William Calizaya

LEVANTANDO OBSERVACIONES PASO2.pdf

My Files

My Files

Universidad Peruana Union

Detalles del documento

Identificador de la entrega
trn:oid:::29566:454972203

Fecha de entrega
2 may 2025, 9:44 a.m. GMT-5

Fecha de descarga
2 may 2025, 9:46 a.m. GMT-5

Nombre de archivo
LEVANTANDO OBSERVACIONES PASO2.pdf

Tamaño de archivo
987.7 KB

18 Páginas

8733 Palabras

45.577 Caracteres

13% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 8 words)

Top Sources

8% Internet sources
3% Publications
10% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Top Sources

8% Internet sources
3% Publications
10% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	
es.slideshare.net		<1%
2	Submitted works	
Universidad Nacional Abierta y a Distancia, UNAD,UNAD on 2024-05-08		<1%
3	Submitted works	
Universidad Internacional de la Rioja on 2023-06-16		<1%
4	Internet	
podcast.unesp.br		<1%
5	Internet	
www.coursehero.com		<1%
6	Submitted works	
Universidad TecMilenio on 2024-05-16		<1%
7	Submitted works	
Universidad Internacional de la Rioja on 2023-05-10		<1%
8	Internet	
repositorio.udec.cl		<1%
9	Internet	
www.jove.com		<1%
10	Submitted works	
Universidad Pontificia de Salamanca on 2023-07-05		<1%
11	Submitted works	
unajma on 2023-10-16		<1%

12 Internet

www.slideshare.net <1%

13 Submitted works

Centro Europeo de Postgrado - CEUPE on 2023-08-14 <1%

14 Submitted works

Colegio Brains on 2024-02-25 <1%

15 Publication

William David Moreno Rendon, Carolina Burgos Anillo, Daniel Jaramillo-Ramirez, ... <1%

16 Internet

repositorio.utea.edu.pe <1%

17 Submitted works

Universidad Católica Boliviana "San Pablo" on 2024-11-04 <1%

18 Submitted works

Universidad Internacional de la Rioja on 2023-06-02 <1%

19 Submitted works

Universidad Internacional de la Rioja on 2024-01-18 <1%

20 Internet

www.researchgate.net <1%

21 Submitted works

Universidad Internacional de la Rioja on 2024-12-22 <1%

22 Submitted works

Universidad Nacional Abierta y a Distancia, UNAD,UNAD on 2024-06-08 <1%

23 Submitted works

Universidad Nacional Abierta y a Distancia, UNAD,UNAD on 2025-03-16 <1%

24 Internet

blog.enterprisedna.co <1%

25 Submitted works

Infile on 2023-05-23 <1%

26 Submitted works
Universidad Andrés Bello on 2024-01-04 <1%

27 Submitted works
Universidad Internacional de la Rioja on 2023-05-07 <1%

28 Submitted works
imfice on 2024-08-21 <1%

29 Submitted works
Universidad Carlos III de Madrid - EUR on 2024-08-29 <1%

30 Submitted works
Universidad Internacional de la Rioja on 2023-05-26 <1%

31 Submitted works
Universidad Internacional de la Rioja on 2024-06-08 <1%

32 Submitted works
Universidad Nacional Abierta y a Distancia, UNAD,UNAD on 2024-06-10 <1%

33 Submitted works
Universidad de Burgos UBUCEV on 2023-07-24 <1%

34 Submitted works
ITESM: Instituto Tecnologico y de Estudios Superiores de Monterrey on 2024-10-20 <1%

35 Publication
Patricia Batista Grau. "Desarrollo de nanoestructuras de ZnO mediante anodizad... <1%

36 Submitted works
Pontificia Universidad Javeriana Cali on 2024-08-14 <1%

37 Submitted works
Universidad de Salamanca on 2021-10-12 <1%

38 Submitted works
Universitat Politècnica de València on 2021-09-15 <1%

39 Internet
experienceleague.adobe.com <1%

40 Internet
global.tdx.cat <1%

41 Internet
repositorio.uss.edu.pe <1%

42 Submitted works
Corporación Universitaria Minuto de Dios,UNIMINUTO on 2024-08-06 <1%

43 Publication
Ekmel Geçer, Hakkı Bağci. " Examining students' attitudes towards online educati... <1%

44 Submitted works
Infile on 2025-01-08 <1%

45 Submitted works
Universidad Francisco de Vitoria on 2024-12-15 <1%

46 Submitted works
Universidad Internacional de la Rioja on 2023-06-02 <1%

47 Submitted works
Universidad Internacional de la Rioja on 2023-06-03 <1%

48 Submitted works
Universidad Internacional de la Rioja on 2025-01-19 <1%

49 Internet
documentop.com <1%

50 Internet
dspace.unila.edu.br <1%

51 Internet
eprints.ucm.es <1%

52 Internet
hvlib.integris-health.com <1%

53 Internet
patents.google.com <1%

54 Internet
repositorioacademico.upc.edu.pe <1%

55 Internet
www.cienciadedatos.net <1%

56 Internet
www.editorial.udg.mx <1%

57 Internet
www.oirmejor.org.ar <1%

58 Publication
María Salcedo, Gaspar González-Morán, Javier Albiñana. "Exploración ortopédica"... <1%

59 Publication
Mónica Chillarón Pérez. "Análisis y desarrollo de algoritmos de altas prestaciones... <1%

60 Submitted works
Universidad Anahuac México Sur on 2025-03-02 <1%

61 Submitted works
Universidad Internacional de la Rioja on 2023-05-10 <1%

62 Submitted works
Universidad Internacional de la Rioja on 2024-06-09 <1%

63 Submitted works
Universidad Internacional de la Rioja on 2025-01-18 <1%

64 Submitted works
Universidad de Cádiz on 2025-04-21 <1%

65 Internet
alicia.concytec.gob.pe <1%

66 Internet
eciperu.net <1%

67 Internet
repository.ucatolicaluisamigo.edu.co <1%

68 Internet

tesis.usat.edu.pe <1%

69 Submitted works

uncedu on 2024-05-09 <1%

70 Internet

www.medmultilingua.com <1%

71 Internet

www.refinitiv.com <1%

Modelo de Machine Learning Basado en los Factores de Riesgo para Predecir el Sobrepeso y la Obesidad

Willian Quispe, Victor Lee Chuquin, Alex A. Azurin

Resumen

Este estudio presenta un modelo de Machine Learning basado en factores de riesgo para predecir el sobrepeso y la obesidad, destacando su importancia en la prevención y control de esta condición. La metodología incluye limpieza de datos, imputación de valores faltantes, selección de características y evaluación de modelos de regresión como RFR, DTR, SVR y XGB. Los resultados muestran que el GBR alcanzó el mejor desempeño, seguido de cerca por XGB, al predecir el índice de masa corporal (IMC) utilizando 20 características clave, como colesterol LDL, perímetro abdominal, medidas diastólicas, presión arterial, hemoglobina y glucosa. La reducción de características mejoró notablemente el rendimiento de los modelos, especialmente los basados en técnicas de boosting. Estos hallazgos destacan la precisión de los modelos de Machine Learning en la predicción del IMC y su potencial para la identificación temprana de riesgos relacionados con la obesidad en la población adulta peruana.

Palabras clave

Machine Learning, Obesidad, Sobrepeso, Modelos de Regresión.

Introducción

La obesidad ha emergido como un desafío de salud pública de proporciones globales, con su prevalencia experimentando un preocupante aumento a lo largo de las últimas décadas.[1] Según la Organización Mundial de la Salud (OMS), desde 1975, la incidencia de la obesidad se ha casi triplicado a nivel mundial, con más de 1900 millones de adultos afectados por sobrepeso en 2016, de los cuales más de 650 millones fueron clasificados como obesos.[2] Este aumento alarmante, no solo ha generado una crisis de salud pública a nivel mundial, sino que también ha exacerbado la incidencia de enfermedades crónicas con alta mortalidad, como la diabetes tipo 2 y enfermedades cardiovasculares.[3][4] Además, la obesidad ha sido identificada como un factor de riesgo significativo en la severidad y las complicaciones asociadas con enfermedades infecciosas, como el COVID-19, con tasas de mortalidad más altas observadas en países con una mayor prevalencia de obesidad.[5][6][7] La carga de la obesidad también se extiende más allá de las implicaciones físicas, ya que diversos estudios han establecido vínculos entre la obesidad y problemas de salud mental, incluida la disminución de la autoestima y el aumento de riesgo de ideación suicida.[8][9] La obesidad en el Perú ha experimentado un aumento alarmante, con un incremento del 19% desde 1975 hasta el año 2016 en adultos mayores de 18 años, según datos de la OMS.[10] En la actualidad se sigue incrementando anualmente en 2.5%.[1] Tras esto el Ministerio de Salud (MINSA) ha implementado programas para abordar esta problemática, con un enfoque particular en frenar el aumento de casos de obesidad, especialmente entre la población femenina.[11] Sin embargo, a pesar de estos esfuerzos, la prevalencia de la obesidad sigue siendo alta, especialmente en regiones como Tacna, Ica y Madre de Dios.[12] Ante este panorama preocupante, se ha utilizado estrategias efectivas de detección temprana y prevención de la obesidad. En este contexto, el uso de técnicas de Machine Learning (ML) ha surgido como una herramienta prometedora que utiliza algoritmos

de inteligencia artificial para identificar diversos factores asociadas con diversas variables relacionadas con la salud y una de ella es el sobrepeso y obesidad.[13] Sin embargo, aunque estos modelos muestran una capacidad promisoria para precisar, clasificar, segmentar, jerarquizar u otros, a menudo carecen de interpretabilidad, siendo percibidos como "cajas negras".[14] A pesar de los avances en este campo, persisten desafíos significativos, como el alto costo computacional y la necesidad de ajustar los hiperparámetros de los modelos a fin de obtener mejores resultados.[15] Pero es preciso mencionar también la importancia de las características. No obstante, el continuo desarrollo de nuevas metodologías y enfoques en el ámbito del ML ofrece un camino prometedor y amplio para abordar este complejo problema de salud pública. La aplicación de técnicas de inteligencia artificial, ha demostrado ser efectiva en la detección de patrones, características, mejora de la precisión y segmentación del sobrepeso y la obesidad.[15] Estos avances sugieren un progreso hacia enfoques más avanzados y complejos en la aplicación en este campo.

Esta investigación revisa el uso de modelos de máquina y aprendizaje profundo para explicar la incidencia del sobrepeso y la obesidad en la población peruana, destacando los desafíos actuales y las perspectivas futuras. Se destaca que la mejora de esta situación debe ir más allá del Índice de Masa Corporal (IMC), considerando otros factores asociados relevantes. [4]

Se reconoce que la precisión y la importancia de las características pueden variar según el contexto y la configuración específica de cada modelo. En este sentido, el objetivo de esta investigación es diseñar y validar un modelo de ML para identificar factores de riesgo y predecir el sobrepeso y la obesidad. Para ello se estudiarán modelos de ML configurables, teniendo en cuenta la arquitectura y los hiperparámetros. [16]

La investigación sobre la determinación de la obesidad y el sobrepeso ha sido amplia y diversa, donde se ha empleado una variedad de algoritmos de ML y técnicas de inteligencia artificial, utilizado modelos ML supervisados, de clasificación y regresión, junto con diferentes configuraciones de hiperparámetros para mejorar la estimación de esta comorbilidad. Los factores que contribuyen a estos son diversos e incluyen la actividad física.[4][14][17][18][19][20] La diabetes.[4][14][21] El hábito de fumar.[14][19] El consumo de alcohol. [16][21][19] El género.[4][17][16] La edad.[17][16] El índice de masa corporal. [4][13][16][22]La alimentación.[3][4][16][18][19][21] Antecedentes familiares de sobrepeso.[13][16][21] Perímetro abdominal.[23] Factores metabólicos.[4][18] Estado socioeconómico.[4][13] Presión arterial.[4][13][19][21] Dislipidemia.[4] glucosa.[13] Perfil lipídico.[13]

Varios estudios utilizan técnicas de aprendizaje automático (ML) para predecir los factores de riesgo de obesidad. Un estudio realizado en Indonesia identificó 21 variables clave, con una precisión del 85% mediante regresión logística.[19] Otro estudio destacó el uso de CatBoost para predecir la obesidad en adultos con sobrepeso, identificando la circunferencia de la cintura, la circunferencia de la cadera, el sexo femenino y la presión arterial sistólica como predictores clave, con alta precisión (AUC entre 0,84 y 0,91).[23] Un tercer estudio desarrolló un modelo basado en redes neuronales, logrando una precisión del 93,06%, e identificó factores como la actividad física, el consumo de alcohol y los alimentos ricos en calorías como influyentes. 06% de precisión, y señaló como influyentes factores como la actividad física, el consumo de alcohol y los alimentos ricos en calorías.[16] Estos estudios resaltan la importancia de abordar múltiples factores de riesgo y el potencial de los modelos de aprendizaje automático para mejorar la precisión y la comprensión de la obesidad, al tiempo que reconocen las limitaciones en su aplicabilidad general y la necesidad

de realizar más investigaciones.

Materiales y Métodos

La interacción de determinados factores en lugar de una única causa, en el caso del sobrepeso y la obesidad pueden ayudar a tener mejores resultados, estos factores pueden incluir elementos genéticos, biológicos, ambientales, sociales y psicológicos.[24] nos pueden ayudar a reconocer problemas complejos. Al considerar esta complejidad, se puede diseñar intervenciones más efectivas que aborden diversos aspectos contribuyentes, como la alimentación, la actividad física, el entorno obeso genético y los factores psicosociales. Este enfoque permite una comprensión más completa del problema y ayuda a ver estrategias para lograr un mejor desarrollo de lo que se quiere lograr en esta investigación para ello se sigue el siguiente proceso basado en la metodología CRISP-DM como se muestra en la Figura 1.

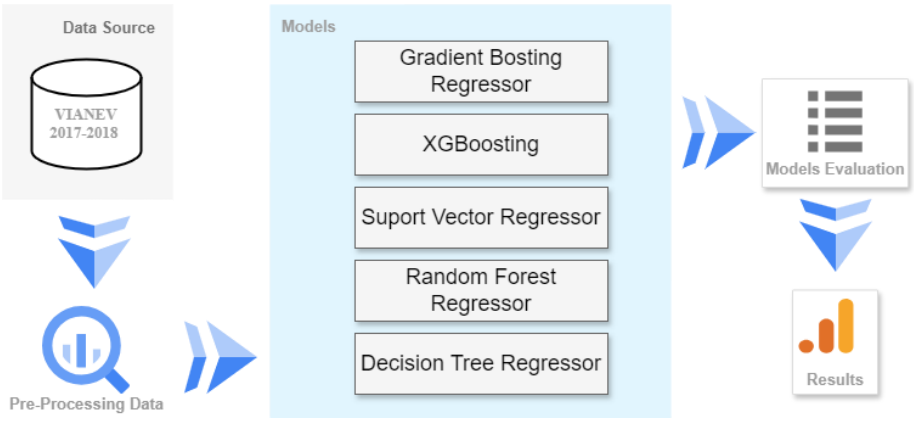

Figura 1. Describe el proceso del trabajo, el procesamiento, la evaluación de los modelos y los resultados.

A. Fuente de Datos

El tipo de muestreo que se hizo para obtener la muestra de investigación fue no probabilístico, por conveniencia los datos para este estudio se utilizaron de la plataforma Web de Datos Abiertos, del gobierno Peruano. Enlace del dataset: https://www.datosabiertos.gob.pe/dataset/estado-nutricional-en-adultos-de-18-59-años-perú-2017-2018. El portal forma parte de las acciones estratégicas comprendidas en la Estrategias Nacional de Datos Gubernamentales sustentada en los principios establecidos en la Carta Internacional de Datos Abiertos adoptada por más de 72 países.[25] La Encuesta de Vigilancia Alimentaria y Nutricional por etapas de Vida-VIANEV 2017-2018.[26] Fue desarrollada por el Centro Nacional de Alimentación y Nutrición (CENAN), El tamaño muestral del estudio fue calculado en 1211 adultos (hombres y mujeres); a nivel nacional, divididas en 03 grandes estratos: Lima Metropolitana (557 adultos), Resto Urbano (256 adultos) y Rural (398 adultos), en coordinaciones también con el INEI tras la entrega de Cartografías.[27]

B. Pre procesamiento de datos

La limpieza de datos abordará posibles valores atípicos, datos faltantes y errores. Se aplicarán técnicas de imputación o eliminación ética de datos ausentes, la transformación adecuada se llevará a cabo para mejorar la calidad de los datos, donde se buscará garantizar la consistencia y

confiabilidad de los resultados, también fomentar la transparencia y la reproducibilidad en el proceso de tratamiento de datos ya que este enfoque meticuloso contribuye significativamente a la robustez y validez de cualquier análisis posterior que se realice con el conjunto de datos.[21] Además, se observa el conjunto de datos con el que se entrenó el algoritmo, si abarca todas las instancias y atributos necesarios para alcanzar el objetivo de entrenamiento, los datos encontrados se dividen en 5 grupos: Salud y Nutrición, Bioquímica, Actividad Física, Enfermedades Crónicas, Hábitos y otros, le los datos están divididos en 1079 filas y 234 entradas de estos 154 son tipo Float64, 68 de tipo Int64 y 12 de tipo Object.

Seguidamente se eliminó las columnas con un 80% de valores igual a 0, obteniendo 105 características las cuales se dividen en Float64(35), Int64(63), Object(5) además se hace el análisis de los valores categóricos, columnas de tipo object, para realizar la transformación de las variables, pero en el proceso no se encuentra una relevancia importante ya que describe las ropas que lleva la persona como: polo algodón, buzo pantalón, Trusa Algodón, Brasier Corto, de igual forma se eliminó las columnas que se muestran como identificadores únicos ID, Conglomerado, Vivienda, hogar, id adulto, hora inicio, min inicio, hora final, min final, 409 codigo1, 409 codigo2, 409 codigo3. Después se insertó nuevas características, el IMC como variable dependiente a partir del peso y talla en cual se clasifica según al valor en sobrepeso y tipo de obesidad, y Peso Final a partir de los valores peso entero y peso decimal descritos como parte de las características en el dataset.

Finalmente se muestra las 20 características más importantes en la Tabla 1, las cuales fueron seleccionadas por el algoritmo de aprendizaje supervisado SVR, configurando la eliminación recursiva de características (RFE), estas características se usaron para finalizar las pruebas y mostrar los resultados correspondientes para cada modelo.

Columna	**Abreviatura**
COLESTEROL (mg/dL)	COL
LDL COLESTEROL (mg/dL)	LDL
407_PESO_BRUTO_ENTERO	PBE
407_PESO_BRUTO_DECIMAL	PBD
409_PESO_PRENDA1	PP1
409_PESO_PRENDA2	PP2
409_PESO_PRENDA3	PP3
410_TALLA_ENTERO	Talla
413_PERIMETRO_ENTERO	PE
414_PRESION_ARTERIAL_AÑO	PA_A
415_PRIMERA_MEDIDA_DIASTOLICA	1MD
416_SEGUNDA_MEDIDA_DIASTOLICA	2MD
418_RESULTADO_FINAL_DIASTOLICA	RFD
419_HEMOGLOBINA_AÑO	Hb_A
422_SANGRE_VENOSA_AÑO	SV_A
424_RESULTADO_PERIMETRO_ABDOMINAL	CC
424_RESULTADO_HB	Hb
424_RESULTADO_PRESION_ARTERIAL	PA
424_RESULTADO_GLUCOSA	Glc
PESO_FINAL	Peso

Tabla 1. caracterizas más relevantes seleccionadas.

También se muestra la correlación entre las variables en las cuales se puede identificar ciertos

factores de riesgo que pueden afectar la variable de interés.[28] Con el fin de proporcionar información sobre cómo las variables están interrelacionadas, y ayudar en la interpretación de resultados, como se muestra en la Figura 2.

En el análisis de la matriz de correlación, se identificaron relaciones importantes entre las variables. Destacan correlaciones positivas fuertes como la de COL y LDL (0.95), relacionada con los niveles de lípidos, y la de Peso e IMC (0.98), dada la dependencia matemática entre estas variables. Asimismo, PA_A y 2MD (0.97) mostraron una relación lineal significativa, relevante para un análisis clínico más profundo. Por otro lado, las correlaciones negativas destacaron entre Hb_A y 2MD (-0.99), así como Glc y 2MD (-0.95), lo que sugiere relaciones inversas significativas con implicaciones metabólicas. Finalmente, algunas variables como PBE y PP1 (0.03) y PA y CC (-0.07) mostraron correlaciones bajas o nulas, indicando una relación lineal débil o inexistente entre ellas. Aunque algunas características no representan una correlación tan fuerte, son relevantes para la precisión de los modelos, debido a que también son factores de riesgo que tienen importancia para este estudio. Este aspecto fue abordado durante la investigación, evaluándolos en modelos de regresión de aprendizaje automático.

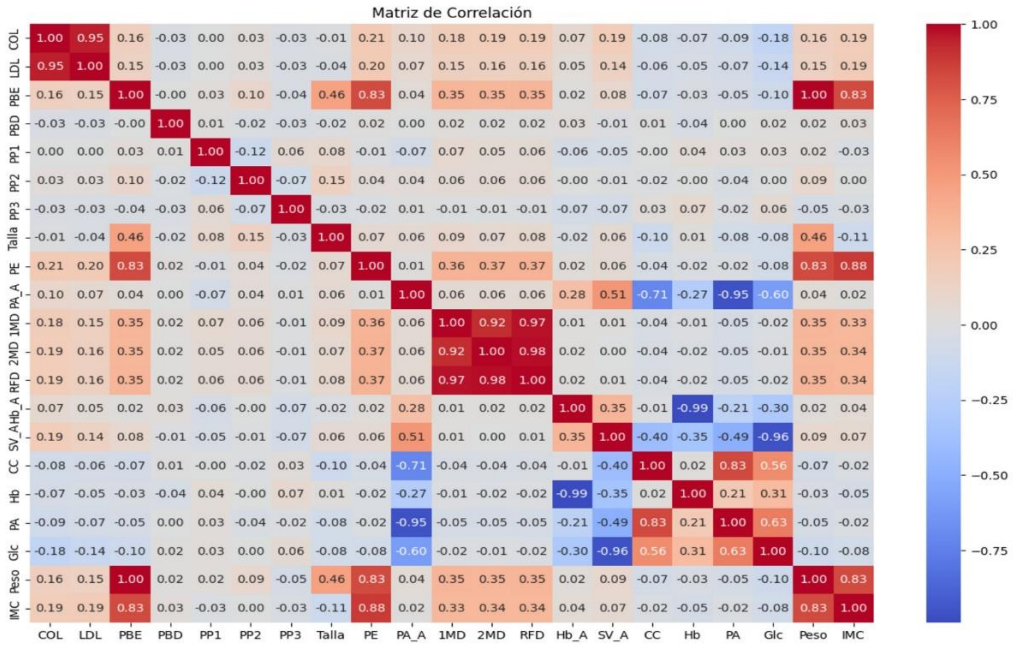

Figura 2. Correlación de las 20 mejores características.

C. Modelos

En este trabajo, se exploró los modelos de Random Forest Regresor, Decision Tree Regresor, Gradient Boosting Regresor, Suport Vector Regresor y XGB, para predecir la variable objetivo. A continuación, se presenta una breve descripción de cada uno de los modelos mencionados.

Random Forest Regresor (RFR).[29] Se basa en conjuntos de árboles de decisión, donde cada árbol hace una predicción independiente lo cual las predicciones de los árboles de decisión, se realizan con una serie de decisiones basadas en características de entrada, en cada nodo del árbol, se eligen

una característica y un umbral. Este proceso se emplea de forma recursiva en cada subconjunto hasta que se alcanza un nodo final, donde se realizan predicciones en función de la media (regresión). El pronóstico forestal aleatorio final se obtiene promediando todos los pronósticos de árboles individuales.
La fórmula matemática para la predicción de RFR como se muestra en la ecuación (1) puede ser expresada como:

$$\hat{y} = \frac{1}{N}\sum_{i=0}^{n} f_{i(x)} \quad (1)$$

En el modelo, la predicción final de una muestra se obtiene promediando las predicciones $f_{i(x)}$ de cada uno de los árboles individuales en el conjunto. Cada árbol i realiza una predicción basada en las características de los datos de entrada, y el valor final de la predicción es el promedio de todas las predicciones realizadas por los N árboles en el bosque. Este enfoque permite reducir la varianza y mejorar la precisión de la predicción al combinar los resultados de múltiples árboles, lo que mitiga el riesgo de sobreajuste que podría ocurrir si se utilizara un solo árbol de decisión.

Decision Tree Regressor (DTR).[30] El algoritmo basado en árboles que exploran características del conjunto de datos mediante evaluaciones secuenciales. Cada evaluación conduce a la formación de una rama en un árbol de decisión, donde al final de cada serie de ramas se encuentran nodos de hojas con etiquetas predichas. Este enfoque no lineal proporciona una estructura más flexible para capturar relaciones complejas en los datos. La construcción de árboles de decisión se presenta como una herramienta efectiva para abordar la no linealidad, ofreciendo versatilidad y mejorando la capacidad predictiva en problemas de ML.

La fórmula matemática del modelo DTR como se muestra en la ecuación (2) se puede expresar como:

$$\hat{y} = \frac{\sum_{i=0}^{n} y_i}{\mathrm{n}} \quad (2)$$

En la predicción de cada muestra se realiza calculando la media de los valores de los datos y_i dentro de la hoja terminal a la que pertenece la observación. Cada hoja terminal contiene un conjunto de n observaciones, y el valor de salida del modelo es el promedio de y_i para todas las muestras en esa hoja. Este enfoque permite que el árbol divida el espacio de características en regiones jerárquicas, donde cada región se caracteriza por una segmentación de los datos que minimiza el error dentro de ella. De esta manera, el modelo asigna una predicción basada en el comportamiento promedio de las observaciones que comparten características similares.

Gradient Boosting Regresor (GBR).[31] Es un enfoque para construir modelos de regresión de manera aditiva, donde se ajusta secuencialmente una función simple y parametrizada los base learners, el procedimiento de aprendizaje ajusta sucesivamente nuevos modelos para proporcionar una estimación más precisa de la variable de respuesta. La idea principal detrás de este algoritmo es construir los nuevos, base learners de manera que estén a menudo correlacionados con el gradiente negativo de la función de pérdida asociada al conjunto completo. Este enfoque de construcción iterativa y secuencial de modelos permite mejorar la precisión del modelo a medida

que se ajusta a la complejidad de los datos y minimiza la función de pérdida.

El objetivo del algoritmo de GBR como se muestra en la ecuación (3) es ajustar un modelo de regresión F(x) que minimice la función de pérdida, que se define como:

$$F_{m(x)} = F_{\{m-1\}}(x) + \eta \cdot \sum_{k=0}^{n} \frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \quad (3)$$

El modelo utiliza un enfoque iterativo en el que se construyen árboles de decisión de manera secuencial para minimizar los errores residuales de las predicciones anteriores. En cada iteración m, el modelo se ajusta utilizando una función de pérdida $L(y_i, F(x_i))$, como el MSE, que mide la discrepancia entre los valores observados (y_i) y las predicciones actuales del modelo ($F(x_i)$). La actualización de los parámetros del modelo está controlada por una tasa de aprendizaje (η), un hiperparámetro que regula la contribución de cada árbol nuevo en el conjunto final. Este proceso permite que los nuevos árboles enfoquen su ajuste en los errores residuales, logrando un modelo altamente preciso y robusto, especialmente en escenarios donde los datos presentan patrones complejos.

Suport Vector Regresor (SVR). [29] Es una técnica de regresión que utiliza hiperplanos para maximizar el margen entre clases. Funciona evaluando los datos de entrada en un espacio de características de alta dimensión, lo que lo hace efectivo para modelar relaciones complejas en problemas de regresión.

El modelo matemático de SVR como se muestra en la ecuación (4) se puede expresar de la siguiente manera:

$$\min \frac{1}{2} ||\omega||^2 + C \sum_{i=1}^{n} \backslash \max\left(0, \left|y_i - (w \cdot x_i + b)\right| - \epsilon\right) \quad (4)$$

El modelo se basa en encontrar un hiperplano óptimo que minimice los errores en las predicciones, mientras maximiza el margen de tolerancia definido por ϵ. Este margen permite ignorar errores pequeños, enfocándose en aquellos valores que se encuentren fuera de este rango. La predicción está determinada por un vector de pesos (w) que define la dirección del hiperplano y un término de sesgo (b) que ajusta su posición en el espacio de características. Además, el modelo incluye un parámetro de penalización (C) que controla el impacto de los errores fuera del margen, logrando un balance entre minimizar los errores y evitar un sobreajuste. Esta estructura matemática permite al SVR ser efectivo tanto para problemas lineales como no lineales, gracias al uso de funciones kernel.

XGBoostingrRegresor (XGB).[32] Implica la adición secuencial de modelos a un conjunto, donde cada modelo evalúa las predicciones de sus predecesores y contribuye a un promedio ponderado. En términos generales, se refiere a una estrategia en la que varios modelos clásicos se combinan para formar un modelo robusto. Cada nuevo modelo débil se esfuerza por corregir las deficiencias de sus predecesores con el objetivo de mejorar la precisión general del conjunto.

Matemáticamente, el modelo XGB como se muestra en la ecuación (5) se puede expresar de la siguiente manera:

Cada árbol f_k se construye de la siguiente manera:

$$L = \sum_{\{i=1\}}^{n} l\left((y_i, \hat{y}_i\right) + \sum_{\{k=1\}}^{K} \Omega(f_k) \quad (5)$$

Y para termino de regularización

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda||\omega||^2$$

El modelo utiliza una función de pérdida l(y_i , $\hat{y}_i$), como el error cuadrático medio (MSE), para evaluar la discrepancia entre los valores observados (y_i) y los predichos ($\hat{y}_i$). Adicionalmente, incorpora un término de regularización Ω(fk) que ayuda a controlar la complejidad del modelo mediante dos componentes principales: el número de hojas en un árbol (T) y un término de regularización L2 (λ). Este enfoque permite penalizar tanto la complejidad del modelo como el número de divisiones en los árboles, con el objetivo de evitar el sobreajuste. Por último, se incluye un factor γ para ajustar la penalización asociada a la adición de nuevas hojas en un árbol. Estas características distintivas del XGB garantizan un equilibrio entre el ajuste preciso a los datos y la capacidad de generalización del modelo, haciéndolo altamente eficiente en problemas de regresión y clasificación.

D. Evaluación

El desarrollo de los modelos se realizó en Google Colab para registrar y visualizar los resultados de los modelos de Machine Learning, se utilizó la plataforma Neptune. La capacidad de Neptune para organizar y comparar fácilmente los resultados de los experimentos facilitó la identificación de los modelos más prometedores. La transparencia proporcionada por la plataforma también permitió una validación rigurosa de los resultados, lo que contribuyó a la confianza en la conclusión de este estudio. Se registraron métricas de rendimiento.

Métricas

Para evaluar los modelos generalmente, un menor Error Cuadrático Medio (MSE) como se muestra en la ecuación (6). indica un mejor modelo. El MSE mide la diferencia promedio entre los valores observados y_i y los valores predichos $\hat{y}_i$ donde n es el número total de observaciones. Un valor de MSE más bajo significa que las predicciones del modelo son más precisas en promedio.[33]

$$MSE = \frac{1}{n}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (6)$$

Sin embargo, es importante considerar otros factores además del MSE al evaluar un modelo. Algunos de estos factores son:

R-cuadrado R^2: Como se muestra en la ecuación (7), mide la cantidad de variación en la variable objetivo que se explica por el modelo. Un valor de R^2 más alto indica que el modelo explica más de la variación en la variable objetivo.[34] Donde y_i son los valores observados, $\hat{y}_i$ son los valores predichos, $\bar{y}$ es la media de los valores observados. El término en el numerador mide la suma de

los errores cuadráticos (residuos) entre los valores observados y los predichos, mientras que el denominador representa la variabilidad total de los datos. Un valor de R^2 cercano a 1 indica que el modelo explica bien la variabilidad de los datos, mientras que un valor cercano a 0 sugiere que el modelo no tiene un buen poder explicativo. Es importante destacar que un valor de R^2 negativo puede ocurrir cuando el modelo se ajusta peor que una simple media de los valores observados

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (7)$$

Error absoluto medio (MAE): Como se muestra en la ecuación (8), mide la diferencia promedio entre las predicciones del modelo y los valores reales. Un valor de MAE más bajo indica que las predicciones del modelo son más precisas en promedio.[35] Donde n es el número total de observaciones, y_i son los valores reales y $\hat{y}_i$ son los valores predichos. El MAE proporciona una estimación del error promedio en las predicciones sin importar la dirección del error. Es una métrica robusta porque no penaliza los errores grandes de manera cuadrática, lo que permite una evaluación más uniforme del desempeño del modelo en diferentes magnitudes de error.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (8)$$

Raíz cuadrada del error cuadrático medio (RMSE): Como se muestra en la ecuación (9), es la raíz cuadrada del MSE. Un valor de RMSE más bajo indica que las predicciones del modelo son más precisas en promedio.[36] Donde n es el número total de observaciones, y_i son los valores reales y $\hat{y}_i$ son los valores predichos. El RMSE penaliza los errores grandes más que el MAE, debido a que los errores se elevan al cuadrado. Esto lo convierte en una métrica útil para aplicaciones donde los errores grandes son especialmente indeseables. Al calcular la raíz cuadrada, el RMSE devuelve el error en las mismas unidades que las variables originales, facilitando su interpretación

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (9)$$

Para la evaluación se tomó en cuenta la importancia de las características ya que influyen directamente en el rendimiento de los modelos de ML. Características informativas y relevantes pueden mejorar la capacidad predictiva del modelo, mientras que características irrelevantes o redundantes pueden introducir ruido y afectar negativamente la calidad de las predicciones.[37] Cabe mencionar que una gran cantidad de características puede llevar a un problema con la dimensionalidad, donde la complejidad del modelo aumenta con la cantidad de características, pero puede resultar en sobreajuste y pérdida de generalización.[38] La identificación y selección de características relevantes ayudan a reducir la dimensionalidad y mejorar la capacidad del modelo para generalizar a nuevos datos.[21] En este punto para obtener las mejores características para el modelo de estudio, se usó el algoritmo de aprendizaje supervisado SVR, configurando la Eliminación Recursiva de Características (RFE), técnica de selección de características utilizada en ML para identificar las características más importantes en un conjunto de datos.

La configuración de los hiperparámetros se ajustó en función del tamaño del conjunto de datos utilizado en el estudio, como se detalla en la Tabla 2.

Modelo	Hiperparámetros
DTR	max_depth: 3, min_samples_leaf: 10
SVR	kernel: rbf
RFR	n_estimators: 100, max_depth: 3
GBR	max_depth: 3, n_estimators: 100, learning_rate: 0.1, min_samples_leaf: 10
XGB	objective: reg:squarederror, max_depth: 3, learning_rate: 0.1, n_estimators: 100, min_child_weight: 10

Tabla 2. Hiperparámetros de los modelos usados.

Se consideró una partición equitativa de los datos para cada modelo, asignando el 80% al entrenamiento y el 20% restante a las pruebas. Para evitar el sobreajuste, se emplearon técnicas de validación basadas en el análisis MSE de la curva de aprendizaje y la validación a medida que aumentaba el tamaño del conjunto de entrenamiento durante el ajuste de los hiperparámetros, lo que garantizaba que los modelos generalizaban adecuadamente a los datos no vistos.

E. Resultados

Se evaluó el rendimiento de los modelos de ML en base a diferentes pruebas y conjuntos de características divididas en cantidades descendentes, para la primera prueba se consideró 88 características, para la segunda prueba se tuvo 60, en la tercera 40 características y en la cuarta prueba se utilizó un mínimo de 20 características.

La primera prueba se consiguió en base a un total de 88 características, exhiben el desempeño sobresaliente del modelo GBR en comparación con los otros modelos evaluados, como se muestra en la Tabla 3, el GBR demuestra el menor MSE de 0.659 y MAE de 0.599, lo cual señala su superior habilidad para realizar predicciones cercanas a los valores reales. Asimismo, presenta un mejor coeficiente de determinación R^2 de 0.975, indicando que es capaz de explicar una mayor proporción de la variabilidad presente en los datos. El RMSE de 0.812 corrobora que las predicciones del modelo se encuentran próximas a los valores reales. El algoritmo XGB sigue de cerca al GBR en términos de rendimiento, aunque con valores ligeramente inferiores en MSE de 0.685, MAE de 0.599 y R^2 de 0.974, y RMSE un poco elevado de 0.828. Por otro lado, el modelo SVR exhibe un desempeño notablemente inferior en comparación con GBR y XGB, evidenciado por el MSE de 5.335 y RMSE de 2.310 significativamente alto, lo que sugiere predicciones más alejadas de los valores reales, junto con un R^2 de 0.799 considerablemente bajo. El modelo RFR muestra un rendimiento menor en comparación con el modelo SVR, presentando MSE 6.251 y RMSE 2.500 y un R^2 de 0765, lo que implica una capacidad de predicción menos precisa y una menor capacidad para explicar la variabilidad en los datos. Por último, el modelo DTR muestra el peor desempeño entre todos los modelos evaluados, con los valores de MSE 7.813, MAE de2.085 y RMSE 2.795, y el coeficiente de determinación R^2 de 0.706, indicando una capacidad de predicción menor a todos los modelos evaluados y una explicación mínima de la variabilidad en los datos.

Modelo	MAE	MSE	R^2	RMSE
GBR	0.599	0.659	0.975	0.812
XGB	0.598	0.685	0.974	0.828
SVR	1.601	5.335	0.799	2.310
RFR	1.885	6.251	0.765	2.500

DTR	2.085	7.813	0.706	2.795

Tabla 3. Resultados de los modelos según métricas de las 88 características

En la segunda prueba, se observa que no hay una mejora en el desempeño de los modelos GBR y XGB al reducir el conjunto de características de 88 a 60 como se muestra en la Tabla 4, pero si hay una mejora significativa en el modelo SVR particularmente en el MAE de 1.404, MSE 4.607 y R^2 de 0.827, lo que sugiere una mayor capacidad predictiva., no obstante los modelos GRB y XGB siguen liderando, la cual se representa en los valores de MSE 0.828, MAE de 0.609 y RMSE 0.828 y un R^2 de 0.974 seguido por el modelo XGB con los valores de MSE 0.685, MAE de 0.609 y RMSE 0.836 y un R^2 de 0.974, en términos de rendimiento, manteniendo las primeras dos posiciones en ambos casos. Por otro lado, los modelos RFR y DTR mantienen posiciones similares en ambos escenarios, evidenciando un rendimiento bajo con valores RMSE DE 2.510 y 2.795 la cual se refleja en el R^2 de 0.763 y 0.706 respectivamente en comparación con los otros modelos.

Modelo	MAE	MSE	R^2	RMSE
GBR	0.609	0.685	0.974	0.828
XGB	0.606	0.700	0.974	0.836
SVR	1.404	4.607	0.827	2.146
RFR	1.888	6.298	0.763	2.510
DTR	2.085	7.813	0.706	2.795

Tabla 4. Resultados de los modelos según métricas de las 60 características

Para la tercera prueba cuyos resultados se visualizan en la Tabla 5, se evidencia una mejora en el desempeño de los modelos al reducir el conjunto de características de 60 a 40. Los modelos GBR y XGB continúan liderando, con el RMSE de 0.806, y 0.818 respectivamente, que se refleja en el R^2 de 0.976, y 0.975 en términos de rendimiento, manteniendo sus posiciones. El modelo SVR continúa mostrando una mejora notable en todos los aspectos con valores del MAE de 1.213, MSE de 4.122, R^2 de 0.845 y RSME de 2.030. Mientras tanto, los modelos RFR y DTR mantienen posiciones similares en los tres casos, con un desempeño bajo de RMSE de 2.506 y 2.795 que da por consecuencia un R^2 de 0.764 y 0.706 respectivamente en comparación con los otros modelos.

Modelo	MAE	MSE	R^2	RMSE
GBR	0.591	0.649	0.976	0.806
XGB	0.594	0.670	0.975	0.818
SVR	1.213	4.122	0.845	2.030
RFR	1.885	6.282	0.764	2.506
DTR	2.085	7.813	0.706	2.795

Tabla 5. Resultados de los modelos según métricas de las 40 características.

En la cuarta prueba, los modelos exhibieron una mejora sustancial en los resultados, como se detalla en la Tabla 6. La reducción del conjunto de características a solo 20, en contraste con los conjuntos de 88, 60 y 40 características anteriores, reveló un desempeño especialmente destacado. Los modelos GBR y XGB continúan dominando en términos de rendimiento, manteniendo la primera y segunda posición, aunque con diferencias más sutiles entre ellos en esta ocasión. El modelo GBR alcanza el RMSE más bajo de 0.799, lo que indica una mayor precisión en sus predicciones en comparación con el modelo XGB de RMSE 0.811. Ambos modelos exhiben el R^2

de 0.976 y 0975 respectivamente con una diferencia mínima, sugieren una capacidad superior para explicar la varianza en los datos en comparación con los modelos SVR, RFR Y DTR. Por otro lado, el modelo SVR se mantiene en el tercer lugar, su desempeño ha mejorado en términos de RMSE de 1.975 y R^2 de 0.853 a medida que se ha reducido el número de características, esto sugiere una capacidad mejorada para generalizar a partir de un conjunto de datos más pequeño y evitar el sobreajuste. En contraste, RFR y DTR muestran consistentemente un rendimiento inferior en todos los casos, ocupando posiciones bajas, independientemente del número de características utilizadas.

Modelo	MAE	MSE	R^2	RMSE
GBR	0.583	0.638	0.976	0.799
XGB	0.585	0.658	0.975	0.811
SVR	1.000	3.899	0.853	1.975
RFR	1.878	6.239	0.765	2.498
DTR	2.085	7.813	0.706	2.795

Tabla 6. Resultados de los modelos según métricas de las 20 características.

Para poder entender mejor el desempeño de los modelos a partir de la cuarta prueba se realizó diagramas de dispersión de los valores reales en el eje "x" y los valores predichos en el eje "y" con aproximación a la línea de referencia de cada modelo estos proporcionan una forma visual de evaluar la precisión del modelo en la predicción del IMC. Al observar la distribución de los puntos y su proximidad a la línea de referencia, se puede entender de qué tan bien está funcionando el modelo en general y en qué áreas específicas puede necesitar ajustes o mejoras.

El modelo DTR tiende a sobreestimar el IMC. como se muestra en la Figura A de la Tabla 7. Esto se observa porque la mayoría de los puntos se encuentran por encima de la línea de referencia. Un modelo de árbol de decisión simple puede no ser capaz de capturar la complejidad de la relación entre características y la obesidad.

El modelo RFR muestras una mejora respecto al modelo anterior como se muestras en la Figura B de la Tabla 7, esta vez los puntos están más cerca de la línea de referencia, aunque hay algunos puntos alejados especialmente en el rango de IMC alto. También se observa una concentración de puntos en la zona de sobrepeso, lo que indica que el modelo puede tener un ligero sesgo hacia esa categoría, en comparación con el modelo anterior, se ve una mejora en todos los niveles del IMC lo que se representa en los puntos dispersos a lo largo de la línea de referencia.

En este punto para el modelo SVR como se muestra en la Figura C de la Tabla 7, hay una aproximación perceptible de los puntos a la línea de referencia, lo que indica que este modelo está mostrando mejores resultados que los modelos anteriores, es así que tenemos una aproximación más precisa para los puntos que representa el sobrepeso a línea de referencia, pero sin encontrar aun mejoras relevantes respecto a los puntos que representan la obesidad 2 y 3 respectivamente.

Dado que los mejores los resultados en las métricas evaluadas anteriormente en los modelos fueron EL GBR y el XGB, hay una evidente mejora de precisión respecto a los otros modelos mencionados como se muestra en la Figura D y E de la Tabla 7 respectivamente para cada modelo, es preciso mencionar que las diferencias entre ella son mínimas, pero en este punto en evaluación de las mejores 20 características seleccionadas, los puntos más próximos a la línea de referencia para todos los niveles descritos para el IMC, es del modelo GBR dando mejores resultados descritos

aunque las diferencia es mínima.

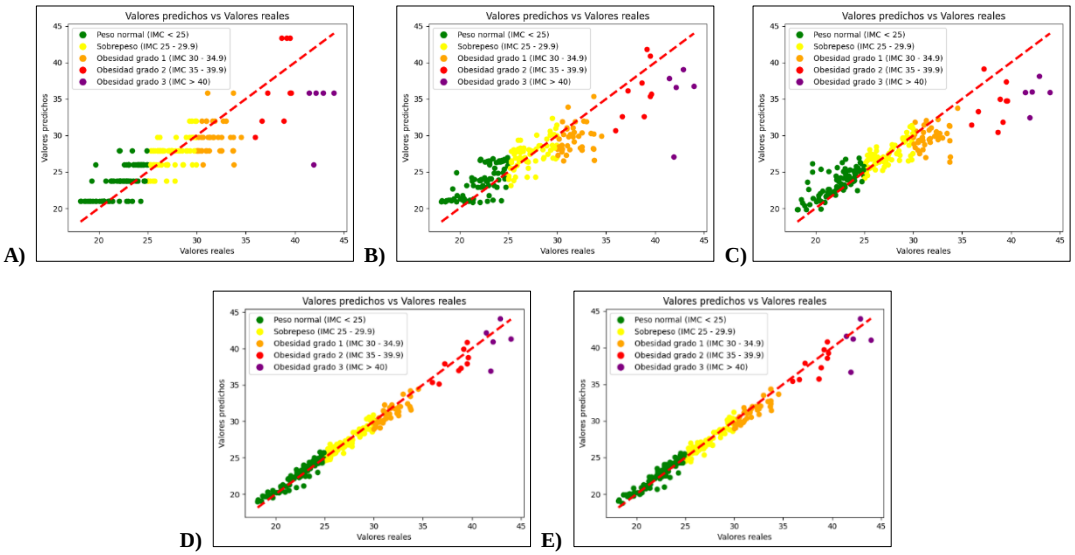

Tabla 7. Comparación de resultados de análisis estadísticos.

Esto también se puede distinguir claramente si se traza los puntos de los valores reales para el IMC, como se muestra en la Tabla 8, en el tamaño de muestra conectando la línea de las predicciones para resaltar la diferencia entre ambas. Si las líneas de predicción siguen de cerca los puntos de los valores reales, indica que el modelo está produciendo predicciones precisas en general.

Si observamos de cerca el modelo DTR, notamos que la diferencia entre las líneas rojas y los puntos azules es significativa para la mayoría de las muestras, como se muestra en la Figura A de la Tabla 8. Esto indica que, aunque el modelo puede ser preciso en algunas instancias, no lo es para la mayoría, de igual forma en el caso del modelo RFR se observa que no está realizando predicciones precisas para la mayoría de las muestras, como se muestra en la Figura B de la Tabla 8, entonces se podría decir que ambos modelos han realizado predicciones con mayor variabilidad, en comparación a la siguientes las cuales abordaremos a continuación.

El modelo SVR realiza predicciones precisas para la mayoría de las muestras, como se muestra en la Figura C de la Tabla 8. Sin embargo, hay algunas muestras en las que las líneas rojas se desvían ligeramente de los puntos azules. Esto significa que el modelo ha realizado predicciones menos precisas para estas muestras. Se observa también que algunas líneas rojas son casi horizontales, lo que indica que el modelo ha realizado predicciones muy precisas para esas muestras.

En el caso de los modelos Bosting que pertenecen a la familia ensemble, las predicciones son precisas en la mayoría de los puntos a lo largo del incremento de conjunto de muestra, en ambos modelos los resultados de desempeño es parecido como se muestra en la Figura D de la Tabla 8, para el modelo GBR y Figura E de la Tabla 8 para el modelo XGB, que no se diferencia en nada respecto al trazado de las líneas de predicción en referencia a los valores reales en el tamaño de muestra, aunque hay una mínima diferencia y si se quiere ser más exactos de que modelo es mejor nótese el resultado de las métricas la cuales se explicaron anteriormente, aunque es preciso mencionar que ambos modelos son eficientes, esto se ha podido ver a lo largo de la comparación

exhaustiva que se realizó en esta investigación.

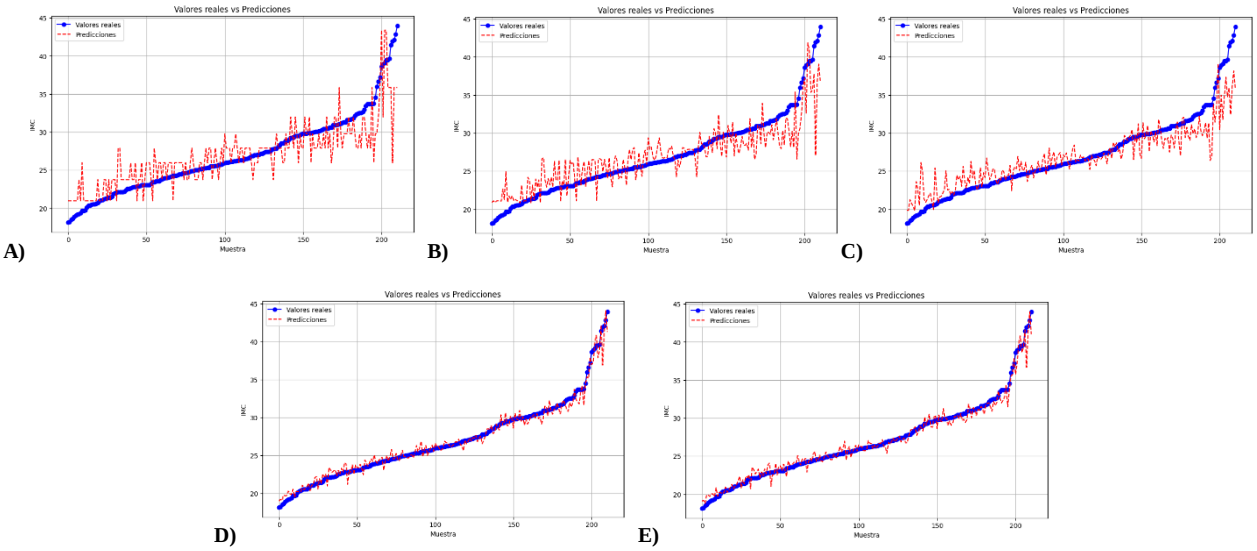

Tabla 8. Comparación de puntos de los valores reales en el tamaño de muestra conectadas a la línea de las predicciones.

F. Discusión

En las pruebas, los resultados mostraron que el modelo de GBR tuvo el mejor desempeño en todas las métricas evaluadas, seguido de cerca por XGB, esto nos indica que los modelos basados en boosting superaron al SVR y al RFR y DTR siendo este último debido a su simplicidad quien mostro los resultados más bajos frente a los otros modelos mencionados, se pudo observar a la vez mejoras en el desempeño de los modelos al reducir el número de características, lo que sugiere que una selección más cuidadosa de características puede resultar en modelos más efectivos para la predicción, es así que modelo de GBR mostro un desempeño sobresaliente en casi todas las evaluaciones , esto se visualiza claramente al realizar la predicciones con las 20 características más importantes , que podrían determinar casos de sobre peso y obesidad que se resalta en el promedio cuantificado de la magnitud del error en las predicciones RMSE de 0.809 promediada de los resultados, que se representan en la variabilidad de los datos alrededor de su media dada por el R^2 de 0.975 también promediada de los resultados, seguido por XGB. RMSE de 0.825 y R^2 de 0.974, ambas promediadas de los resultados.

En el estudio realizado, se observa que tras la evaluación de diferentes modelos de regresión utilizando distintas cantidades de características (90, 60 y 40), se identificó un conjunto de 20 características que demostraron ser las más relevantes para predecir el objetivo de interés, al comparar los resultados de los modelos entrenados con estas, con aquellos entrenados con un mayor número de características, se evidencia que hay una mejora en los resultados. Esto sugiere que estas 20 características son altamente informativos y proporcionan una representación suficientemente completa de los datos para realizar predicciones precisas, por lo tanto, estos hallazgos resaltan la importancia estos en la predicción del Sobre Peso y Obesidad. Su capacidad para mantener un rendimiento predictivo sólido que sugiere que son fundamentales en la determinación de los resultados de salud o biométricos estudiados.

Este análisis subraya la relevancia clínica y práctica de estas características seleccionadas, la cual se podría determinar como Factores, principalmente los Colesterol LDL, el Perímetro Abdominal, Las Medidas Diastólicas, La Presión Arterial, el valor de Hemoglobina, La Glucosa en la sangre, podría tener implicaciones importantes en la identificación y seguimiento de condiciones de salud o toma de decisiones clínicas basadas en modelos predictivos para determinar casos de sobre peso y obesidad, También es importante señalar las limitaciones de este estudio, siendo el tamaño de la muestra una de las más relevantes. Por ello, en futuras investigaciones, es fundamental trabajar con muestras más grandes para analizar con mayor precisión las diferencias en la robustez de los modelos, especialmente en los modelos Boosting, y evaluar su capacidad de generalización a nuevos datos. Otra de las limitaciones fueron el no poder comparar el desempeño de los modelos de ML con modelos de Redes Neuronales, de regresión esto también debido al tamaño de los datos en evaluación, ya que los modelos de ANN Y DNN son más complejos.[3] Para otros estudios se recomienda la evaluación con estos modelos, incluyendo un número mayor de datos para no entrar en problemas de sobre ajuste. Como recomendación para futuros estudios en temas de obesidad en la realidad peruana aplicando ML, se aconseja tomar en cuenta, la composición nutricional de los alimentos, factores clave, entre ellos los micronutrientes, las calorías y el contenido de grasa y otros, ya que el Perú es considerado y reconocido por su variedad culinaria a nivel mundial.[39] Por ende podría ser un aporte interesante que ayude a contrarrestar este problema.

G. Conclusión

En base a las evaluaciones realizadas en la investigación sobre el estado nutricional en adultos en Perú, se destaca la importancia de la selección cuidadosa de factores en la predicción de la obesidad. Se observó que, al reducir el número de factores, se mejoró el rendimiento de los modelos de regresión, especialmente aquellos basados en boosting como Gradient Boosting y XGB. La identificación de 20 factores relevantes para predecir el IMC resalta la necesidad de abordar múltiples factores de riesgo en la lucha contra la obesidad. A pesar de las limitaciones como el tamaño de la muestra y la imputación de datos, los resultados obtenidos muestran que los modelos de Gradient Boosting y XGB evaluados tienen un buen rendimiento en la predicción. Estos hallazgos subrayan la importancia de considerar enfoques avanzados de análisis de datos y abordar la obesidad como un desafío global de salud pública. Además, se destaca la relevancia de utilizar modelos de ML basados en factores de riesgo para predecir el sobrepeso y la obesidad, reconociendo la necesidad de prevenir y controlar esta condición de salud. De modelos evaluados, el Gradien Boosting, demostró tener una relevante precisión en la predicción del índice de masa corporal (IMC) por sobre los otros modelos con un coeficiente de determinación de 97%, lo que sugiere su utilidad en la identificación temprana de riesgos relacionados con la obesidad en la población adulta peruana.

Referencias

[1] T. Lobstein, R. Jackson-Leach, J. Powis, H. Brinsden, and M. Gray, “World Obesity Atlas 2023,”

Mar. 2023, Accessed: Dec. 02, 2023. [Online]. Available: https://data.worldobesity.org/publications/?cat=19

[2] "Obesity and overweight." Accessed: Dec. 13, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

[3] H. Kim, D. H. Lim, and Y. Kim, "Classification and Prediction on the Effects of Nutritional Intake on Overweight/Obesity, Dyslipidemia, Hypertension and Type 2 Diabetes Mellitus Using Deep Learning Model: 4–7th Korea National Health and Nutrition Examination Survey," *International Journal of Environmental Research and Public Health 2021, Vol. 18, Page 5597*, vol. 18, no. 11, p. 5597, May 2021, doi: 10.3390/IJERPH18115597.

[4] M. Safaei, E. A. Sundararajan, M. Driss, W. Boulila, and A. Shapi'i, "A systematic literature review on obesity: Understanding the causes & consequences of obesity and reviewing various machine learning approaches used to predict obesity," *Comput Biol Med*, vol. 136, p. 104754, Sep. 2021, doi: 10.1016/J.COMPBIOMED.2021.104754.

[5] D. Petrova, E. Salamanca-Fernández, M. Rodríguez Barranco, P. Navarro Pérez, J. J. Jiménez Moleón, and M.-J. Sánchez, "Obesity as a risk factor in COVID-19: Possible mechanisms and implications | La obesidad como factor de riesgo en personas con COVID-19: posibles mecanismos e implicaciones," *Aten Primaria*, vol. 52, no. 7, pp. 496–500, 2020, doi: 10.1016/j.aprim.2020.05.003.

[6] E. Pérez-Cruz, J. A. Castañón-González, S. Ortiz-Gutiérrez, J. Garduño-López, and Y. Luna-Camacho, "Impact of obesity and diabetes mellitus in critically ill patients with SARS-CoV-2," *Obes Res Clin Pract*, vol. 15, no. 4, pp. 402–405, 2021, doi: 10.1016/j.orcp.2021.05.001.

[7] L. Palaiodimos *et al.*, "Severe obesity is associated with higher in-hospital mortality in a cohort of patients with COVID-19 in the Bronx, New York," *Metabolism*, vol. 108, 2020, doi: 10.1016/j.metabol.2020.154262.

[8] J. Mattar, V. Chauvin, J. Marleau, K. Kavalidou, and A. J. Romain, "Association between physical multimorbidity and suicidal ideation in young adults with obesity," *J Affect Disord Rep*, vol. 14, 2023, doi: 10.1016/j.jadr.2023.100661.

[9] C. E. Graham and M. L. Frisco, "The Mental 'Weight' of Discrimination: The Relationship between Perceived Interpersonal Weight Discrimination and Suicidality in the United States," *J Health Soc Behav*, vol. 64, no. 4, pp. 610–625, 2023, doi: 10.1177/00221465231200634.

[10] "Prevalencia estandarizada por edad de obesidad entre adultos (18+ años)." Accessed: Dec. 16, 2023. [Online]. Available: https://data.who.int/es/indicators/i/BEFA58B

[11] F. Diez-Canseco and L. Saavedra-Garcia, "Programas sociales y reducción de la obesidad en el Perú: reflexiones desde la investigación," *Rev Peru Med Exp Salud Publica*, vol. 34, no. 1, pp. 105–112, Jan. 2017, doi: 10.17843/RPMESP.2017.341.2772.

[12] J. Pajuelo Ramírez, L. Torres Aparcana, R. Agüero Zamora, and I. Bernui Leo, "El sobrepeso, la obesidad y la obesidad abdominal en la población adulta del Perú," *Anales de la Facultad de Medicina*, vol. 80, no. 1, pp. 21–27, Mar. 2019, doi: 10.15381/anales.v80i1.15863.

[13] A. Chatterjee, M. W. Gerdes, and S. G. Martinez, "Identification of Risk Factors Associated with Obesity and Overweight—A Machine Learning Overview," *Sensors 2020, Vol. 20, Page 2734*, vol. 20, no. 9, p. 2734, May 2020, doi: 10.3390/S20092734.

[14] B. Allen, "An interpretable machine learning model of cross-sectional U.S. county-level obesity prevalence using explainable artificial intelligence," *PLoS One*, vol. 18, no. 10, p. e0292341, Oct. 2023, doi: 10.1371/JOURNAL.PONE.0292341.

[15] R. An, J. Shen, and Y. Xiao, "Applications of Artificial Intelligence to Obesity Research: Scoping Review of Methodologies," *J Med Internet Res 2022;24(12):e40589 https://www.jmir.org/2022/12/e40589*, vol. 24, no. 12, p. e40589, Dec. 2022, doi: 10.2196/40589.

[16] F. H. Yagin *et al.*, "Estimation of Obesity Levels with a Trained Neural Network Approach optimized by the Bayesian Technique," *Applied Sciences (Switzerland)*, vol. 13, no. 6, Mar. 2023, doi: 10.3390/APP13063875.
[17] J. E. Wong, M. Yamaguchi, N. Nishi, M. Araki, and L. H. Wee, "Predicting Overweight and Obesity Status Among Malaysian Working Adults With Machine Learning or Logistic Regression: Retrospective Comparison Study.," *JMIR Form Res*, vol. 6, no. 12, p. e40404, Dec. 2022, doi: 10.2196/40404.
[18] J. Jeon, S. Lee, and C. Oh, "Age-specific risk factors for the prediction of obesity using a machine learning approach," *Front Public Health*, vol. 10, p. 998782, Jan. 2023, doi: 10.3389/FPUBH.2022.998782/BIBTEX.
[19] S. A. Thamrin, D. S. Arsyad, H. Kuswanto, A. Lawi, and S. Nasir, "Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research 2018," *Front Nutr*, vol. 8, p. 669155, Jun. 2021, doi: 10.3389/FNUT.2021.669155/BIBTEX.
[20] P. Forte *et al.*, "A Deep Learning Neural Network to Classify Obesity Risk in Portuguese Adolescents Based on Physical Fitness Levels and Body Mass Index Percentiles: Insights for National Health Policies," *Behavioral Sciences*, vol. 13, no. 7, 2023, doi: 10.3390/bs13070522.
[21] H. Siddiqui *et al.*, "A Survey on Machine and Deep Learning Models for Childhood and Adolescent Obesity," *IEEE Access*, vol. 9, pp. 157337–157360, 2021, doi: 10.1109/ACCESS.2021.3131128.
[22] Y. Cao, M. Raoof, E. Szabo, J. Ottosson, and I. Näslund, "Using Bayesian Networks to Predict Long-Term Health-Related Quality of Life and Comorbidity after Bariatric Surgery: A Study Based on the Scandinavian Obesity Surgery Registry," *Journal of Clinical Medicine 2020, Vol. 9, Page 1895*, vol. 9, no. 6, p. 1895, Jun. 2020, doi: 10.3390/JCM9061895.
[23] W. Lin, S. Shi, H. Huang, J. Wen, and G. Chen, "Predicting risk of obesity in overweight adults using interpretable machine learning algorithms," *Front Endocrinol (Lausanne)*, vol. 14, p. 1292167, Jan. 2023, doi: 10.3389/fendo.2023.1292167.
[24] Y. N. Flores *et al.*, "Increased prevalence of psychosocial, behavioral, and socio-environmental risk factors among overweight and obese youths in Mexico and the United States," *Int J Environ Res Public Health*, vol. 16, no. 9, p. 1534, May 2019, doi: 10.3390/ijerph16091534.
[25] "Datos Abiertos | Plataforma Nacional de Datos Abiertos." Accessed: Mar. 11, 2024. [Online]. Available: https://www.datosabiertos.gob.pe/datos-abiertos-0
[26] "ESTADO NUTRICIONAL EN ADULTOS DE 18 A 59 AÑOS, PERÚ: 2017 – 2018 | Plataforma Nacional de Datos Abiertos." Accessed: Jan. 06, 2024. [Online]. Available: https://www.datosabiertos.gob.pe/dataset/estado-nutricional-en-adultos-de-18-59-a%C3%B1os-per%C3%BA-2017-%E2%80%93-2018
[27] Carolina Tarqui Mamani *et al.*, "Estado nutricional en adultos de 18 a 59 años,VIANEV 2017-2018," pp. 1–191, Mar. 2023, Accessed: Dec. 02, 2023. [Online]. Available: https://cdn.www.gob.pe/uploads/document/file/4527291/Informe%20Tecnico-%20Estado%20nutricional%20en%20adultos%20de%2018%20a%2059%20a%C3%B1os%2CVIANEV%202017-2018G6aBA.pdf?v=1683566457
[28] D. D. Solomon *et al.*, "Hybrid Majority Voting: Prediction and Classification Model for Obesity," *Diagnostics 2023, Vol. 13, Page 2610*, vol. 13, no. 15, p. 2610, Aug. 2023, doi: 10.3390/DIAGNOSTICS13152610.
[29] R. Sujatha, J. M. Chatterjee, S. S. Thorunitha, and S. Nadhiya, "Evaluation of Dietary Habits in Relation to Covid-19 Mortality Rate Using Machine Learning Techniques," *Journal of System and Management Sciences*, vol. 12, no. 2, pp. 174–194, Jan. 2022, doi: 10.33168/JSMS.2022.0208.

[30] N. Mahesh, P. B. Pati, K. Deepa, and S. Yanan, "Body Fat Prediction using Various Regression Techniques," *Proceedings of the 2nd IEEE International Conference on Advances in Computing, Communication and Applied Informatics, ACCAI 2023*, p. Chennai, doi: 10.1109/ACCAI58221.2023.10200647.

[31] D. A. T. Boncan, Y. Yu, M. Zhang, J. Lian, and V. Vardhanabhuti, "Machine learning prediction of hepatic steatosis using body composition parameters: A UK Biobank Study," *npj Aging*, vol. 10, no. 1, Dec. 2024, doi: 10.1038/s41514-023-00127-z.

[32] S. Aarthi, P. Kavishka, G. Meena, and S. Chidambaram, "UNA NUEVA TÉCNICA DE PREDICCIÓN PARA IDENTIFICAR LA ENFERMEDAD RENAL CRÓNICA EN ENFOQUES DE ANÁLISIS DE DATOS," *2023 Conferencia internacional IEEE sobre metodologías de investigación en gestión del conocimiento, inteligencia artificial e ingeniería de telecomunicaciones, RMKMATE 2023*, p. Chennai, doi: 10.1109/RMKMATE59243.2023.10369646.

[33] M. Al-Hassan, B. Abu-Salih, E. Alshdaifat, A. Aloqaily, and A. Rodan, "An Improved Fusion-Based Semantic Similarity Measure for Effective Collaborative Filtering Recommendations," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, p. 45, Dec. 2024, doi: 10.1007/s44196-024-00429-4.

[34] C. Ogwo, W. Osisioma, D. I. Okoye, and J. Patel, "Predicting dental anxiety in young adults: classical statistical modelling approach versus machine learning approach," *BMC Oral Health*, vol. 24, no. 1, p. 313, Dec. 2024, doi: 10.1186/s12903-024-04012-3.

[35] W. Zhou, Z. Yan, and L. Zhang, "A comparative study of 11 non-linear regression models highlighting autoencoder, DBN, and SVR, enhanced by SHAP importance analysis in soybean branching prediction," *Sci Rep*, vol. 14, no. 1, p. 5905, Dec. 2024, doi: 10.1038/s41598-024-55243-x.

[36] J. Okita *et al.*, "Development and validation of a machine learning model to predict time to renal replacement therapy in patients with chronic kidney disease," *BMC Nephrol*, vol. 25, no. 1, p. 101, Dec. 2024, doi: 10.1186/s12882-024-03527-9.

[37] F. Zhang, M. Petersen, L. Johnson, J. Hall, and S. E. O'Bryant, "Comorbidities Incorporated to Improve Prediction for Prevalent Mild Cognitive Impairment and Alzheimer's Disease in the HABS-HD Study," *Journal of Alzheimer's Disease*, vol. 96, no. 4, pp. 1529–1546, Dec. 2023, doi: 10.3233/JAD-230755.

[38] X. Pang, C. B. Forrest, F. Lê-Scherban, and A. J. Masino, "Prediction of early childhood obesity with machine learning and electronic health record data," *Int J Med Inform*, vol. 150, p. 104454, Jun. 2021, doi: 10.1016/J.IJMEDINF.2021.104454.

[39] R. Matta, "Posibilidades y límites del desarrollo en el patrimonio inmaterial. El caso de la cocina peruana," *Apuntes: Revista de Estudios sobre Patrimonio Cultural - Journal of Cultural Heritage Studies*, vol. 24, no. 2, pp. 196–207, 2011, Accessed: Mar. 27, 2024. [Online]. Available: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1657-97632011000200006&lng=en&nrm=iso&tlng=es