *Información institucional. Plataforma digital única del Estado peruano — Gob.pe*. https://www.gob.pe/institucion/dirislimacentro/institucional

Dirección de Redes Integradas de Salud Lima Centro [DIRIS Lima Centro]. (s. f.). *Organigrama institucional [PDF]. Plataforma digital única del Estado peruano — Gob.pe*. Recuperado el 18 de octubre de 2025, de https://www.gob.pe/rails/active_storage/blobs/redirect/eyJfcmFpbHMiOnsiZGF0YSI6MTQ0MTY0LCJwdXIiOiJibG9iX2lkIn19--bb81595b392460425ad993878eded682e0b7d0b0/Organigrama%20Dirs%20Lima%20Centro.pdf

Dirección de Redes Integradas de Salud Lima Centro [DIRIS Lima Centro]. (s. f.). *Organización. Plataforma digital única del Estado peruano — Gob.pe*. https://www.gob.pe/institucion/dirislimacentro/organizacion

Dirección de Redes Integradas de Salud Lima Centro. (s. f.). *Institucional*. Recuperado el 14 de junio de 2017, de https://www.gob.pe/institucion/dirislimacentro/institucional.

Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (3rd ed.). O'Reilly Media.

Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep Learning*, The MIT Press.

Harrington, P. (2012). *Machine Learning in Action*. Manning.

Mahajan, P. (2022). *Artificial Intelligence in Healthcare*.

Mesa de Concertación para la Lucha contra la Pobreza & Ministerio de Salud del Perú. (2024, 18 de abril). *Balance de los Avances en la Vacunación Nacional 2023 y Desafíos para el año 2024* [Documento de trabajo]. Recuperado de https://www.mesadeconcertacion.org.pe/storage/documentos/2024-04-23/minsa-reunion-mclcp-18-de-abril.pdf

Ministerio de Economía y Finanzas [MEF]. (2024, 15 de noviembre). *Actualización a la versión 24.02.02.u1 del Sistema Integrado de Gestión Administrativa – SIGA MEF. Gob.pe*. https://www.gob.pe/institucion/mef/noticias/1058191-actualizacion-a-la-version-24-02-02-u1-del-sistema-integrado-de-gestion-administrativa-siga-mef

Ministerio de Salud [MINSA]. (16 de junio de 2017). *Resolución Ministerial N.° 467-2017/MINSA: Aprueban el Manual de Operaciones de las Direcciones de Redes Integradas de Salud* [Norma legal]. https://www.gob.pe/institucion/minsa/normas-legales/189346-467-2017-minsa

Ministerio de Salud [MINSA]. (30 de julio de 2024). *Vacunas del esquema nacional de vacunación en el Perú. Plataforma digital única del Estado peruano — Gob.pe*. https://www.gob.pe/22037-vacunas-del-esquema-nacional-de-vacunacion-en-el-peru

Ministerio de Salud del Perú. (2005). *Directiva Nº 033-MINSA/DGSP* [PDF]. Recuperado e https://www.minsa.gob.pe/Recursos/OTRANS/01InformacionInst/archivolegaldigital/Directiva2005/D033-MINSA-DGSP.PDF

Ministerio de Salud. (2017). *Resolución Ministerial N° 467-2017/MINSA: Crean la Dirección de Redes Integradas de Salud (DIRIS) Lima Centro*. Recuperado de https://www.gob.pe/institucion/minsa/normas-legales/189346-467-2017-minsa.

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill Science.

Organización Mundial de la Salud (2024). *Vaccines and immunization: What is vaccination?* Recuperado de https://www.who.int/news-room/questions-and-answers/item/vaccines-and-immunization-what-is-vaccination

Pan American Health Organization. (s. f.). *Immunization*. Recuperado de https://www.paho.org/en/topics/immunization

Perú, Ministerio de Economía y Finanzas [MEF]. (2019, 13 de marzo). *Decreto Supremo N.° 082-2019-EF, que aprueba el Texto Único Ordenado de la Ley N.° 30225, Ley de Contrataciones del Estado. Plataforma digital única del Estado peruano — Gob.pe*. https://www.gob.pe/institucion/mef/normas-legales/266672-082-2019-ef Gobierno del Perú

Perú, Ministerio de Economía y Finanzas [MEF]. (31 de diciembre de 2018). *Decreto Supremo N.° 344-2018-EF, que aprueba el Reglamento de la Ley N.° 30225, Ley de Contrataciones del Estado. Plataforma digital única del Estado peruano — Gob.pe*. https://www.gob.pe/institucion/mef/normas-legales/235964-344-2018-ef

Poder Judicial del Perú. (s. f.). *Metodología para la implementación de la Gestión por Procesos* [PDF]. https://www.pj.gob.pe/wps/wcm/connect/0d999d80408093a7aba8ef9515c1560a/2.%2BMetodologia_de_GxP.pdf?MOD=AJPERES

Presidencia del Consejo de Ministros [PCM]. (13 de octubre de 2025). *Gestión por Procesos en Entidades Públicas. Plataforma digital única del Estado peruano — Gob.pe*.

https://www.gob.pe/22194-gestion-por-procesos-en-entidades-publicas Gobierno del Perú

Russel, S. J., Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.

ANEXO N° 01

Solicitud acceso a la base de datos de inmunizaciones (2020 – Agosto 2025)

PERÚ | Ministerio de Salud | Dirección de Redes Integradas de Salud Lima Centro

"Decenio de la Igualdad de Oportunidades para Mujeres y Hombres"
"Año de la recuperación y consolidación de la economía peruana"

Solicitud: Solicito acceso a la base de datos del de Inmunizaciones (2020 - 31 de agosto del 2025)

DR. PEDRO ALEJANDRO CRUZADO PUENTE

DIRECTOR GENERAL

DIRECCIÓN DE REDES INTEGRADAS DE SALUD LIMA CENTRO

ATENCIÓN : OFICINA DE EPIDEMIOLOGÍA, INTELIGENCIA SANITARIA Y DOCENCIA E INVESTIGACIÓN

De: Antonio Ruis Pelaez Flores identificado con Código de Estudiante: 20100654, Alocén Tito Claudia Vanessa, Código de Estudiante: 20200051, Díaz Espíritu Jorge Manuel, Código de Estudiante: 20100658:

Fecha: Lima, 22 de septiembre de 2025

Asunto: Solicitud de acceso a la base de datos de Inmunizaciones

Estimado Dr. Cruzado:

Me dirijo a usted con el debido respeto para solicitar acceso a la base de datos de Inmunizaciones correspondiente al período comprendido entre el año 2020 y el 31 de agosto de 2025, las siguientes variables:

N°	Variables
1	RIS_RN
2	EESS_PN
3	DIRECCION_PN
4	EESS_HIS_O_EESS_CPRO1
5	DIREC_HIS_DIREC_CPRO1
6	FECHA_NACIMIENTO_PN
7	GENERO
8	GRUPO_DE_EDAD
9	CELULAR
10	IPV_1¬ dosis
11	IPV_2¬ dosis
12	IPV_3¬ dosis
13	PENTA_1¬ dosis
14	PENTA_2¬ dosis
15	PENTA_3¬ dosis
16	NEUMOCOCO_1¬ dosis
17	NEUMOCOCO_2¬ dosis
18	ROTAVIRUS_1¬ dosis
19	ROTAVIRUS_2¬ dosis
20	INFLUENZA_3-_1¬ dosis
21	INFLUENZA_3-_2¬ dosis
22	SPR_1¬ dosis
23	SPR_2¬ dosis
24	DPT_1¬ Dosis refuerzo
25	IPV_1¬ Dosis refuerzo
26	INFLUENZA_3+_Dosis ·nica

PERÚ | Ministerio de Salud | Dirección de Redes Integradas de Salud Lima Centro

"Decenio de la Igualdad de Oportunidades para Mujeres y Hombres"
"Año de la recuperación y consolidación de la economía peruana"

27	HAV_Dosis ·nica
28	VARICELA_Dosis ·nica
29	DPT_2¬ Dosis refuerzo
30	IPV_2¬ Dosis refuerzo
31	esquema_completo
32	CNVLC_Ubigeo_LugarNacido
33	CNVLC_DPTO_EESS
34	CNVLC_PROV_EESS
35	CNVLC_DIST_EESS
36	CNVLC_Ipress
37	CNVLC_CO_LOCAL
38	CNVLC_Nombre_EESS
39	CNVLC_Diresa_Diris
40	CNVLC_Institucion
41	CNVLC_Categoria
42	CNVLC_PERIODO
43	CNVLC_FE_NACIDO
44	CNVLC_PESO_NACIDO
45	CNVLC_TALLA_NACIDO
46	CNVLC_APGAR_5_NACIDO
47	CNVLC_DUR_EMB_PARTO
48	CNVLC_Condicion_Parto
49	CNVLC_sexo_nacido
50	CNVLC_Tipo_Parto
51	CNVLC_Financiador_Parto
52	CNVLC_Ligadura_corte
53	CNVLC_Malformacion_Congenita
54	CNVLC_Lactancia_Precoz
55	CNVLC_PERCEF
56	CNVLC_PERTOR
57	CNVLC_Lugar_Nacido
58	CNVLC_Estado_Civil
59	CNVLC_Nivel_Intruccion_Madre
60	CNVLC_Num_embar_madre
61	CNVLC_Hijos_vivo_madre
62	CNVLC_Dpto_Madre
63	CNVLC_Prov_Madre
64	CNVLC_Dist_Madre
65	CNVLC_Ubigeo_DOM_Madre
66	CNVLC_Edad_Madre

En el ámbito de la Dirección de Redes Integradas de Salud Lima Centro. Esta solicitud se realiza en el marco de una investigación académica de tesis (TSP) y artículo científico que estamos desarrollando en la Universidad ESAN.

Agradeceríamos que la información solicitada fuera remitida en formato Excel a los siguientes correos electrónicos: 20100654@ue.edu.pe, 20200051@ue.edu.pe, 20100658@ue.edu.pe.

PERÚ | Ministerio de Salud | Dirección de Redes Integradas de Salud Lima Centro

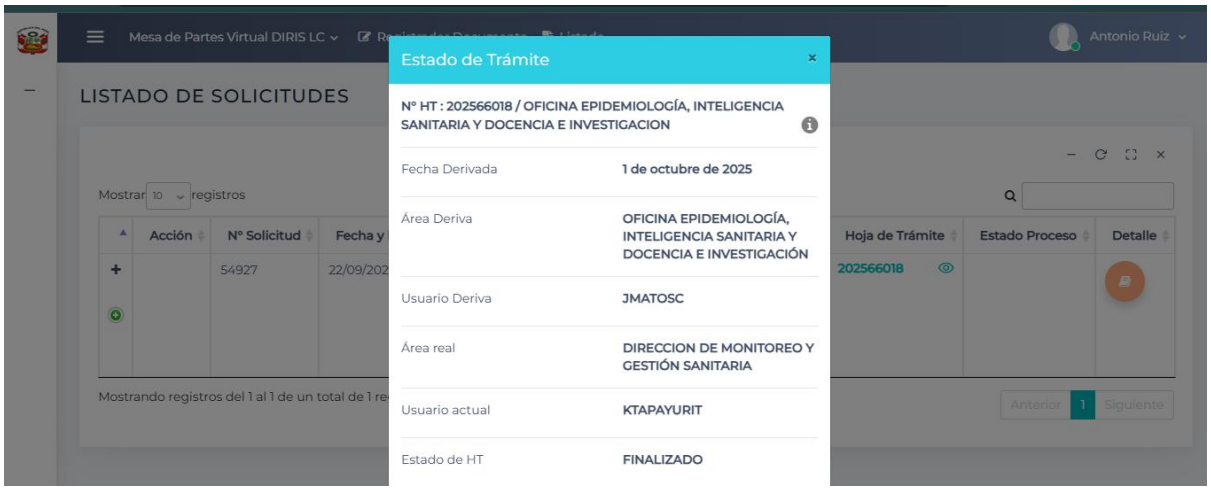
"Decenio de la Igualdad de Oportunidades para Mujeres y Hombres"
"Año de la recuperación y consolidación de la economía peruana"

Por lo expuesto, confiamos en que se autoricen las coordinaciones necesarias para el acceso a dicha información y quedamos atentos a cualquier requerimiento adicional.

Agradecemos de antemano su atención y apoyo en este importante proceso de investigación.

Atentamente,

Antonio Ruis Pelaez Flores
DNI: 48488815
Celular: 982349650
Correo electrónico: cosme_pelaez@hotmail.com, 20100654@ue.edu.pe


Mesa de Partes Virtual DIRIS LC
Antonio Ruiz
LISTADO DE SOLICITUDES
Mostrar 10 registros
Acción
Nº Solicitud
Fecha y
Hoja de Trámite
Estado Proceso
Detalle
54927
22/09/202
202566018
Mostrando registros del 1 al 1 de un total de 1 re
Anterior
1
Siguiente
Estado de Trámite
Nº HT : 202566018 / OFICINA EPIDEMIOLOGÍA, INTELIGENCIA SANITARIA Y DOCENCIA E INVESTIGACION
Fecha Derivada
1 de octubre de 2025
Área Deriva
OFICINA EPIDEMIOLOGÍA, INTELIGENCIA SANITARIA Y DOCENCIA E INVESTIGACIÓN
Usuario Deriva
JMATOSC
Área real
DIRECCION DE MONITOREO Y GESTIÓN SANITARIA
Usuario actual
KTAPAYURIT
Estado de HT
FINALIZADO

DECLARACIÓN DE VALIDACIÓN DEL ETIQUETADO DEL DATASET

Yo, **Lic. Kerly Rocío Bazalar Gonzales**, Coordinadora de Desarrollo Infantil Temprano del Hospital San Juan de Lurigancho del Ministerio de Salud, en pleno uso de mis facultades y funciones, declaro haber revisado y validado el **correcto etiquetado de los registros sobre el estado de vacunación infantil**, realizado por los Bachilleres en Ingeniería de Sistemas de la Universidad ESAN: Claudia Vanessa Alocén Tito, Jorge Manuel Díaz Espíritu y Antonio Ruis Pelaez Flores.

Dicha labor de etiquetado constituye un insumo fundamental para el desarrollo del Trabajo de Suficiencia Profesional titulado **"Modelo de Clasificación del Estado de Vacunación Infantil en el Primer Año de Vida en la Jurisdicción de la DIRIS Lima Centro utilizando Técnicas de Machine Learning"**, garantizando la confiabilidad y consistencia de los datos empleados.

Lima, 25 de septiembre del 2025

MINISTERIO DE SALUD
Dirección de Redes Integradas de Salud Lima Centro
HOSPITAL SAN JUAN DE LURIGANCHO
LIC. KERLY BAZALAR GONZALES
COORDINADORA DE CRED/ESNI
C.E.P. 80424

Lic. Kerly Rocío Bazalar Gonzales

DNI N° 47635331

ANEXO N° 02

Problema General	Objetivo General	Hipótesis General	Variables de estudio	Instrumentos
¿Es posible clasificar el estado de vacunación infantil en el primer año de vida en la jurisdicción de una dirección de Redes Integradas de Salud de Lima Metropolitana Utilizando técnicas de Machine Learning?	Desarrollar un modelo de clasificación del estado de vacunación infantil en el primer año de vida en la jurisdicción de una Dirección de Redes Integradas de Salud de Lima Metropolitana Utilizando técnicas de Machine Learning.	Las técnicas de machine learning permiten clasificar de manera precisa el estado de vacunación infantil en el primer año de vida en la jurisdicción de una Dirección de Redes Integradas de Salud de Lima Metropolitana.	Variable Independiente: Técnicas de Machine Learning	Base de datos
Problemas Específicos	**Objetivos Específicos**	**Hipótesis Específicas**	**Variables de estudio**	
PE1: ¿Existen suficientes datos en el sistema His Minsa y Padrón Nominal sobre el estado de vacunación para la obtención del dataset? PE2: ¿Realizar el preprocesamiento que permite un mejor entrenamiento de un modelo? PE3: ¿Qué técnicas de selección de características resultan útiles para la clasificación? PE4: ¿Qué técnicas de machine learning se pueden aplicar para construir un modelo de clasificación del estado de vacunación infantil en el primer año de vida? PE5: ¿Qué métricas existen para evaluar desempeño de un modelo de clasificación del estado de vacunación infantil en el primer año de vida?	OE1: Obtener un dataset de infantil en el primer año de vida sobre estado de vacunación. OE2: Analizar el impacto del preprocesamiento en el entrenamiento de un modelo para clasificar el estado de vacunación. OE3: Aplicar técnicas de selección de características para mejorar la clasificación del estado de vacunación infantil en el primer año de vida. OE4: Utilizar técnicas de clasificación de machine learning del estado de vacunación infantil en el primer año de vida. OE5: Utilizar métricas para evaluar el desempeño del modelo de clasificación del estado de vacunación infantil en el primer año de vida.	HE1: La disponibilidad de datos en el sistema HIS y padrón nominal que permita construir un dataset adecuado para clasificar el estado de vacunación. HE2: Realizar el preprocesamiento influye significativamente en el entrenamiento efectivo de un modelo de clasificación de estado de vacunación. HE3: La aplicación de técnicas de selección de características mejora la capacidad de generalización de los modelos utilizados para clasificar el estado de vacunación infantil en el primer año de vida. HE4: Las técnicas avanzadas de Machine Learning permiten desarrollar un modelo preciso para clasificar el estado de vacunación infantil en el primer año de vida. HE4: La utilización de métricas específicas y relevantes mejora la evaluación del desempeño del modelo de clasificación del estado de vacunación.	Variable Dependiente: Estado de vacunación infantil	