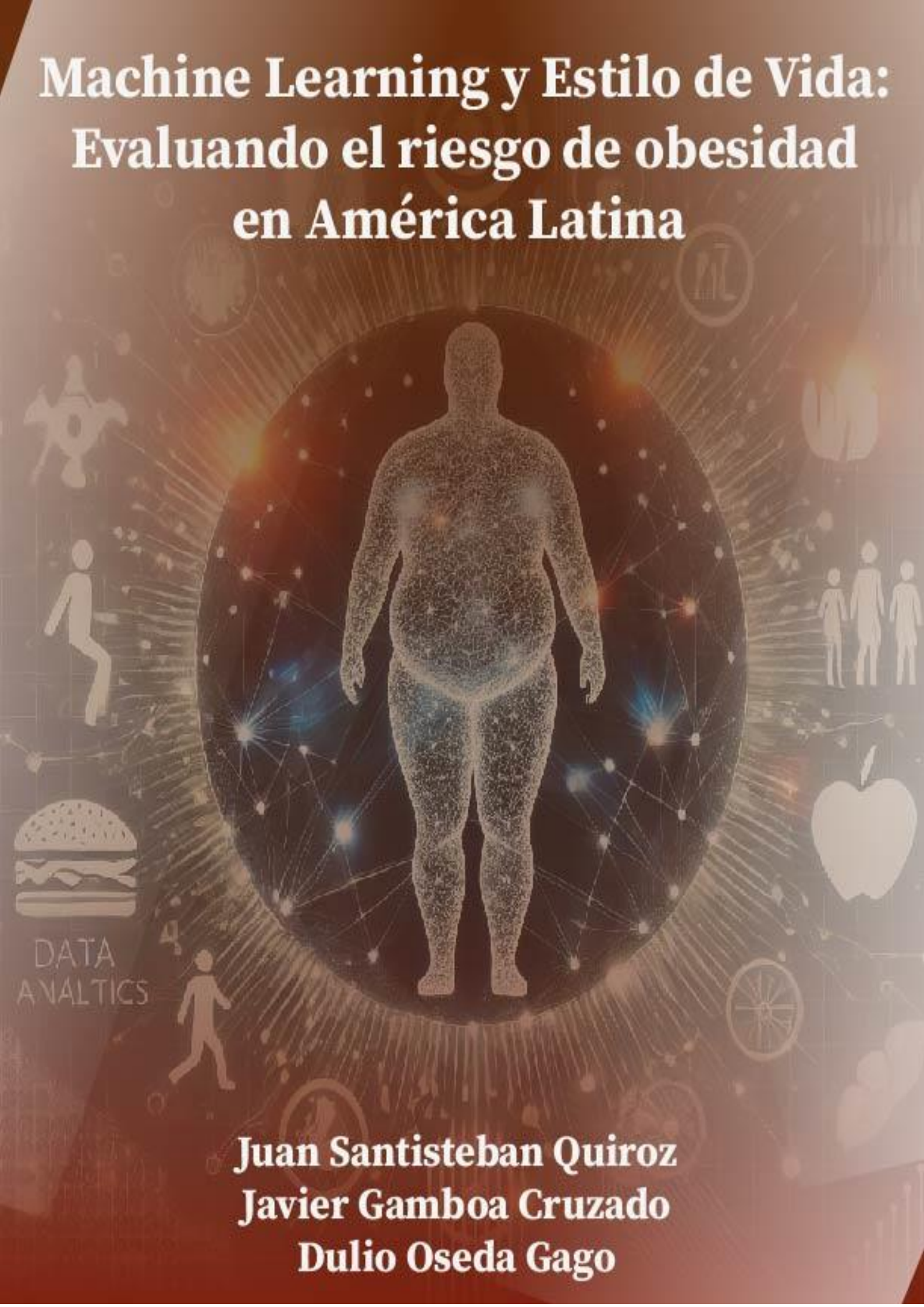
Machine Learning y Estilo de Vida: Evaluando el riesgo de obesidad en América Latina



Juan Santisteban Quiroz
Javier Gamboa Cruzado
Dulio Oseda Gago

MACHINE LEARNING Y ESTILO DE VIDA: EVALUANDO EL RIESGO DE OBESIDAD EN AMÉRICA LATINA

AUTORES:

Juan Santisteban Quiroz

Javier Gamboa Cruzado

Dulio Oseda Gago

La presente obra fue revisada por 2 pares académicos externos ciegos conforme al proceso editorial del Centro de Investigación Latinoamericano para el Desarrollo e Innovación CILADI.

Los rigurosos procedimientos editoriales de CILADI garantizan la selección de manuscritos por sus aportes significativos al conocimiento y cualidades científicas.

Todas las obras publicadas por CILADI cuentan con ISBN y se encuentran disponibles en la web (www.ciladi.org)

-Editorial-
CILADI
Centro de Investigación Latinoamericano
para el Desarrollo e Innovación

Centro de Investigación Latinoamericano

para el Desarrollo e Innovación

Guayaquil- Ecuador

https://ciladi.org/

AÑO 2024

ISBN: 978-9942--7292-5-5

Prólogo

"Machine Learning y Estilo de Vida: Evaluando el Riesgo de Obesidad en América Latina" es una obra que traza un camino innovador en la intersección entre tecnología y salud pública. A través de la combinación de técnicas avanzadas de aprendizaje automático y datos epidemiológicos, el libro ofrece una perspectiva transformadora sobre cómo el estilo de vida afecta el riesgo de obesidad en las poblaciones de Colombia, México y Perú. Con un enfoque pionero, esta investigación establece un modelo que no solo profundiza en la comprensión de factores de riesgo críticos, sino que también redefine la precisión y accesibilidad en evaluaciones de salud.

La metodología presentada aquí es mucho más que un avance académico; es una herramienta práctica diseñada para optimizar los recursos en entornos donde la necesidad de soluciones eficientes es primordial. Al emplear algoritmos de última generación, este libro no solo brinda resultados precisos, sino que también establece un estándar para futuras investigaciones y aplicaciones en el ámbito de la salud y la nutrición.

Este libro se convierte en un referente en el uso de la tecnología para abordar los desafíos de salud pública, inspirando tanto a académicos como a profesionales de la salud a integrar enfoques innovadores en su trabajo diario. Al explorar el potencial del machine learning en un contexto real y urgente, esta obra no solo ilumina el camino hacia soluciones más eficaces, sino que también fomenta un cambio positivo hacia una mayor comprensión y prevención de la obesidad en América Latina.

PhD. Antonio Poveda G.
Editor

ÍNDICE GENERAL

RESUMEN

En esta investigación se exploró la implementación y eficacia de una solución de Machine Learning para optimizar la estimación de la influencia del estilo de vida en el riesgo de obesidad en poblaciones de Colombia, México y Perú. A través de una metodología nueva y robusta, denominada DORA, se desarrolló una solución de Machine Learning que no solo logró incrementar significativamente la eficiencia de la estimación, sino que también redujo tanto el tiempo como el costo asociado a este proceso.

La solución de Machine Learning, implementada mediante tecnologías avanzadas como Java 18 y Spring Boot para el Back-end y React para el Front-end, y hospedada en plataformas de alta disponibilidad como Render y Vercel, demostró ser una herramienta valiosa en el contexto de la salud pública y la epidemiología.

Los hallazgos de esta investigación no solo validan las hipótesis propuestas, sino que también abren puertas a futuras investigaciones y aplicaciones prácticas en el ámbito de la salud y el bienestar poblacional, especialmente en entornos caracterizados por recursos y presupuestos limitados.

Palabras clave: Machine Learning, Estilo de Vida, Riesgo de Obesidad, Metodología DORA, Eficiencia de la Estimación, Tiempo de Estimación, Costo de Estimación

ABSTRACT

This research explored the implementation and efficacy of a Machine Learning solution to optimize the estimation of the influence of lifestyle on obesity risk in populations from Colombia, Mexico, and Peru. Through a novel and robust methodology, named DORA, an Machine Learning solution was developed which not only significantly enhanced estimation efficiency but also reduced both the time and cost associated with this process.

The Machine Learning solution, implemented using cutting-edge technologies like Java 18 and Spring Boot for the Back-end and React for the Front-end, and hosted on high-availability platforms such as Render and Vercel, proved to be a valuable tool in the context of public health and epidemiology.

The findings of this research not only validate the proposed hypothesis but also pave the way for future research and practical applications in the realm of health and population well-being, especially in environments characterized by limited resources and budgets.

Keywords: Machine Learning, Lifestyle, Obesity Risk, DORA Methodology, Estimation Efficiency, Estimation Time, Estimation Cost.

CAPÍTULO I: INTRODUCCIÓN

1.1. Situación Problemática

La obesidad se refiere a la acumulación excesiva de tejido adiposo en el cuerpo (Ferdowsy, Rahi, Jabiullah, y Habib, 2021, p. 1). Esta condición representa un desafío crítico para la salud pública y ocupa el quinto lugar en las causas principales de mortalidad a nivel mundial (The European Association for the Study of Obesity, s.f.). Es una enfermedad que afecta tanto a hombres como a mujeres, y en las últimas décadas ha mostrado un preocupante aumento. De acuerdo con las estimaciones de la Organización Mundial de la Salud (OMS), se proyecta que para el año 2030, más del 40% de la población mundial padecerá sobrepeso, y más de una quinta parte de esa población será diagnosticada con obesidad (Ramírez, Aparcana, Zamora, y Leo, 2019, p. 22).

La Organización Panamericana de la Salud (2020) informó lo siguiente:

> Sudamérica tuvo una alta prevalencia de sobrepeso infantil en 2019, alcanzando un 7.9%. En países como Chile, México, Perú y Uruguay, se están intensifican- do esfuerzos gubernamentales para reducir el consumo de alimentos no saludables mediante etiquetado frontal.
>
> Por otro lado, en Perú, Colombia y México, los índices de obesidad y sobrepeso siguen en aumento. En Perú, la mala alimentación afecta más a las zonas desfavorecidas de la sierra y la selva, con áreas como Callao, Ica, Lima, Moquegua y Tacna presentando prevalencias de sobrepeso entre 12.5% y 16%.
>
> En Colombia, áreas como Arauca, Cauca, Guaviare, Meta, San Andrés y Providencia, Valle y Vaupés enfrentan altas prevalencias de sobrepeso en menores de 5 años, con problemas simultáneos de retraso en el crecimiento. En México, regiones como Sonora, Chihuahua, Baja California, Nuevo León y Yucatán también experimentan altos índices de sobrepeso (Organización Panamericana de la Salud, 2020, pp. 24-61).

Dado que la malnutrición, el sobrepeso y la obesidad frecuentemente coexisten, resulta fundamental proporcionar soluciones holísticas. Investigadores han dedicado considerables esfuerzos a la identificación temprana de los factores que influyen en la génesis de la obesidad, incluso desarrollando herramientas en línea como el cálculo del Índice de masa corporal (IMC) (World health Organization, s.f.), mediante la cual es posible evaluar el grado de obesidad de un individuo. No obstante, estas herramientas se limitan al cálculo del IMC, excluyendo otros aspectos relevantes como: antecedentes familiares de obesidad, tiempo dedicado a la actividad física, perfiles de expresión genética, pautas alimenticias, medicación, estatus socioeconómico, influencias psicológicas y sociales, entre otros elementos. Los autores Safaei, Sundararajan, Driss, Boulila, y Shapi'i (2021) concluyen que con frecuencia estos factores conducen al sobrepeso y la obesidad. Asimismo, el autor Salahuddin (2018) sostiene que debido al impacto del estilo de vida y la influencia de internet, las personas tienden a consumir en mayor medida alimentos rápidos y otros productos poco saludables (p. 1).

El sobrepeso y la obesidad constituyen destacadas patologías relacionadas con el estilo de vida, llevando a diversas complicaciones de salud y contribuyendo al desarrollo de numerosas enfermedades crónicas, entre ellas cáncer, diabetes, síndrome metabólico y enfermedades cardiovasculares (Safaei et al., 2021, p. 1). En virtud de la información previa, no sería exagerado afirmar que la obesidad se perfila como una amenaza principal para la humanidad.

En consecuencia, resulta imperativo disponer de una herramienta capaz de detectar y diagnosticar tempranamente la obesidad, que permita evaluar el grado de influencia de diversos factores del estilo de vida que inciden en esta enfermedad. Esta herramienta sería de gran utilidad tanto para las autoridades de salud pública como para los profesionales médicos y la población en general, en su lucha conjunta por prevenir y atenuar esta problemática.

1.2. Formulación del Problema

1.2.1. *Problema General*

¿De qué manera la aplicación de una solución de Machine Learning contribuye en la estimación de la influencia del Estilo de Vida en el riesgo de obesidad de la población de Colombia, México y Perú?

1.2.2. *Problemas Específicos*

- ¿De qué manera la aplicación de una solución de Machine Learning incrementa la eficiencia de la estimación de la influencia del Estilo de Vida en el riesgo de obesidad de la población de Colombia, México y Perú?
- ¿De qué manera la aplicación de una solución de Machine Learning reduce el tiempo de la estimación de la influencia del Estilo de Vida en el riesgo de obesidad de la población de Colombia, México y Perú?
- ¿De qué manera la aplicación de una solución de Machine Learning reduce el costo de la estimación de la influencia del Estilo de Vida en el riesgo de obesidad de la población de Colombia, México y Perú?

1.3. Justificación

1.3.1. *Justificación Teórica*

El empleo de técnicas de aprendizaje automático brinda una capacidad predictiva considerablemente más robusta en contraste con enfoques simples como la regresión lineal, o métodos estadísticos como la fórmula del Índice de masa corporal. Las técnicas de aprendizaje automático, tales como redes neuronales, árboles de decisión, bosques aleatorios y aprendizaje profundo, representan algunos de los métodos más efectivos para posibilitar la identificación temprana y la gestión clínica de la obesidad.

1.3.2. *Justificación Práctica*

El interés en la prevención y mitigación de la obesidad en las poblaciones de Colombia, México y Perú es considerable. Contar con información pertinente acerca de la influencia de los factores del Estilo de Vida en el riesgo de obesidad adquiere suma relevancia, ya que beneficia la toma de decisiones tanto de las autoridades de salud pública como de los profesionales médicos y la población en general, con respecto a las estrategias para la prevención y el control de esta enfermedad.

1.4. Objetivos

1.4.1. *Objetivo General*

Desarrollar una solución de Machine Learning que permita optimizar la estimación de la influencia del Estilo de Vida en el riesgo de obesidad de la población de Colombia, México y Perú.

1.4.2. *Objetivos Específicos*

- Incrementar la eficiencia de la estimación de la influencia del Estilo de Vida en el riesgo de obesidad de la población de Colombia, México y Perú.
- Disminuir el tiempo de la estimación de la influencia del Estilo de Vida en el riesgo de obesidad de la población de Colombia, México y Perú.
- Disminuir el costo de la estimación de la influencia del Estilo de Vida en el riesgo de obesidad de la población de Colombia, México y Perú.

1.5. Hipótesis

1.5.1. *Hipótesis General*

La aplicación de una solución de Machine Learning contribuye significativamente en la estimación de la influencia del Estilo de Vida en el riesgo de obesidad de la población de Colombia, México y Perú.

1.5.2. *Hipótesis Específicas*

- El uso de una solución de Machine Learning incrementa significativamente la eficiencia de la estimación de la influencia del Estilo de Vida en el riesgo de obesidad de la población de Colombia, México y Perú.

- El uso de una solución de Machine Learning disminuye el tiempo de la estimación de la influencia del Estilo de Vida en el riesgo de obesidad de la población de Colombia, México y Perú.

- El uso de una solución de Machine Learning disminuye el costo de la estimación de la influencia del Estilo de Vida en el riesgo de obesidad de la población de Colombia, México y Perú.

CAPÍTULO II: MARCO TEÓRICO

2.1. Antecedentes de Investigación

2.1.1. *Nacionales*

En la búsqueda de abordar la problemática de la obesidad desde diversas perspectivas, varios estudios nacionales han contribuido al entendimiento de los factores asociados y las técnicas de predicción. Un ejemplo relevante es el estudio titulado «*Comparación de algoritmos de clasificación para la predicción de casos de obesidad infantil*» (Velando, Córdova, Condori Castro, Cayra, y Sulla-Torres, 2016), el cual se enfocó en identificar a personas de 6 a 17 años en riesgo de obesidad a partir de registros médicos de crecimiento, mediante algoritmos de clasificación. Para llevar a cabo esta investigación, se utilizó un conjunto de datos compuesto por, 5962 registros médicos de individuos en el rango de edad mencionado. Estos registros contenían 16 atributos relevantes para el análisis. Cuatro técnicas de aprendizaje automático fueron entrenadas en el proceso: el árbol de decisión C4.5, la máquina de vectores de soporte (SVM), el clasificador Naive Bayes y el algoritmo de propagación hacia atrás (BP). La evaluación de las técnicas se basó en criterios de exactitud, sensibilidad, especificidad y curva ROC. Los resultados obtenidos demostraron que el árbol de decisión destacó con una exactitud del 97.23%, posicionándose como la técnica más adecuada para predecir casos de obesidad a partir de los registros médicos. Estos hallazgos establecieron la relevancia de las técnicas de clasificación en la identificación temprana de personas en riesgo de obesidad. En otra investigación, Trujillo Aspilcueta (2018) se propuso identificar los factores asociados a la obesidad en colaboradores de una institución de salud. La tesis «*Factores asociados a sobrepeso y obesidad en trabajadores de una Institución Pública de Salud. Lima, Perú*» se basó en 715 registros sobre información alimenticia, actividad física y características socio-demográficas de los colaboradores. El objetivo era tomar medidas preventivas y correctivas en la salud de los colaboradores. Los resultados de esta investigación destacaron que el consumo de frutas y verduras jugaba un papel determinante en la disminución de sobrepeso y obesidad. Además, se concluyó que la frecuencia de actividad física no mostró un rol relevante en el incremento de estos problemas de salud. Estos

resultados aportaron valiosa información para enfocar estrategias de prevención en el ámbito laboral y mejorar la salud de los trabajadores. Otra tesis de relevancia en la comprensión de la obesidad es «*Encuentro clínico de los médicos con pacientes con sobrepeso y obesidad en consulta externa de un hospital público de Lima*» (Torres Nolasco, 2019). En esta investigación se exploraron las características del proceso de consejería médica en sobrepeso y obesidad en un hospital general. A través del análisis de 40 registros de audio de consultas, se examinaron aspectos como la duración de las consejerías, los factores tratados en las consultas y las recomendaciones médicas. Los resultados obtenidos de esta investigación reflejaron que el tiempo promedio de la consejería fue de 3 minutos y 9 segundos. Además, se observó que el 65% de los médicos se centró en temas relacionados con el peso y la nutrición, mientras que el 35% abordó aspectos vinculados a la actividad física. Estos hallazgos proporcionaron ideas valiosas sobre la práctica médica en relación con la obesidad y sugieren la necesidad de un abordaje más completo en las consultas. En línea con el enfoque de modelos computacionales, Sulla Torres (2018) presentó la tesis «*Modelo híbrido de árbol de decisión difusa con optimización por enjambre de partículas para clasificación de Obesidad Escolar*». Esta investigación propuso un modelo de inteligencia artificial que fusionó árbol de decisión, lógica difusa y optimización por enjambre de partículas (PSO) para clasificar la obesidad escolar. Los datos antropométricos de una base de datos de escolares se utilizaron para analizar la relación entre la interpretabilidad y la exactitud del modelo. Los resultados destacaron que el modelo híbrido logró una exactitud del 84% en la clasificación de obesidad en el sexo masculino y del 89% en el sexo femenino. Estos hallazgos subrayan el potencial de los enfoques híbridos en la predicción de la obesidad, demostrando la viabilidad de incorporar diferentes técnicas de análisis para obtener resultados más precisos y significativos. En resumen, estos estudios nacionales contribuyeron significativamente a la comprensión de los factores asociados a la obesidad, así como a la identificación y evaluación de técnicas de predicción y análisis de datos. Los hallazgos obtenidos de estas investigaciones sirven como base sólida para la presente investigación, permitiendo aprovechar y expandir el conocimiento acumulado en la lucha contra la obesidad.

2.1.2. *Internacionales*

En el ámbito de la investigación sobre la obesidad y su predicción mediante el uso de técnicas de aprendizaje automático, diversos autores han abordado esta problemática desde enfoques y metodologías variadas. Ferdowsy et al. (2021) realizaron un estudio titulado «*A machine learning approach for obesity risk prediction*», con el propósito de aumentar la conciencia sobre los factores de riesgo de obesidad en la población de Bangladesh. Utilizando datos de 1100 registros, recolectaron información detallada sobre estilo de vida, actividades diarias, rutinas alimentarias, altura y peso. La investigación involucró nueve técnicas de aprendizaje automático, desde k vecinos más cercanos (k-NN) hasta potenciación del gradiente (GB), evaluando

su desempeño en términos de métricas como exactitud, sensibilidad y F1-score. Los resultados destacaron la regresión logística (LR) como la técnica con mejor desempeño en la clasificación del riesgo de obesidad. En una línea similar, Cervantes y Palacio (2020) llevaron a cabo el estudio «*Estimation of obesity levels based on computational intelligence*», que empleó técnicas supervisadas y no supervisadas de minería de datos para abordar la detección de niveles de obesidad. Basándose en un conjunto de datos de 178 estudiantes, evaluaron la presencia de obesidad mediante técnicas como árbol de decisión (DT), máquina de vectores de soporte (SVM) y Simple K-Means. La metodología híbrida DT + Simple K-Means demostró ser eficaz al lograr una alta precisión del 98.5% y un área bajo la curva ROC del 99.5%, contribuyendo así a orientar a individuos y profesionales de la salud hacia un estilo de vida más saludable. Por otro lado, De-La-Hoz-Correa, Mendoza Palechor, De-La-Hoz-Manotas, Morales Ortega, y Sánchez Hernández (2019) desarrollaron el estudio «*Obesity Level Estimation Software based on Decision Trees*», que se centró en la estimación de la obesidad a través de un software basado en árboles de decisión y minería de datos SEMMA. Los datos recopilados en Colombia, México y Perú de 712 estudiantes universitarios permitieron la implementación de técnicas como árboles de decisión J48, redes bayesianas y regresión logística. La técnica de árbol de decisión demostró ser la más precisa, con 97.4% de precisión y exhaustividad de 97.8%. La aplicación de este modelo en un software de escritorio ofreció una herramienta práctica para estimar el nivel de obesidad en individuos. Explorando factores de riesgo intergeneracionales en obesidad infantil, Lee, Bang, Moon, y Kim (2019) llevaron a cabo el estudio «*Risk Factors for Obesity Among Children Aged 24 to 80 months in Korea: A Decision Tree Analysis*». Utilizando una amplia base de datos con más de un millón de registros, examinaron factores socio-económicos y relacionados con los padres e hijos para predecir la prevalencia de obesidad y sobrepeso en niños. A través de técnicas de aprendizaje automático, como el árbol de decisión, identificaron factores de riesgo clave como madres obesas antes de la concepción y padres obesos. Estos resultados tienen implicaciones en la gestión de la salud pública y en el diseño de intervenciones tempranas para prevenir la obesidad infantil. La investigación realizada por Pang, Forrest, Lê-Scherban, y Masino (2021), titulada «*Prediction of early childhood obesity with machine learning and electronic health record data*», se enfocó en la predicción de la obesidad infantil temprana utilizando historias clínicas electrónicas. Mediante siete técnicas de aprendizaje automático, incluyendo XGBoost y redes neuronales, se evaluó la capacidad de los modelos para predecir la incidencia de obesidad en niños de 2 a 7 años. El XGBoost se destacó con un AUC de 0.81 y una exactitud del 64.14%, lo que respalda su utilidad para la clasificación de la incidencia de obesidad en la población infantil. Sun, Wang, y Sun (2020) realizaron el estudio «*Estimating neighbourhood-level prevalence of adult obesity by socio-economic, behavioural and built environment factors in New York City*», que abordó la modelización de la prevalencia de la obesidad en adultos a nivel local. Mediante modelos de regresión espacial y no espacial, analizaron factores socioeconómicos, de comportamiento y del entorno construido en la ciudad de Nueva York. Destacaron la importancia de la reducción de la ingesta de bebidas azucaradas

como una estrategia efectiva para disminuir la prevalencia de la obesidad en adultos. En el ámbito de la infancia, Rossman et al. (2021) llevaron a cabo el estudio «*Prediction of Childhood Obesity from Nationwide Health Records*», en el cual desarrollaron un modelo inteligente para predecir el riesgo de obesidad en niños utilizando registros médicos electrónicos. A través de la técnica Gradient Boosting Trees, estimaron el riesgo de obesidad en niños de 5 a 6 años utilizando datos de los primeros 2 años de vida. El modelo alcanzó un auROC de 0.803 y auPR de 0.312, lo que demuestra su capacidad para predecir el riesgo de obesidad infantil antes de que aumente el IMC. Por otra parte, Chiong, Fan, Hu, y Chiong (2021) se enfocaron en el desarrollo de un modelo predictivo del porcentaje de grasa corporal en su estudio «*Using an improved relative error support vector machine for body fat prediction*». Utilizando algoritmos de aprendizaje automático y dos conjuntos de datos, abordaron la predicción de la composición corporal. Destacaron la eficacia del algoritmo IRE-SVM, que logró un MAE de 3.6261 y un MAC de 0.9554 en la predicción del porcentaje de grasa corporal. En un enfoque relacionado, Shao (2022) investigaron la influencia de hábitos alimentarios y patrones de comportamiento de la obesidad, en su estudio «*Comparison of prediction of obesity status based on different machine learning approaches with different factor quantities*», basándose en cinco enfoques de aprendizaje automático: árbol de decisión, máquina de vectores de soporte, XGBoost, bos- que aleatorio y árboles extremadamente aleatorios. Los algoritmos fueron evaluados con tres métricas: exactitud, puntuación Kappa y coeficiente de correlación de Matthews. Concluyeron que el modelo XGBoost tuvo el mejor efecto de clasificación en 14 factores, con una precisión del 97.16%. En conjunto, estas investigaciones internacionales abordan de manera integral la predicción y detección de la obesidad a través de diversas técnicas de aprendizaje automático y análisis de datos. Desde la aplicación de algoritmos específicos hasta la consideración de factores socioeconómicos y de comportamiento, estas investigaciones contribuyen al avance en la comprensión y abordaje de la obesidad en diferentes contextos y grupos poblacionales.

2.2. Bases Teóricas

Existen estudios que buscan identificar la influencia de diferentes factores económicos, sociales, genéticos y del comportamiento en la generación de sobrepeso y obesidad en la población, los cuales usan modelos matemáticos, estadística descriptiva y algoritmos basados en aprendizaje automático. Esta sección se enfoca en mostrar las diferentes teorías sobre los factores del estilo de vida que influyen y/o provocan la obesidad y las técnicas de aprendizaje automático utilizadas para predecir el riesgo de obesidad.

2.2.1. *Machine Learning*

Diferentes estudios han investigado cómo las técnicas de aprendizaje automático pueden predecir la obesidad en adultos, contribuyendo en la identificación de: posibles parámetros existentes que influyen y provocan la obesidad en niños y adultos; las principales enfermedades, afecciones y otros efectos negativos para la salud relacionados con la obesidad y el sobrepeso en adultos e identificar las técnicas de aprendizaje automático que se utilizan actualmente en la predicción y/o identificación automática de la obesidad en adultos (Safaei et al., 2021, p. 2).

La Tabla 1 describe algunas de las técnicas de aprendizaje automático utilizadas para la predicción de obesidad, según lo informado por investigaciones anteriores.

Tabla 1

Técnicas de aprendizaje automático utilizadas para predecir la obesidad.

Autor(es)	Técnica(s) de aprendizaje automático
Cervantes y Palacio (2020)	• Árbol de decisión (DT) • Simple k-Means
De-La-Hoz-Correa et al. (2019)	• Árbol de decisión (DT)
Cheng et al. (2020)	• Regresión logística (LR)
Ferdowsy et al. (2021)	• Regresión logística (LR) • Naïve Bayes
Lee et al. (2019)	• Árbol de decisión (DT)
Montanez et al. (2017)	• Máquina de vector de soporte (SVM) • Perceptrón multicapa (MLP)
Pang et al. (2021)	• XGBoost (XGB)
Suca et al. (2016)	• Árbol de decisión (DT) • Máquina de vector de soporte (SVM) • Redes neuronales (NN)
Lazarou et al. (2012)	• Árbol de decisión (DT)

De acuerdo a DeGregory et al. (2018), el aprendizaje automático engloba un conjunto de algoritmos capaces de caracterizar, adaptar, aprender, predecir y analizar datos, potenciando la comprensión de la obesidad y mejorando la precisión predictiva de manera sin precedentes. Esto ha llevado a un incremento en la aplicación del aprendizaje automático en la investigación de la obesidad, abriendo nuevas perspectivas en este campo (p. 1).

Árbol de Decisión (DT)

Según Serengil (2020), el Árbol de Decisión está relacionado con la teoría de la probabilidad, especialmente con el proceso de decisión Bayesiano, y se caracterizan por su estructura similar a un árbol horizontal. La raíz del árbol representa el punto de inicio de la toma de decisiones, y las acciones disponibles para el tomador de decisiones se presentan en orden cronológico, clasificándose en dos categorías: puntos de decisión o puntos de decisión con incertidumbre (p. 1).

En el proceso de construcción de un Árbol de Decisión, como menciona Brownlee (2021), en cada división se busca una característica que divida los datos etiquetados de manera que los nodos secundarios sean más homogéneos que el nodo padre del que provienen. En otras palabras, a medida que avanzamos en el árbol, los nodos se vuelven más puros, reflejando una mayor homogeneidad en sus categorías (p. 1).

Bosques Aleatorios (RF)

Según Martín Bueno (2017), se denomina bosques aleatorios o Random Forest a un tipo de modelo de predicción basado en árboles. En lugar de emplear un único árbol, utiliza un conjunto de ellos. Estos árboles pueden ser de tipo CART o CI. Su desarrollo fue impulsado por el objetivo de reducir la variabilidad y, como consecuencia, aumentar la confiabilidad y robustez del modelo (p. 28).

Adicionalmente, según Jeffares (2018), el término «aleatorio» proviene del muestreo aleatorio del conjunto de entrenamiento. Dado que se utiliza una colección de árboles, se les llama bosques aleatorios. En la construcción de cada nodo del árbol, se selecciona aleatoriamente un subconjunto de entidades y se busca el punto de corte óptimo para determinar la división de la función. El nodo raíz se establece utilizando la característica del subconjunto que produce la división más homogénea (p. 1).

Light Gradient Boosting Machine (LightGBM)

Según Serengil (2018), el algoritmo denominado Light Gradient Boosting Machine, acrónimo de «Máquinas Ligeras de Aumento de Gradiente», representa una técnica en la que los árboles de decisión se construyen secuencialmente mediante el enfoque del aumento de gradiente. Cada árbol se desarrolla en función del error del árbol previo, y las predicciones se obtienen mediante la acumulación de las contribuciones de todos estos árboles. Una particularidad del LightGBM es su enfoque en el crecimiento de árboles basado en hojas o nodos. Originado por Microsoft, este algoritmo ha ganado amplia aceptación en la comunidad de aprendizaje automático debido a su velocidad y rendimiento, siendo capaz de ejecutarse hasta seis veces más rápido que su contraparte, el algoritmo Extreme Gradient Boosting (XGBoost).

Extreme Gradient Boosting (XGBoost)

Según Serengil (2019), el término «Extreme Gradient Boosting», conocido también como XGBoost, describe un algoritmo ampliamente empleado en el ámbito del aprendizaje automático (Machine Learning) para abordar problemas de clasificación. Extreme Gradient Boosting (XGBoost) representa una implementación de árboles de decisión fortalecidos mediante gradiente, diseñados con enfoque en velocidad y rendimiento. Esta herramienta de software, creada por Tianqi Chen, brinda una implementación del modelo y ofrece soporte para tres enfoques fundamentales de aumento de gradiente: algoritmo de aumento de gradiente, aumento de gradiente estocástico y aumento de gradiente regularizado. La característica distintiva de Extreme Gradient Boosting (XGBoost) radica en su estrategia de crecimiento de árboles basada en el nivel del árbol (p. 1).

Extremely Randomized Trees (ET)

Según Ceballos (2019), el término «Extremely Randomized Trees», también conocido como ET, describe un algoritmo del campo de aprendizaje automático (Machine Learning) empleado para resolver desafíos de clasificación. Este enfoque se basa en nodos y conexiones, con el propósito de establecer la división de un conjunto principal de datos en subconjuntos de características afines que se seleccionan de manera aleatoria. Dado que las divisiones son elegidas aleatoriamente para cada característica en el clasificador de árboles adicionales, se traduce en un menor costo computacional en comparación con Bosques Aleatorios (RF) (p. 1).

Regresión Logística (LR)

Según Pant (2019), la es un algoritmo fundamental en el análisis predictivo de aprendizaje automático, se emplea para abordar cuestiones de clasificación, fundamentándose en el concepto de probabilidad. Este método utiliza una función de costo conocida como función Sigmoide o función logística, representada por la ecuación (1), con el propósito de mapear predicciones en probabilidades. La hipótesis subyacente en la regresión logística busca restringir la función de costo en el intervalo entre 0 y 1 (p. 1).

$$f(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

Fórmula de una función sigmoide

A continuación, se presentan las principales métricas para evaluar el rendimiento de las técnicas de clasificación empleadas en el ámbito del aprendizaje automático.

Exactitud

La exactitud, también conocida como accuracy, se define como el porcentaje de la muestra total que ha sido clasificada correctamente. Puede cuantificarse utilizando la ecuación (2):

$$\text{Exactitud} = \frac{MCC}{ME} \times 100\% \tag{2}$$

Fórmula de la exactitud

Donde:

- MCC, representa el número total de muestras clasificadas correctamente.
- ME, denota el número total de muestras evaluadas.

Área Bajo la Curva ROC (AUC-ROC)

El Área Bajo la Curva ROC, también conocida como AUC-ROC, es una medida que cuantifica el área bajo la curva de probabilidad que representa la relación entre la tasa de verdaderos positivos (TPR) o sensibilidad y la tasa de falsos positivos (FPR) a diferentes umbrales. Esta

curva es conocida como Curva ROC. Los valores de Área Bajo la Curva ROC (AUC-ROC) oscilan entre 0 y 1, siendo un indicador de la capacidad del modelo para discriminar entre las clases positiva y negativa.

$$\text{Sensibilidad (TPR)} = \frac{VP}{(VP + FN)} \times 100\% \quad (3)$$

Fórmula de la sensibilidad

$$\text{Especificidad} = \frac{VN}{(FP + FN)} \times 100\% \quad (4)$$

Fórmula de la Especificidad

$$\text{FPR} = 1 - \text{Especificidad} \quad (5)$$

Fórmula de la proporción de falsos positivos

Donde:

- VP es el total de verdaderos positivos.
- VN es el total de verdaderos negativos.
- FP es el total de falsos positivos.
- FN es el total de falsos negativos.

Precisión

La precisión se define como el porcentaje del total de muestra positiva que fue clasificada positivamente, puede ser medida de acuerdo con la siguiente ecuación (6):

$$\text{Precisión} = \frac{VP}{(VP + FP)} \times 100\% \quad (6)$$

Fórmula de la precisión

Donde:

- VP es el total de verdaderos positivos.
- FP es el total de falsos positivos.

Exhaustividad

Se entiende por exhaustividad o recall como el porcentaje del total de la muestra positiva que fue correctamente clasificada como positiva.

$$\text{Exhaustividad} = \frac{VP}{(VP + FN)} \times 100\% \tag{7}$$

Fórmula de la exhaustividad

Donde:

- VP es el total de verdaderos positivos.
- FN es el total de falsos negativos.

F_1-Score

El Valor-F o F_1-score se define como la medida de la media armónica entre la precisión y la exhaustividad (recall).

$$F_1\text{-score} = \frac{2 \cdot P \cdot R}{P + R} \times 100\% \tag{8}$$

Fórmula del Valor-F

Donde:

- P es la precisión.
- R es la exhaustividad o recall.

2.2.2. *Estimación de la Influencia del Estilo de Vida en el Riesgo de Obesidad de la Población*

El sobrepeso y la obesidad representan enfermedades predominantes del estilo de vida que generan una serie de problemáticas de salud, contribuyendo al desarrollo de diversas enfermedades crónicas como el cáncer, la diabetes, el síndrome metabólico y las enfermedades cardiovasculares (Safaei et al., 2021, p. 1).

Ciertos factores de carácter genético y del estilo de vida influyen en la probabilidad de que una persona sea susceptible a la obesidad. Los patrones destacados de obesidad observados en contextos geográficos y regiones específicas resaltan aún más la influencia de estos factores en la manifestación de la enfermedad. La comprensión de las causas y determinantes de la obesidad constituye un paso fundamental hacia la formulación de políticas efectivas y el diseño de programas de prevención viables, considerando las mencionadas complicaciones adicionales (Safaei et al., 2021, p. 2).

Investigaciones llevadas a cabo por Felső, Lohner, Hollódy, Erhardt, y Molnár (2017) concluyen que la duración del sueño impacta en el aumento de peso en niños, además de confirmar que elementos adicionales como el estilo de vida sedentario, una alimentación poco saludable y la resistencia a la insulina pueden predisponer a los niños a padecer trastornos del sueño y, por consiguiente, incrementos no saludables en el peso corporal (p. 760).

En general, la mayor parte de la literatura sobre obesidad se enfoca en explorar los posi- bles parámetros que causan, impactan y/o empeoran la obesidad en adultos considerando las muestras representativas (Safaei et al., 2021, p. 3). La Tabla 2 describe algunos de los factores influyentes que determinan el sobrepeso u obesidad en adultos, según lo informado por investigaciones anteriores.

Tabla 2

Potenciales factores del estilo de vida que influyen y/o provocan la obesidad.

Autor(es)	Factor(es)
Cervantes y Palacio (2020) De-La-Hoz-Correa et al. (2019)	• Ingesta de comida chatarra • Ingesta de vegetales • Tabaquismo • Ingesta de comida entre horas • Ingesta de agua

Tabla 2. Continuación

Autor(es)	Factor(es)
	• Actividad física
	• Ingesta de alcohol
	• Medio de transporte
Cheng et al. (2020)	• Tabaquismo materno durante el embarazo
	• Calificaciones educativas
	• Actividad física
Ferdowsy et al. (2021)	• Dieta
	• Actividad física
	• Depresión / estrés
	• Tiempo de ocio en redes sociales
	• Tabaquismo
	• Ingesta de comida chatarra
	• Ingesta de comida saludable
Keramat et al. (2021)	• Estado civil
	• Educación
	• Ingesta de frutas y verduras
	• Ingesta de alcohol
	• Actividad física
Lee et al. (2019)	• Tabaquismo
	• Depresión
	• Familiares con sobrepeso
Pang et al. (2021)	• Temperatura corporal
	• Ubicación geográfica
	• Frecuencia respiratoria
Suca et al. (2016)	• Nivel socio-económico
Lazarou et al. (2012)	• Ingesta de bebidas gaseosas
	• Ingesta de comida chatarra

Tabla 2. Continuación

Autor(es)	Factor(es)
	• Ingesta de frutas y verduras • Consumo de lácteos y proteínas • Consumo de leche, pan y cereales

Eficiencia de la Estimación de la Influencia del Estilo de Vida en el Riesgo de Obesidad

La capacidad de un modelo para clasificar correctamente el nivel de obesidad basado en diversos factores del estilo de vida es esencial para la efectividad de la metodología. Estos niveles de obesidad abarcan desde peso insuficiente hasta diferentes grados de obesidad y sobrepeso, específicamente: peso insuficiente, peso normal, obesidad tipo I, obesidad tipo II, obesidad tipo III, sobrepeso nivel I y sobrepeso nivel II. Además de la precisión en la clasificación, es crucial determinar el impacto o grado de importancia de cada factor del estilo de vida. Comprender estas influencias puede ofrecer insights valiosos para intervenciones dirigidas y estrategias de prevención.

Tiempo de la Estimación de la Influencia del Estilo de Vida en el Riesgo de Obesidad

El tiempo de estimación se refiere al período necesario para completar el proceso de evaluación del riesgo de obesidad en función del estilo de vida de un individuo. Esta evaluación comienza una vez que se ha obtenido y procesado el formulario que detalla las prácticas y hábitos de vida del individuo. Posteriormente, el profesional de la salud lleva a cabo una estimación informada sobre el nivel de obesidad potencial del individuo. Es esencial optimizar este tiempo para proporcionar respuestas rápidas y eficientes, especialmente en entornos clínicos donde el tiempo es esencial.

Costo de la Estimación de la Influencia del Estilo de Vida en el Riesgo de Obesidad

El costo de la estimación engloba todos los gastos asociados con el proceso de evaluación del riesgo de obesidad de un individuo, basándose en su estilo de vida. Una vez que se ha completado el formulario detallando el estilo de vida del individuo, se inicia el proceso de estimación. Este costo no solo cubre los gastos operativos, como la administración y procesamiento del formulario, sino también la compensación del profesional de la salud encargado de interpretar los

resultados y ofrecer una evaluación precisa. Es vital considerar estos costos para garantizar un proceso de estimación sostenible y accesible para todos los pacientes.

2.2.3. *Metodología para Machine Learning*

En general, todo proyecto de investigación relacionado con aprendizaje automático sigue los lineamientos de una metodología para minería de datos. La Tabla 3 describe algunas de las metodologías para minería de datos utilizadas para la predicción de obesidad, según lo informado por investigaciones anteriores.

Tabla 3

Metodologías para Machine Learning utilizadas para predecir el riesgo de obesidad.

Metodología para Machine Learning	Autor(es)
SEMMA	• De-La-Hoz-Correa et al. (2019)
CRISP-DM	• -
KDD	• Sulla Torres (2018)
No especifica	• Cervantes y Palacio (2020) • Lee et al. (2019) • Pang et al. (2021) • Suca et al. (2016) • Lazarou et al. (2012) • Trujillo Aspilcueta (2018) • Sun et al. (2020) • Rossman et al. (2021) • Chiong et al. (2021)

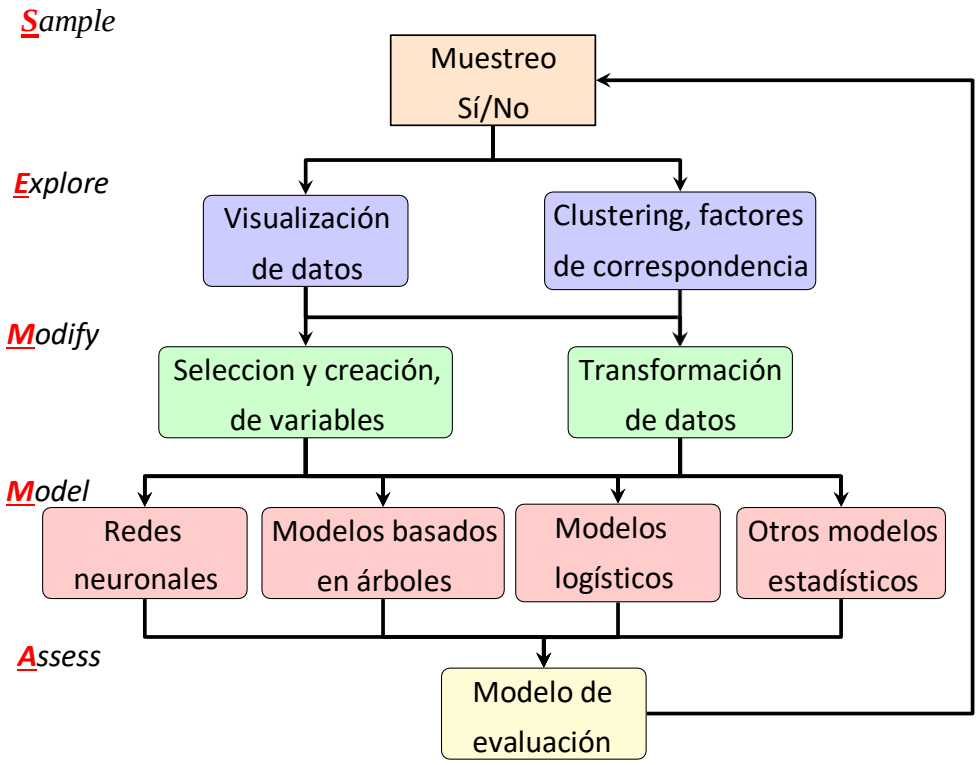
Metodología SEMMA

La metodología SEMMA fue desarrollada por SAS Institute en 1997 y presentada como parte de SAS Enterprise Miner, una solución integral para el desarrollo de minería de datos (ComputerWorld, 1997); en la cual dividieron el proceso de minería de datos en 5 etapas, repre-

sentadas por el acrónimo «Sampling, Exploring, Modifying, Modeling, and Assessing (SEMMA)» tal como se muestra en Figura 1.

Figura 1

Fases de la metodología SEMMA



Nota: Datos tomados de SAS Institute Inc. (2018, p. 322)

Matignon y SAS Institute. (2007) define lo siguiente:

> La metodología SEMMA vuelve más sencillo el trabajo de los analistas del negocio, apoya en la aplicación de técnicas de visualización y exploración estadística, selección y transformación de las variables más significativas y la validación de la exactitud de los modelos (p. 6).

La Tabla 4 detalla las etapas que conforman la metodología SEMMA. Dichas etapas establecen un marco estructurado para guiar a los profesionales en el proceso de análisis y modelado de datos, asegurando una implementación sistemática y coherente.

Tabla 4

Etapas de la metodología SEMMA

Etapa	Descripción
Sample (Muestreo)	Se obtienen los datos de una o más fuentes de información. La muestra debe ser lo suficientemente grande para contener información relevante y lo suficientemente pequeña para que pueda ser procesada rápidamente.
Explore (Exploración)	Se exploran los datos para tener un mejor entendimiento sobre qué información se va a analizar. Se trata de identificar de forma anticipada alguna relación o tendencia entre características y eliminar anomalías.
Modify (Modificación)	Se seleccionan los datos y transforman las variables con las que se alimentarán al modelo.
Model (Modelado)	Se aplican algoritmos de minería de datos en búsqueda de una combinación de los datos que prediga de forma confiable un resultado deseado.
Assess (Evaluación)	Se evalúa la utilidad y la confiabilidad de los resultados del proceso de minería de datos.

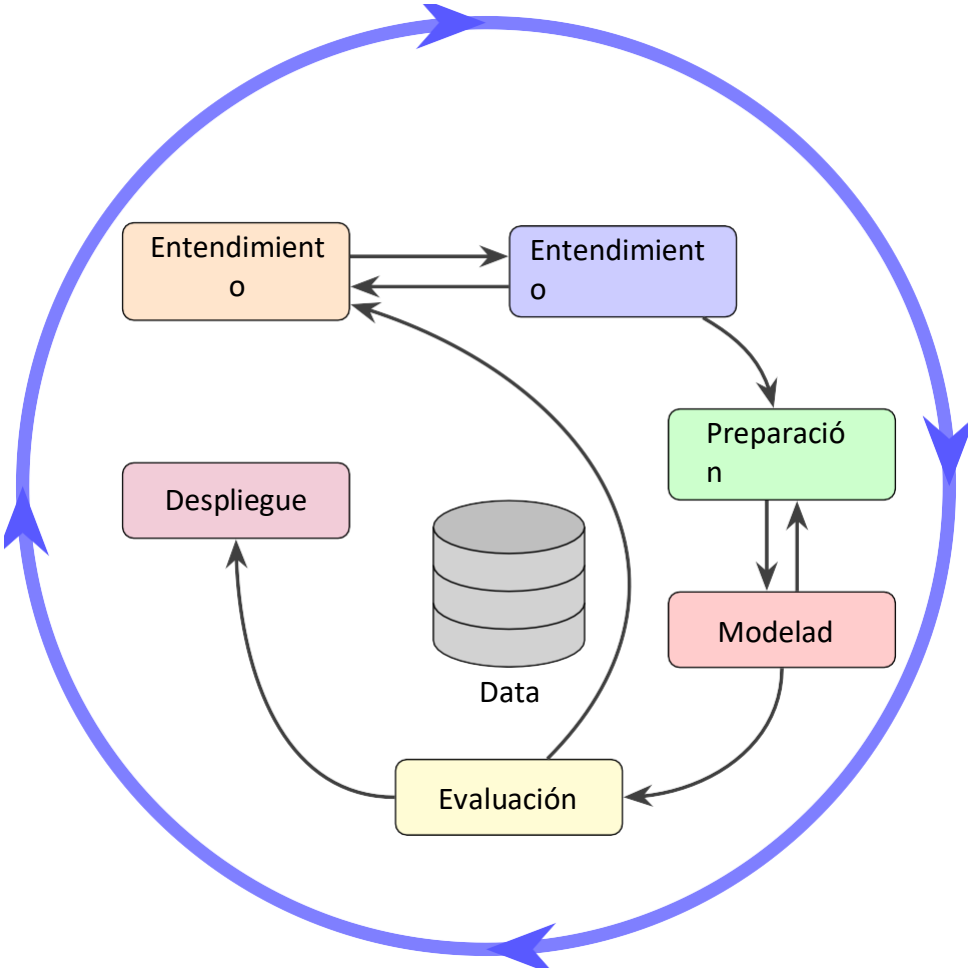
Nota: *Datos tomados de Matignon y SAS Institute. (2007, p. 6)*

Metodología CRISP-DM

La metodología CRISP-DM, cuyo acrónimo proviene de «Cross Industry Standard Process for Data Mining», emergió en 1996 como resultado del esfuerzo conjunto de tres destacados expertos en el ámbito de la minería de datos. Esta metodología fue concebida con el objetivo principal de estandarizar y optimizar el proceso de extracción de conocimiento a partir de grandes volúmenes de datos. Los creadores de CRISP-DM estructuraron cuidadosamente el proceso en seis etapas fundamentales, asegurando que cada una de ellas abordara aspectos críticos y esenciales de la minería de datos (IBM, 2011, p. 1). La representación gráfica de estas seis etapas, que ilustra el flujo y las interacciones entre ellas, se puede visualizar en la Figura 2.

Figura 2

Fases de la metodología CRISP-DM



Nota: Datos tomados de IBM (2011, p. 1)

La Tabla 5 describe detalladamente las etapas que conforman la metodología CRISP-DM. Estas etapas delinean el flujo de trabajo estructurado para llevar a cabo proyectos de minería de datos de manera efectiva y eficiente.

Tabla 5

Etapas de la metodología CRISP-DM

Etapa	Descripción
Business understanding (Entendimiento del negocio)	Esta etapa implica entender los objetivos y requisitos del proyecto desde una perspectiva empresarial, y luego convertir este conocimiento en una definición de problema de minería de datos y un plan preliminar.

Tabla 5. Continuación

Etapa	Descripción
Data understanding (Entendimiento de los datos)	Aquí se comienza a familiarizarse con los datos, identificando la calidad de los mismos, descubriendo primeras perspectivas, o detectando subconjuntos interesantes para formar hipótesis sobre los datos ocultos.
Data preparation (Preparación de los datos)	Esta fase cubre todas las actividades necesarias para construir el conjunto de datos final que se usará en el modelado, desde la selección de tablas, registros y atributos, hasta la limpieza y transformación de los datos.
Modeling (Modelado)	En esta etapa se seleccionan y aplican diversas técnicas de modelado utilizando los datos preparados. También se calibran los parámetros del modelo a óptimos, y se realizan los pasos necesarios según la herramienta de modelado seleccionada.
Evaluation (Evaluación)	Una vez construidos los modelos, es vital evaluarlos en detalle para determinar la calidad del modelo y su adecuación a los objetivos empresariales establecidos en la primera etapa. Es esencial considerar todas las decisiones empresariales que se tomarán basadas en el modelo.
Deployment (Despliegue)	La fase de implementación puede ser tan simple como generar un informe o tan compleja como implementar un modelo en un sistema empresarial en tiempo real. Aquí se presenta el modelo a la empresa y se planifica su uso práctico.

Nota: *Datos tomados de Shearer et al. (2000, p. 13-22)*

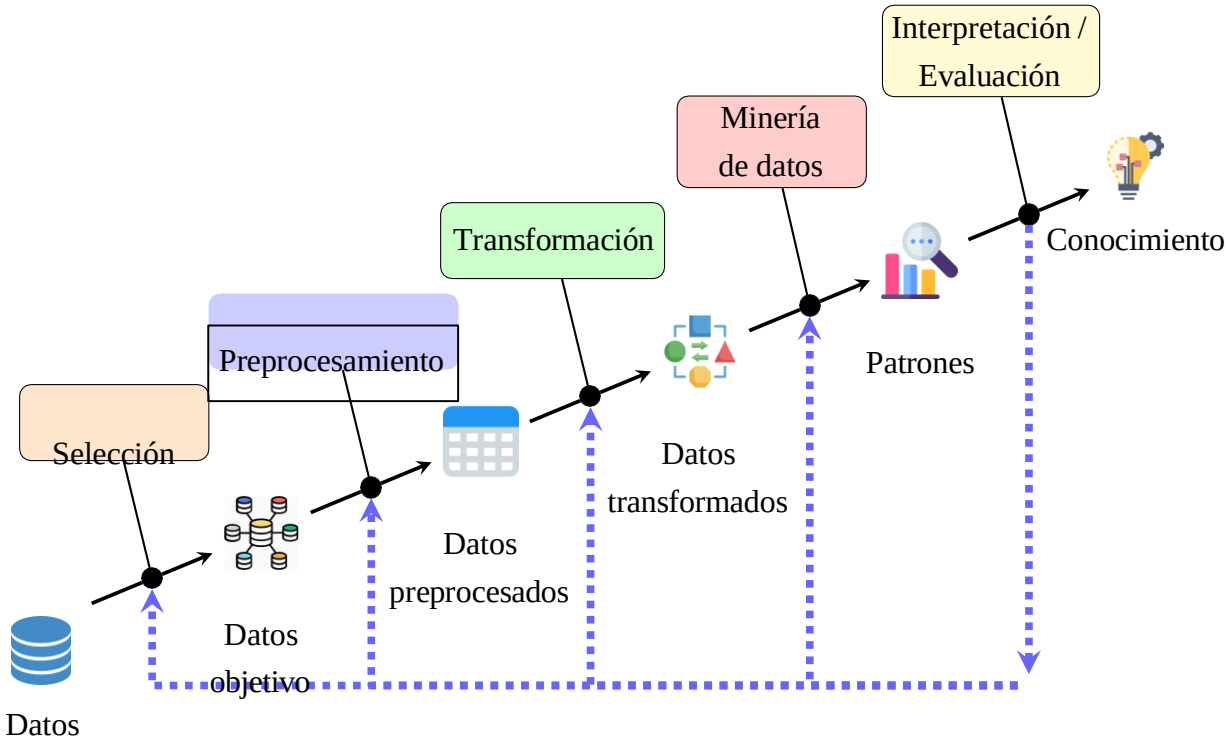
Metodología KDD

Originada en la década de 1990, la metodología KDD se consolidó como un dominio interdisciplinario, abarcando áreas como las estadísticas, la inteligencia artificial, el aprendizaje automático y los sistemas de bases de datos. El acrónimo KDD hace referencia a «Knowled- ge Discovery in Databases», cuya traducción al español es «Descubrimiento de Conocimiento en Bases de Datos». Esta metodología tiene como propósito primordial el descubrimiento de patrones significativos y el extracto de conocimientos valiosos y fácilmente interpretables de extensos volúmenes de datos (Fayyad, Piatetsky-Shapiro, y Smyth, 1996, p. 37). El proceso de KDD se estructura en cinco etapas esenciales. La representación gráfica de estas etapas, que

destaca la dinámica y conexiones entre ellas, se encuentra ilustrada en la Figura 3.

Figura 3

Fases de la metodología KDD



Nota: Datos tomados de Fayyad, Piatetsky-Shapiro, y Smyth (1996, p. 41)

La Tabla 6 describe detalladamente las etapas que conforman la metodología KDD. Estas etapas representan el proceso integral de extracción de conocimientos útiles y valiosos a partir de grandes volúmenes de datos.

Tabla 6

Etapas de la metodología KDD

Etapa	Descripción
Selection (Selección)	En esta etapa, se identifica y selecciona el conjunto de datos que será objeto de estudio. Esto podría involucrar la extracción de ciertos datos de una base de datos más grande, basándose en ciertos criterios.
Preprocessing (Preprocesamiento)	Una vez seleccionados los datos, se lleva a cabo un proceso de limpieza y transformación. Esto puede incluir la eliminación de datos faltantes, la corrección de errores en los datos y la transformación de los datos en un formato adecuado para las siguientes etapas.

Tabla 6. Continuación

Etapa	Descripción
Transformation (Transformación)	Los datos preprocesados se transforman en una forma adecuada para la minería de datos. Esto podría involucrar operaciones como la reducción de la dimensionalidad, la normalización, y el cálculo de ciertas agregaciones.
Data mining (Minería de datos)	Esta es la etapa central del proceso KDD, donde se aplican técnicas específicas para extraer patrones de los datos. Estos patrones podrían ser clusters, árboles de decisión, reglas de asociación, entre otros.
Interpretation / Evalution (Interpretación / Evaluación)	Los patrones encontrados en la etapa de minería de datos se evalúan e interpretan en términos del problema del mundo real. Esta etapa busca determinar si los patrones descubiertos son realmente útiles y significativos en el contexto del problema que se está tratando de resolver.

Nota: *Datos tomados de Fayyad et al. (1996, p. 37-54)*

2.3. Marcos Conceptuales

En esta sección, se detallan términos fundamentales que son relevantes para la comprensión y contextualización de la presente investigación.

Estilo de Vida

En el ámbito médico, se refiere a la incorporación de prácticas saludables para reducir riesgos de enfermedades crónicas y respaldar su tratamiento, fundamentado en evidencia científica robusta que aboga por la salud óptima (Rippe, 2019). Esta noción engloba componentes como una dieta adecuada, actividad física, sueño saludable y abstención de tabaco, entre otros (Hâncu, 2021).

Obesidad

Según Mechanick, Farkouh, Newman, y Garvey (2020), es una enfermedad crónica basa- da en adiposidad (ABCD) que se manifiesta en cuatro etapas. Estas abarcan desde factores

genéticos y ambientales hasta el diagnóstico basado en medidas antropométricas, culminando en complicaciones cardio-metabólicas. Esta enfermedad se interrelaciona con otras afecciones crónicas, como la enfermedad cardio-metabólica crónica.

Machine Learning

Según Mitchell (1997), es una rama de la inteligencia artificial que se centra en la construcción de sistemas que pueden aprender de los datos. En lugar de ser programados explícitamente para realizar una tarea, estos sistemas utilizan algoritmos y modelos estadísticos para analizar y extraer patrones de los datos.

Metodología

Según Kothari (2004), se refiere a un sistema de prácticas, técnicas, procedimientos y reglas utilizado por aquellos que trabajan en una disciplina. Una metodología a menudo implica un enfoque teórico y práctico para abordar problemas y preguntas en una investigación o estudio particular.

Algoritmo

Según Rayward-Smith, Cormen, Leiserson, y Rivest (1991), es un conjunto finito de instrucciones bien definidas y ordenadas que permiten llevar a cabo una tarea o resolver un problema. Los algoritmos son esenciales para que los programadores creen programas de computadora que resuelvan problemas específicos.

Aprendizaje Supervisado

Según Bishop (2007), es una técnica de Machine Learning donde un algoritmo aprende a partir de datos etiquetados, y hace predicciones basadas en ese conocimiento. En este enfoque, tanto los datos de entrada como la salida deseada se proporcionan, y el algoritmo itera a través de los datos para encontrar patrones y ajustar su modelo

CAPÍTULO III: METODOLOGÍA

3.1. Tipo de Investigación

Según Sánchez Carlessi, Reyes Romero, y Mejia Sáenz (2018), la investigación puede ser del tipo básica o aplicada. Siendo la investigación básica la que busca nuevos conocimientos, como principios y leyes científicas; y la investigación aplicada la que aprovecha conocimientos de la investigación básica para la solución de problemas inmediatos (p. 79). Por tanto, se puede inferir que la investigación fue del **tipo básica y aplicada** con enfoque cuantitativo.

3.1.1. *Nivel de Investigación*

Según Hernández Sampieri, Fernández Collado, y Baptista Lucio (2014), la investigación cuantitativa puede tener 4 niveles o alcances de investigación: exploratorio, descriptivo, correlacional o explicativo. Siendo el alcance descriptivo el que tiene como finalidad medir o recoger información sobre las variables y no indicar la relación entre estas. Por tanto, se puede inferir que la investigación fue de **nivel descriptivo**.

3.1.2. *Diseño de Investigación*

Según Hernández Sampieri et al. (2014), el diseño de investigación puede ser no experimental o experimental, siendo a la vez, el diseño experimental de tres tipos: pre-experimental, cuasi experimental o experimental puro. El diseño experimental puro es aquel que cuenta con grupos de comparación (manipulación de la variable independiente) y equivalencia de los grupos, pudiendo utilizar prepruebas y pospruebas para analizar la evolución de los grupos antes y después del tratamiento experimental (p. 127,141). Por tanto, se puede inferir que el diseño de investigación fue **experimental puro con grupo experimental (Ge) y grupo de control (Gc)** (1).

$$\begin{array}{cccc} G_e & A & X & O_1 \\ G_c & A & & O_2 \end{array} \quad (1)$$

Fórmula del diseño posprueba con grupo de control (Sánchez Carlessi et al., 2018, p. 54)

Donde:

- G_e es el grupo experimental, al que se aplicará la solución de Machine Learning.
- G_c es el grupo de control, al que no se aplicará la solución de Machine Learning.
- A es la elección aleatorizada de los sujetos.
- O_1 y O_2 es el posprueba para los grupos experimental y control, respectivamente.

3.1.3. *Método de Investigación*

Según Jiménez y Jacinto (2017), los métodos de investigación más reconocidos son el analítico-sintético, el inductivo-deductivo, el método de analogías, el hipotético-deductivo, el histórico-lógico, el sistémico-estructural-funcional, la sistematización, el genético y la modelación. El método hipotético-deductivo, parte del planteamiento de hipótesis inferidas de principios o leyes o sugeridas por los datos empíricos, y aplicando reglas de la deducción. La hipótesis se prueba mediante la experimentación por grupos (grupo experimental y control) y finalmente se refuta o verifica la hipótesis (p. 8,12). En esta investigación, se plantearon hipótesis general y específicas que fueron demostradas en relación con la problemática planteada. Por lo tanto, se puede inferir que el método de investigación utilizado fue **hipotético-deductivo**.

3.2. Población y Muestra

3.2.1. *Unidad de Análisis*

La unidad de análisis fue el Proceso de estimación de la influencia de factores del Estilo de Vida en el riesgo de obesidad de la población de Colombia, México y Perú.

3.2.2. *Población de Estudio*

Todos los procesos de estimación de la influencia de factores del Estilo de Vida en el riesgo de obesidad de la población de Colombia, México y Perú.

3.2.3. *Muestra*

La muestra consta de 30 procesos de estimación de la influencia de factores del Estilo de Vida en el riesgo de obesidad de la población de Colombia, México y Perú en establecimiento de salud. Nivel de confianza de 95% y margen de error de 5%.

3.2.4. *Muestreo*

Según Luna Espinoza, Hernández Suaárez, y Tinoco Zermeño (2009), el muestreo fue **aleatorio estratificado**, dado que, la población es muy heterogénea y las consideraciones de costo limitan el tamaño de la muestra.

3.3. Operacionalización de Variables

3.3.1. *Identificación de Variables*

La Tabla 7 ofrece una presentación de las variables empleadas en este estudio, incluyendo tanto su definición precisa como los indicadores asociados y una descripción completa de los mismos.

Tabla 7

Variables de la investigación

Variable independiente	*Machine Learning*
Definición operacional	Modelo producto del entrenamiento de técnicas de aprendizaje automático.
Indicador	Presencia-ausencia.

Tabla 7. Continuación

Variable dependiente	*Estimación de la influencia del Estilo de Vida en el riesgo de obesidad de la población de Colombia, México y Perú.*
Definición operacional	Es el desempeño que tiene una solución de Machine Learning al clasificar el riesgo de obesidad.
Dimensiones	• Eficiencia • Tiempo • Costo
Indicadores	• Eficiencia generada en cada proceso de estimación: Es la cantidad de registros correctamente clasificados. • Tiempo requerido en cada proceso de estimación: Es el tiempo requerido para realizar el proceso de estimación. • Costo generado en cada proceso de estimación: Es el costo demandado para realizar el proceso de estimación.

3.3.2. *Tabla de Operacionalización de Variables*

La Tabla 8 presenta una síntesis meticulosa de la operacionalización de las variables empleadas en esta investigación, detallando cómo cada variable ha sido definida y cómo se procederá a su medición.

Tabla 8

Matriz de Operacionalización de Variables

Variables	Dimensiones	Indicadores	Índices	Escala de medición
Variable independiente Machine Learning	-	Presencia - Ausencia	Sí No	-
Variable dependiente Estimación de la influencia del Estilo de Vida en el riesgo de obesidad	Eficiencia	Eficiencia generada en cada proceso de estimación	[0 - 100]	%/ operación
	Tiempo	Tiempo requerido en cada proceso de estimación	[0 - 8]	min / operación
	Costo	Costo generado en cada proceso de estimación	[1 - 10]	$ / operación

3.4. Técnicas e Instrumentos de Recolección de Datos

3.4.1. *Técnicas*

- **Observación**, según Hernández Sampieri et al. (2014) esta técnica implica el registro sistemático, válido y confiable de comportamientos y situaciones observables, a través de un conjunto de categorías y subcategorías (p. 252).

3.4.2. *Instrumentos*

- Ficha de observación.

3.5. Procedimientos

- Se realizó la revisión de la literatura en búsqueda de antecedentes de investigación en los que fueron identificados algoritmos de Machine Learning usados, conjunto de datos, métricas y vacíos en las investigaciones.
- Se creó una solución de Machine Learning para la Clasificación de Obesidad usando la metodología de minería de datos Data-Driven Obesity Risk Analysis (DORA).
- Se realizó la contrastación de hipótesis a través de la prueba t-Student utilizando el software Minitab, permitiendo al investigador obtener conclusiones y demostrar la hipótesis en relación con la problemática planteada.

CAPÍTULO IV: RESULTADOS Y DISCUSIÓN

4.1. Elaboración de la Nueva Metodología

Este capítulo presenta la concepción y desarrollo de la metodología Data-Driven Obesity Risk Analysis (DORA), cuya traducción al español es «Análisis de riesgo de obesidad basado en datos», diseñada específicamente para estimar la influencia del estilo de vida en el riesgo de obesidad de la población. Surgiendo de un análisis detenido de las metodologías SEMMA, CRISP-DM y KDD, la metodología DORA integra los principios esenciales de estas aproximaciones con innovaciones propias del investigador, buscando optimizar la precisión y aplicabilidad de las estimaciones.

4.1.1. *Etapas de la Metodología DORA*

La Tabla 9 presenta un desglose exhaustivo de las etapas que integran la metodología Data-Driven Obesity Risk Analysis (DORA). Diseñada como una hoja de ruta holística, esta metodología guía de principio a fin en la realización de proyectos de minería de datos, garantizando resultados efectivos y un canal de comunicación fluido con el usuario final.

Tabla 9

Etapas de la metodología DORA

Etapa	Descripción
Selección de Datos (KDD)	Identificar y obtener el conjunto de datos más relevante y apropiado. Extracción de datos de diversas fuentes, evaluación de la calidad y relevancia de los datos, y selección de subconjuntos de datos específicos basados en criterios predefinidos.

Tabla 9. Continuación

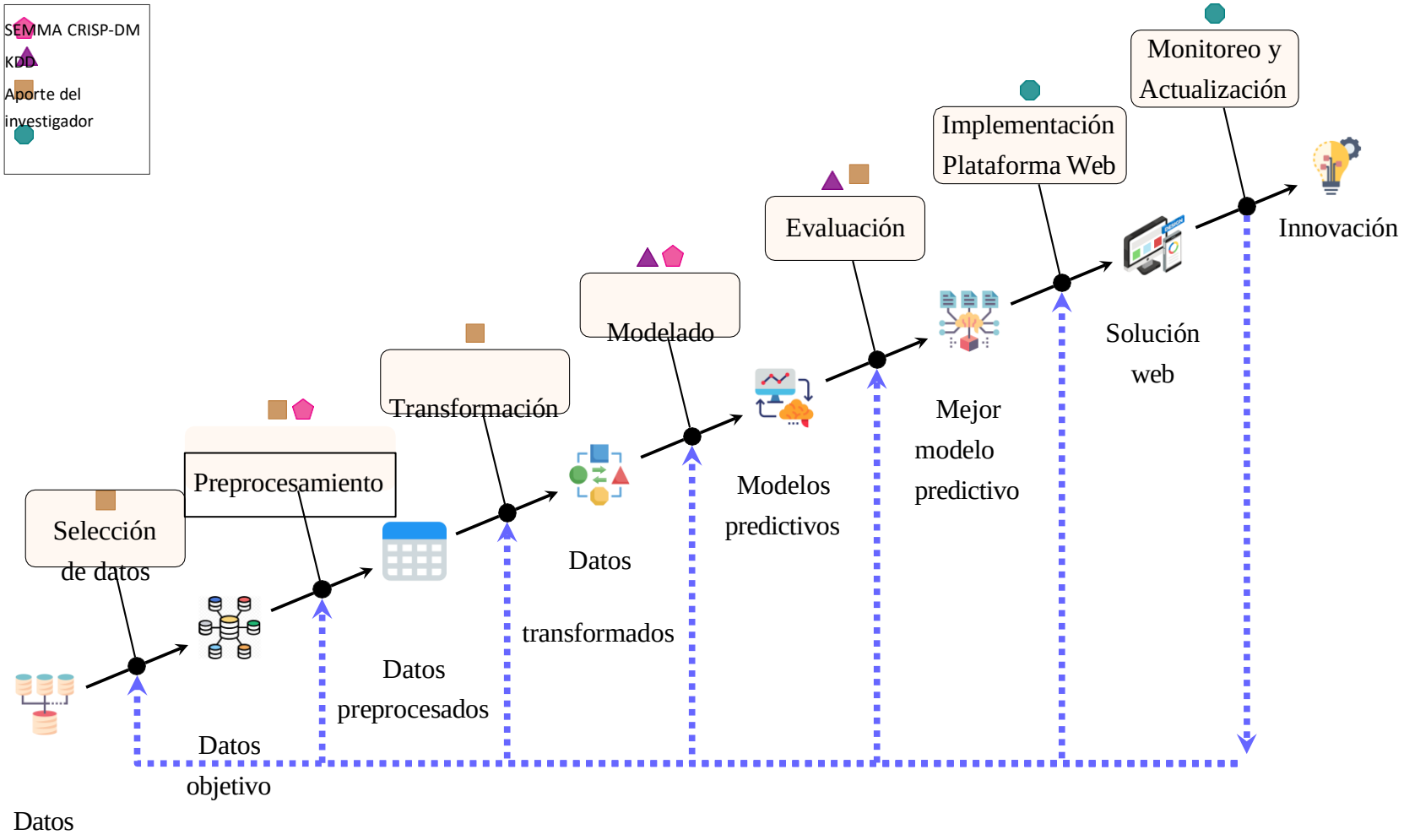
Etapa	Descripción
Preprocesamiento (KDD, SEMMA)	Asegurar que los datos estén limpios y en un formato adecuado para el análisis. Limpiar datos faltantes o erróneos, tratar outliers, normalizar o estandarizar valores, y codificar variables categóricas.
Transformación (KDD)	Convertir los datos en un formato óptimo para la minería de datos. Reducción de dimensionalidad, creación de variables derivadas, y transformaciones específicas como la codificación one-hot o la normalización.
Modelado (CRISP-DM, SEMMA)	Desarrollar modelos predictivos que puedan clasificar el riesgo de obesidad. Selección de algoritmos de aprendizaje automático, entrenamiento de modelos, validación cruzada, y evaluación de métricas de rendimiento.
Evaluación (CRISP-DM, KDD)	Determinar la precisión y robustez del modelo desarrollado. Uso de conjuntos de datos de prueba, evaluación de métricas como precisión, recall, F_1-score, y ROC-AUC, e interpretación de resultados en el contexto clínico.
Implementación de Plataforma Web	Preparar el modelo para su implementación en una plataforma en línea. Crear una infraestructura que pueda manejar solicitudes y realizar predicciones en tiempo real. Proporcionar una interfaz amigable para los usuarios finales.
Monitoreo y Actualización	Asegurar que la plataforma y el modelo sigan siendo relevantes y precisos a lo largo del tiempo. Monitoreo del desempeño del modelo en datos en vivo, recopilación de feedback de los usuarios, y actualización periódica del modelo y la plataforma según sea necesario.

4.1.2. *Flujograma y Estructura de la Metodología DORA*

La metodología DORA está estructurada en una serie de etapas interconectadas, diseñadas para guiar sistemáticamente el proceso de análisis de datos. La Figura 4 ilustra visualmente este proceso, destacando la secuencia y relación entre cada una de las fases. Este flujograma proporciona una perspectiva clara de cómo se desarrolla la metodología, facilitando su comprensión y permitiendo una implementación más efectiva.

Figura 4

Fases de la metodología DORA



4.2. Desarrollo de la Solución de Machine Learning para la Clasificación de Obesidad

En este capítulo, se desarrollará el modelo de clasificación de obesidad basado en factores del estilo de vida, siguiendo la metodología «Data-Driven Obesity Risk Analysis (DORA)» para minería de datos. Se elaborará, evaluará y perfeccionará el modelo con el propósito de alcanzar los objetivos de la investigación.

4.2.1. *Selección de Datos*

Durante esta fase, se procedió a la selección del conjunto de datos fundamental para el desarrollo del modelo de clasificación. El conjunto de datos empleado proviene de la investigación realizada por Palechor y Manotas (2019), el cual abarca 16 factores relevantes para determinar la presencia de obesidad.

La recopilación de datos se realizó a través de cuestionarios diseñados con preguntas específicas, los cuales fueron administrados a diversos grupos en Colombia, México y Perú. En la Tabla 10 se ofrece una descripción detallada de las características obtenidas mediante este cuestionario.

Tabla 10

Descripción de las características de la muestra estudiada

Nombre	Descripción	Tipo de dato
Gender	Género	Categórico: Female, Male
Age	Edad	Numérico
Height	Altura	Numérico
Weight	Peso	Numérico
FHWO	Familiares con sobrepeso	Categórico: Yes, No
FAVC	Consume frecuentemente alimentos altos en calorías	Categórico: Yes, No
FCVC	Número de comidas donde suele comer verduras	Numérico
NCP	Número de comidas principales al día	Numérico

Tabla 10. Continuación

Nombre	Descripción	Tipo de dato
CAEC	Ingiere comida entre horas	Categórico: Always, Frequently, Sometimes, No
SMOKE	Fuma frecuentemente	Categórico: Yes, No
CH2O	Litros de agua que bebe al día	Numérico
SCC	Monitorea las calorías que consume a diario	Categórico: Yes, No
FAF	Frecuencia de días por semana que a menudo tiene actividad física	Numérico
TUE	Tiempo de uso de dispositivos tecnológicos a diario	Numérico
CALC	Frecuencia de ingesta de alcohol	Categórico: Always, Frequently, Sometimes, No
MTRANS	Medio de transporte que usa habitualmente	Categórico: Automobile, Bike, Motorbike, Public transport, Walking
NObeyesdad	Índice de masa corporal	Categórico: Insufficient Weight, Normal Weight, Obesity Type I, Obesity Type II, Obesity Type III, Overweight Level I, Overweight Level II

Los detalles específicos, incluyendo el tamaño de la muestra y las características esenciales de los datos, se presentan en la Tabla 11. Este análisis proporciona una visión general crucial para la comprensión de la información con la que se trabajará en el proceso de construcción del modelo de clasificación de obesidad.

Tabla 11

Características de la muestra estudiada

Característica	Frecuencia(i)	Porcentaje (%)
Género		

Tabla 11. Continuación

Característica	Frecuencia(i)	Porcentaje (%)
Masculino	1,068	49.41
Femenino	1,043	50.59
Total	2,111	100.00
Edad (años)		
14-23	1,239	58.69
24-33	664	31.45
34-43	185	8.76
44-53	16	0.76
54-63	7	0.33
Total	2,111	100.00
Estatura (metros)		
1.45-1.65	652	30.89
1.65-1.85	1,346	63.76
1.85-2.05	113	5.35
Total	2,111	100.00
IMC		
Peso insuficiente	272	12.88
Peso normal	287	13.60
Obesidad tipo I	351	16.63
Obesidad tipo II	297	14.07
Obesidad tipo III	324	15.35
Sobrepeso nivel I	290	13.74
Sobrepeso nivel II	290	13.74
Total	2,111	100.00

4.2.2. *Preprocesamiento*

En esta etapa, se realizó una transformación de las características categóricas del conjunto de datos, convirtiéndolas en valores numéricos enteros para facilitar el procesamiento del modelo. Los detalles de esta codificación se presentan en la Tabla 12.

Tabla 12

Descripción de las características categóricas y su valor numérico equivalente

Nombre	Descripción	Valor categórico	Valor numérico equivalente
Gender	Género	Female	0
		Male	1
FHWO	Familiares con sobrepeso	No	0
		Yes	1
FAVC	Consume frecuentemente alimentos altos en calorías	No	0
		Yes	1
CAEC	Ingiere comida entre horas	Always	0
		Frequently	1
		Sometimes	2
		No	3
SMOKE	Fuma frecuentemente	No	0
		Yes	1
SCC	Monitorea las calorías que consume a diario	No	0
		Yes	1
CALC	Frecuencia de ingesta de alcohol	Always	0
		Frequently	1
		Sometimes	2
		No	3
MTRANS	Medio de transporte que usa habitualmente	Automobile	0
		Bike	1
		Motorbike	2
		Public transport	3
		Walking	4

Tabla 12. Continuación

Nombre	Descripción	Valor categórico	Valor numérico equivalente
NObeyesdad	Índice de masa corporal	Insufficient Weight	0
		Normal Weight	1
		Obesity Type I	2
		Obesity Type II	3
		Obesity Type III	4
		Overweight Level I	5
		Overweight Level II	6

La Figura 5 muestra el código Python implementado para convertir las características categóricas en representaciones numéricas.

Figura 5

Conversión de variables a valores categóricos.

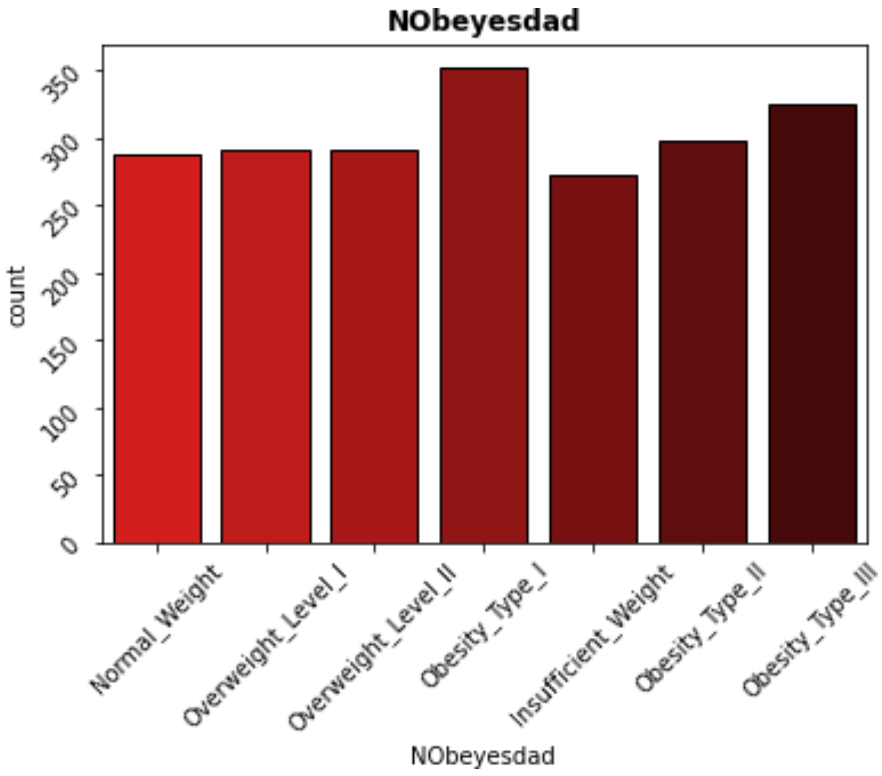
```
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
from sklearn import preprocessing
lbl = preprocessing.LabelEncoder()

df = pd.read_csv('../input/obesity-levels/ObesityDataSet.csv', encoding='ISO-8859-2')

df['Gender_cat'] = lbl.fit_transform(df['Gender'].astype(str))
df['FHWO_cat'] = lbl.fit_transform(df['family_history_with_overweight'].astype(str))
df['FAVC_cat'] = lbl.fit_transform(df['FAVC'].astype(str))
df['CAEC_cat'] = lbl.fit_transform(df['CAEC'].astype(str))
df['SMOKE_cat'] = lbl.fit_transform(df['SMOKE'].astype(str))
df['SCC_cat'] = lbl.fit_transform(df['SCC'].astype(str))
df['CALC_cat'] = lbl.fit_transform(df['CALC'].astype(str))
df['MTRANS_cat'] = lbl.fit_transform(df['MTRANS'].astype(str))
df['NObeyesdad_cat'] = lbl.fit_transform(df['NObeyesdad'].astype(str))
```

4.2.3. *Transformación*

En esta etapa, se observó que el conjunto de datos carecía de valores faltantes, sin embargo, se identificó un desequilibrio en los datos con respecto a la característica objetivo, el nivel de obesidad, como se aprecia en la Figura 6.

Figura 6

Distribución de frecuencia de la característica nivel de obesidad



Con el objetivo de abordar este desbalance, se aplicó la técnica de sobre-muestreo Synthetic Minority Over-Sampling Technique (SMOTE), tal como se representa en la Figura 7; lo que permitió obtener un conjunto de datos balanceado, como se ilustra en la Figura 8. Este proceso de modificación resultó esencial para garantizar la integridad y eficacia del modelo de clasificación de obesidad a desarrollar.

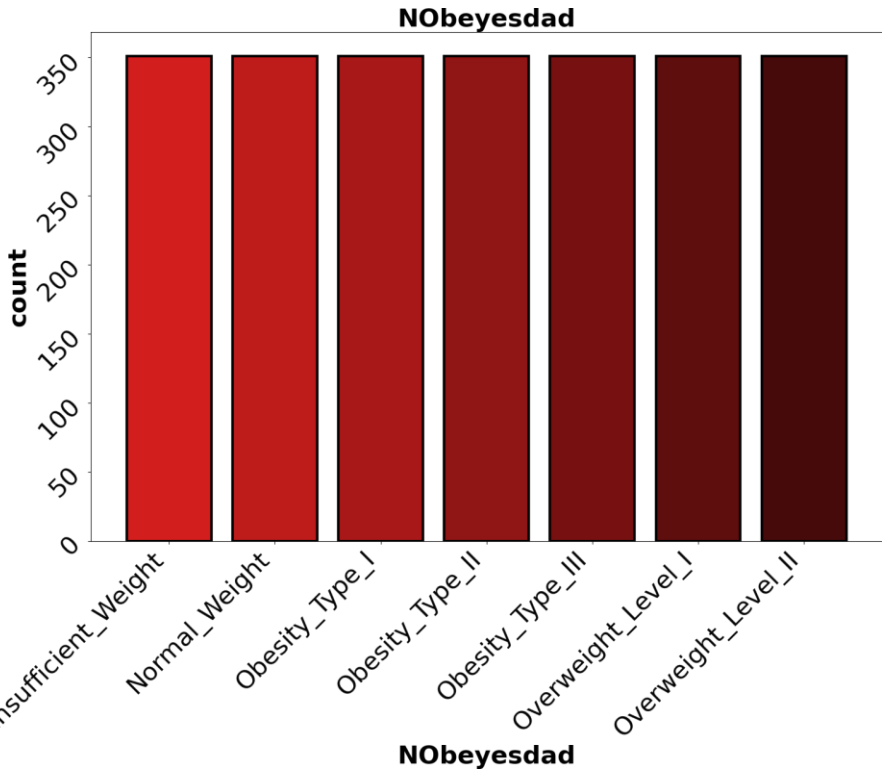
Figura 7

Sobre-muestreo usando SMOTE

```
from imblearn.over_sampling import SMOTE

oversample = SMOTE()

X = df[['Gender_cat','Age','Height','Weight','FHWO_cat', 'FAVC_cat', 'FCVC', 'NCP',
↪  'CAEC_cat','SMOKE_cat', 'CH2O', 'SCC_cat', 'FAF','TUE', 'CALC_cat','MTRANS_cat' ]]
y = df['NObeyesdad_cat']

X, y = oversample.fit_resample(X, y)
```

Figura 8

Distribución balanceada de frecuencia de la característica nivel de obesidad



4.2.4. *Modelado*

En esta etapa, se procedió al entrenamiento de diversos algoritmos de clasificación de Machine Learning, con el objetivo de desarrollar un modelo eficaz para la tarea de clasificación de obesidad. Los algoritmos entrenados fueron los siguientes:

- Extreme Gradient Boosting (XGBoost)
- Light Gradient Boosting Machine (LightGBM)
- Bosques Aleatorios (RF)
- Árbol de Decisión (DT)
- Extremely Randomized Trees (ET)
- Regresión Logística (LR)

Definición de Variables Globales

Las variables globales utilizadas durante el entrenamiento de los algoritmos de clasificación se detallan en la Figura 9.

Figura 9

Definición de las variables globales

```
import numpy as np
result = np.array([]).reshape(0, 13)

iteration = None
iteration_fold = 0

X_train = X_test = None
y_train = y_test = None
X_val = y_val = None

clf = clf_name = None

y_pred = y_pred_val = None
accuracy_metric = accuracy_metric_val = None
precision_metric = precision_metric_val = None
recall_metric = recall_metric_val = None
f1_metric = f1_metric_val = None
roc_auc = roc_auc_val = None
roc_auc_metric = roc_auc_metric_val = None

cm = cm_val = None
feature_imp = None
```

Definición de Algoritmos de Clasificación

Los algoritmos de clasificación empleados para dicho entrenamiento se describen en la Figura 10.

Figura 10

Definición de los modelos entrenados

```
headers = {'nro': [], 'model': [], 'accuracy_metric': [],
           'accuracy_metric_val': [], 'precision_metric': [],
           'precision_metric_val': [], 'recall_metric': [],
           'recall_metric_val': [], 'f1_metric': [],
           'f1_metric_val': [], 'roc_auc_metric': [],
           'roc_auc_metric_val': [], 'nro_fold': [],
          }
result = pd.DataFrame()

from enum import Enum
class Models(Enum):
    XG_BOOST = 'XGBoost'
    LIGTH_GBM = 'LightGBM'
    RANDOM_FOREST = 'RandomForest'
    DECISION_TREE = 'DecisionTree'
    EXTRA_TREES = 'ExtraTrees'
    LOGISTIC_REGRESSION = 'LogRegres'
    RIDGE_CLASSIFIER = 'RidgeClassifier'
    SDG_CLASSIFIER = 'SGDClassifier'

folds = 5
scores = {}
cnt = None
```

División del Conjunto de Datos

El conjunto de datos fue dividido asignando el 60% para entrenamiento, 20% para pruebas y el 20% restante para validación, como se muestra en la Figura 11. Esta estrategia aseguró una evaluación meticulosa y permitió una comparación precisa del rendimiento predictivo de los modelos.

Figura 11

División del conjunto de datos.

```
    def split_data():
        from sklearn.model_selection import train_test_split
        global X
        global y

        global X_train
        global X_test
        global y_train
        global y_test

        global X_train
        global X_val
        global y_train
        global y_val

        X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2,
        '→  random_state = None, stratify = y)
        X_val = X_test
        y_val = y_test
```

Además, se empleó la técnica de validación cruzada K-Fold, dividiendo el conjunto de entrenamiento en 5 iteraciones, según se detalla en la Figura 12.

Figura 12

Configuración de la validación cruzada.

```
from sklearn.model_selection import KFold, StratifiedKFold, cross_val_score,
↪  cross_validate
from sklearn import linear_model, tree, ensemble
from sklearn.metrics import make_scorer, accuracy_score, f1_score, roc_auc_score,
↪  precision_score, recall_score

kf = StratifiedKFold(n_splits = folds, shuffle=True, random_state=42)

scoring = {"accuracy": make_scorer(accuracy_score),
           "f1_score": make_scorer(f1_score, average = 'micro'),
           "precision": make_scorer(precision_score, average = 'micro'),
           "recall": make_scorer(recall_score, average = 'micro')
          }

def cross_validation():
    global scores
    global clf
    global X_train
    global y_train
    global kf
    global scoring

    scores = cross_validate(clf, X_train, y_train, cv = kf,
                            scoring = scoring,
                            return_train_score = True,
                           return_estimator = True)
```

Entrenamiento de los Algoritmos

La Figura 13 detalla los parámetros y la configuración adoptada para entrenar los algoritmos.

Figura 13

Entrenamiento de los algoritmos

```
for x in range (0, 1):
    global clf_name
    global iteration

    iteration = x
    split_data()
    train_lgbm_model()
    train_xgboost_model()
    train_random_forest_model()
    train_decision_tree_model()
    train_extra_tree_model()
    train_log_regres_model()
    train_ridge_class_model()
    train_sdg_class_model()

save_result()
summary()
```

4.2.5. *Evaluación*

La etapa de evaluación es fundamental para determinar la eficacia de los algoritmos en la clasificación y predicción. En este proceso, se utilizó un conjunto de cinco métricas para medir su rendimiento. Estas métricas fueron elegidas debido a su relevancia en la evaluación de modelos de clasificación y su capacidad para proporcionar una visión integral del comportamiento del algoritmo. Las métricas seleccionadas son:

- Exactitud (Accuracy): Mide la proporción de predicciones correctas.
- Precisión: Evalúa la precisión de las predicciones positivas.
- Exhaustividad (Recall): Indica la proporción de positivos reales que fueron correctamente identificados.
- Valor F: Una métrica que combina precisión y exhaustividad.
- Área bajo la Curva ROC (ROC AUC): Evalúa la capacidad del modelo para discriminar entre clases.

Para una comprensión detallada de cómo se desempeñaron los algoritmos con base en estas métricas, se puede consultar la Tabla 13 donde se presentan los resultados obtenidos en el experimento:

Tabla 13

Comparación de los resultados obtenidos en los algoritmos entrenados.

Algoritmo	Exactitud (%)	Precisión (%)	Exhaustividad (%)	Valor F (%)	AUC ROC (%)
Extreme Gradient Boosting (XGBoost)	97.04	97.06	97.06	97.03	99.87
Light Gradient Boosting Machine (LightGBM)	97.36	97.39	97.37	97.36	99.90
Bosques aleatorios (RF)	95.72	95.92	95.72	95.73	99.77
Árbol de Decisión (DT)	93.23	93.23	93.24	93.17	96.06
Extremely Randomized Trees (ET)	95.72	94.80	94.63	94.64	99.61
Regresión logística (LR)	82.84	82.68	82.89	82.55	97.32

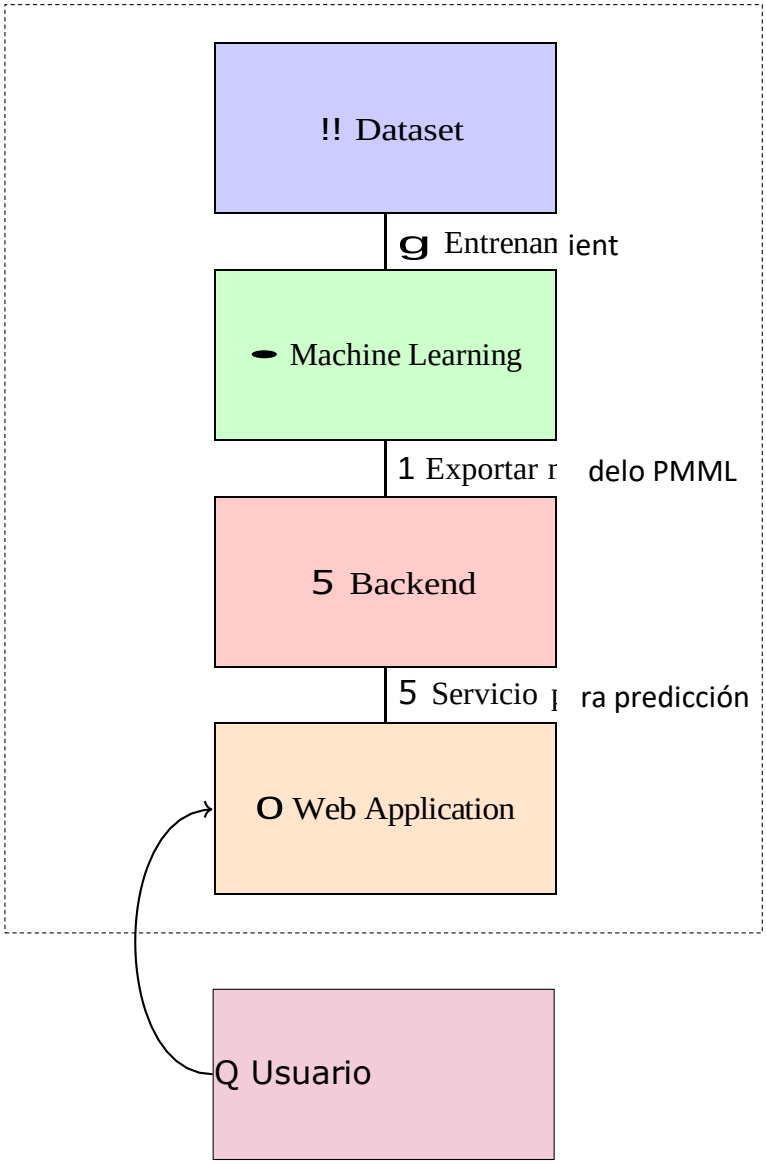
4.2.6. *Implementación de Plataforma Web*

Arquitectura

En la siguiente Figura 14, se presenta el diagrama de flujo que ilustra la arquitectura de la solución de aprendizaje automático propuesta. Comprende un flujo desde la obtención de datos (!!) hasta la interfaz del usuario (Q), pasando por el proceso de Machine Learning **(•)** y los componentes de backend (5) y frontend (0).

Figura 14

Diagrama de Flujo de la Arquitectura de la Solución de Machine Learning



Requerimientos

En esta sección, se detallan los diversos requisitos necesarios para la implementación exitosa de la solución propuesta. Estos requisitos abarcaron aspectos relacionados con el software, el hardware, así como los requisitos funcionales y no funcionales que guiaron el desarrollo y despliegue de la solución. La comprensión completa de estos requerimientos resultó fundamental para lograr la eficiencia y efectividad en la consecución de los objetivos planteados.

Requerimientos de Software A continuación, se detallan los requerimientos de software utilizados en el desarrollo de la solución:

- Python 3.8 o superior: Lenguaje de programación utilizado para implementar y entrenar los modelos de clasificación.
- Bibliotecas de Machine Learning: Scikit-Learn, XGBoost, LightGBM, RandomForest, entre otras, para el desarrollo de los modelos de clasificación.
- React y JavaScript: Framework y lenguaje de programación utilizados para el desarrollo del frontend.
- Spring Boot y Java 18: Framework y lenguaje de programación utilizados para el desarrollo del backend.
- Servicio de Hosting en la Nube: Vercel para el frontend y Render para el backend, para el despliegue y puesta en producción de la solución web.
- Servicio en la nube para la gestión de repositorios de código: Bitbucket, optimizando el control de versiones y facilitando la colaboración en equipo de manera eficiente.

Requerimientos de Hardware A continuación se presentan los requerimientos de hardware necesarios para el desarrollo y ejecución de la solución:

- Computador con al menos 16GB de RAM y procesador Intel Core i7 de 10ª generación o equivalente (por ejemplo, Apple M1) para el desarrollo y entrenamiento de los modelos de clasificación.
- Accesorios de cómputo estándar como teclado, mouse y monitor para facilitar la interacción y el trabajo eficiente en el desarrollo.
- Conexión a Internet estable y de alta velocidad para la descarga de bibliotecas, herramientas y para el despliegue en la nube.
- Espacio de almacenamiento suficiente para el almacenamiento de datos, código fuente y archivos de proyecto.
- Recursos adicionales pueden ser necesarios durante las fases intensivas de entrenamiento de modelos, especialmente para modelos más complejos.

Requerimientos Funcionales A continuación, se detallan los requerimientos funcionales que guían la implementación y funcionamiento de la solución web para la evaluación de riesgo de obesidad. Estos requisitos aseguran la accesibilidad, precisión y usabilidad de la plataforma interactiva.

1. **Acceso a la Solución Web:** Los usuarios pueden acceder a la solución web a través de un enlace público, sin necesidad de iniciar sesión.

2. **Formulario de Entrada:** La solución proporciona un formulario interactivo que permite a los usuarios ingresar información relevante para la evaluación de riesgo de obesidad. En la Tabla 14 se especifican los campos de los que consta el formulario:

Tabla 14

Campos del formulario interactivo

#	Campo	Tipo	Descripción
1	Edad	Numérico	Edad del usuario (mayor de 18 años).
2	Estatura	Numérico	Altura del usuario en metros.
3	Peso	Numérico	Peso del usuario en kilogramos.
4	Género	Selector	Género del usuario (Masculino o Femenino).
5	Comidas principales	Numérico	Número diario de comidas principales.
6	Comidas con verduras	Numérico	Número diario de comidas con verduras.
7	Consumo diario de agua	Numérico	Litros de agua bebidos al día.
8	Actividad física semanal	Numérico	Días de actividad física por semana.
9	Familiares con sobrepeso	Booleano	Presencia de familiares con sobrepeso (Sí o No).
10	Uso de tecnología	Numérico	Horas diarias usando dispositivos tecnológicos.
11	Alimentos altos en calorías	Booleano	Consumo frecuente de alimentos altos en calorías (Sí o No).

Tabla 14. Continuación

#	Campo	Tipo	Descripción
12	Comida entre horas	Selector	Frecuencia de ingesta de comida entre horas (No, A veces, Frecuentemente, Siempre).
13	Fumar	Booleano	Frecuencia de fumar (Sí o No).
14	Consumo de alcohol	Selector	Frecuencia de consumo de alcohol (No, A veces, Frecuentemente, Siempre).
15	Monitoreo de calorías	Booleano	Monitoreo diario de calorías (Sí o No).
16	Medio de transporte	Selector	Medio de transporte habitual (Automóvil, Motocicleta, Bicicleta, Transporte público, Caminar).

3. **Evaluación y Resultado:** Después de completar el formulario, el usuario puede hacer clic en el botón «Evaluar». Los datos del formulario se envían al backend para su procesamiento. El modelo de Machine Learning analiza la información proporcionada y devuelve el nivel de obesidad predicho. El resultado se presenta al usuario en la interfaz de la solución web como parte de la respuesta visual.

Requerimientos No Funcionales A continuación, se detallan los requerimientos no funcionales para la solución de clasificación de obesidad, abarcando aspectos de interfaz, confiabilidad, seguridad, rendimiento, mantenimiento y cumplimiento de estándares, garantizando así una solución robusta y eficiente.

1. **Interfaz de Usuario:**

   - **Usabilidad:** La interfaz debe ser intuitiva y de fácil navegación, permitiendo a los usuarios completar el formulario sin dificultades.

   - **Accesibilidad:** La solución web debe ser accesible para personas con discapacidades visuales y motoras, cumpliendo con las pautas de accesibilidad WCAG 2.0.

2. **Confiabilidad:**

   - **Tolerancia a Fallas:** La solución debe manejar errores de conexión y fallos tempo-

rales sin pérdida de datos del usuario.

- **Disponibilidad:** El sistema debe estar disponible las 24 horas del día, los 7 días de la semana, con tiempos de inactividad planificados mínimos.

3. **Seguridad:**

   - **Privacidad de Datos:** Los datos del usuario no se almacenarán en la solución y se manejarán de manera confidencial durante su procesamiento.

4. **Rendimiento:**

   - **Tiempo de Respuesta:** La solución debe proporcionar una respuesta rápida al usuario después de enviar el formulario.

   - **Escalabilidad:** La solución debe ser capaz de manejar múltiples solicitudes concurrentes sin degradación significativa del rendimiento.

5. **Mantenimiento:**

   - **Facilidad de Mantenimiento:** La solución, desarrollada utilizando tecnologías como Python, React y Spring Boot, debe ser diseñada y codificada de manera que sea fácil de entender y modificar. Permitiendo a los desarrolladores de software mantenerla de manera efectiva y eficiente, sin requerir recursos extensos.

6. **Estándares:**

   - **Cumplimiento de Normas:** La solución debe seguir los estándares y mejores prácticas de desarrollo web y seguridad de datos.

   - **Compatibilidad del Navegador:** La solución debe ser compatible con los navegadores web más utilizados, como Chrome, Firefox, Safari y Edge.

Desarrollo de la Solución Web

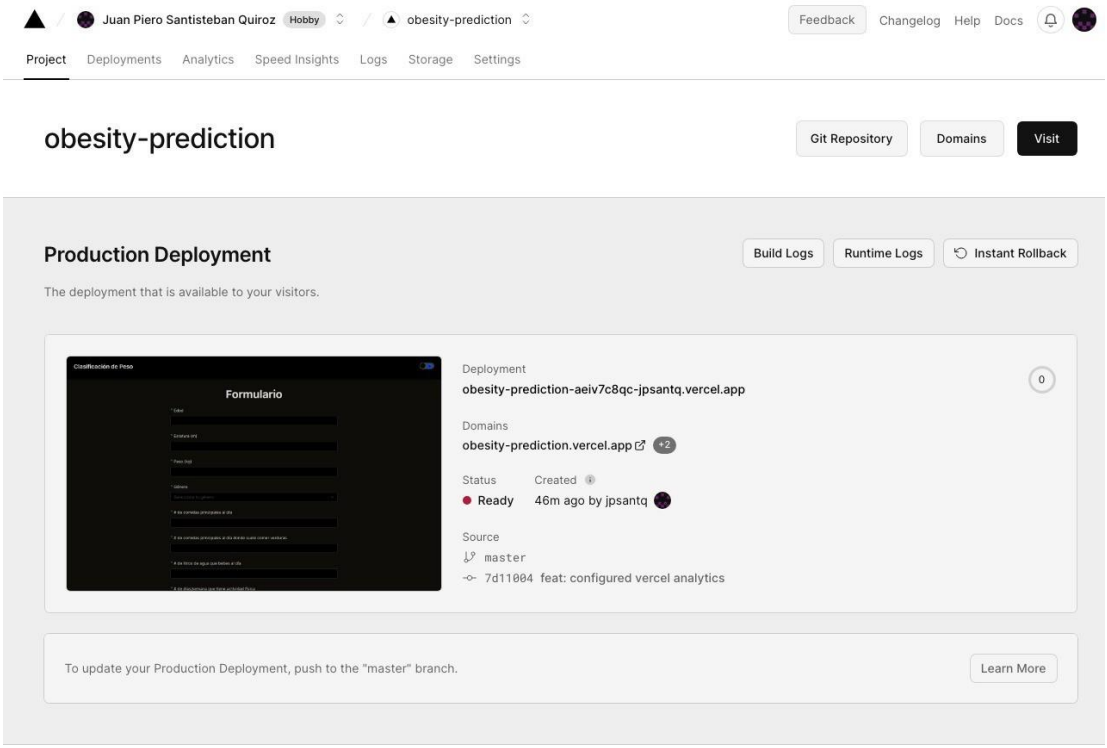
El desarrollo de la solución de Machine Learning se llevó a cabo siguiendo cuidadosamente los requerimientos que se detallaron en la sección anterior. Este proceso se dividió en dos componentes principales: el Front-end y el Back-end.

Para el desarrollo del Front-end, se eligió la biblioteca React. Esta elección se basó en su

popularidad y eficiencia, ya que React es una de las bibliotecas más utilizadas en el desarrollo de aplicaciones web modernas debido a su facilidad de uso y capacidad de crear interfaces de usuario dinámicas. La implementación y puesta en marcha del Front-end en un entorno de producción se realizó utilizando la plataforma Vercel. Esta plataforma es conocida por su facilidad de despliegue y escalabilidad, como se ilustra en la Figura 15.

Figura 15

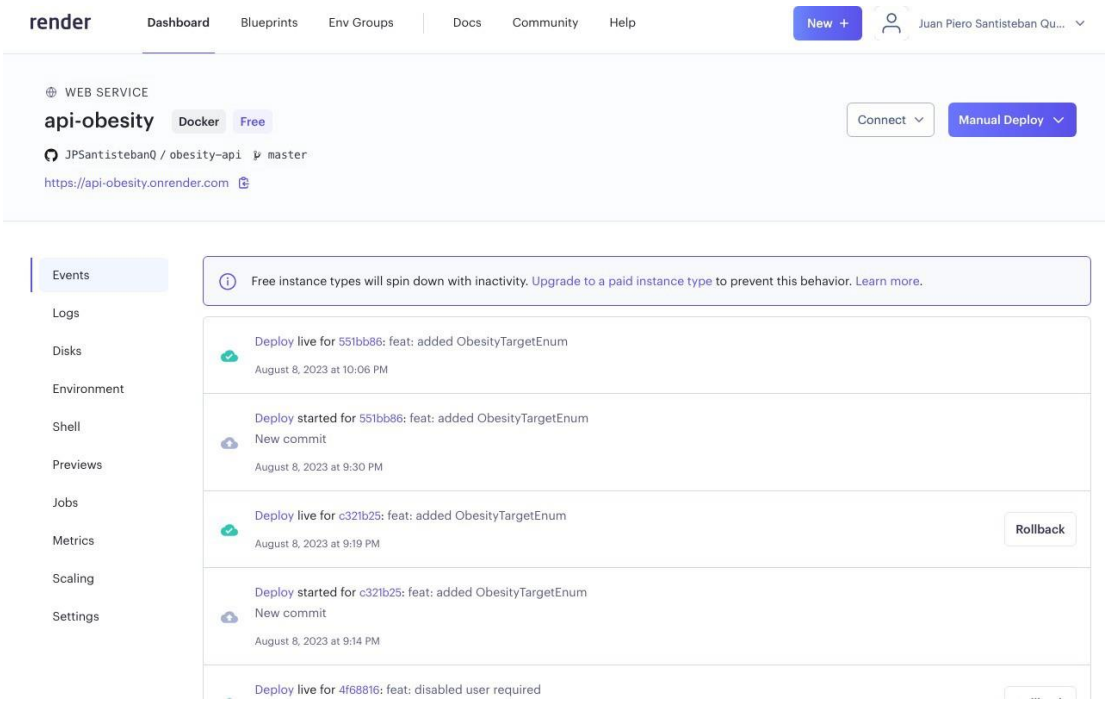
Visualización del despliegue en producción del Front-end en Vercel



Por otro lado, para el Back-end, se utilizó el lenguaje de programación Java en su versión 18, junto con el framework Spring Boot. Spring Boot es reconocido por ser uno de los frameworks más robustos y confiables en la actualidad, especialmente cuando se trata de desarrollar aplicaciones empresariales y de gran escala. La implementación y despliegue del Back-end se llevó a cabo en la plataforma Render, un servicio altamente escalable y confiable, tal como se refleja en la Figura 16.

Figura 16

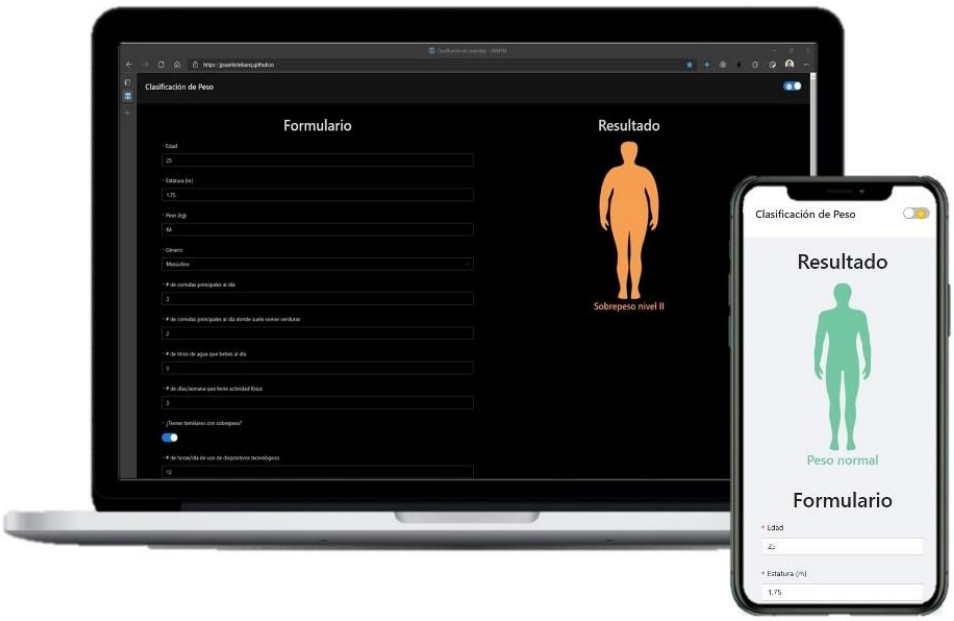
Visualización del despliegue en producción del Back-end en Render



Una vez finalizado el desarrollo, la solución completa se puede visualizar desde diferentes dispositivos, como un ordenador de escritorio y dispositivos móviles. La Figura 17 muestra la apariencia y la interfaz de la solución en dichos dispositivos.

Figura 17

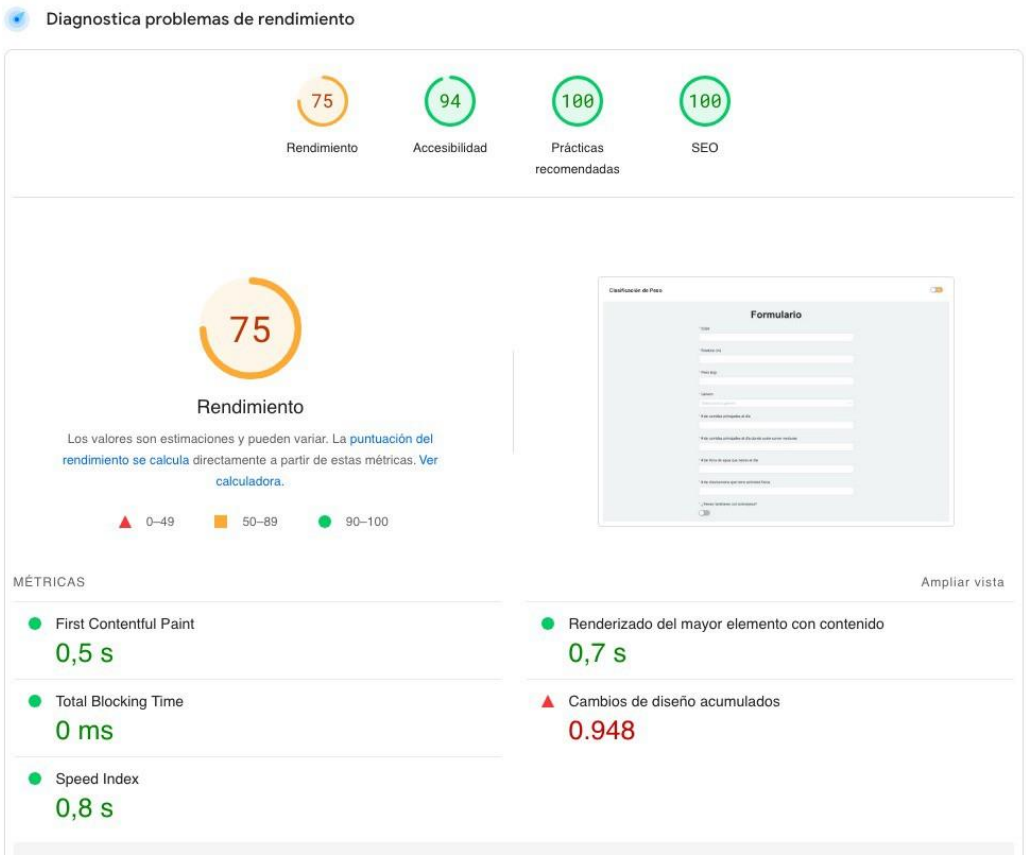
Vista previa de la interfaz de la solución web en diferentes dispositivos



Para asegurarnos de que nuestra solución no solo funciona, sino que también ofrece una experiencia de usuario óptima, se llevó a cabo un análisis de rendimiento utilizando Google PageSpeed. Esta herramienta está diseñada específicamente para evaluar y ofrecer sugerencias para mejorar el rendimiento de las páginas web. Los resultados obtenidos, como se muestra en la Figura 18, son prometedores:

Figura 18

Informe de rendimiento de la solución web según Google PageSpeed Insights



- El rendimiento general de la página obtuvo una puntuación del 75%, lo que indica una buena velocidad de carga y una experiencia de usuario fluida.
- En cuanto a la Accesibilidad, se logró un impresionante 94%, lo que significa que la solución es accesible para una amplia variedad de usuarios, incluyendo aquellos con discapacidades.
- Las Prácticas recomendadas y el SEO, ambos con una puntuación perfecta del 100%, demuestran que la solución sigue las mejores prácticas de la industria y está optimizada para motores de búsqueda.

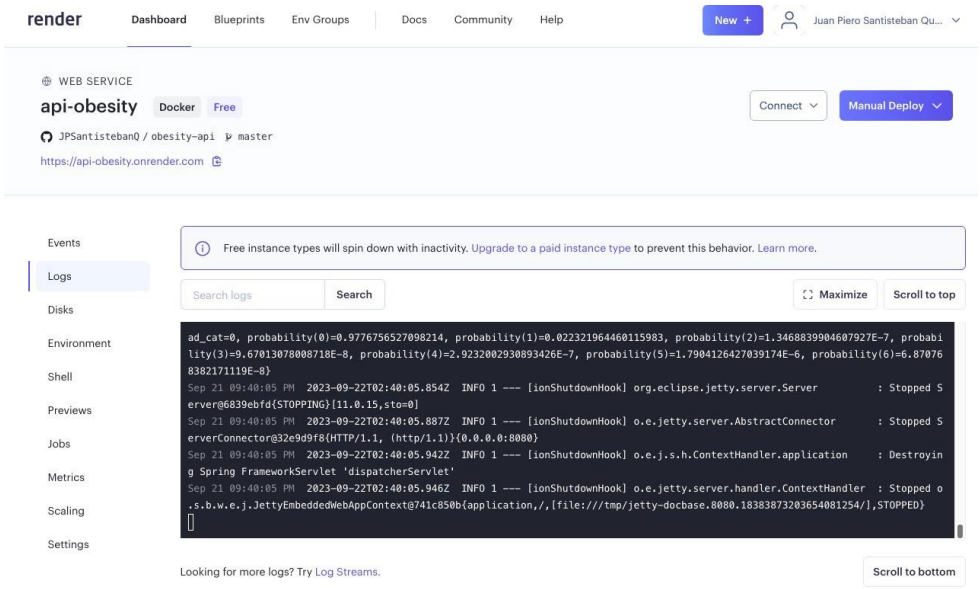
4.2.7. *Monitoreo y Actualización*

Mantener un seguimiento constante y actualizar la solución de Machine Learning es esencial para garantizar su eficacia y eficiencia. A medida que los datos cambian y evolucionan, es crucial que la solución se adapte y evolucione con ellos.

En el Back-end, hemos implementado medidas de monitoreo para asegurar que el sistema funcione de manera óptima. Una de esas medidas es el seguimiento de logs a través de la plataforma Render. Estos logs ofrecen información detallada sobre las operaciones, errores y otras actividades relevantes del sistema, lo que facilita la identificación y resolución de problemas. Una visualización detallada de estos logs se puede encontrar en la Figura 19.

Figura 19

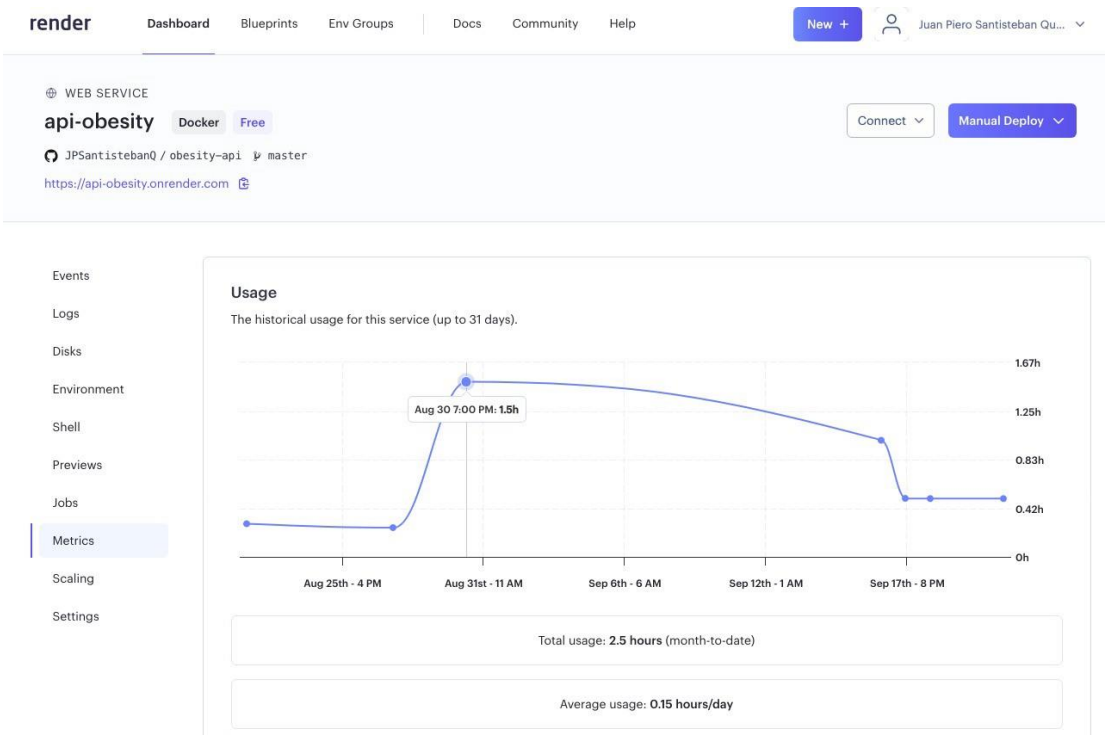
Registro detallado de logs en la plataforma Render para la solución de Machine Learning



Además, para obtener una visión más amplia del rendimiento del sistema, se habilitó un panel de métricas en la misma plataforma. Este panel proporciona datos en tiempo real sobre diferentes aspectos del sistema, como el uso de recursos, tiempos de respuesta y otros indicadores clave de rendimiento. Estas métricas se visualizan en la Figura 20.

Figura 20

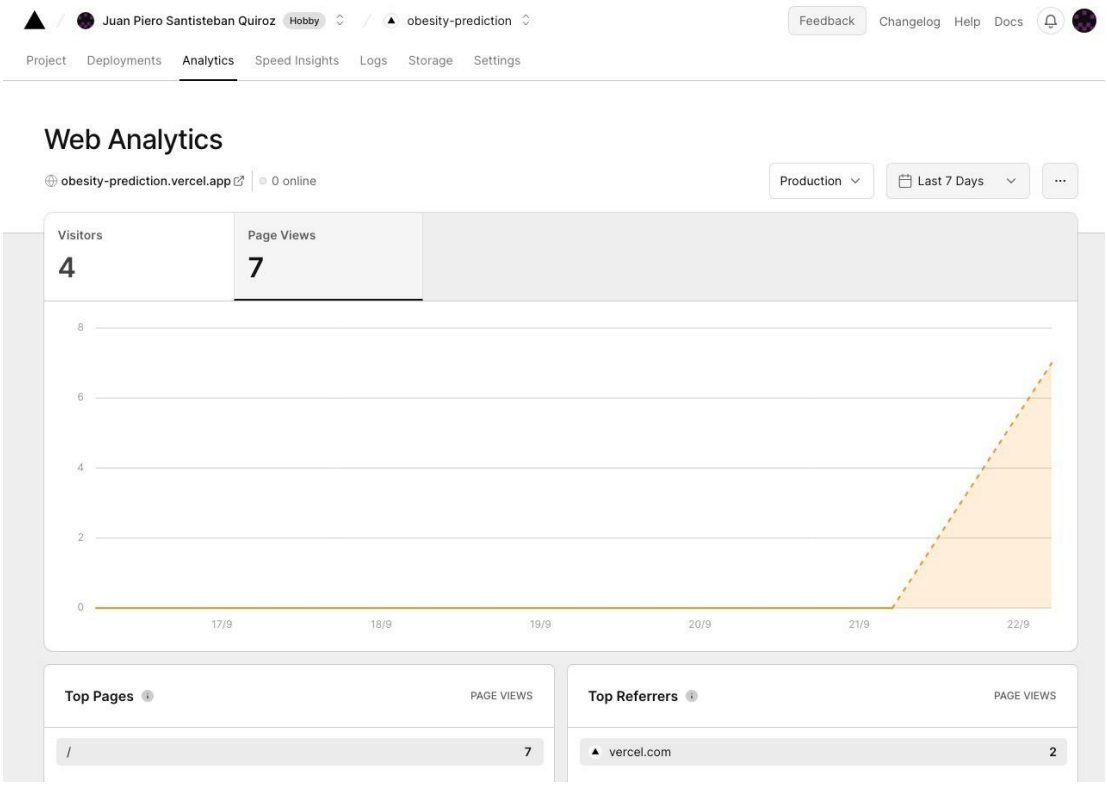
Panel de métricas de uso de recursos en la plataforma Render para la solución de Machine Learning



En cuanto al Front-end, es vital comprender cómo los usuarios interactúan con la solución y asegurarse de que la experiencia del usuario sea fluida. Por ello, se ha habilitado el análisis web en la plataforma Vercel. Este análisis ofrece insights sobre la actividad del usuario, páginas más visitadas, tiempo de permanencia, entre otros. La información recopilada es esencial para realizar mejoras y optimizaciones basadas en el comportamiento real del usuario. Una representación gráfica de estos análisis se encuentra en la Figura 21.

Figura 21

Análisis web y métricas de interacción del usuario en la plataforma Vercel para la solución de Machine Learning



La combinación de estas herramientas y métricas proporciona una visión completa de la salud y el rendimiento de la solución, permitiendo intervenciones oportunas y actualizaciones regulares para mantenerla a la vanguardia de las necesidades del usuario y los avances tecnológicos.

4.3. Resultados

Esta sección está dedicada a la exposición detallada de los resultados cuantitativos, Tabla 15, derivados de la experimentación realizada en una empresa dedicada a brindar consultorías nutricionales, su análisis subsiguiente está destinado a ofrecer insights profundos y conclusiones estadísticamente robustas.

Tabla 15

Resultados sobres los indicadores de la investigación.

N	PosPrueba del Grupo de control - G_c			PosPrueba del Grupo experimental - G_e		
	Eficiencia (%)	Tiempo (min)	Costo ($)	Eficiencia (%)	Tiempo (min)	Costo ($)
1	86.67	3.00	2.25	99.17	2.03	1.62
2	80.00	3.45	2.59	98.33	2.34	1.87
3	88.33	2.60	1.95	86.99	1.66	1.33
4	93.33	3.94	2.96	90.41	2.96	2.37
5	76.67	3.58	2.69	97.13	2.76	2.21
6	81.67	3.90	2.93	95.87	2.90	2.32
7	85.00	2.75	2.06	93.12	1.99	1.59
8	88.33	2.88	2.16	89.54	1.76	1.41
9	88.33	3.60	2.70	86.34	2.74	2.19
10	88.33	4.11	3.08	96.93	3.42	2.73
11	95.00	3.90	2.92	99.67	2.66	2.13
12	95.00	3.04	2.28	97.13	2.13	1.70
13	91.67	3.44	2.58	95.87	2.75	2.20
14	91.67	3.02	2.27	93.12	1.96	1.57
15	88.33	2.93	2.20	89.54	1.76	1.41
16	95.00	3.31	2.49	92.48	2.34	1.87
17	85.00	2.90	2.17	91.46	1.74	1.39
18	86.67	3.45	2.59	88.62	2.16	1.73
19	95.00	3.80	2.85	98.17	2.76	2.21
20	93.33	2.89	2.17	98.86	1.65	1.32
21	93.33	3.47	2.61	99.08	2.35	1.88
22	83.33	3.09	2.32	87.35	2.00	1.60
23	91.67	2.92	2.19	95.85	1.75	1.40
24	95.00	3.43	2.58	81.57	2.32	1.85
25	81.67	3.11	2.33	86.27	1.99	1.59

Tabla 15. Continuación

N	PosPrueba del Grupo de control - G_c			PosPrueba del Grupo experimental - G_e		
	Eficiencia (%)	Tiempo (min)	Costo ($)	Eficiencia (%)	Tiempo (min)	Costo ($)
26	83.33	2.88	2.16	98.81	1.66	1.33
27	91.67	3.65	2.74	93.21	2.53	2.03
28	75.00	2.81	2.11	79.93	1.74	1.39
29	91.67	3.35	2.51	89.76	2.15	1.72
30	85.00	3.09	2.32	90.78	2.06	1.65

4.4. Análisis y Discusión de Resultados

Esta sección ofrecerá un análisis pormenorizado de cada indicador evaluado. A través de gráficos, tablas y discusiones, se busca proporcionar una comprensión profunda de los hallazgos y su relevancia en el contexto de la investigación.

4.4.1. *Indicador 1: Eficiencia de la Estimación*

El indicador «Eficiencia de la Estimación» proporciona una medida cuantitativa del rendimiento y precisión con la que la solución de Machine Learning puede predecir o estimar resultados basados en los datos ingresados. Los resultados detallados para este indicador se pueden encontrar en la Tabla 16.

Tabla 16

Resultados sobre el indicador Eficiencia de la estimación.

N	PosPrueba del Grupo de control - G_c (%)	PosPrueba del Grupo experimental - G_e (%)
μ	88.17	92.71
$Nro. >= \mu$	18	16

Tabla 16. Continuación

N	PosPrueba del Grupo de control - G_c (%)	PosPrueba del Grupo experimental - G_e (%)
%	60	53

Es importante señalar que los valores presentados en la Tabla 16 representan el promedio de la eficiencia de la estimación a lo largo de un periodo de 3 semanas. Este periodo se seleccionó para obtener una comprensión más robusta de la usabilidad y consistencia de la solución de Machine Learning en diferentes momentos y condiciones.

- Al analizar los resultados, se destaca que el «Grupo experimental» demostró una eficiencia superior, con un promedio del 92.71%, en comparación con el «Grupo de control» que alcanzó una eficiencia del 88.17%. Esta diferencia indica que las intervenciones o técnicas aplicadas al «Grupo experimental» tuvieron un impacto positivo en su desempeño.

- En el «Grupo de control», es notable que el 60% de las observaciones individuales superaron su propia eficiencia promedio del 88.17%. Esto sugiere que, aunque el promedio general de eficiencia fue menor en comparación con el «Grupo experimental», hubo un número significativo de casos en el «Grupo de control» que tuvieron un desempeño por encima de lo esperado.

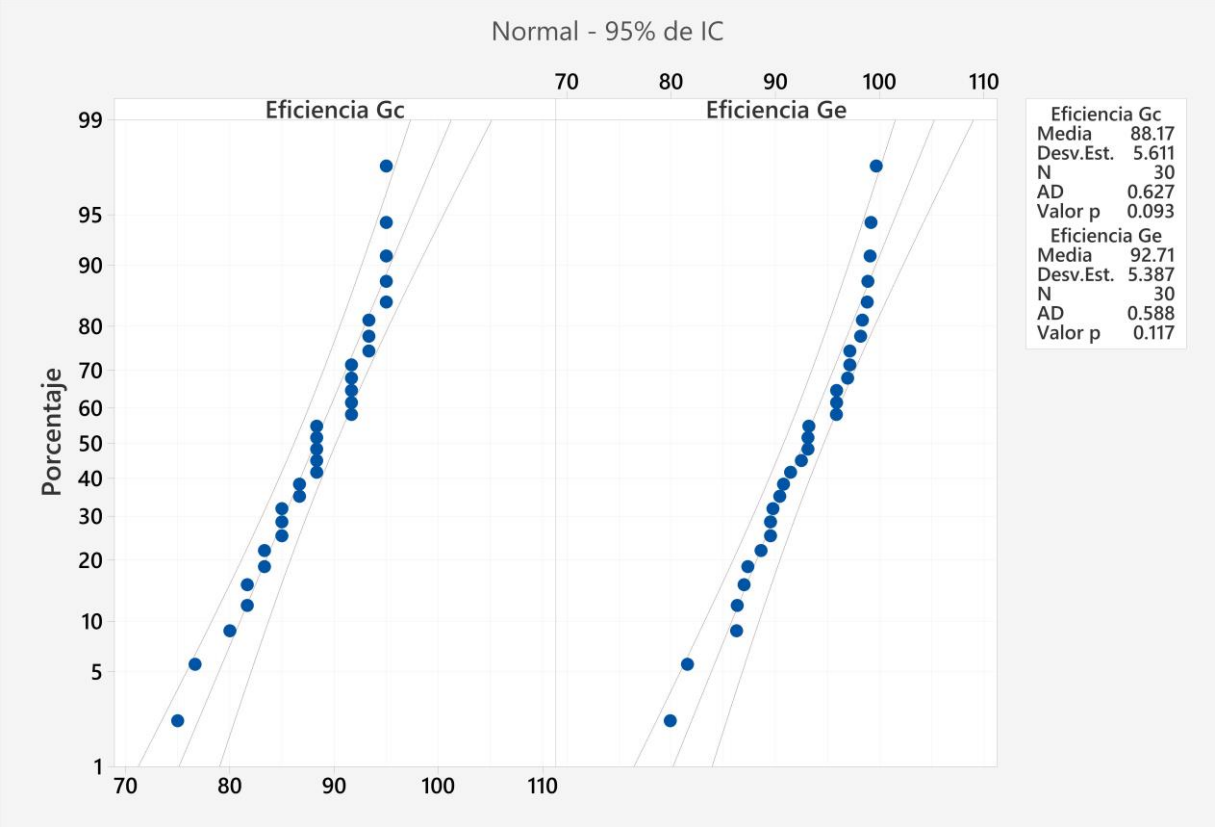
- Por otro lado, en el «Grupo experimental», el 53% de las observaciones individuales superaron su promedio de eficiencia del 92.71%. A pesar de que este porcentaje es menor que el del «Grupo de control», sigue siendo significativo, especialmente teniendo en cuenta que el promedio de eficiencia del «Grupo experimental» ya era más alto.

Prueba de Normalidad

Para evaluar la distribución de los datos y determinar si se ajustan a una distribución normal, se llevó a cabo una prueba de normalidad. Como parte de este proceso, se generó un gráfico de probabilidad, el cual se puede visualizar en Figura 22. En este gráfico, se puede apreciar claramente que tanto los datos del «Grupo de control» como los del «Grupo experimental» siguen una distribución que se asemeja a la normal. Una indicación clave de esto es que los valores p obtenidos en el análisis son superiores a 0.05, lo que, considerando un índice de confianza del 95%, respalda la hipótesis de normalidad.

Figura 22

Prueba de normalidad: Eficiencia de la estimación

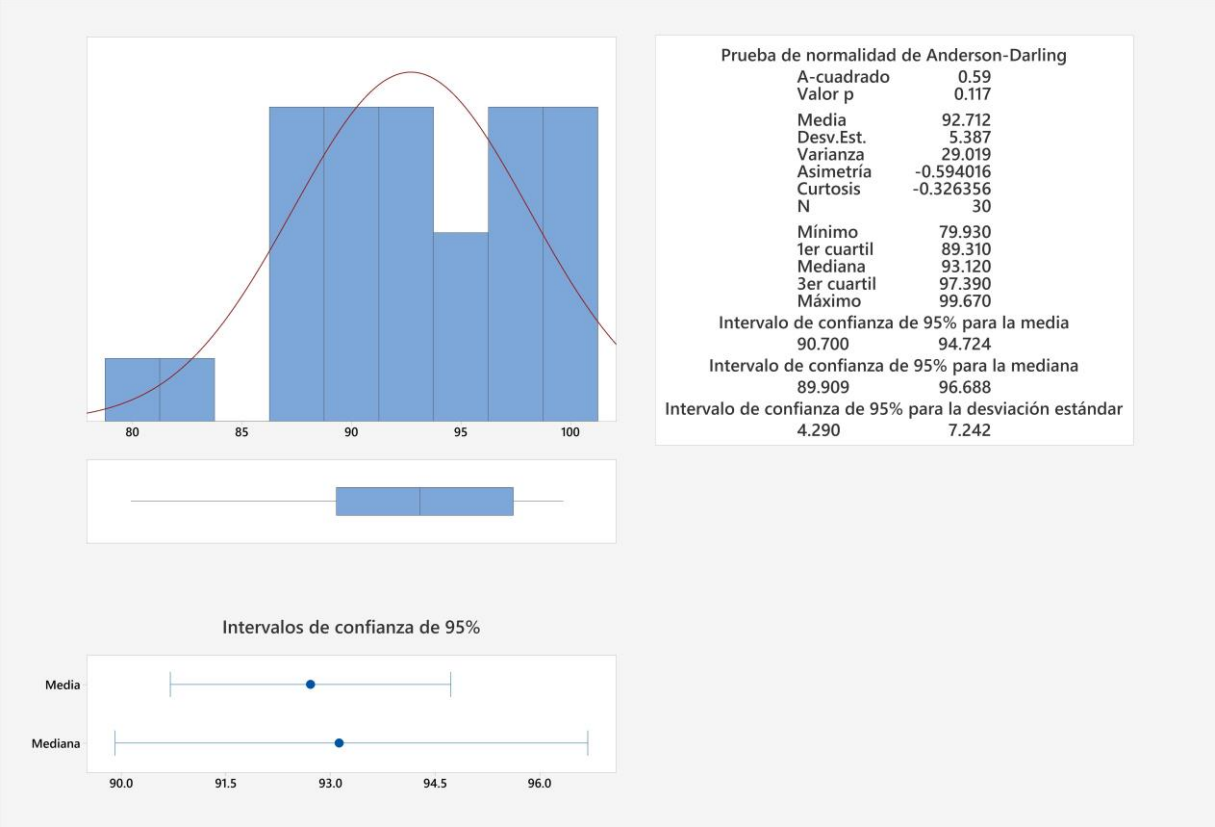


Nota: Generado con el software Minitab 20.3

Adicionalmente, para obtener un análisis más detallado y complementario del «Grupo experimental», se elaboró un informe resumen que presenta diversos estadísticos descriptivos y resultados de pruebas de normalidad. En este informe, que se presenta en Figura 23, se em- pleó específicamente la prueba de normalidad «Anderson Darling», que es una de las pruebas estadísticas más robustas y ampliamente utilizadas para evaluar la normalidad de una muestra.

Figura 23

Informe de resumen: Eficiencia de la estimación



Nota: Generado con el software Minitab 20.3

a) A partir de los resultados, se nota que el valor $p = 0.117$ es superior al nivel de significancia $\alpha = 0.05$. Esto sugiere que no hay evidencia suficiente para rechazar la hipótesis nula, por lo que se confirma que los datos siguen una distribución normal y son adecuados para su posterior análisis.

b) Al analizar la media de la «Eficiencia de la estimación», se encuentra que con un 95% de confianza, su valor está comprendido entre el 90.70% y 94.724%. Además, es interesante notar que la mediana se encuentra en un rango que traslapa a la media, específicamente entre el 89.909% y 96.688% para la misma métrica.

c) En cuanto a la dispersión de los datos, la «Eficiencia de la estimación» más alejada de la media, basada en la desviación estándar, es del 5.387%. Sin embargo, es importante considerar que estos valores pueden fluctuar en un rango que va desde el 4.29% hasta el 7.242%.

d) La asimetría registrada es de -0.594016, lo que indica una asimetría negativa. Esto sugiere que hay una acumulación de valores ligeramente inferiores en la distribución de datos

con respecto a la «Eficiencia de la estimación».

e) La curtosis, con un valor de -0.326356, señala que la distribución tiene colas ligeramente más pesadas y un pico menos pronunciado que una distribución normal, lo que indica una menor concentración de datos alrededor de la media.

f) De acuerdo con el dato del primer cuartil, se puede inferir que el 25% de las observaciones de la «Eficiencia de la estimación» son iguales o inferiores al 89.31%.

g) Al considerar el segundo cuartil, que coincide con la mediana, se deduce que la mitad de las observaciones de la «Eficiencia de la estimación» son iguales o inferiores al 93.12%.

h) Finalmente, el tercer cuartil nos indica que el 75% de las observaciones de la «Eficiencia de la estimación» son iguales o menores al 97.39%.

4.4.2. *Indicador 2: Tiempo de la Estimación*

El tiempo necesario para llevar a cabo la estimación es un aspecto crucial para evaluar la eficiencia y usabilidad de cualquier solución basada en Machine Learning. En este contexto, el indicador «Tiempo de la Estimación» se refiere al lapso promedio necesario para realizar una estimación usando la solución propuesta.

Tabla 17

Resultados sobre el indicador Tiempo de la estimación.

N	PosPrueba del Grupo de control - G_c (Minutos)	PosPrueba del Grupo experimental - G_e (Minutos)
μ	3.28	2.23
$Nro. <= \mu$	15	17
%	50	57

Para ofrecer una visión clara y detallada de este indicador, los resultados se han consolidado y presentado en la Tabla 17. Es importante señalar que los valores reflejados en esta tabla representan el promedio del tiempo requerido para realizar las estimaciones a lo largo de un periodo de 3 semanas. Estos datos proporcionan insights valiosos sobre la eficiencia de la solución y su aplicabilidad en escenarios del mundo real, donde el tiempo es a menudo un recurso limitado.

- Al analizar los tiempos promedio de estimación, se destaca que el grupo experimental tiene un tiempo medio de 2.23 minutos, lo cual, curiosamente, es inferior al tiempo promedio del grupo de control que es de 3.28 minutos. Esto sugiere que la solución aplicada al grupo experimental podría ser más eficiente en términos de tiempo.

- Una observación interesante es que en el grupo de control, el 50% de las estimaciones se completaron en un tiempo inferior a su promedio de 3.28 minutos. Esto indica una tendencia centralizada hacia estimaciones más rápidas en la mitad de los casos dentro de este grupo.

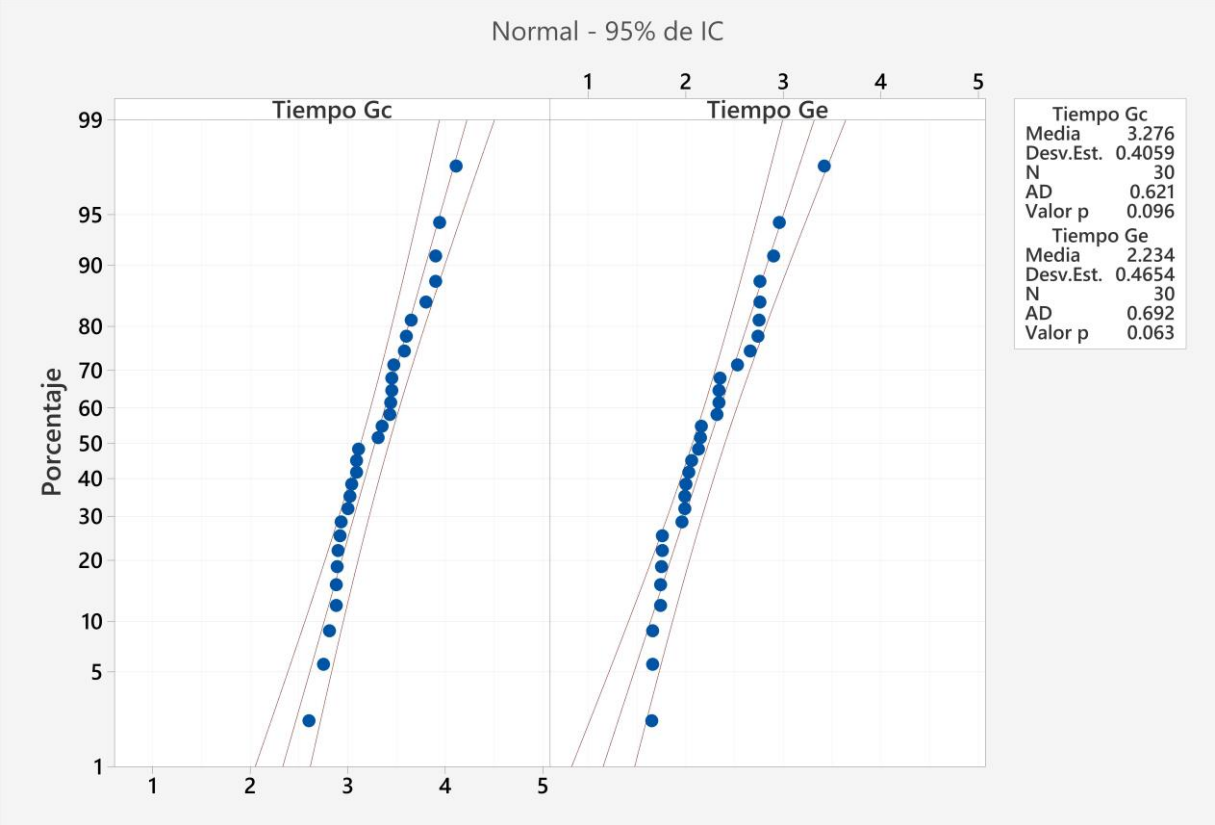
- Similarmente, en el grupo experimental, el 57% de las estimaciones se completaron en un lapso menor al promedio de 2.23 minutos. Esto refuerza la idea de que la mayoría de las estimaciones en este grupo tienden a ser más rápidas que su promedio general, sugiriendo una alta eficiencia en la mayoría de los casos.

Prueba de Normalidad

Para evaluar la distribución de los datos y determinar si se ajustan a una distribución normal, se llevó a cabo una serie de pruebas de normalidad. Una herramienta visual esencial para este propósito es el gráfico de probabilidad. Como se puede apreciar en el Figura 24, tanto el grupo de control como el experimental parecen seguir una distribución normal en sus datos. Este comportamiento se corrobora con valores p que superan el umbral de 0.05, lo que indica que podemos estar razonablemente seguros, con un 95% de confianza, de la normalidad de los datos.

Figura 24

Prueba de normalidad: Tiempo de la estimación

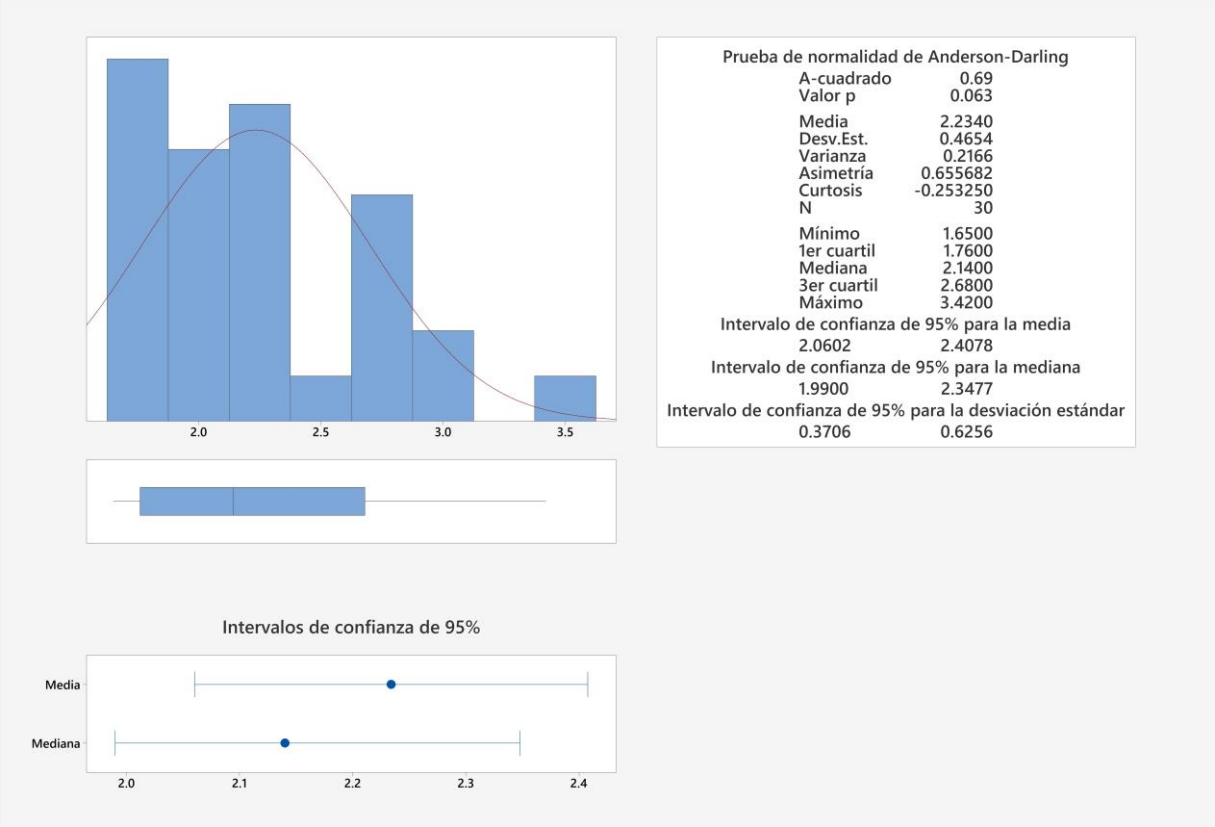


Nota: Generado con el software Minitab 20.3

Adicionalmente, para obtener una comprensión más profunda y cuantitativa de la normalidad en el grupo experimental, se generó un informe resumen que se ilustra en la Figura 25. Este informe utiliza la prueba de normalidad «Anderson Darling», que es una de las pruebas estadísticas más confiables para evaluar la normalidad de una distribución. Los resultados proporcionan una base sólida para los análisis subsiguientes, asegurando que se cumplan los supuestos necesarios para las pruebas estadísticas utilizadas.

Figura 25

Informe de resumen: Tiempo de la estimación



Nota: Generado con el software Minitab 20.3

a) Basándonos en la prueba de normalidad, se observa que el valor $p = 0.063$ supera el umbral crítico $\alpha = 0.05$. Esto confirma que la distribución de los datos se ajusta razonablemente bien a una distribución normal, lo que valida su uso en análisis estadísticos que asumen la normalidad.

b) Con un nivel de confianza del 95%, la media estimada del «Tiempo de la estimación» se encuentra en el rango de 2.0602 minutos a 2.4078 minutos. Es interesante notar que el intervalo para la mediana se solapa con el de la media, variando entre 1.99 minutos y 2.3477 minutos.

c) Tomando en cuenta la desviación estándar, el tiempo de estimación más alejado del promedio es de 0.4654 minutos. No obstante, este valor puede oscilar entre 0.3706 minutos y 0.6256 minutos.

d) La asimetría calculada es positiva con un valor de 0.655682, sugiriendo que la distribución tiene una cola más larga hacia la derecha. Esto indica que hay un sesgo hacia tiempos de estimación más cortos.

e) La curtosis resultó en -0.25325, lo que sugiere que la distribución presenta picos menos pronunciados en comparación con una distribución normal.

f) El primer cuartil nos informa que el 25% de las observaciones del «Tiempo de la estimación» son iguales o menores a 1.76 minutos.

g) El segundo cuartil, que coincide con la mediana, señala que la mitad de las observaciones tienen un «Tiempo de la estimación» de 2.14 minutos o menos.

h) Por último, el tercer cuartil indica que el 75% de las observaciones tienen un «Tiempo de la estimación» que no supera los 2.68 minutos.

4.4.3. *Indicador 3: Costo de la Estimación*

El Costo de la estimación es un factor crucial para determinar la viabilidad y eficiencia económica de cualquier solución basada en Machine Learning. Estos costos pueden influir en decisiones sobre la adopción y adaptación de la tecnología, y proporcionar una comprensión más profunda de la inversión requerida en relación con los beneficios obtenidos.

Tabla 18

Resultados sobre el indicador Costo de la estimación.

N	PosPrueba del Grupo de control - G_c (Dólares)	PosPrueba del Grupo experimental - G_e (Dólares)
μ	2.46	1.79
$Nro. <= \mu$	15	17
%	50	57

Los resultados obtenidos sobre este indicador están detallados en la Tabla 18. Estos resultados representan el promedio de los costos asociados a la estimación, recopilados durante un período de 3 semanas. Esta recopilación de datos tiene como objetivo analizar la usabilidad y eficiencia económica de la solución de Machine Learning propuesta, y así proporcionar una perspectiva clara sobre la inversión requerida en comparación con el valor añadido que ofrece al proceso de estimación.

- A partir de los resultados obtenidos, se destaca que el grupo experimental registró un costo

promedio de 1.79 dólares, lo cual es evidentemente más económico en comparación con el costo promedio de 2.46 dólares del grupo de control. Esta diferencia de costos sugiere que la implementación de ciertas técnicas o procedimientos en el grupo experimental pudo haber resultado en ahorros significativos.

- Al examinar los datos del grupo de control, se detecta que la mitad de las estimaciones se llevó a cabo a un costo que resultó ser inferior al promedio de 2.46 dólares. Esto podría indicar que, aunque el promedio general es más alto, existen varias instancias donde el costo de la estimación fue más eficiente.

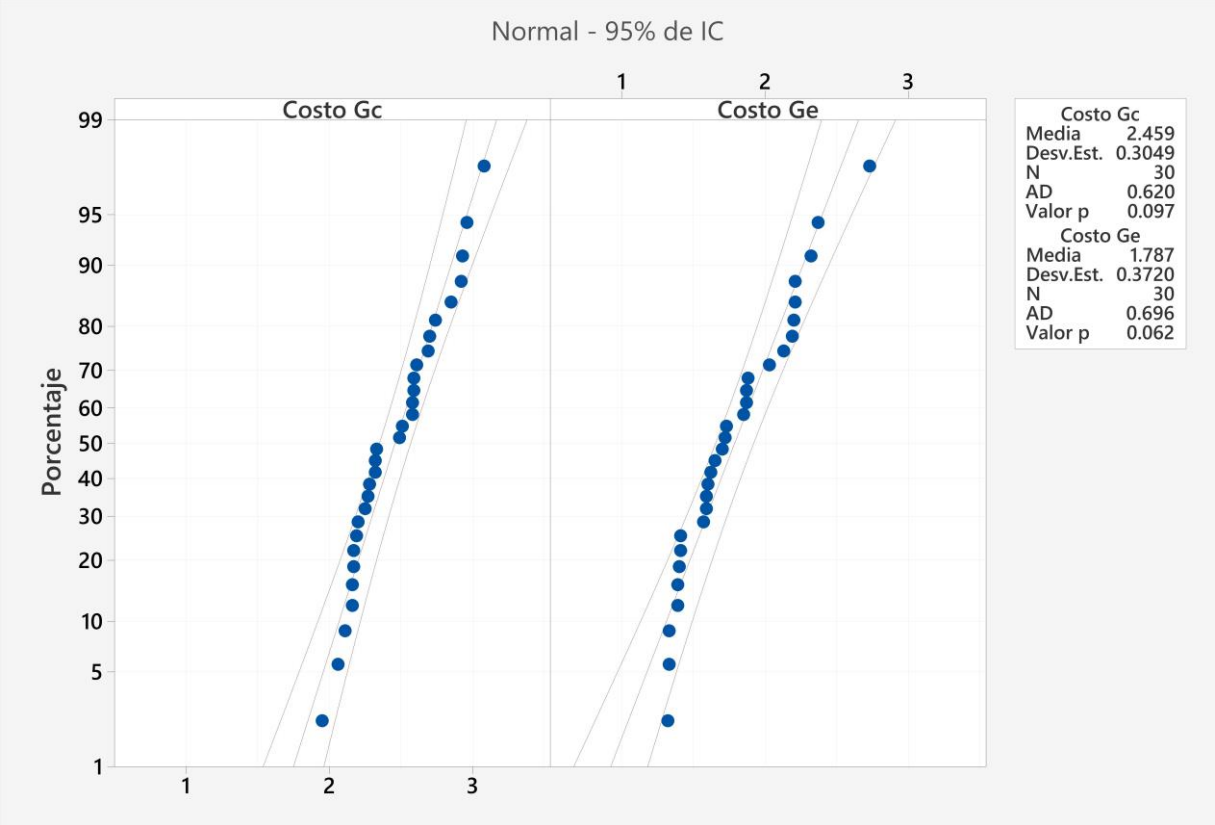
- De manera análoga, en el grupo experimental, el 57% de las estimaciones se efectuó a un costo que fue menor al promedio de 1.79 dólares. Esto refuerza la idea de que una amplia mayoría de las estimaciones en este grupo se realizó de manera más rentable y eficaz, consolidando aún más la eficiencia del grupo experimental en términos de costos.

Prueba de Normalidad

Para evaluar la distribución de los datos y determinar si se ajustan a una distribución normal, se elaboró un gráfico de probabilidad que se presenta en la Figura 26. En dicho gráfico, se puede visualizar claramente que tanto los datos del grupo de control como los del grupo experimental se alinean estrechamente con una distribución normal, lo que es esencial para la validez de muchos procedimientos estadísticos. Al considerar un índice de confianza del 95%, los valores p obtenidos son superiores al umbral de 0.05, lo que corrobora la normalidad de los datos.

Figura 26

Prueba de normalidad: Costo de la estimación

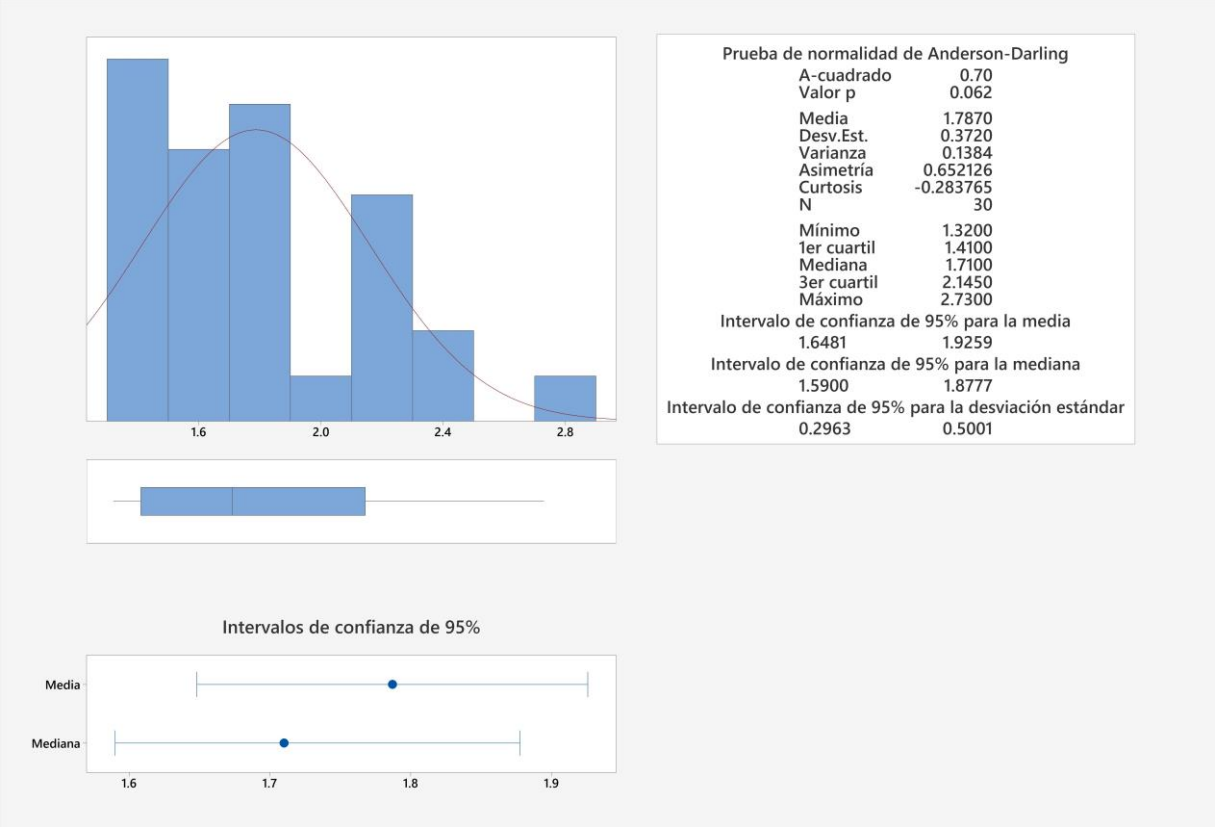


Nota: Generado con el software Minitab 20.3

Como parte del análisis en profundidad, se generó un informe detallado enfocado en el grupo experimental, cuyo resumen se puede apreciar en la Figura 27. En este informe, se empleó la prueba de normalidad conocida como «Anderson Darling», que es ampliamente reconocida por su capacidad para detectar desviaciones de la normalidad. Este informe proporciona información adicional y refuerza la evidencia sobre la naturaleza de la distribución de los datos.

Figura 27

Informe de resumen: Costo de la estimación



Nota: Generado con el software Minitab 20.3

a) Basado en los resultados de la prueba de normalidad, se observa que el valor p = 0.062 es superior al nivel de significancia α establecido en 0.05. Esto confirma la hipótesis de que los datos siguen una distribución normal, lo que es fundamental para la validación de los procedimientos estadísticos que se aplicarán posteriormente.

b) Al considerar un nivel de confianza del 95%, se estima que la «Media» del «Costo de la estimación» se encuentra en el rango de 1.6481 dólares a 1.9259 dólares. Es interesante notar que la «Mediana», que es el valor que divide la distribución en dos partes iguales, traslapa a la media en un intervalo que va desde 1.59 dólares hasta 1.8777 dólares.

c) La variabilidad en el «Costo de la estimación» es medida por la «Desviación Estándar», que se sitúa en 0.3720 dólares. Sin embargo, es importante tener en cuenta que este valor puede oscilar entre 0.2963 dólares y 0.5001 dólares.

d) La «Asimetría» de la distribución es positiva y se valora en 0.652126. Esto sugiere que la distribución tiene una cola más larga hacia la derecha, lo que indica una presencia de valores positivos bajos en relación con el «Costo de la estimación».

e) La «Curtosis» tiene un valor de -0.283765, lo que indica que la distribución tiene colas más livianas y menos picos pronunciados en comparación con una distribución normal.

f) Del análisis de los cuartiles, el primer cuartil nos muestra que el 25% de las observaciones tienen un «Costo de la estimación» de 1.41 dólares o menos.

g) El segundo cuartil, que coincide con la «Mediana», nos indica que la mitad de los datos tiene un «Costo de la estimación» de 1.71 dólares o menos.

h) Finalmente, el tercer cuartil revela que el 75% de los datos se encuentra por debajo de los 2.145 dólares en cuanto al «Costo de la estimación».

4.5. Contrastación de las Hipótesis

Este capítulo proporciona un análisis detallado y discusión acerca de los resultados obtenidos en el experimento, poniendo un énfasis particular en la contrastación y validación de las hipótesis planteadas en la investigación. Se analiza el impacto de la implementación de una solución de Machine Learning en la eficiencia de la estimación de la influencia del estilo de vida en el riesgo de obesidad, basándose en la data recolectada de las poblaciones de Colombia, México y Perú.

4.5.1. *Contrastación de H1*

Planteamiento de la Hipótesis

La hipótesis 1 (H1) postula que: El empleo de una solución basada en Machine Learning **incrementa de manera significativa la eficiencia** en la estimación de la influencia del estilo de vida sobre el riesgo de obesidad en las poblaciones de Colombia, México y Perú. La eficiencia aquí se entiende como la capacidad de la solución propuesta para producir resultados precisos y confiables en comparación con métodos tradicionales.

H_0: La solución de Machine Learning no incrementa significativamente la eficiencia de la estimación.

H_a: La solución de Machine Learning incrementa significativamente la eficiencia de la estimación.

Para expresar matemáticamente la hipótesis:

μ_1 = Media de la eficiencia de la estimación PosPrueba en el G_c (Grupo Control)

μ_2 = Media de la eficiencia de la estimación PosPrueba en el G_e (Grupo Experimental)

Donde las hipótesis nula (H_0) y alternativa (H_a) se definen como:

$$H_0 : \mu_1 \geq \mu_2$$
$$H_a : \mu_1 < \mu_2$$

Nivel de Significancia

El nivel de significancia se ha establecido en $\alpha = 0.05$, lo que implica que estamos dispuestos a aceptar un 5% de probabilidad de rechazar la hipótesis nula H_0 cuando esta es verdadera. Este nivel se selecciona comúnmente en la investigación científica, ya que proporciona un balance sólido entre la confiabilidad y la rigurosidad de los resultados, operando bajo un nivel de confianza del 95%.

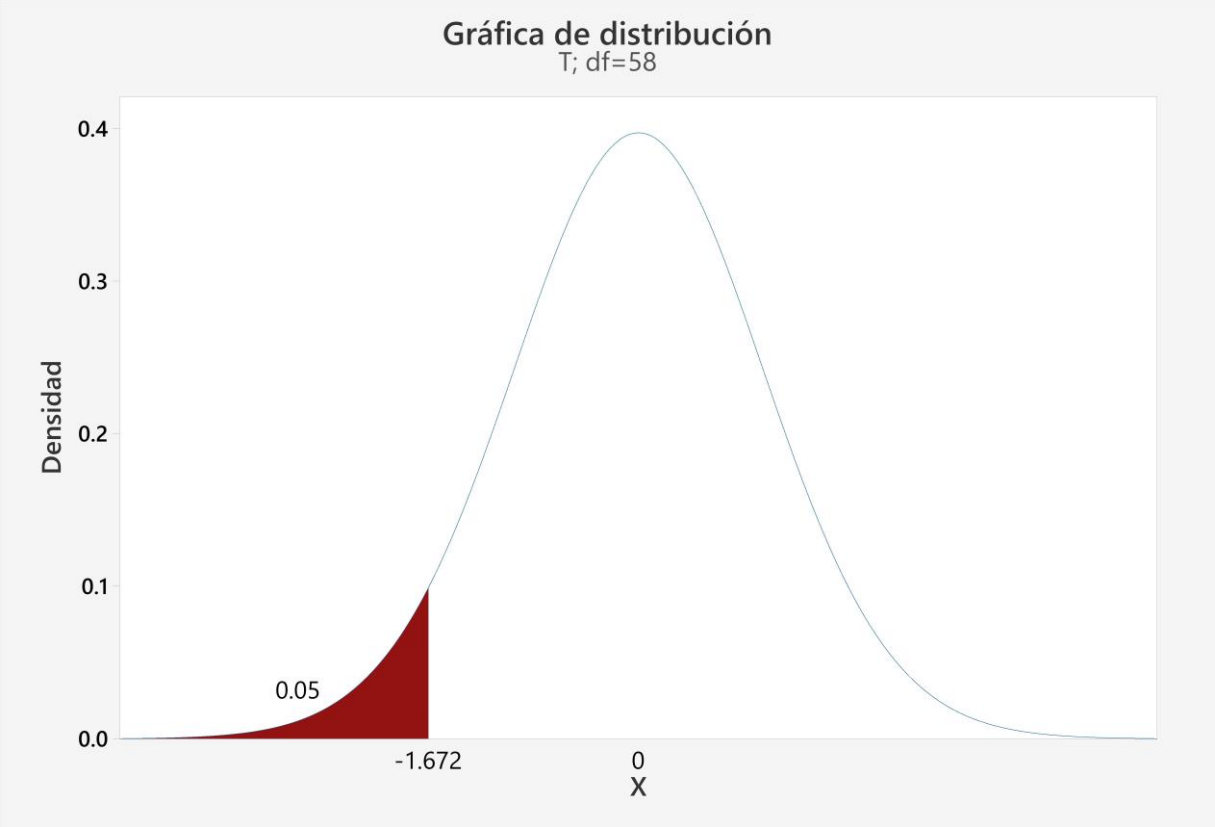
Valores de la Región de Aceptación y Rechazo

En esta sección se ofrecen los resultados detallados de los valores estadísticos derivados de las muestras de la PosPrueba tanto para el Grupo Control (G_c) como para el Grupo Experimental (G_e). Es crucial destacar que el valor p juega un papel vital al decidir si los resultados experimentales caen dentro de la región de aceptación o de rechazo, proporcionando un criterio robusto para la validación de la hipótesis.

Visualización de la Distribución t En la Figura 28, se visualiza la gráfica de la distribución t, para un grado de libertad de 58, ilustrando claramente las regiones de aceptación y rechazo de la hipótesis H_1.

Figura 28

Definición de las Regiones de Aceptación y Rechazo para la Hipótesis H1



Nota: Generado con el software Minitab 20.3

Análisis Descriptivo En la Tabla 19, se presentan los estadísticos descriptivos (media, desviación estándar y error estándar de la media) de las muestras PosPrueba de G_c y G_e, específicamente en lo que respecta a la eficiencia de la estimación.

Tabla 19

Estadísticas descriptivas de la PosPrueba del G_c y G_e de la H1

Muestra	N	Media	Desviación estándar	Error estándar de la media
PosPrueba G_c - Eficiencia	30	88.17	5.611	1.02
PosPrueba G_e - Eficiencia	30	92.71	5.387	0.984

Nota: *Generado con el software Minitab 20.3*

Diferencias entre las Muestras A continuación, en la Tabla 20, se detallan las diferencias de las medias entre ambas muestras, proporcionando un vistazo cuantitativo a la variación experimentada entre los dos grupos.

Tabla 20

Diferencia de las muestras de la H1

Diferencia de la Media	Diferencia de la Desviación estándar	Diferencia del Error estándar de la media	Límite superior del 95% para la diferencia
-4.55	0.224	0.036	-2.17

Nota: *Generado con el software Minitab 20.3*

Evaluación de la Prueba t La Tabla 21 muestra la prueba t calculada para H_1, proporcionando una visión precisa de la significancia estadística de los resultados experimentales.

Tabla 21

Estadísticas de la Prueba t de la H1

Hipótesis nula	H_o: $\mu_1 - \mu_2 \geq 0$
Hipótesis alterna	H_a: $\mu_1 - \mu_2 < 0$
Valor t	-3.20
Valor p	0.001
Grados de libertad (GL)	58

Nota: *Generado con el software Minitab 20.3*

Análisis del Resultado Con grados de libertad definidos como $(G_c-1)+(G_e-1) = 58$ para el conjunto total de las muestras, y un valor crítico de -1.672 (que delimita la zona de aceptación con un 95%de confianza y define una región de rechazo en la cola izquierda con $\alpha = 0.05$), se puede observar que el valor calculado $T = -3.20$ se ubica firmemente dentro de la zona de rechazo. Además, el valor $p = 0.001 < \alpha = 0.05$, indica que existe evidencia suficiente para considerar la prueba de hipótesis como significativa, validando así nuestra hipótesis alternativa.

Determinación de la Decisión y Conclusión Estadística

Tras un análisis exhaustivo de los datos recabados y evaluación de los resultados de las pruebas estadísticas, se llega a una decisión respecto a las hipótesis planteadas:

Se rechaza la hipótesis nula, H_0: $\mu_1 \geq \mu_2$, y por ende, se acepta la hipótesis alternativa, H_a: $\mu_1 < \mu_2$.

Con una confianza del 95%, y permitiéndonos un margen de error del 5%, se concluye que se debe rechazar la hipótesis nula, H_0, que sostiene que la eficiencia de la estimación de la influencia del estilo de vida en el riesgo de obesidad de la población es mayor o igual en la PosPrueba del Grupo Control, G_c, comparada con la del Grupo Experimental, G_e. La evidencia estadística respalda firmemente la aceptación de la hipótesis alternativa, indicando que la eficiencia de la estimación es significativamente menor en la PosPrueba de G_c en comparación con G_e.

En definitiva, se puede afirmar con una certeza estadística del 95% que la Hipótesis 1 (H1) es válida y los resultados obtenidos son significativos. Este resultado implica que el uso de una solución de Machine Learning ha demostrado ser efectivo al incrementar significativamente la eficiencia en la estimación de la influencia del estilo de vida en el riesgo de obesidad para las poblaciones de Colombia, México y Perú.

4.5.2. *Contrastación de H2*

Planteamiento de la Hipótesis

La Hipótesis 2 (H2) postula que: La implementación de una solución basada en Machine Learning **reducirá significativamente el tiempo** necesario para estimar la influencia del estilo de vida en el riesgo de obesidad en las poblaciones de Colombia, México y Perú.

H_0: La solución de Machine Learning no disminuye significativamente el tiempo de la estimación.

H_a: La solución de Machine Learning disminuye significativamente el tiempo de la estimación.

Se define como:

μ_1 = Media del tiempo de estimación en la PosPrueba del G_c (Grupo Control)

μ_2 = Media del tiempo de estimación en la PosPrueba del G_e (Grupo Experimental)

Donde las hipótesis nula (H_0) y alternativa (H_a) se definen como:

$$H_0 : \mu_1 \leq \mu_2$$

$$H_a : \mu_1 > \mu_2$$

Nivel de Significancia

El nivel de significancia adoptado para esta prueba hipotética es $\alpha = 0.05$, lo que implica que estamos dispuestos a aceptar un 5% de probabilidad de rechazar la hipótesis nula, H_0, cuando esta es verdadera, basándonos en nuestro nivel de confianza del 95%.

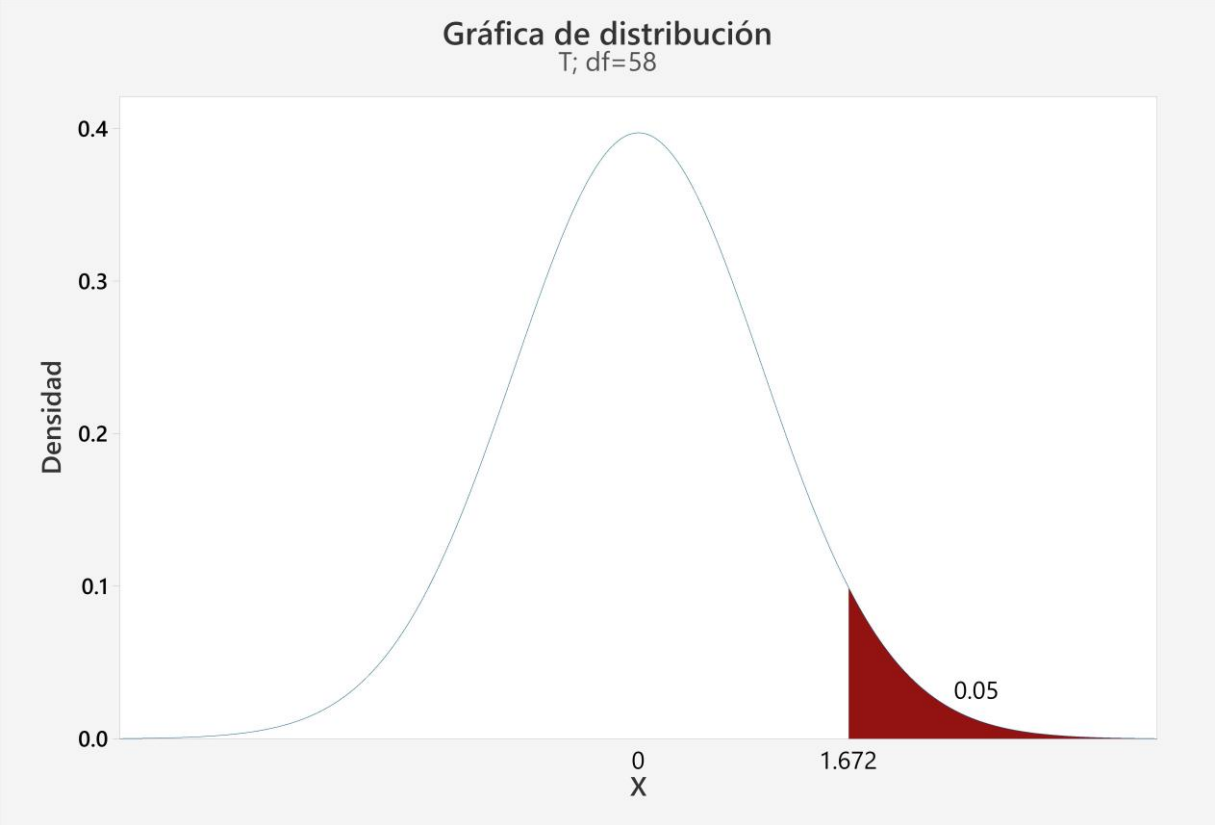
Valores de la Región de Aceptación y Rechazo

Este segmento se dedica a presentar y analizar los valores estadísticos derivados de las diferencias entre las muestras de la PosPrueba de los grupos de control, G_c, y experimental, G_e. En particular, se presta atención al valor p para discernir entre las regiones de aceptación y rechazo de la hipótesis.

Visualización de la Distribución t La Figura 29 ilustra la distribución t, estableciendo así las regiones de aceptación y rechazo para la hipótesis H2 con un grado de libertad de 58.

Figura 29

Definición de las Regiones de Aceptación y Rechazo para la Hipótesis H2



Nota: Generado con el software Minitab 20.3

Análisis Descriptivo En la Tabla 22, se exhiben las estadísticas descriptivas, que incluyen la media, la desviación estándar y el error estándar de la media, de las muestras tomadas en la PosPrueba de ambos G_c y G_e, las cuales están intrínsecamente relacionadas con el tiempo de estimación.

Tabla 22

Estadísticas descriptivas de la PosPrueba del G_c y G_e de la H2

Muestra	N	Media	Desviación estándar	Error estándar de la media
PosPrueba G_c - Tiempo	30	3.276	0.4059	0.0741
PosPrueba G_e - Tiempo	30	2.234	0.4654	0.0850

Nota: *Generado con el software Minitab 20.3*

Diferencias entre las Muestras Las diferencias entre las medias de estas muestras se exploran en la Tabla 23, proporcionando así una perspectiva cuantitativa sobre el grado de variación experimentado entre los dos grupos bajo estudio.

Tabla 23

Diferencia de las muestras de la H2

Diferencia de la Media	Diferencia de la Desviación estándar	Diferencia del Error estándar de la media	Límite inferior del 95% para la diferencia
1.042	-0.0595	-0.0109	0.854

Nota: *Generado con el software Minitab 20.3*

Evaluación de la Prueba t A continuación, en la Tabla 24, se presenta la prueba t de H2, la cual es esencial para determinar la significancia de las diferencias observadas.

Tabla 24

Estadísticas de la Prueba t de la H2

Hipótesis nula	H_o: $\mu_1 - \mu_2 \leq 0$
Hipótesis alterna	H_a: $\mu_1 - \mu_2 > 0$
Valor t	9.25
Valor p	0.000
Grados de libertad (GL)	58

Nota: *Generado con el software Minitab 20.3*

Análisis del Resultado Dado que el total de grados de libertad para ambas muestras es de $(G_c - 1) + (G_e - 1) = 58$, se deriva un valor crítico de 1.672, que define la frontera de la zona de aceptación al 95% y marca el inicio de la región de rechazo en una cola a la derecha de 0.05. El valor observado de $T = 9.25$ se encuentra firmemente dentro de la zona de rechazo y, además, dado que el valor $p = 0.000 < \alpha = 0.05$, se concluye que existe evidencia suficientemente fuerte para considerar la prueba de hipótesis como significativa. Este análisis se discutirá y contextualizará más a fondo en las siguientes secciones, abordando sus implicaciones y potencial impacto en el ámbito de estudio.

Determinación de la Decisión y Conclusión Estadística

El proceso de toma de decisiones en el contexto estadístico se basa en una evaluación crítica y metodológica de los resultados experimentales obtenidos. En este caso, la hipótesis nula, H_0: $\mu_1 \leq \mu_2$, y la hipótesis alternativa, H_a: $\mu_1 > \mu_2$, se someten a un escrutinio basado en los datos recabados y los análisis estadísticos realizados.

Se establece que, bajo un nivel de significancia del 5%, se rechaza la hipótesis nula H_0 que postula que el Tiempo de la estimación de la influencia del estilo de vida en el riesgo de obesidad de la población es menor o igual en la PosPrueba del grupo de control G_c que en la PosPrueba del grupo experimental G_e. Esta decisión se basa en la evidencia estadística recabada y analizada en las secciones anteriores, la cual sugiere un soporte significativo para la hipótesis alternativa H_a.

La hipótesis alternativa, H_a: $\mu_1 > \mu_2$, plantea que el Tiempo de la estimación de la influencia del estilo de vida en el riesgo de obesidad de la población es, de hecho, mayor en la PosPrueba del G_c que en la PosPrueba del G_e. A la luz de los datos analizados y los valores p obtenidos, esta hipótesis es aceptada, indicando que los resultados son estadísticamente significativos y que la implementación de la solución de Machine Learning influyó positivamente en la reducción del tiempo de estimación en el grupo experimental.

Por ende, se confirma que la hipótesis H_2 es verdadera, respaldada por resultados estadísticamente significativos. La validez de esta afirmación y sus posibles implicaciones para la investigación en general y para las futuras aplicaciones de soluciones de Machine Learning en este campo en particular, se explorarán y discutirán en profundidad en las secciones subsiguientes.

4.5.3. *Contrastación de H3*

Planteamiento de la Hipótesis

La evaluación del costo es fundamental para determinar la viabilidad y eficiencia económica de la solución propuesta y su potencial aplicación a escala. La hipótesis 3 (H3) postula que: La implementación de una solución de Machine Learning **reduce significativamente el costo** asociado a la estimación de la influencia del estilo de vida en el riesgo de obesidad de las poblaciones mencionadas.

H_0: La solución de Machine Learning no disminuye significativamente el costo de la estimación.

H_a: La solución de Machine Learning disminuye significativamente el costo de la estimación.

Donde:

μ_1 = Media del costo de la estimación en la PosPrueba del grupo de control G_c (Grupo Control)

μ_2 = Media del costo de la estimación en la PosPrueba del grupo experimental G_e (Grupo Experimental)

Donde las hipótesis nula (H_0) y alternativa (H_a) se definen como:

$$H_0 : \mu_1 \leq \mu_2$$

$$H_a : \mu_1 > \mu_2$$

Nivel de Significancia

El nivel de significancia establecido para este análisis es de $\alpha = 0.05$, lo que indica un 5% de tolerancia para el error tipo I, es decir, la probabilidad de rechazar la hipótesis nula cuando esta es verdadera. Este valor, ampliamente utilizado en la ciencia de datos, se elige para garantizar que los resultados obtenidos no se deban simplemente a variaciones aleatorias en los datos, permitiendo así un 95% de confianza en las decisiones tomadas a partir de este análisis.

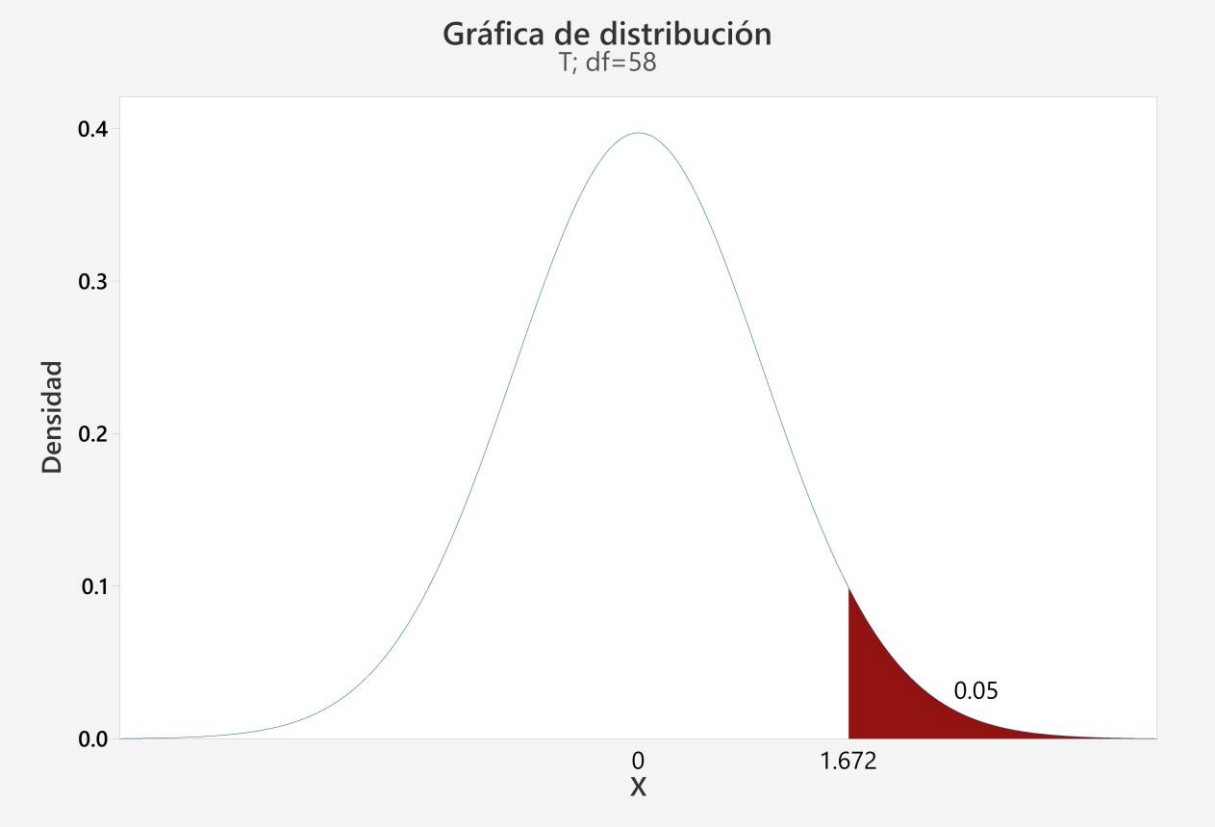
Valores de la Región de Aceptación y Rechazo

Es fundamental explorar y comprender las diferencias observadas entre las muestras de la PosPrueba de los grupos de control (G_c) y experimental (G_e) para validar la hipótesis propuesta.

Visualización de la Distribución t La Figura 30 visualiza la distribución t, indicando las regiones de aceptación y rechazo de la hipótesis H3 y proporcionando una representación gráfica de los resultados de la prueba t y su relación con el nivel de significancia establecido.

Figura 30

Definición de las Regiones de Aceptación y Rechazo para la Hipótesis H3



Nota: Generado con el software Minitab 20.3

Análisis Descriptivo La Tabla 25 presenta una visión general de las características fundamentales de las muestras, incluyendo media, desviación estándar y error estándar de la media, proporcionando una base sólida para una interpretación más detallada de los datos.

Tabla 25

Estadísticas descriptivas de la PosPrueba del G_c y G_e de la H3

Muestra	N	Media	Desviación estándar	Error estándar de la media
PosPrueba G_c - Costo	30	2.459	0.305	0.056
PosPrueba G_e - Costo	30	1.787	0.372	0.068

Nota: *Generado con el software Minitab 20.3*

Diferencias entre las Muestras La Tabla 26 detalla las diferencias de las medias entre las muestras, ofreciendo una visión cuantitativa de la variación observada y proporcionando una base para evaluar la significancia de estos cambios.

Tabla 26

Diferencia de las muestras de la H3

Diferencia de la Media	Diferencia de la Desviación estándar	Diferencia del Error estándar de la media	Límite inferior del 95% para la diferencia
0.6717	-0.067	-0.012	0.5249

Nota: *Generado con el software Minitab 20.3*

Evaluación de la Prueba t La prueba t, que se detalla en la Tabla 27, se realiza para comparar las medias de los dos grupos y determinar si las diferencias observadas son estadísticamente significativas.

Tabla 27

Estadísticas de la Prueba t de la H3

Hipótesis nula	H_o: $\mu_1 - \mu_2 \leq 0$
Hipótesis alterna	H_a: $\mu_1 - \mu_2 > 0$
Valor t	7.65
Valor p	0.000
Grados de libertad (GL)	58

Nota: *Generado con el software Minitab 20.3*

Análisis del Resultado Con un grado de libertad definido como $(G_c - 1) + (G_e - 1) = 58$, se calcula un valor crítico de 1.672 que delimita la zona de aceptación al 95% y una zona de rechazo a la derecha del 5%. Dado que el valor $T = 7.65$ se sitúa dentro de la zona de rechazo, y que el valor $p = 0.000 < \alpha = 0.05$, se concluye que hay evidencia suficiente para rechazar la hipótesis nula, indicando que los resultados de la prueba de hipótesis son estadísticamente significativos.

Estas detalladas exploraciones y representaciones visuales permiten no solo validar las hipó-

tesis desde un punto de vista estadístico, sino también facilitar la interpretación y comprensión de los datos y resultados obtenidos, siendo crucial para la toma de decisiones informadas y la planificación de futuras investigaciones y aplicaciones.

Determinación de la Decisión y Conclusión Estadística

Dadas las pruebas estadísticas aplicadas y los resultados obtenidos, se decide rechazar la hipótesis nula $H_0 : \mu_1 \leq \mu_2$ y, por ende, aceptar la hipótesis alternativa $H_a : \mu_1 > \mu_2$.

La decisión estadística anterior conduce a la conclusión de que, bajo un nivel de significancia del 5%, existe una base sólida para rechazar la hipótesis nula H_0, la cual sostiene que el Costo de la estimación de la influencia del estilo de vida en el riesgo de obesidad de la población, evaluado en la PosPrueba, es menor o igual para el grupo de control G_c en comparación con el grupo experimental G_e.

Más aún, la evidencia estadística recabada apoya la hipótesis alternativa, afirmando que el Costo de la estimación es, de hecho, mayor en la PosPrueba del G_c que en la PosPrueba del G_e.

Por ende, se confirma que la hipótesis H_3 es verdadera, respaldada por resultados estadísticamente significativos.

4.6. Discusión de Resultados

En esta investigación se desarrolló una solución basada en Machine Learning para estimar la influencia del estilo de vida en el riesgo de obesidad de la población de Colombia, México y Perú. Este modelo de Machine Learning requiere información de 16 factores del estilo de vida que ya se recopilan como parte de la práctica clínica habitual. La integración de estas variables en un modelo computacional avanzado ofrece una nueva perspectiva en el abordaje de la obesidad, permitiendo una comprensión más profunda de cómo las variadas facetas del estilo de vida contribuyen al riesgo.

4.6.1. *Discusión del Indicador 1 - Eficiencia de la Estimación*

El modelo de Machine Learning desarrollado en este estudio demostró ser excepcionalmente eficiente, logrando una exactitud del 97.36%, precisión del 97.39%, exhaustividad del 97.37%, valor F de 97.36 y un impresionante AUC ROC de 99.90%. Estos resultados no solo evidencian

la capacidad del modelo para realizar predicciones precisas y confiables, sino que también destacan su superioridad en comparación con enfoques anteriores. Por ejemplo, la precisión obtenida supera la del modelo de Cervantes y Palacio (2020), resaltando la mejora en la identificación correcta de los casos de riesgo de obesidad. Al comparar nuestro modelo con investigaciones anteriores, es evidente que hemos alcanzado un nuevo hito. Por ejemplo, los modelos de Ferdowsy et al. (2021) y Sulla Torres (2018) mostraron una exactitud considerable, pero nuestro modelo no solo iguala este rendimiento, sino que también mejora en términos de equilibrio entre precisión y exhaustividad. Asimismo, en comparación con el modelo de De-La-Hoz-Correa et al. (2019), nuestro enfoque demuestra una mejora notable en la identificación de casos de riesgo de obesidad, lo que es esencial para la prevención y manejo efectivos de esta condición.

Además, estos resultados señalan un avance significativo en la predicción de la obesidad utilizando Machine Learning. La combinación de una alta exactitud con una impresionante área bajo la curva ROC sugiere que el modelo no solo es capaz de identificar correctamente a los individuos en riesgo, sino que también minimiza los falsos positivos y negativos. Esto es particularmente crucial en el contexto de la salud pública, donde la identificación precisa de los individuos en riesgo puede conducir a intervenciones más efectivas y oportunas.

Esta mejora en el rendimiento puede atribuirse a varios factores, incluyendo la sofisticación del algoritmo utilizado, la calidad y diversidad de los datos de entrada, y la eficacia de los procesos de entrenamiento y validación del modelo. La inclusión de un conjunto amplio y diverso de factores del estilo de vida permite una comprensión más holística y matizada del riesgo de obesidad, lo cual es crucial en países con diferentes culturas y estilos de vida como Colombia, México y Perú.

Los hallazgos de este estudio tienen implicaciones significativas para la práctica clínica y la salud pública. La capacidad de predecir con precisión el riesgo de obesidad basándose en una variedad de factores del estilo de vida puede facilitar la identificación temprana de individuos en riesgo y la implementación de estrategias de prevención personalizadas. Esto es particularmente relevante en el contexto de la lucha contra la obesidad, una de las principales preocupaciones de salud pública en el mundo actual. Al proporcionar una herramienta de predicción precisa y confiable, podemos mejorar las estrategias de prevención y manejo de la obesidad, lo que a su vez puede tener un impacto positivo en la reducción de las tasas de obesidad y en la mejora de la salud y el bienestar de las poblaciones en riesgo.

4.6.2. *Discusión del Indicador 2 - Tiempo de la Estimación*

El tiempo promedio de la estimación, antes de aplicar la solución de Machine Learning, fue de 3 minutos y 16 segundos. Este resultado muestra una notable concordancia con los hallazgos de Torres Nolasco (2019), quienes reportaron un tiempo promedio de 3 minutos y 9 segundos en un estudio similar. La similitud en estos tiempos refleja la consistencia en las metodologías convencionales de estimación del riesgo de obesidad. Sin embargo, la verdadera innovación de nuestro estudio se manifiesta en la implementación de la solución de Machine Learning, la cual optimizó significativamente el tiempo de estimación a una media de 2 minutos 14 segundos.

La reducción en el tiempo de estimación, al aplicar la solución de Machine Learning, no solo mejora la eficiencia del proceso, sino que también tiene implicaciones directas en la práctica clínica. En entornos donde el tiempo es un recurso crítico, como en consultas médicas rápidas o en situaciones de alta demanda de servicios de salud, esta mejora en la eficiencia puede significar la diferencia entre la detección a tiempo o tardía de riesgos de obesidad. Por lo tanto, la implementación de este modelo de Machine Learning podría traducirse en una mejor atención al paciente y en la optimización de los recursos sanitarios.

La agilización del tiempo de estimación no solo beneficia la logística de las prácticas clínicas, sino que también potencia la toma de decisiones en tiempo real. Con la capacidad de obtener estimaciones rápidas y precisas, los profesionales de la salud pueden tomar decisiones informadas y oportunas, lo que es crucial para la implementación de intervenciones preventivas y estrategias de manejo personalizadas para pacientes en riesgo de obesidad.

Al comparar nuestro modelo con el de Torres Nolasco (2019), es evidente que, aunque ambos enfoques son similares en términos de precisión y metodología, nuestro modelo de Machine Learning sobresale en términos de eficiencia temporal. Este avance es un testimonio de la evolución de las tecnologías de Machine Learning y su aplicación en el campo de la salud pública. Además, sugiere que la inversión en desarrollo tecnológico, específicamente en inteligencia artificial y aprendizaje automático, es una estrategia viable y fructífera para mejorar los sistemas de salud.

En conclusión, la mejora del tiempo de estimación mediante la aplicación de soluciones de Machine Learning no solo es un avance técnico, sino que también tiene un impacto directo y significativo en la eficacia de las prácticas de salud pública. Este hallazgo abre las puertas a futuras investigaciones que podrían explorar la implementación de tales tecnologías en una variedad más amplia de contextos clínicos y de salud pública, con el potencial de mejorar aún más los procesos y resultados de atención al paciente en diversas poblaciones y entornos.

4.6.3. *Discusión del Indicador 3 - Costo de la Estimación*

El costo promedio de la estimación, antes de aplicar la solución de Machine Learning, fue de $2.46. Este dato es crucial al contrastarlo con el estudio de Andersson, Eliasson, y Steen Carlsson (2022), un referente en la investigación de los costos asociados a la obesidad. El estudio de Andersson et al. (2022) enfatiza que los costos de tratar la obesidad no se limitan a los gastos médicos directos, sino que incluyen también costos indirectos como la pérdida de productividad. Nuestro estudio demuestra cómo la implementación de soluciones basadas en Machine Learning puede mitigar estos costos. Concretamente, la aplicación de nuestro modelo de Machine Learning resultó en una disminución del costo promedio a $1.79, destacando la eficiencia en términos de costos laborales y operativos.

La precisión del modelo de Machine Learning en la identificación de riesgos de obesidad implica un beneficio adicional en términos de reducción de costos a largo plazo. Al permitir intervenciones más tempranas y precisas, se puede disminuir la necesidad de tratamientos más costosos y prolongados asociados a complicaciones avanzadas de la obesidad. Esto es particularmente relevante en el contexto de sistemas de salud con recursos limitados, donde la prevención eficiente puede tener un impacto significativo en la gestión de los gastos sanitarios.

Además de la reducción de costos directos, la implementación de nuestro modelo de Machine Learning puede tener un impacto positivo en la calidad de vida de los pacientes y en la productividad general. Al facilitar diagnósticos más rápidos y precisos, se potencia la capacidad de los individuos para gestionar mejor su salud, lo que a su vez puede llevar a una menor ausencia laboral y a una mayor productividad. Esta ventaja es particularmente relevante en el contexto de los países en estudio, donde la carga económica de la obesidad puede ser considerable.

Mientras que los resultados de nuestro estudio son prometedores, es crucial considerar los desafíos asociados con la adopción de tecnologías Machine Learning en el sector salud. Estos desafíos incluyen la necesidad de infraestructura adecuada, la formación del personal de salud en el uso de estas tecnologías, y la garantía de la privacidad y seguridad de los datos de los pacientes. Sin embargo, dados los beneficios potenciales en términos de reducción de costos y mejora en la calidad de la atención, la inversión en estas tecnologías parece ser una decisión estratégica acertada para los sistemas de salud en Colombia, México y Perú.

En conclusión, la incorporación de soluciones basadas en Machine Learning para la estimación del riesgo de obesidad no solo presenta una oportunidad para mejorar la eficiencia y reducir costos, sino que también ofrece un camino hacia una atención sanitaria más proactiva y centrada en el paciente. Los resultados de este estudio subrayan la importancia de continuar explorando y desarrollando tecnologías de Machine Learning en el ámbito de la salud pública, con el fin

de enfrentar de manera más efectiva los desafíos asociados a la obesidad y otras condiciones crónicas.

CONCLUSIONES

a) La **Eficiencia de la Estimación** fue significativamente potenciada mediante la implementación de una solución de Machine Learning, fundamentada en la recién introducida metodología DORA. Este incremento no solo valida la aplicabilidad de la solución propuesta, sino que también enfatiza la relevancia de incorporar técnicas de Machine Learning en el análisis de la influencia del estilo de vida en el riesgo de obesidad en los contextos de Colombia, México y Perú.

b) El **Tiempo de la Estimación** experimentó una notable reducción, subrayando una vez más la eficacia de la solución de Machine Learning que se apoya en la metodología DORA. Esta disminución en el tiempo necesario para realizar estimaciones puede te- ner profundas implicaciones en la rapidez con la que se pueden implementar estrategias de intervención y en la agilidad de las respuestas ante emergentes problemas de salud pública relacionados con la obesidad.

c) El **Costo de la Estimación** también se vio significativamente reducido, lo que no solo valida la hipótesis propuesta, sino que también introduce una perspectiva de viabilidad económica en la implementación de soluciones de Machine Learning en este contexto específico. Este descenso en los costos asociados a la estimación realza la sostenibilidad de tales intervenciones a largo plazo y en diferentes escalas geográficas y demográficas.

d) La implementación de soluciones basadas en Machine Learning se perfila como una **estrategia viable y económicamente eficiente**, especialmente en lo que concierne a tareas de estimación en el ámbito de la salud pública. La evidencia proveniente de esta investigación, por lo tanto, no solo refuerza la literatura existente que vincula positivamente a las soluciones de Machine Learning con la eficiencia en las estimaciones en el sector de la salud, sino que también plantea la necesidad de explorar más a fondo sus aplicaciones en otros ámbitos y contextos.

e) La metodología Data-Driven Obesity Risk Analysis (DORA) emergió como un marco de trabajo robusto y eficiente para el desarrollo de proyectos de Machine Learning, ofreciendo una ruta estructurada y replicable para futuras investigaciones y aplicaciones prác-

ticas en el campo. Además, esta metodología podría explorarse y adaptarse en futuras investigaciones para discernir su aplicabilidad y eficacia en otros contextos y para otros problemas de salud.

RECOMENDACIONES

a) **Ampliación de la aplicación metodológica:** Dada la eficacia demostrada por la metodología DORA, se recomienda su aplicación en una variedad de proyectos que también se adentren en el uso de Machine Learning, permitiendo así evaluar su versatilidad y aplicabilidad en una diversidad de contextos y problemáticas.

b) **Exploración de factores de influencia:** Es recomendable llevar a cabo investigaciones adicionales para discernir los factores adicionales que pudieran influir en los resultados observados, y cómo estos pueden ser integrados de manera efectiva en la solución de Machine Learning para enriquecer y afinar aún más las estimaciones y predicciones generadas.

c) **Análisis de sensibilidad:** Proponer un análisis de sensibilidad de los modelos utilizados en la solución de Machine Learning para determinar cómo diferentes valores de las variables de entrada afectan a las salidas del modelo, lo que puede ofrecer una visión más detallada de las variaciones y precisión de las estimaciones proporcionadas por el modelo.

d) **Estudios de caso comparativos:** Realizar estudios de caso en los que la metodología DORA se aplique en diferentes contextos (tanto geográficos como demográficos), permitiendo no solo validar la aplicabilidad y eficacia de la metodología en una variedad de escenarios, sino también identificar cualquier ajuste o adaptación necesaria para atender a peculiaridades específicas de diferentes contextos.

e) **Análisis de escalabilidad:** Considerando los resultados positivos obtenidos en este estudio, se sugiere explorar la escalabilidad de la solución de Machine Learning, evaluando su performance y eficacia en proyectos de mayor envergadura y complejidad, así como en diferentes ámbitos y sectores.

f) **Integración de tecnologías emergentes:** Se recomienda explorar la posibilidad de integrar tecnologías emergentes, como el Internet de las Cosas (IoT) y la inteligencia artificial (IA), en la solución de Machine Learning para posibilitar la recolección y análisis de datos en tiempo real, lo que podría ofrecer una perspectiva más actualizada y dinámica sobre

las tendencias y patrones observados.

g) **Investigación sobre factores de riesgo emergentes:** Con el constante cambio y evolución de los estilos de vida y los factores de riesgo asociados, sería pertinente explorar cómo los nuevos elementos (por ejemplo, nuevos patrones alimenticios, hábitos de ejercicio, o incluso factores socioeconómicos) pueden ser integrados en los modelos para reflejar de manera más precisa las realidades actuales y emergentes.

h) **Énfasis en la ética de datos:** Asegurar que todas las fases del trabajo con Machine Learning, desde la recopilación de datos hasta la implementación de modelos, estén alineadas con las normativas éticas y legales correspondientes, garantizando la privacidad y seguridad de los datos manejados y generando confianza en los stakeholders y participantes del estudio.

i) **Exploración de nuevos algoritmos de clasificación:** Se sugiere explorar la aplicación de diferentes algoritmos de clasificación, especialmente aquellos que puedan haber surgido recientemente en el campo de Machine Learning. Comparar su desempeño y eficacia con los algoritmos usados en la investigación podría revelar nuevas perspectivas o enfoques para abordar la problemática en estudio.

j) **Desarrollo de aplicaciones prácticas:** Basado en los hallazgos y modelos desarrollados, explorar el desarrollo de aplicaciones prácticas o herramientas que puedan ser utilizadas por profesionales de la salud, formuladores de políticas o incluso individuos, para informar, guiar y apoyar estrategias y decisiones relacionadas con la gestión y mitigación del riesgo de obesidad.

REFERENCIAS BIBLIOGRÁFICAS

Andersson, E., Eliasson, B., y Steen Carlsson, K. (2022, 6). Current and fu- ture costs of obesity in Sweden. *Health Policy*, *126*(6), 558–564. Des- cargado de https://www.sciencedirect.com/science/article/pii/S0168851022000665 doi: 10.1016/J.HEALTHPOL.2022.03.010

Bishop, C. (2007). Pattern Recognition and Machine Learning (Information Science and Statistics), 1st edn. 2006. corr. 2nd printing edn. *Springer, New York*.

Brownlee, J. (2021, 2). *A Gentle Introduction to XGBoost for Applied Machine Learning*. Descargado de https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

Ceballos, F. (2019, 7). *An Intuitive Explanation of Random Forest and Extra Trees Classifiers*. Descargado de https://towardsdatascience.com/ an-intuitive-explanation-of-random-forest-and-extra-trees-classifiers-8507ac21d54b

Cervantes, R. C., y Palacio, U. M. (2020, 1). Estimation of obesity levels based on computational intelligence. *Informatics in Medicine Unlocked*, *21*, 100472. doi: 10.1016/J.IMU.2020.100472

Cheng, H., Montgomery, S., Green, A., y Furnham, A. (2020, 8). Biomedical, psycho- logical, environmental and behavioural factors associated with adult obesity in a na- tionally representative sample. *Journal of Public Health*, *42*(3), 570–578. Descar- gado de https://academic.oup.com/jpubhealth/article/42/3/570/5364179 doi: 10.1093/PUBMED/FDZ009

Chiong, R., Fan, Z., Hu, Z., y Chiong, F. (2021, 1). Using an improved relative error support vector machine for body fat prediction. *Computer Methods and Programs in Biomedicine*, *198*, 105749. doi: 10.1016/J.CMPB.2020.105749

ComputerWorld. (1997). *SAS Institute presenta una solución completa para da- ta mining*. Descargado de https://www.computerworld.es/archive/sas-institute-presenta-una-solucion-completa-para-data-mining

DeGregory, K. W., Kuiper, P., DeSilvio, T., Pleuss, J. D., Miller, R., Roginski, J. W., . . . Thomas, D. M. (2018, 5). A review of machine learning in obesity. *Obesity reviews : an official journal of the International Association for the Study of Obesity*, *19*(5), 668–685. Descargado de https://pubmed.ncbi.nlm.nih.gov/29426065/ doi: 10.1111/OBR.12667

De-La-Hoz-Correa, E., Mendoza Palechor, F., De-La-Hoz-Manotas, A., Morales Ortega, R., y Sánchez Hernández, A. B. (2019). Obesity level estimation software based on decision trees. *Universidad de la Costa*. Descargado de https://repositorio.cuc.edu.co/handle/11323/4176

Fayyad, U., Piatetsky-Shapiro, G., y Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3).

Felső, R., Lohner, S., Hollódy, K., Erhardt, y Molnár, D. (2017, 9). Relationship between sleep

duration and childhood obesity: Systematic review including the potential underl- ying mechanisms. *Nutrition, Metabolism and Cardiovascular Diseases*, *27*(9), 751–761. doi: 10.1016/J.NUMECD.2017.07.008

Ferdowsy, F., Rahi, K. S. A., Jabiullah, M. I., y Habib, M. T. (2021, 11). A machine learning approach for obesity risk prediction. *Current Research in Behavioral Sciences*, *2*, 100053. doi: 10.1016/J.CRBEHA.2021.100053

Hâncu, A. M. (2021, 10). Lifestyle Factors and Obesity. *Role of Obesity in Human Health and Disease*. Descargado de undefined/state.item.id doi: 10.5772/INTECHOPEN .100254

Hernández Sampieri, R., Fernández Collado, C., y Baptista Lucio, M. d. P. (2014). *Metodología de la investigación* (Vol. 6). McGraw-Hill Interamericana.

IBM. (2011). IBM SPSS Modeler CRISP-DM Guide. *Career: Data and Analytics.*

Jeffares, A. (2018, 7). *Decision Trees: A Complete Introduction.* Descargado de https:// towardsdatascience.com/decision-trees-60707f06e836

Jiménez, A. R., y Jacinto, A. O. P. (2017, 7). Métodos científicos de indagación y de construcción del conocimiento. *Revista Escuela de Administración de Negocios*(82), 175–195. Descargado de https://journal.universidadean.edu.co/index.php/ Revista/article/view/1647 doi: 10.21158/01208160.N82.2017.1647

Keramat, S. A., Alam, K., Gow, J., y Biddle, S. J. (2021, 1). Impact of Disadvantaged Neighborhoods and Lifestyle Factors on Adult Obesity: Evidence From a 5-Year Cohort Study in Australia. *American Journal of Health Promotion*, *35*(1), 28–37. Descargado de https://journals.sagepub.com/doi/10.1177/0890117120928790 doi: 10.1177/0890117120928790

Kothari, C. (2004). *Research methodology: methods and techniques*. doi: http://196.29.172.66: 8080/jspui/bitstream/123456789/2574/1/Research{\%}20Methodology.pdf

Lazarou, C., Karaolis, M., Matalas, A. L., y Panagiotakos, D. B. (2012, 11). Dietary pat- terns analysis using data mining method. An application to data from the CYKIDS study. *Computer Methods and Programs in Biomedicine*, *108*(2), 706–714. doi: 10.1016/ J.CMPB.2011.12.011

Lee, I., Bang, K. S., Moon, H., y Kim, J. (2019, 5). Risk Factors for Obesity Among Children Aged 24 to 80 months in Korea: A Decision Tree Analysis. *Journal of Pediatric Nursing*, *46*, e15-e23. doi: 10.1016/J.PEDN.2019.02.004

Luna Espinoza, I., Hernández Suaárez, C. M., y Tinoco Zermeño, M. A. (2009). Muestreo estadistico : tamano de muestra y estimacion de parametros.

Martín Bueno, B. (2017). *Predicción semanal de precios de la energía eléctrica utilizando bosques aleatorios - Archivo Digital UPM* (Tesis Doctoral, Universidad Politécnica de Madrid, Madrid). Descargado de https://oa.upm.es/47648/

Matignon, R., y SAS Institute. (2007). *Data mining using SAS Enterprise miner*. Wiley-Interscience.

Mechanick, J. I., Farkouh, M. E., Newman, J. D., y Garvey, W. T. (2020, 2). Cardiometabolic-Based Chronic Disease, Addressing Knowledge and Clinical Practice Gaps: JACC State- of-the-Art Review. *Journal of the American College of Cardiology*, *75*(5), 539–555. Descargado de https://pubmed.ncbi.nlm.nih.gov/32029137/ doi: 10.1016/

J.JACC.2019.11.046

Mitchell, T. M. (1997). Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, *45*.

Montanez, C. A. C., Fergus, P., Hussain, A., Al-Jumeily, D., Abdulaimma, B., Hind, J., y Radi, N. (2017, 6). Machine learning approaches for the prediction of obesity using publicly available genetic profiles. *Proceedings of the International Joint Conference on Neural Networks*, *2017-May*, 2743–2750. doi: 10.1109/IJCNN.2017.7966194

Organización Panamericana de la Salud. (2020). *Panorama de la seguridad alimentaria y nutricional en América Latina y el Caribe 2020*. FAO, OPS, WFP and UNICEF. Descargado de https://iris.paho.org/handle/10665.2/53143https://www.ifad.org/documents/38714170/42182280/panorama2020.pdf/2a9c3dac-6730-4ccf-07da-ec53dac91f9f doi: 10.4060/CB2242ES

Palechor, F. M., y Manotas, A. d. l. H. (2019, 8). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data in Brief*, *25*, 104344. doi: 10.1016/J.DIB.2019.104344

Pang, X., Forrest, C. B., Lê-Scherban, F., y Masino, A. J. (2021, 6). Prediction of early childhood obesity with machine learning and electronic health record data. *International Journal of Medical Informatics*, *150*, 104454. doi: 10.1016/J.IJMEDINF.2021.104454

Pant, A. (2019, 1). *Introduction to Logistic Regression*. Descargado de https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148

Ramírez, J. P., Aparcana, L. T., Zamora, R. A., y Leo, I. B. (2019, 3). El sobrepeso, la obesidad y la obesidad abdominal en la población adulta del Perú. *Anales de la Facultad de Medi- cina*, *80*(1), 21–27. Descargado de https://revistasinvestigacion.unmsm.edu.pe/index.php/anales/article/view/15871 doi: 10.15381/anales.v80i1.15863

Rayward-Smith, V. J., Cormen, T. H., Leiserson, C. E., y Rivest, R. L. (1991). Introduction to Algorithms. *The Journal of the Operational Research Society*, *42*(9). doi: 10.2307/2583667

Rippe, J. M. (2019). *Lifestyle medicine*. CRC Press. Descargado de http://portal.amelica.org/ameli/jatsRepo/234/2341110008/movil/index.html

Rossman, H., Shilo, S., Barbash-Hazan, S., Artzi, N. S., Hadar, E., Balicer, R. D., … Segal, E. (2021, 6). Prediction of Childhood Obesity from Nationwide Health Records. *The Journal of Pediatrics*, *233*, 132–140. doi: 10.1016/J.JPEDS.2021.02.010

Safaei, M., Sundararajan, E. A., Driss, M., Boulila, W., y Shapi'i, A. (2021, 9). A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity. *Computers in Biology and Medicine*, *136*, 104754. doi: 10.1016/J.COMPBIOMED.2021.104754

Salahuddin, T. (2018, 9). *Obesity is increasing among the younger generation in Ban- gladesh | undefined*. Descargado de https://www.thedailystar.net/health/obesity-increasing-in-bangladesh-younger-generation-1637107

Sánchez Carlessi, H., Reyes Romero, C., y Mejia Sáenz, K. (2018). Ma- nual

de términos en investigación cientifica, tegnológica y humanista. *Jour- nal of Chemical Information and Modeling*, *500*, 1689–1699. Descarga- do de https://isbn.cloud/9786124735141/manual-de-terminos-en-investigacion-cientifica-tecnologica-y-humanistica/

SAS Institute Inc. (2018). *SAS® Enterprise Miner™ 15.1: Reference Help* (1.ª ed., Vol. 1; SAS Institute Inc., Ed.). North Carolina: SAS Institute Inc. Descargado de https://documentation.sas.com/doc/en/emref/15.1/n061bzurmej4j3n1jnj8bbjjm1a2.htm

Serengil, S. I. (2018, 10). *A Gentle Introduction to LightGBM for Applied Machine Learning.* Descargado de https://sefiks.com/2018/10/13/a-gentle-introduction-to-lightgbm-for-applied-machine-learning/

Serengil, S. I. (2019, 11). *A Gentle Introduction to XGBoost for Applied Machi- ne Learning.* Descargado de https://sefiks.com/2019/11/03/a-gentle-introduction-to-xgboost-for-applied-machine-learning/

Serengil, S. I. (2020, 4). *Feature Importance in Decision Trees.* Des- cargado de https://sefiks.com/2020/04/06/feature-importance-in-decision-trees/

Shao, G. (2022, 12). Comparison of prediction of obesity status based on different machine learning approaches with different factor quantities. *https://doi.org/10.1117/12.2660726*, *12458*, 881–888. Descargado de https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12458/124583U/Comparison-of-prediction-of-obesity-status-based-on-different-machine/10.1117/12.2660726.fullhttps://www.spiedigitallibrary.org/conference-proceedings-of-spie/12458/124583U/Comparison-of-prediction-of-obesity-status-based-on-different-machine/10.1117/12.2660726.short doi: 10.1117/12.2660726

Shearer, C., Watson, H. J., Grecich, D. G., Moss, L., Adelman, S., Hammer, K., y Herdlein, S. a. (2000). The CRISP-DM model: The New Blueprint for Data Mining. *Journal of Data Warehousing*.

Suca, C., Córdova, A., Condori, A., Cayra, J., y Sulla, J. (2016, 3). (PDF) COMPARACIÓN DE ALGORITMOS DE CLASIFICACIÓN PARA LA PREDICCIÓN DE CASOS DE OBESIDAD INFANTIL. *ResearchGa- te*. Descargado de https://www.researchgate.net/publication/301567339_COMPARACION_DE_ALGORITMOS_DE_CLASIFICACION_PARA_LA_PREDICCION_DE_CASOS_DE_OBESIDAD_INFANTIL

Sulla Torres, J. A. (2018). Modelo hibrido de árbol de decisión difusa con optimización por enjambre de partículas para clasificación de Obesidad Escolar. *Universidad Nacional de San Agustín de Arequipa*. Descargado de http://repositorio.unsa.edu.pe/handle/UNSA/6154

Sun, Y., Wang, S., y Sun, X. (2020, 9). Estimating neighbourhood-level prevalence of adult obesity by socio-economic, behavioural and built environment factors in New York City. *Public Health*, *186*, 57–62. doi: 10.1016/J.PUHE.2020.05.003

The European Association for the Study of Obesity. (s.f.). *Statistics - EASO*. Descargado de

https://easo.org/media-portal/statistics/

Torres Nolasco, M. F. (2019). *Encuentro clínico de los médicos con pacientes con sobrepeso y obesidad en consulta externa de un hospital público de Lima* (Tesis Doctoral, Universidad Peruana Cayetano Heredia, Lima). Descargado de https://repositorio.upch.edu.pe/handle/20.500.12866/6383?show=full

Trujillo Aspilcueta, H. (2018). *Factores asociados a sobrepeso y obesidad en trabajadores de una Institución Pública de Salud. Lima, Perú* (Tesis Doctoral, Universidad Nacional Federico Villarreal, Lima). Descargado de https://alicia.concytec.gob.pe/vufind/Record/RUNF_33d9e98cdf583d3f14b15e1d5adf1536/ Details

Velando, C., Córdova, A., Condori Castro, A., Cayra, J., y Sulla-Torres, J. (2016). COMPARACIÓN DE ALGORITMOS DE CLASIFICACIÓN PARA LA PREDICCIÓN DE CASOS DE OBESIDAD INFANTIL. *ResearchGate*, 1–9. Descargado de https://www.researchgate.net/publication/301567339_COMPARACION_DE_ALGORITMOS_DE_CLASIFICACION_PARA_LA_PREDICCION_DE_CASOS_DE_OBESIDAD_INFANTIL

World health Organization. (s.f.). *WHO EMRO | Body mass index calculator | Information resources | Nutrition.* Descargado de http://www.emro.who.int/nutrition/information-resources/bmi-calculator.html

-Editorial-
CILADI
Centro de Investigación Latinoamericano
para el Desarrollo e Innovación

ISBN: 978-9942-7292-5-5