SAMSUNG

2024
Innovation
Campus

UDEM

PROYECTO SIC:

Modelado Predictivo de Enfermedades Cardiometabólicas Mediante Técnicas de Aprendizaje Automático

Omar Aldair Hernández Velasco

José Ernesto Velázquez Martínez

Oscar Alejandro Velazquez Alvarado

Isai Aram Perez Flores

Abril 2025

Índice

1. Introducción

México es un país que históricamente ha sufrido de enfermedades cardiometabólicas, en especial de tres de ellas: **la obesidad**, **la diabetes** y **la hipertensión**. Estos son padecimientos que afectan a muchos mexicanos. Algunas estadísticas son alarmantes:

- Según la Encuesta Nacional de Salud y Nutrición (ENSANUT) 2022, el 75.2 % de las personas mayores de 20 años presentan sobrepeso y obesidad, con una prevalencia mayor en mujeres (76.8 %) que en hombres. [9]
- La obesidad infantil también es preocupante. Se estima que para 2035, el 56 % de la niñez mexicana sufrirá sobrepeso, lo que podría reducir su esperanza de vida en comparación con generaciones anteriores.[6]
- La prevalencia total de diabetes en adultos mexicanos es del 18.3 %, según datos de 2022. De estos, el 12.6 % corresponde a casos diagnosticados y el 5.8 % a casos no diagnosticados.[9]
- El 22.1 % de la población adulta presenta prediabetes, condición que aumenta el riesgo de desarrollar diabetes tipo 2. [3]
- En México, el 29.9 % de las personas adultas viven con hipertensión arterial. [8]
- Se estima que más de 30 millones de mexicanos padecen hipertensión arterial; sin embargo, el 46 % de ellos desconoce su condición. [1]

Además, hay una relación causa-consecuencia entre estas enfermedades. Las personas que viven con obesidad tienen 1.7 veces más riesgo de padecer diabetes, 3.6 veces más riesgo de desarrollar hipertensión arterial y 2.3 veces más riesgo de presentar alteraciones en el perfil de lípidos, en comparación con aquellas con un índice de masa corporal normal [2].

Ante esta situación, nos dimos a la tarea de analizar bases de datos relacionadas con estas enfermedades para aplicar algoritmos de machine learning e inteligencia artificial con el objetivo de obtener información más detallada. Construimos modelos capaces de diferenciar entre personas que padecen o no estas enfermedades, e incluso modelos que clasifican a las personas en distintos niveles de obesidad.

Además, realizamos regresiones para predecir valores de ciertas medidas, como el índice de masa corporal (IMC) o la presión arterial, en función de otras características y hábitos diarios. Por último, aplicamos técnicas de clusterización para identificar grupos con características en común, lo que nos permitió detectar posibles tendencias ocultas.

1.1. Revisión de literatura

Con el objetivo de realizar un buen proyecto, nos dimos a la tarea de revisar literatura relacionada con nuestro trabajo. Buscamos artículos científicos en los que se aplicaran modelos de machine learning e inteligencia artificial para estudiar las enfermedades que analizamos. Estudiamos los siguientes artículos:

- **A machine learning approach for obesity risk prediction** [4]: Este estudio analiza el riesgo de obesidad mediante algoritmos de machine learning, aplicando nueve modelos diferentes a un conjunto de datos de más de 1100 personas con y sin obesidad. Se evaluaron clasificadores como k-NN, random forest, regresión logística, SVM y gradient boosting, entre otros. El objetivo del estudio es predecir el riesgo de obesidad y comprender sus causas para fomentar la prevención, especialmente en la población de Bangladesh.
- **Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research 2018** [10]: Este estudio compara modelos de machine learning (regresión logística, CART y Naïve Bayes) para predecir la obesidad en adultos con datos de salud pública de Indonesia. La regresión logística obtuvo el mejor desempeño, y se utilizó SMOTE para abordar el desbalance de clases. Se identificaron factores clave como hábitos alimenticios, actividad física, tabaquismo e hipertensión, proporcionando información útil para mejorar estrategias de prevención.
- **Multiomics and eXplainable artificial intelligence for decision support in insulin resistance early diagnosis: A pediatric population-based longitudinal study** [11] : Este estudio propone un sistema de inteligencia artificial explicable para predecir la resistencia a la insulina en niños con obesidad, usando datos clínicos y biológicos de 90 niños prepúberes. El sistema mostró una alta precisión en las predicciones (AUC y G-mean de 0.92). Identificaron biomarcadores importantes, como el IMC, la relación entre leptina y adiponectina, y ciertos patrones genéticos. Los resultados destacan la importancia de usar datos diversos y métodos de IA explicables para mejorar el diagnóstico temprano.
- **Machine Learning Approaches for Predicting Hypertension and Its Associated Factors Using Population-Level Data From Three South Asian Countries** [5] : Este estudio utiliza enfoques de machine learning para predecir la hipertensión y sus factores asociados en una base de datos de tres países del sur de Asia. Compara el rendimiento de modelos de machine learning con métodos estadísticos tradicionales y destaca que los modelos de machine learning superan las predicciones clínicas en términos de precisión.

2. Metodología

Para hacer el proyecto, utilizamos tres datasets, cada uno de ellos destinado a una enfermedad específica, ya que es difícil conseguir una base de datos que junte parámetros de estas cuatro enfermedades. Cada uno de los datasets tuvo sus dificultades, por lo que se trataron de forma distinta.

Metodología aplicada por dataset

- **Obesidad:** Realizamos un análisis completo que incluyó clasificación, regresión y clusterización.
- **Diabetes:** Nos centramos en clasificación y clusterización, priorizando la identificación de grupos con características similares.
- **Hipertensión:**Nuestro enfoque principal fue la clasificación, donde evaluamos múltiples modelos para seleccionar el más óptimo.

2.1. Obesidad

Para analizar esta enfermedad, utilizamos el dataset "Dataset for Estimation of Obesity Levels Based on Eating Habits and Physical Condition in Individuals from Colombia, Peru, and Mexico". En él se recoge información relacionada con los hábitos alimenticios y la actividad física en México, Colombia y Perú. Se puede encontrar información detallada acerca del dataset en el artículo del mismo nombre [7].

El dataset consta de 2,111 registros y 17 variables, entre las cuales destacan: peso, altura, nivel de obesidad, métodos de transporte e historial familiar, entre otras. Hay 8 variables numéricas, 5 categóricas y 4 booleanas. El flujo de trabajo seguido fue, a grandes rasgos, el siguiente:

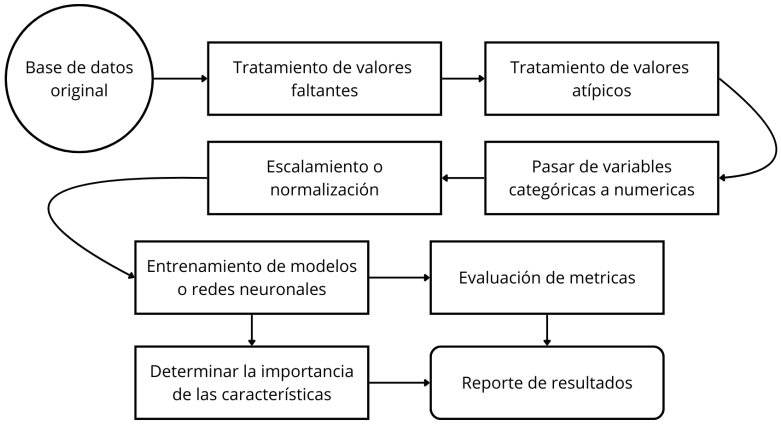

Figura 1: Flujo de trabajo utilizado en el análisis del dataset de obesidad.

2.1.1. Preprocesamiento

Comenzamos con el tratamiento de valores faltantes. En el análisis exploratorio inicial, nos dimos cuenta de que el dataset no tenía valores faltantes, por lo que pudimos pasar al siguiente paso sin problemas. Para detectar y tratar los valores atípicos, utilizamos el método del rango intercuartílico. Cuatro variables tenían valores atípicos: Age (Edad), Height (Altura), Weight (Peso) y NCP (Número de comidas al día).

En las variables Age y NCP (Número de comidas completas al día), la eliminación de datos supondría una reducción significativa en el número de registros. Estos valores se consideran atípicos no por un error de registro o por ser valores extremadamente raros, sino porque hay una cantidad mucho mayor de registros con valores menores, como se puede ver en Figura 2 y Figura 3. Consideramos que para estas dos variables no es necesario eliminar los valores atípicos, pues representan la naturaleza de un grupo de individuos dentro del dataset. Aunque son minoría, son registros totalmente válidos: las personas mayores serían eliminadas por completo y también las personas de bajos recursos que comen una vez al día o aquellas que consumen más de tres comidas diarias, lo que implicaría la pérdida de información valiosa. En conclusión, estos valores permanecerán en el análisis.

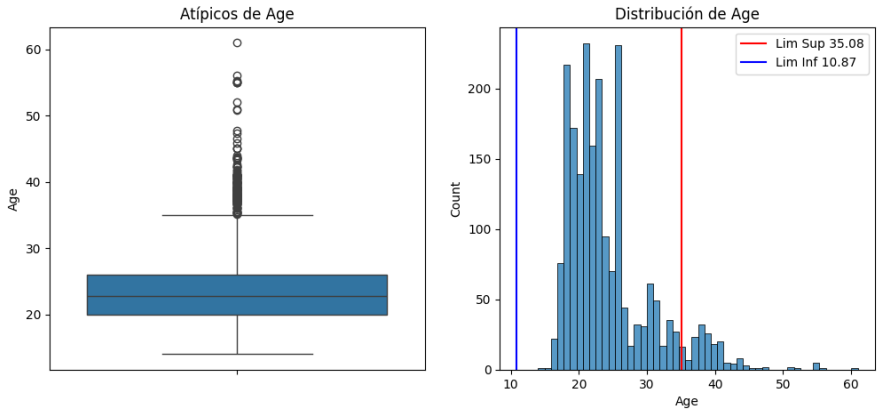

Figura 2: Gráfica de bigotes e histograma de la variable edad.

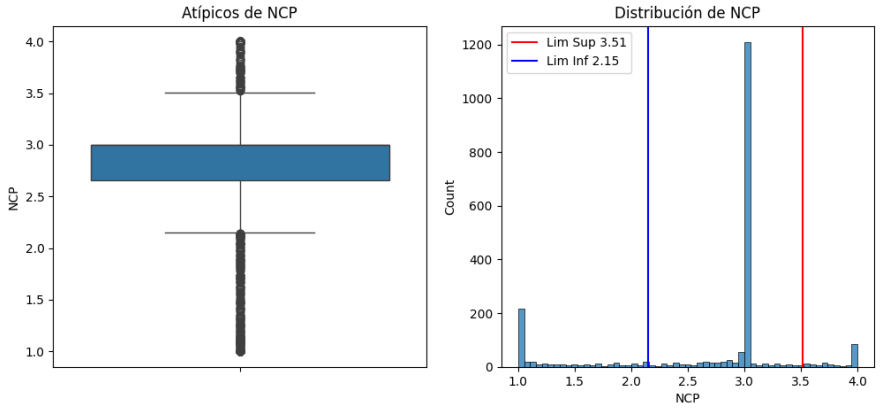

Figura 3: Gráfica de bigotes e histograma de la variable NCP (Número de comidas diarias).

La variable Height (Altura) tiene un límite superior de 1.97 m, lo cual se puede considerar atípico, ya que el promedio de altura en los países de donde se recolectó la información está bastante alejado de este valor. Esto podría deberse a errores de medición o a casos extremadamente raros.

Para la variable Weight (Peso), se consideran atípicos los valores por encima de 170 kg. Esta cifra podría tiene ocurrencias muy poco frecuentes, por lo que puede considerarse un valor atípico.

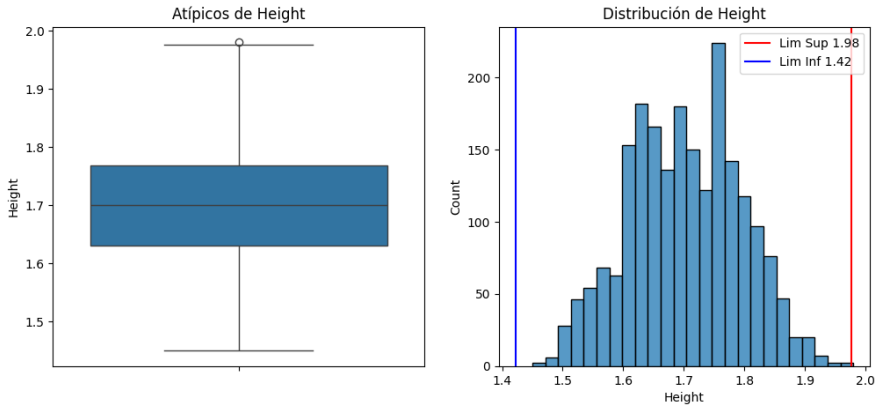

Figura 4: Gráfica de bigotes e histograma de la variable Height (Altura).

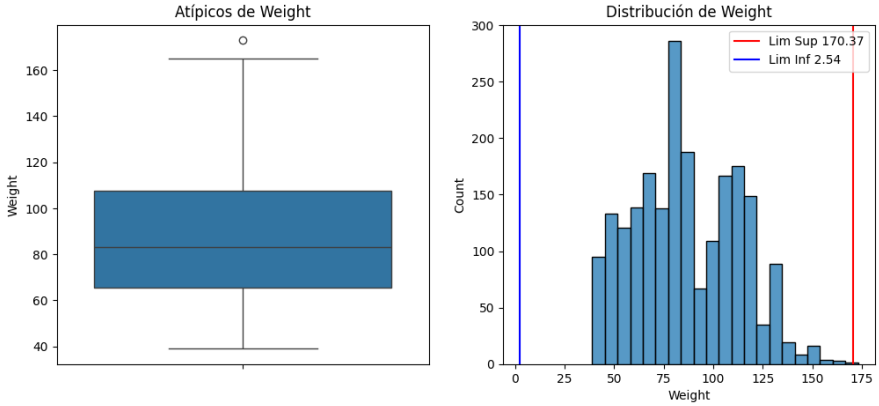

Figura 5: Gráfica de bigotes e histograma de la variable Weight (Peso).

Después de eliminar los valores atípicos, obtuvimos un dataset de 2109 registros, un decremento bastante pequeño. El siguiente paso es el manejo de las variables categóricas. Este dataset tenía variables nominales y ordinales.

Tratamos las variables nominales con **One Hot Encoding**. Estas eran: **Gender** (Género), **family history with overweight** (Historial familiar con obesidad), **SMOKE** (Fumador), **SCC** (Monitorea las calorías que consume al día), **FAVC** (Consume comida altamente calórica con frecuencia) y **MTRANS** (Método de transporte). A excepción de **MTRANS**, todas eran booleanas. **MTRANS** tiene cinco respuestas correspondientes a distintos medios de transporte; como no hay una jerarquía clara, decidimos utilizar One Hot Encoding.

Eliminamos la primera columna generada de cada variable para evitar problemas de multicolinealidad.

Para las variables ordinales **CAEC** (¿Consumes algo entre comidas principales?) y **CALC** (¿Con qué frecuencia consumes alcohol?) utilizamos Label Encoding. Estas variables tienen como respuestas los adverbios de frecuencia: *No*, *A veces*, *Frecuentemente* y *Siempre*. Dado que existe una jerarquía en estos valores, asignamos la siguiente codificación:

- No: 0
- A veces: 1
- Frecuentemente: 2
- Siempre: 3

La variable **NObeyesdad** nos dice el nivel de obesidad de la persona, sobre ella también hicimos label encoding:

- Insufficient weight (Peso insuficiente): 0
- Normal weight (Peso normal): 1
- Overweight level 1 (Sobrepeso nivel 1): 2
- Overweight level 2 (Sobrepeso nivel 2): 3
- Obesity level 1 (Obesidad nivel 1): 4
- Obesity level 2 (Obesidad nivel 2): 5
- Obesity level 3 (Obesidad nivel 3): 6

Para este dataset utilizamos estandarización Z-Score ya que nuestras variables se distribuían aproximadamente de manera normal. Esta parte es importante ya que usaremos algoritmos sensibles a las diferentes escalas en los datos.

Después del preprocesamiento obtuvimos un dataframe con 2109 registros y 20 variables.

2.1.2. Modelos y redes neuronales

Algoritmos de clasificación. En este dataset efectuamos tareas de clasificación sobre la variable **NObeyesdad**. Esta clase está bien balanceada, por lo que es una buena señal para confiar en la métrica de *accuracy*, aunque sin dejar de lado las otras métricas.

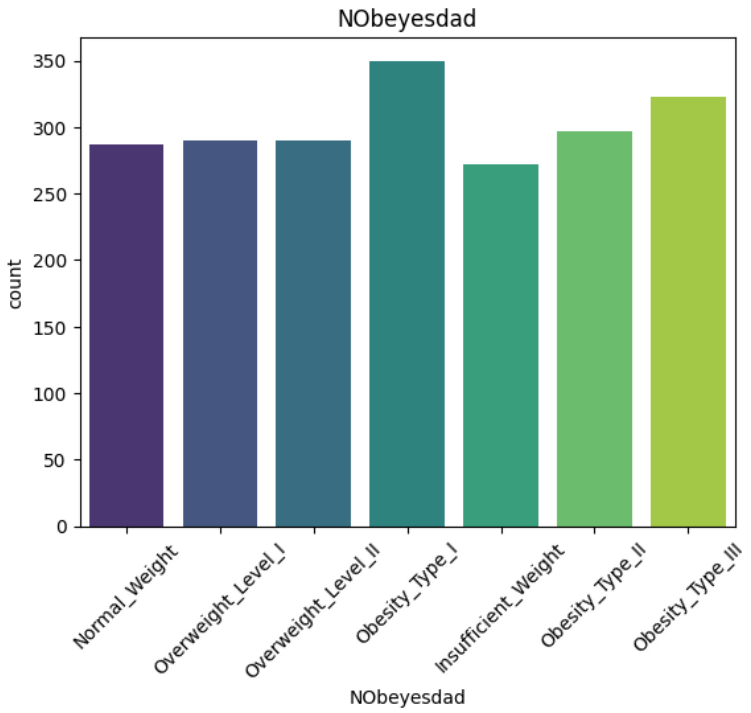

Figura 6: Gráfico de barras en donde podemos ver la cantidad de registros de cierta clase.

Para la clasificación decidimos utilizar tres algoritmos de *machine learning*. Los tres algoritmos fueron: KNN, `XGBClassifier` y *Support Vector Machine* (SVM). Antes de comenzar a ajustar los parámetros, decidimos hacer una selección de características para KNN y SVM mediante la implementación de un algoritmo genético, en el cual la función objetivo a maximizar era la *accuracy*.

Para el KNN utilizamos una implementación del algoritmo genético en donde se tomaba en cuenta una búsqueda en cuadrícula (*grid search*) con el parámetro k variando en el rango de 5 a 14 (incluyendo el 14), además de una validación cruzada de 5 pliegues con una partición 80 y 20. Esto significa que la función objetivo a maximizar era el promedio de las métricas de *accuracy* obtenidas en la validación cruzada. Las dos mejores métricas fueron de 0.81, y las obtuvimos con la combinación de variables: Age, Weight, Gender Male, Family History with Overweight Yes, SMOKE Yes, MTRANS Bike y MTRANS Public Transportation. Ambas combinaciones lograron dicho resultado con el parámetro $k = 6$.

En este proyecto nos interesa conocer la importancia de las características en los algoritmos, por lo que decidimos ejecutar nuevamente el algoritmo genético, pero ahora sin búsqueda en cuadrícula, centrándonos únicamente en $k = 6$. De este modo, buscamos identificar cuáles son las combinaciones de características que aparecen con mayor frecuencia en los mejores resultados y, con base en ello, calcular su importancia.

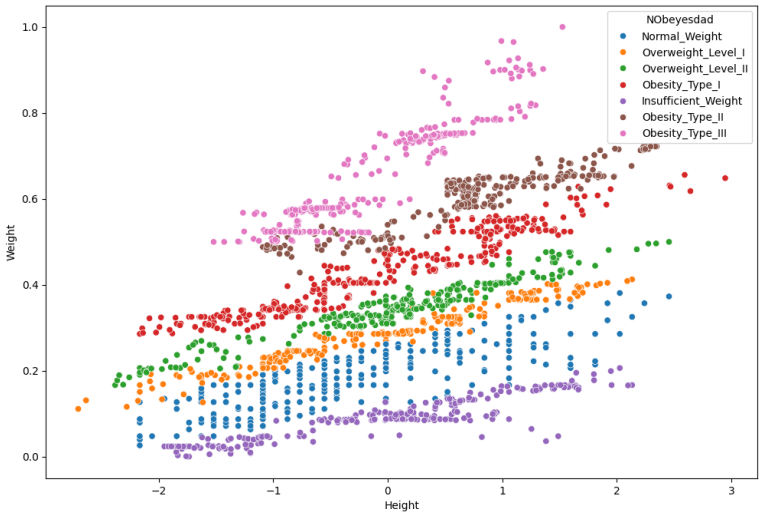

Figura 7: Las variables peso y estatura permiten visualizar claramente la distribución de los datos por clases.

Tras este análisis, encontramos el modelo final: un KNN con $k = 6$, utilizando las características Age, Weight, Gender Male, Family History with Overweight Yes, SCC Yes, MTRANS Bike y MTRANS Public Transportation, **alcanzando un *accuracy* de 0.83**. En las cinco mejores combinaciones de características (todas con valores superiores a 0.796), la frecuencia de las variables fue la siguiente:

Podemos observar que el género, el historial familiar de obesidad, el hábito de fumar, la edad y el peso son algunas de las características más importantes. De hecho, es llamativo que la variable *peso* no aparezca con la misma frecuencia que las tres primeras. Esto podría deberse a que es una variable que puede ser explicada, en mayor o menor medida, en función de otras (como veremos en regresión). De hecho, en las cinco mejores combinaciones en las que el peso no aparece, hay una mayor cantidad de variables relacionadas con hábitos alimenticios, las cuales, en conjunto, podrían explicar la variable peso.

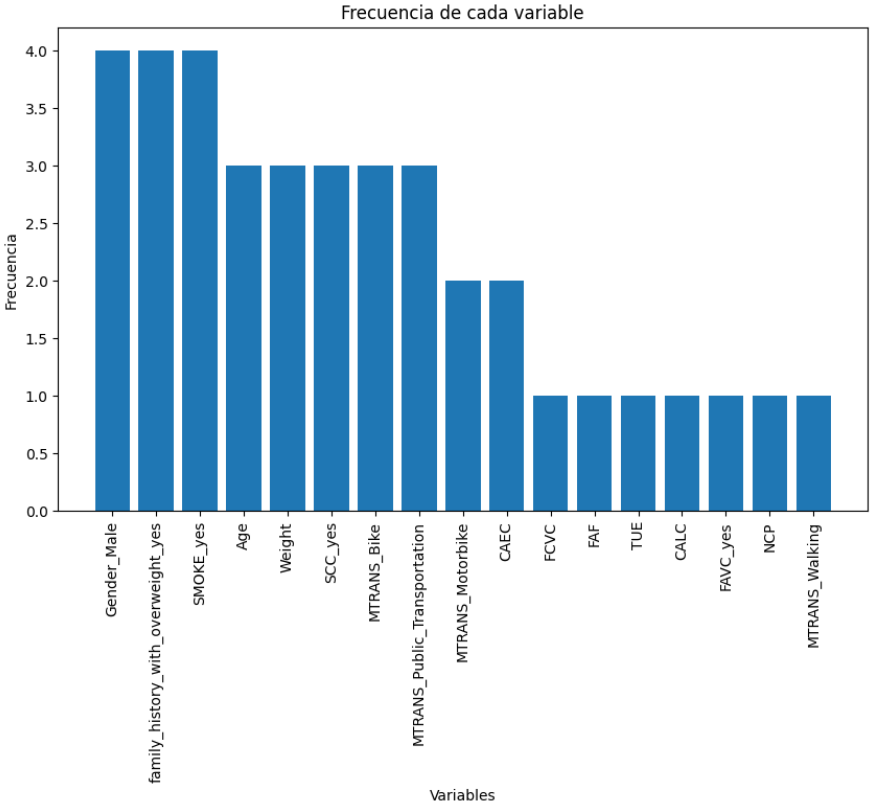

Figura 8: Frecuencia de aparición de las variables en los 5 mejores resultados.

Para el SVM, realizamos una selección de características con un algoritmo genético sin realizar un *grid search*, pero con 5 pliegues para la validación cruzada. La función a maximizar fue el promedio de *accuracy*, y lo corrimos con un kernel lineal y $C = 2$. Obtuvimos muy buenos resultados, ya que **alcanzamos un *accuracy* de 0.9585** con la combinación de variables: Age, Height, Weight, FCVC, NCP, CAEC, CH2O, FAF, TUE, CALC, Gender_Male, family_history_with_overweight_yes, SMOKE_yes, SCC_yes, FAVC_yes, MTRANS_Bike, MTRANS_Motorbike, MTRANS_Public_Transportation, MTRANS_Walking. Además, obtuvimos las características más repetidas en las 5 mejores combinaciones, todas ellas alcanzaron métricas superiores a 0.95.

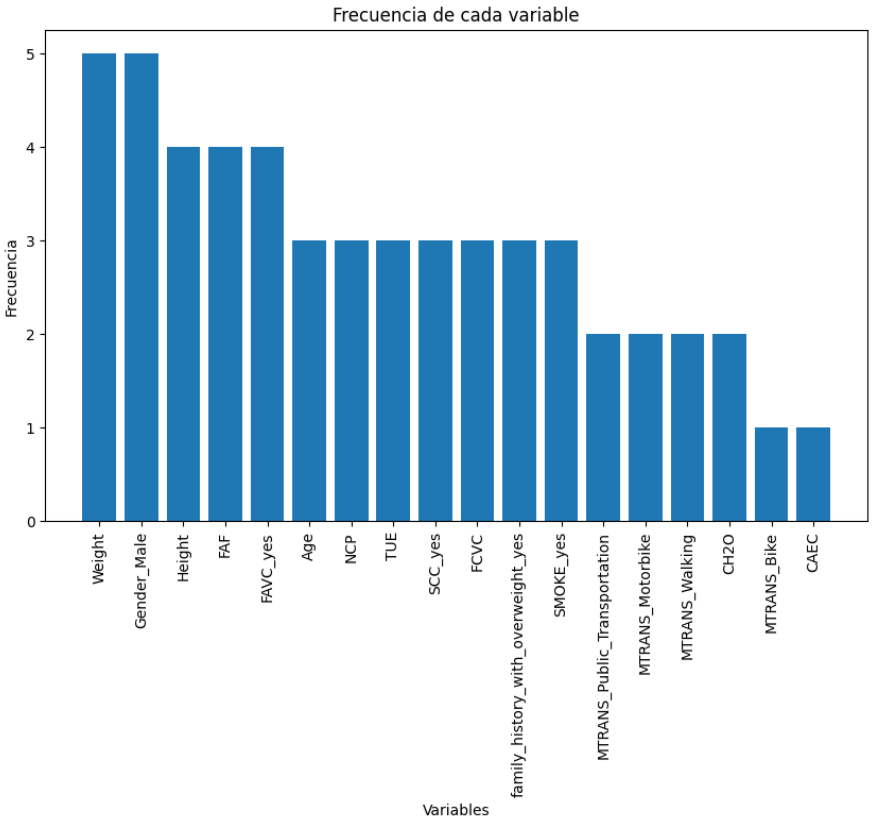

Figura 9: Frecuencia de aparición de variables en las 5 mejores combinaciones con el SVM

Sin embargo, el modelo que tuvo un mejor rendimiento y con diferencia fue el XGBClassifier. Este método no necesita tanta selección de variables, ya que por defecto tiene la capacidad de calcular la importancia de las características. Con la configuración de parámetros: objective=multi:softmax (ya que es una clasificación multiclase) y métrica de evaluación como mlogloss, obtuvimos un *accuracy* medio de 0.9682. El modelo con el que nos quedamos tiene los mismos parámetros y un *accuracy* de 0.971.

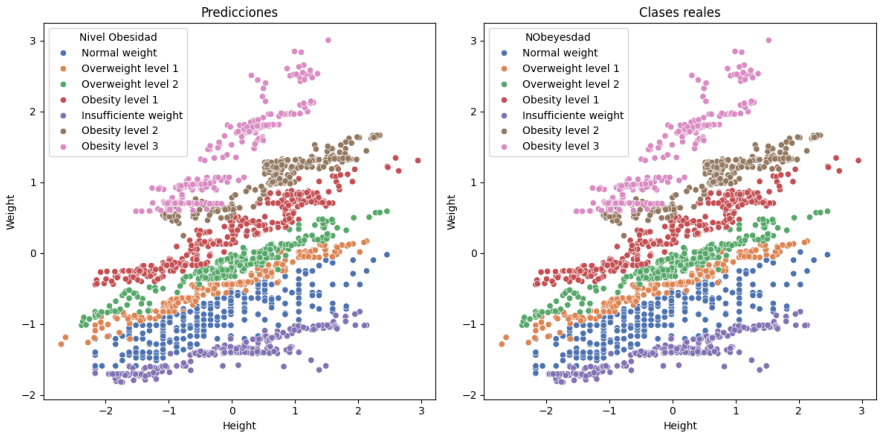

Figura 10: El XGBClassifier abstrae muy bien las relaciones, logrando una clasificación casi perfecta

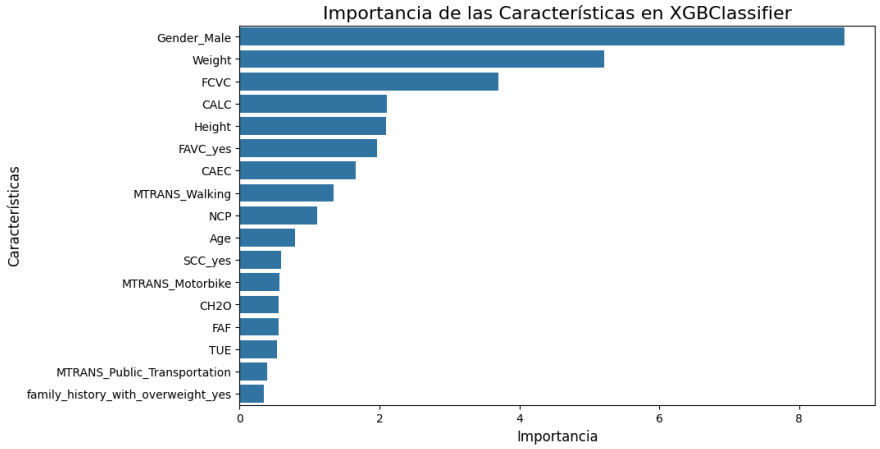

Figura 11: Importancia de las características en el modelo XGBClassifier

A diferencia del SVM y KNN, en los que solo analizábamos la importancia para el modelo por la frecuencia, el XGBClassifier, al ser un árbol, puede calcular la importancia de cada característica dependiendo de cuánto ayuda a minimizar la función de error. En este dataset, podemos ver que el género, el peso, la frecuencia con la que se comen verduras, la frecuencia con la que se consume alcohol y la altura son las características más importantes.

Regresión. Para los algoritmos de regresión probamos dos: la regresión lineal clásica y una regresión SVR. Nuestro objetivo era lograr hacer predicciones de la variable peso con el resto de las variables, a excepción del nivel de obesidad. Al hacer los modelos, notamos que no estaban entendiendo el comportamiento de los pesos más grandes, por lo que decidimos probar con una red neuronal.

Utilizamos una regresión lineal con validación cruzada de 5 pliegues, con los siguientes resultados: R^2 promedio: 0.551 y MSE promedio: 0.448. Estos son malos resultados, lo que nos dice que lo más probable es que no haya una relación lineal entre las variables del dataframe y el objetivo. La siguiente gráfica nos permite visualizar la efectividad del modelo. Cuanto más cerca estén los puntos de la línea roja, mejor será su desempeño.

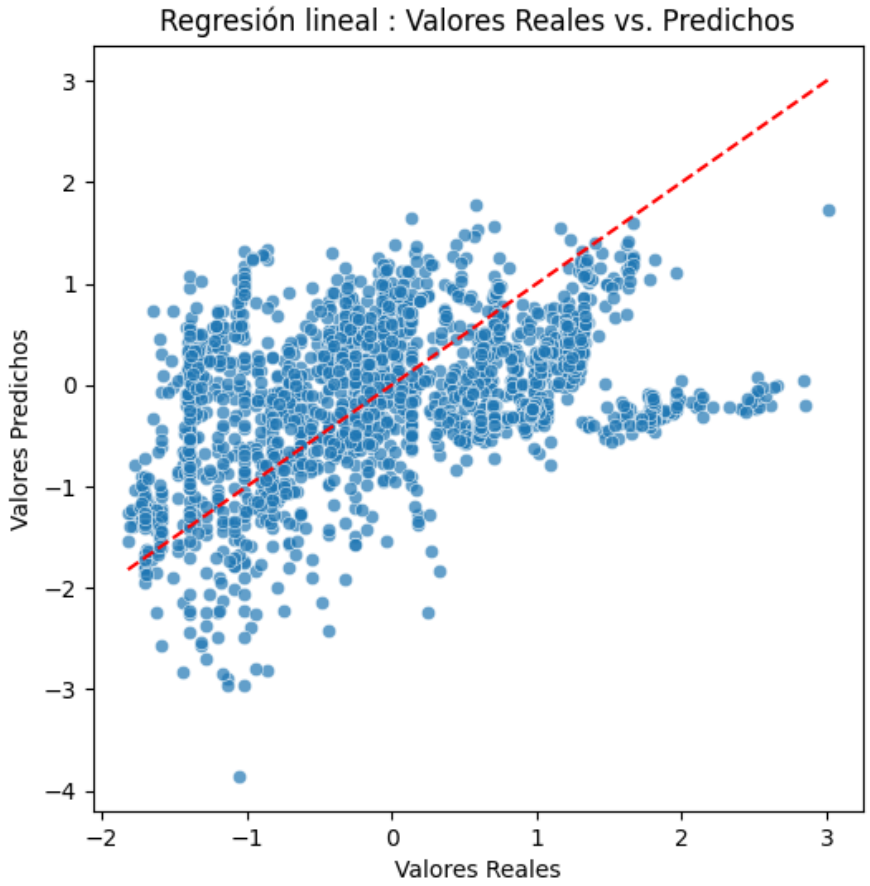

Figura 12: Valores reales vs. predichos usando regresión lineal

Pasamos a utilizar un SVR y lo manejamos de dos formas: con kernel polinomial y con kernel RBF para ver si alguno de ellos era capaz de darnos mejores resultados. Comenzamos con el kernel polinomial. Después de hacer pruebas con los parámetros, usamos $C = 1$, $\epsilon = 0{,}1$ y grado $= 4$, obteniendo un R^2 promedio de 0.725 y un MSE promedio de 0.272. Usando el kernel RBF con los parámetros $C = 2{,}5$ y $\epsilon = 0{,}1$, obtuvimos un R^2 promedio de 0.87 y un MSE promedio de 0.128. Estos resultados fueron mucho mejores que los de la regresión lineal. El SVR con RBF alcanzó métricas bastante competentes. El principal problema que notamos fue que, en los pesos más altos, los modelos

no estaban aprendiendo bien. Esto debe ser porque las relaciones son diferentes a partir de cierto valor y no logran comprenderlas.

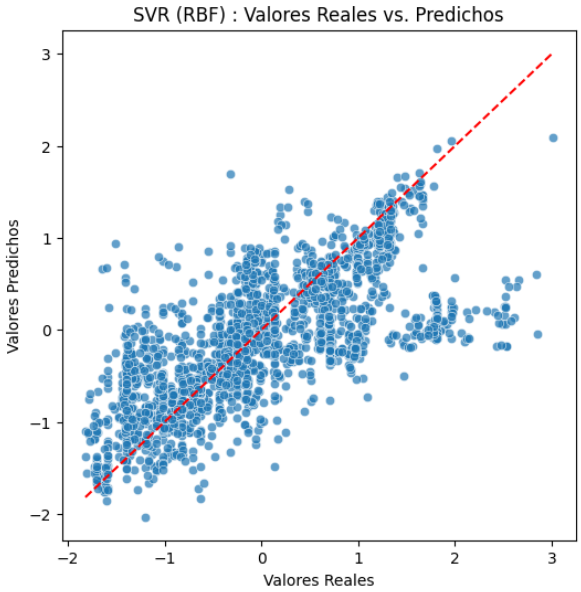

Figura 13: Valores reales vs. predichos usando SVR con kernel RBF

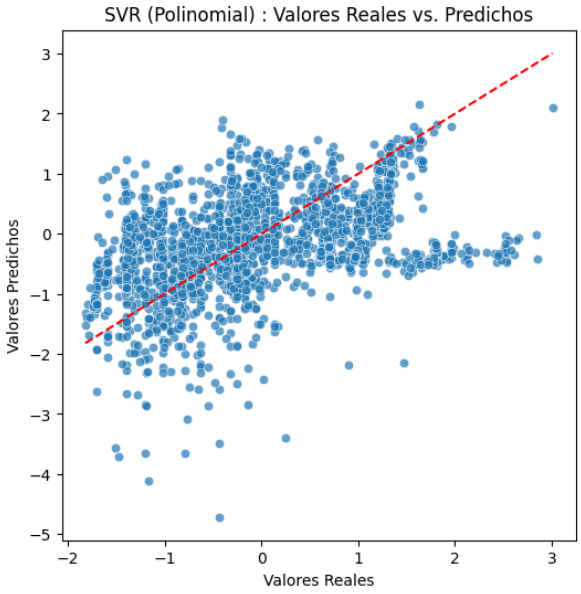

Figura 14: Valores reales vs. predichos usando SVR con kernel polinomial

Propusimos (después de probar con varias arquitecturas) la siguiente

- Capa de entrada con **1024 neuronas**, activación `ReLU`.
- Capa oculta con **512 neuronas**, activación `ReLU`.
- Capa oculta con **256 neuronas**, activación `ReLU`.
- Capa oculta con **128 neuronas**, activación `ReLU`.
- Capa oculta con **128 neuronas**, activación `ReLU`.
- Capa de salida con **1 neurona**.
- Utilizamos un `earlystopping` que se detuviera cuando el `val_loss` dejara de mejorar en un intervalo de 20 épocas.

Para entrenar la red neuronal, primero utilizamos la herramienta `SHAP`, que es una biblioteca que nos ayuda a entender cómo afectan las variables en el resultado. Con ella, podemos ver qué tan relevante es cada característica para la red. Por lo tanto, primero entrenamos la red neuronal y, en base a ella, obtenemos la importancia de cada variable según `SHAP`. Decidimos eliminar las variables relacionadas con transporte (`MTRANS_Bike`, `MTRANS_Motorbike`, `MTRANS_Walking`, `MTRANS_Public_Transportation`), pues no nos estaban proporcionando mucha información sobre el peso. Luego, entrenamos la red otra vez con 16 variables. **Una vez hecho esto, logramos que nuestra red neuronal alcanzara un MSE de 0.094 y un R^2 de 0.908**. De hecho, en la imagen Figura 15, podemos ver cómo los valores más grandes están más cerca de la recta ideal, lo que indica que la red logró capturar esos comportamientos.

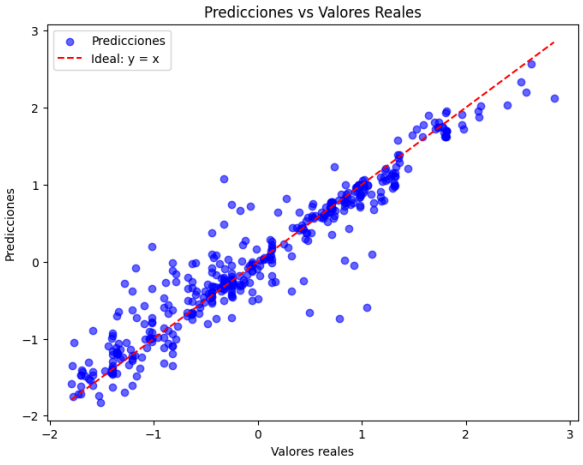

Figura 15: Valores reales vs. predichos usando red neuronal

En la siguiente gráfica podemos observar aspectos interesantes. Esta se interpreta de la siguiente forma: las características están ubicadas en el lado izquierdo, y cada una tiene un cúmulo de puntos delante. Estos cúmulos tienen colores que van del rojo al azul. Los puntos representan registros; un

punto rojo indica que el valor numérico de la característica en ese registro es muy alto, mientras que un punto azul representa un valor numérico bajo. La posición de los puntos respecto al eje x también tiene interpretación. Dependiendo de qué tan a la izquierda esté el punto, predecirá un valor bajo o muy bajo; si está a la derecha, predice valores más grandes.

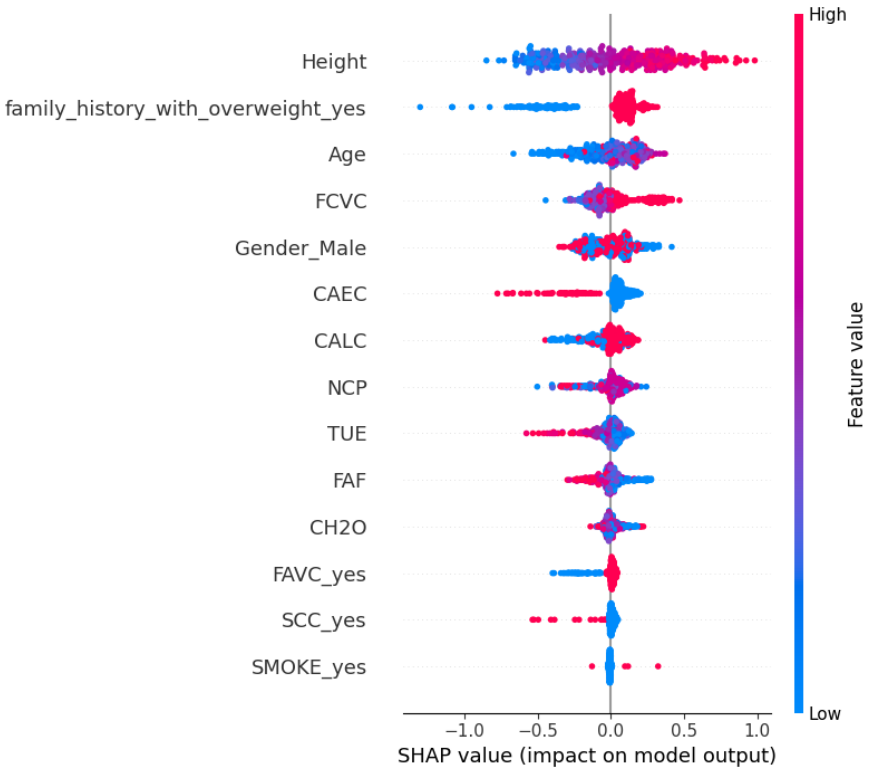

Figura 16: Gráfica de SHAP de importancia e impacto de las características

La variable altura es de las más importantes para la regresión. Además, lo que podemos observar es que a menor altura, menor peso, y a mayor altura, mayor peso. El historial de obesidad en la familia también tiene un comportamiento similar e incluso más polarizado, ya que prácticamente todos los registros con obesidad en la familia tienden a tener valores más altos de peso, y lo contrario ocurre con aquellos que no tienen antecedentes de obesidad. Algo interesante es la variable CAEC (¿Qué tan seguido comes entre tus comidas principales?), que nos indica que quienes comen más entre comidas suelen obtener valores de predicción más bajos. Esto es algo raro, ya que la lógica nos diría que debería ser al revés.

Clusterización. Con la intención de analizar la relación entre **Weight** (Peso) y **CAEC** (frecuencia de consumo entre comidas principales), identificada en el análisis de características de la figura Figura 16, aplicamos una clusterización con **K-Means** sobre estas variables. Nuestro objetivo es determinar si podemos identificar grupos que revelen patrones no observados previamente.

Dado que **K-Means** requiere definir un número de clusters k, utilizamos el **método del codo** para determinar el valor óptimo. Los pasos para aplicarlo son los siguientes:

- Ejecutar **K-Means** con diferentes valores de k (por ejemplo, de 1 a 10).
- Registrar la **inercia** (suma de los cuadrados de las distancias de cada punto a su centroide).
- Graficar la inercia en función de k.
- Identificar el punto donde la reducción de la inercia deja de ser significativa (formando un "codo" en la gráfica).
- Seleccionar el valor de k en el "codo" como el número óptimo de clusters.

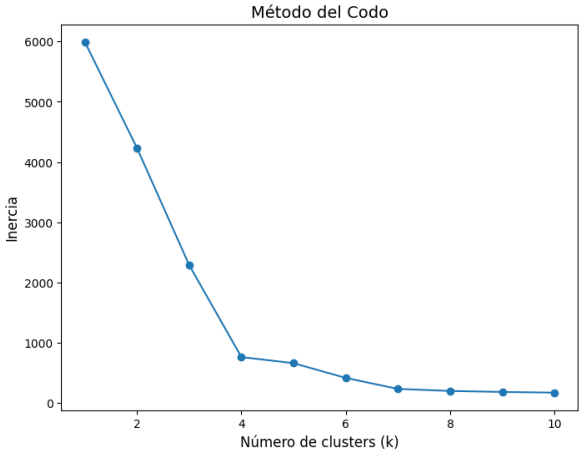

Figura 17: El codo o punto de inflexión se encuentra en $k = 4$.

Dado que el punto de inflexión se encuentra en $k = 4$, seleccionamos este valor y procedimos a ejecutar **K-Means** para visualizar los resultados.

Los cuatro grupos identificados a través de la clusterización son:

- **Personas que no comen entre comidas y pesan poco:** Este grupo **no** come entre comidas principales y tiene un peso bajo. Es un grupo cuya existencia era esperada según la lógica.
- **Personas que no comen entre comidas y pesan mucho:** Aunque anticipábamos la existencia de este grupo, esperábamos que fuera más reducido. Sin embargo, los resultados muestran que hay personas con un peso elevado a pesar de no comer entre comidas.
- **Personas que comen entre comidas y pesan mucho:** Se trata de otro grupo cuya existencia era esperada. Estas personas comen entre comidas con cierta frecuencia. Curiosamente, los que

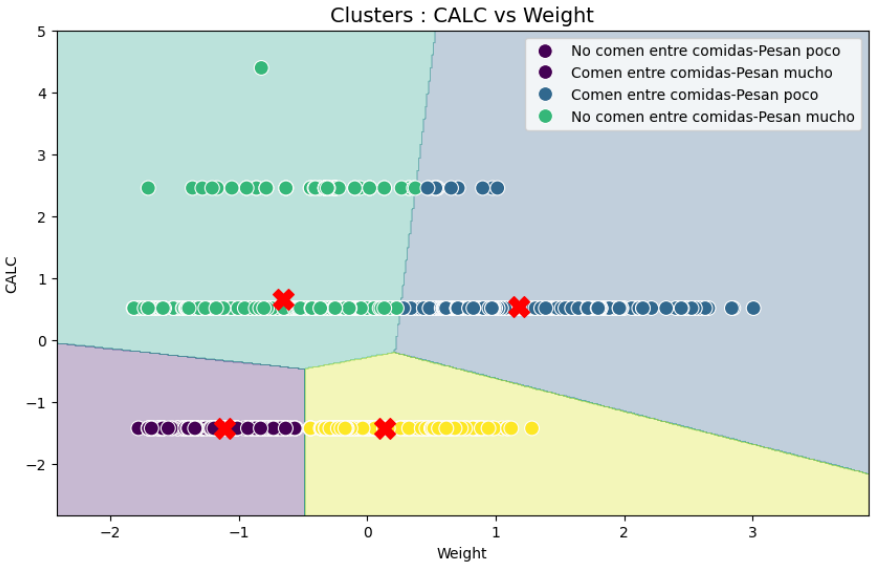

Figura 18: Clusterización con $k = 4$ sobre las variables CAEC y Weight

reportaron comer ”a veces” entre comidas son los que alcanzan los pesos más altos. Esto podría deberse a un sesgo en las respuestas de la encuesta: se esperaría que las personas con mayor peso fueran las que comen ”siempre” entre comidas, pero no hay registros que indiquen esto. Es posible que algunos encuestados, por pena o miedo a relacionar su peso con sus hábitos alimenticios, no hayan respondido con total sinceridad. No obstante, también podría tratarse de un comportamiento real, aunque para confirmar esto sería necesario analizar en mayor profundidad la población encuestada.

- **Personas que comen entre comidas y pesan poco:** No esperábamos encontrar este grupo o, en caso de existir, asumíamos que sería poco numeroso. Sin embargo, la clusterización revela que hay una cantidad considerable de personas en esta categoría. De hecho, las personas que reportaron comer ”siempre” o ´´frecuentemente´´ entre comidas pertenecen mayormente a este grupo de bajo peso. Este hallazgo, en conjunto con el grupo anterior, nos ayuda a interpretar mejor la relación de CAEC observada en la Figura 16. Analizando los grupos encontrados por K-Means, estos patrones comienzan a tener más sentido.

2.2. Diabetes

Para analizar la enfermedad de la Diabetes usamos el conjunto de datos "CDC Diabetes Health Indicators"el cual contiene estadísticas sobre atención médica e información recopilada a través de encuestas sobre el estilo de vida de las personas, junto con sus diagnósticos de diabetes. El Dataset fue formado a traves de encuestas telefonicas financiado por The Behavioral Risk Factor Surveillance System (BRFSS) un sistema de vigilancia en salud publica es los Estados Unidos, en particular trabajamos con el conjunto de datos limpios diabetes_binary_health _ indicators_BRFSS2015.csv es un conjunto de datos con 253,680 respuestas de encuestas a BRFSS2015 de los CDC **Centers for Disease Control and Prevention** .

La variable objetivo, `Diabetes_binary`, tiene dos clases:

- **0**: No tiene diabetes.
- **1**: Prediabetes o diabetes.

El conjunto de datos contiene **21 variables de características**, proporcionando información detallada sobre factores de riesgo, condiciones médicas previas y hábitos de salud. Sin embargo, es importante señalar que los datos no están completamente equilibrados, lo que puede influir en los resultados de los modelos de aprendizaje automático. El estudio de la diabetes es crucial debido a su impacto en la salud pública y la creciente prevalencia de la enfermedad. En este contexto, nos basamos en la información proporcionada por el **Centro de Atención Integral del Paciente con Diabetes (CAIPaDi)**, que destaca la importancia de la detección temprana y el manejo adecuado de la diabetes en México [1].

2.2.1. Preprocesamiento

Primero se cargaron los datos de una fuente pública, una vez cargados los datos eliminamos la columna ID, ya que esta no proporciona información que sea útil para nuestros objetivos, también se eliminó la Variable Diabetes_binary antes de aplicar clustering. Se verificó que no existieran valores faltantes en el dataset, se decidió no eliminar valores atípicos ya que podían ser una fuente importante de información para el análisis, se hicieron gráficas de barras para visualizar la distribución de en cada una de las características,se exploraron otras variables relevantes como el BMI,la edad y nivel de colesterol, notando que los datos tienen un desbalance de clases con un 13,93 % de los registros que tienen diabetes ademas el 75,6 % hace ejercicio regularmente y el 42,4 % de los pacientes tienen colesterol alto. Después de haber hecho ese análisis exploracional, se continuo con el mismo numero de registros.

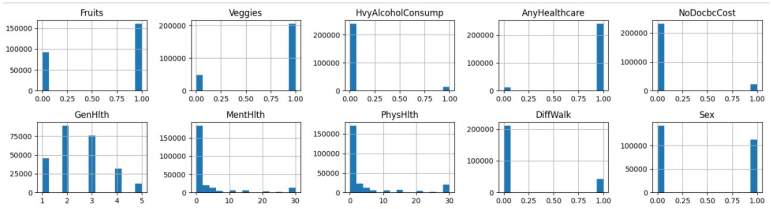

Figura 19: Grafica de proporciones

Factores de riesgo y aspectos de salud

- (HighBP): El 42.9 % de los pacientes presentan hipertensión.
- Colesterol alto (HighChol): El 42.4 % de los pacientes tienen colesterol alto.
- (PhysActivity): El 75.6 % de los pacientes realiza ejercicio regularmente.
- (BMI): La media es de 28.38, lo que indica sobrepeso.

Otros aspectos de salud

- (HeartDiseaseorAttack): El 9.4 % de los pacientes han sufrido enfermedades cardíacas.
- (GenHlth): La media de salud general auto calificada es 2.51, donde 5 indica peor salud.
- (MentHlth y PhysHlth): Los valores de estos indicadores se encuentran en el 75 % superior de la distribución.
- (DiffWalk): El 16.8 % de los pacientes reportan dificultad para caminar.

Variables demográficas

- (Sex): El 44 % de los pacientes son hombres y el 56 % son mujeres.
- (Age): La edad varía entre 1 y 13 clases, categorizadas en intervalos de edad.

Código	Rango de edad
1	18-24 años
2	25-29 años
3	30-34 años
4	35-39 años
5	40-44 años
6	45-49 años
7	50-54 años
8	55-59 años
9	60-64 años
10	65-69 años
11	70-74 años
12	75-79 años
13	80 años o más

Figura 20: Rango

- (Education): La mayoría de los pacientes tiene entre 4 y 6 en el nivel educativo
- (Income): La media de ingresos es 6.05 en una escala de 1 a 8.

2.2.2. Modelos

Clusterización, primero vamos a reducir la dimensionalidad utilizando PCA(Analisis de Componentes Principales) para mejorar la eficiencia computacional y visualizar los resultados del clustering de forma mas clara sin perder mucha información relevante,PCA nos permite identificar las combinaciones de las características mas relevantes.

Como trabajaremos con K-Means y es un dataset algo grande , el tiempo de ejecución se puede volver muy largo , así tendrá que trabajar con menos características.
Aplicamos el método del codo para para elegir un numero de clusters.

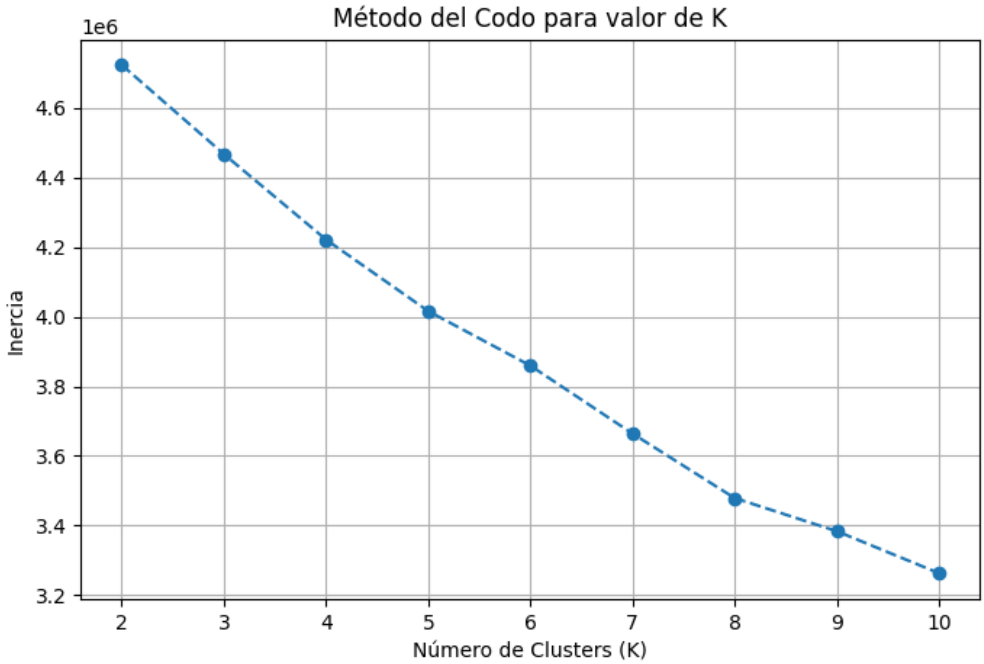

Figura 21: Cluster Optimo

Notemos que no hay un número claro que represente un gran cambio o un punto de inflexión, entonces usar el método del codo no debería ser suficiente , en consecuencia implementamos el método de la silueta como se muestra en la siguiente imagen.

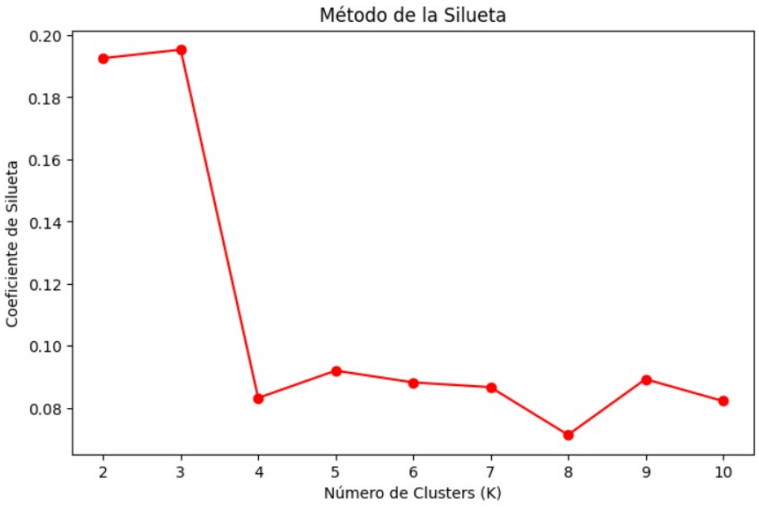

Figura 22: Silueta

El método de silueta es una técnica que se utiliza evaluar los resultados y nos ayuda a determinar el numero optimo de clusters.
El método de la silueta es una métrica que se utiliza para evaluar la calidad del agrupamiento basada en el coeficiente de silueta $s(i)$, definido como:

$$s(i) = \frac{b(i) - a(i)}{\text{máx}(a(i), b(i))}$$

$a(i)$ es la distancia promedio de un punto i a los otros puntos de su clúster, $b(i)$ es la distancia promedio al clúster más cercano. El índice está entre -1 y 1, los valores cercanos a 1 indican clusters bien definidos, mientras que los cercanos a 0 sugieren superposición de clusters. Para determinar el número óptimo de clusters K, calcula el promedio de $s(i)$ para todos los puntos y se selecciona el K que maximiza este valor. Del gráfico vemos que el coeficiente de la silueta disminuye conforme crecen el número de clusters, pero cae significativamente de 3 para 4, lo cual podría indicar que número de clusters adecuados es 3.

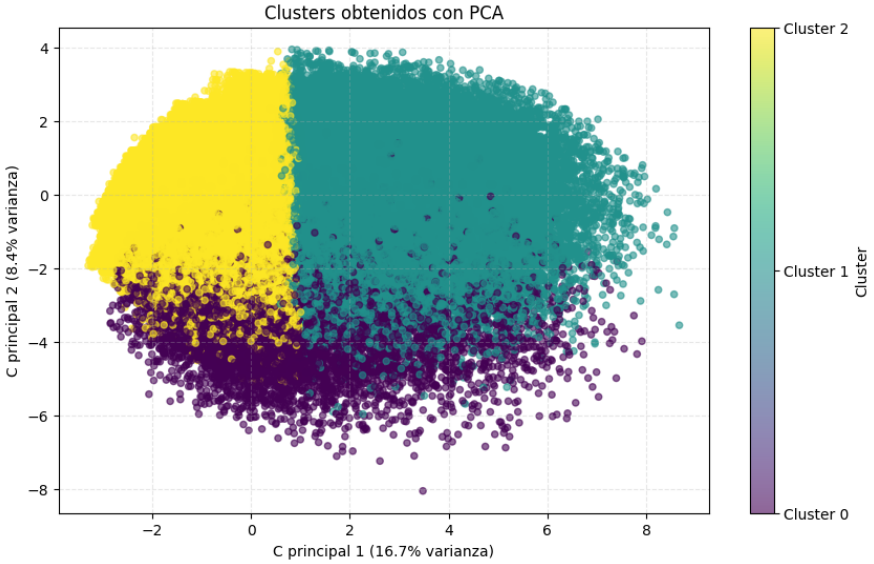

Figura 23: Agrupamiento

Pasemos a analizar cada uno de los cluster, la información se muestra en la siguiente imagen.

Cluster	Age	BMI	HighBP	HighChol	PhysActivity	\
0	6.158056	28.897486	0.342792	0.329776	0.691244	
1	9.497484	30.909098	0.733189	0.638760	0.495210	
2	7.636209	27.438658	0.325718	0.353578	0.854986	

Cluster	HeartDiseaseorAttack	Stroke	Smoker	Diabetes_binary	\
0	0.067265	0.030237	0.493815	0.112620	
1	0.272863	0.128921	0.607027	0.310647	
2	0.031865	0.009548	0.380768	0.079644	

Cluster	Diabetes_binary	Age	0	1	Education	PhysActivity	\
0	0.112620	6.158056	6389	5980	4.520414	0.691244	
1	0.310647	9.497484	38713	25070	4.474938	0.495210	
2	0.079644	7.636209	96872	80656	5.294128	0.854986	

Cluster	Smoker	MentHlth	GenHlth	AnyHealthcare	NoDocbcCost
0	0.493815	4.871453	2.695853	0.000000	0.369795
1	0.607027	7.021213	3.593324	0.999247	0.144004
2	0.380768	1.688883	2.109819	1.000000	0.042782

Figura 24: Silueta

El análisis de clustering aplicado al conjunto de datos reveló tres grupos distintos, cada uno con características únicas. El primer grupo "Jóvenes con Riesgo de Hábitos No Saludables", representado por el cluster 0, tiene un promedio de edad de la clase 6, un índice de masa corporal (BMI) de 28.90, y una prevalencia de diabetes del 11.26 %. Este grupo tiene una proporción significativa de personas que presentan hipertensión alta y colesterol elevado. Además, la actividad física es moderada, con un valor de 0.69, y la tasa de fumadores es del 49.38 %. En cuanto a la salud mental, este grupo tiene una puntuación alta en problemas de salud mental (4.87) y una salud general promedio de 2.70.

El segundo grupo "Alto Riesgo de Diabetes y Enfermedades" Cardiovasculares, el cluster 1, tiene un promedio de edad de la clase 10 , un BMI de 30.91, y una prevalencia de diabetes del 31.06 %. Este grupo presenta un mayor riesgo de hipertensión (73.32 %) y colesterol elevado (63.88 %), mientras que la actividad física es baja, con un índice de 0.49. La tasa de fumadores es también más alta, con un 60.70 %, y la salud mental es considerablemente peor, con una puntuación de 7.02, mientras que la salud general es de 3.59.

El tercer grupo "Estilo de Vida Saludable y Bajo Riesgo", el cluster 2, tiene un promedio de edad de la clase 8, un BMI de 27.44, y la menor prevalencia de diabetes, con solo un 7.96 %. La hipertensión y el colesterol elevado están presentes en menor medida en este grupo, con un índice de 32.57 % y 35.36 %, respectivamente. La actividad física es alta (0.85), mientras que la tasa de fumadores es la más baja, con un 38.08 %. Además, este grupo tiene una puntuación baja en salud mental (1.69) y una salud general de 2.11, lo que indica una mejor condición en comparación con los otros dos grupos. La tasa de acceso a atención médica es la más alta de los tres clusters, con el 100 % de los miembros. En el siguiente grafico se muestra la importancia de cada caracteristica cubierta.

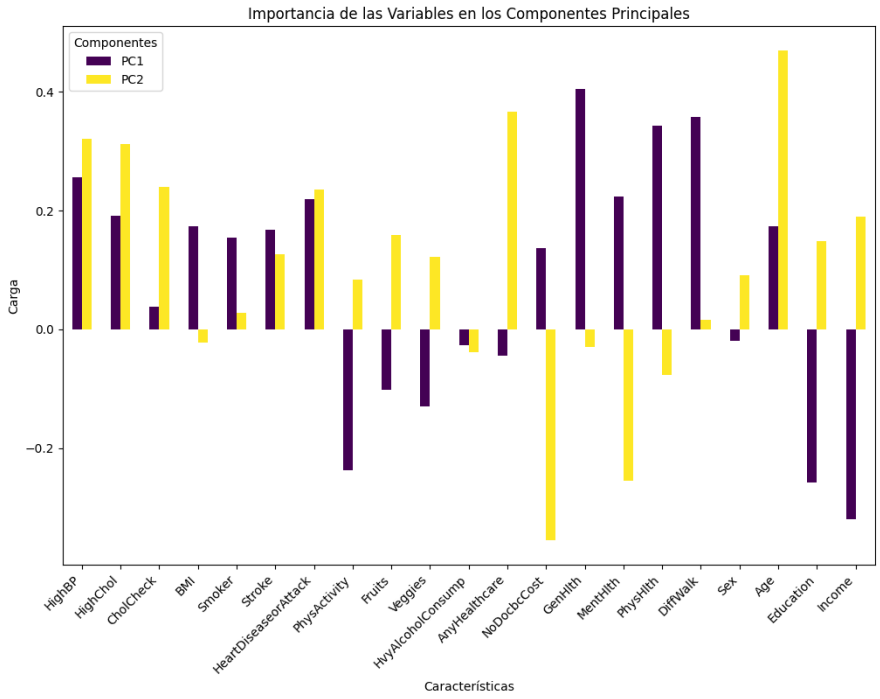

Figura 25: *Importancia*

- **GenHlth (0.405 en PC1)**: Estado general de salud, esta variable tiene una carga alta en el primer componente,asi que la salud general es uno factores importante para comprender la variabilidad de los datos en este componente.
- **PhysHlth (0.344 en PC1)**: Salud física, también, sugiriendo que las condiciones de salud física son importantes para distinguir entre los diferentes grupos.
- **DiffWalk (0.358 en PC1)**: Dificultad para caminar, la dificultad para caminar es importante ya que implica que las personas con problemas de movilidad podrían diferenciarse de otros grupos.
- **Age (0.469 en PC2)**: La edad tiene una carga significativa en el segundo componente, ya que se relaciona con el acceso a la atención médica y la prevalencia de enfermedades.
- **AnyHealthcare (0.367 en PC2)**: Acceso a servicios de salud, esta caracteristica resalta la importancia del acceso a atención médica en la salud general.
- **HighBP (0.321 en PC2)**: La presión arterial alta es relevante, indicando que esta condición está fuertemente relacionada con otros factores de salud.
- **Income (-0.320 en PC1)**:Aunque tiene un signo negativo, la variable de ingreso sigue siendo importante en el primer componente, sugiriendo que los factores socioeconómicos son esenciales para entender la variabilidad en los datos.

Estas características son importantes para diferenciar entre grupos,al tener en valor absoluto mayor magnitud influyen en la formación de clusters.

Para clasificación usamos 3 modelos de clasificación: Random Forest,XGBoost y LGBM. Estos trabajaron con un dataset desbalanceado, mostramos los resultados en la siguiente gráfica.

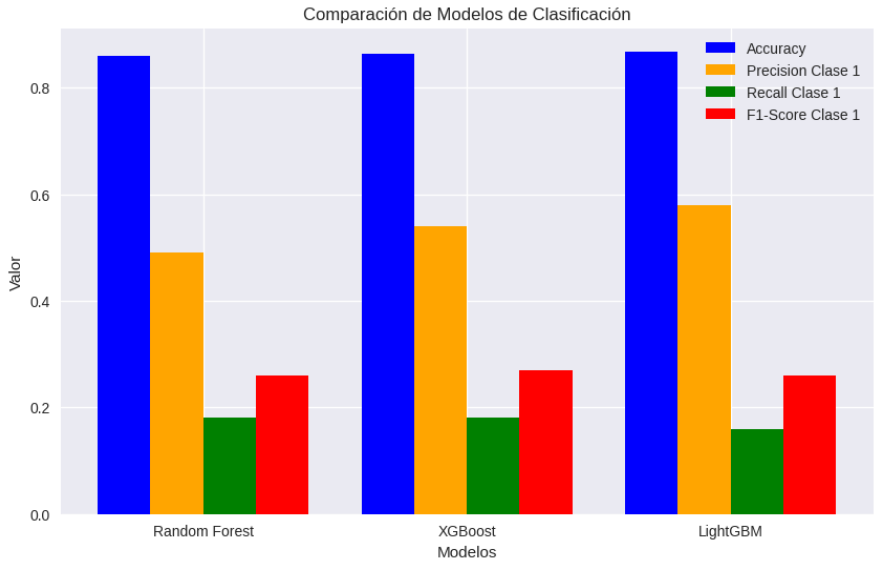

Figura 26: Resultado de clasificación

Con los resultados obtenidos de los tres modelos de clasificación,para predecir si las personas tienen diabetes, podemos notar que alcanzaron una buena precisión, valores entre el 85.94 % y el 86.82 %. Pero aunque es de buena precisión, los modelos fallaron de manera muy notoria para predecir correctamente los casos de diabetes (positiva). Esto lo vemos en el valor del recall y F1-score para la clase 1, clasificando incorrectamente a pacientes con diabetes. Como el dataset es desbalanceado entre las clases, con más personas sin diabetes (clase 0) que con diabetes (clase 1).Los modelos tienden a predecir más fácilmente la clase mayoritaria, la baja tasa de recall para la clase 1 nos dice que estos modelos son mejores para identificar a las personas sin diabetes en comparación con las que si tienen la enfermedad. Veamos las Matrices de confusión en el siguiente gráfico

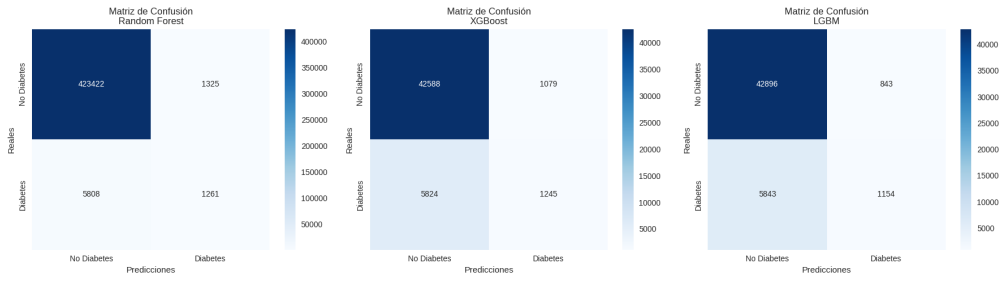

Figura 27: Matrices

Haciendo una comparación de los modelos, LGBM obtuvo el mejor desempeño global con una precisión de 86.82 % pero todos tienen dificultades para identificar bien a los pacientes con diabetes,

aunque los modelos muestran un buen desempeño en términos de precisión global, predecir la clase 1 sigue siendo un desafío importante debido al desbalance.

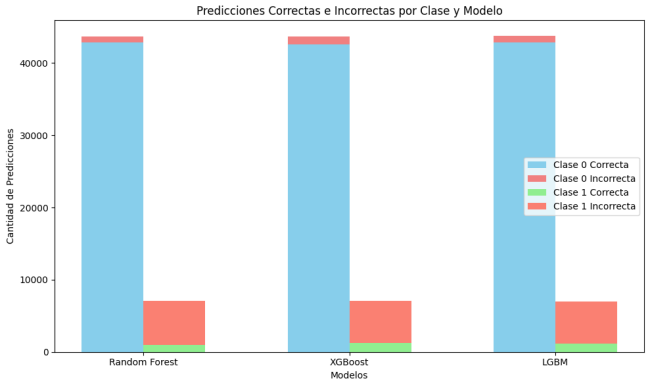

Figura 28: Clases y errores

En Este gráfico podemos notar que los 3 modelos en la clase 0 tiene un excelente desempeño, pero en la clase 1 están teniendo problemas ya que tienen mas errores que el numero de aciertos,los 3 modelos teniendo un desempeño similar.

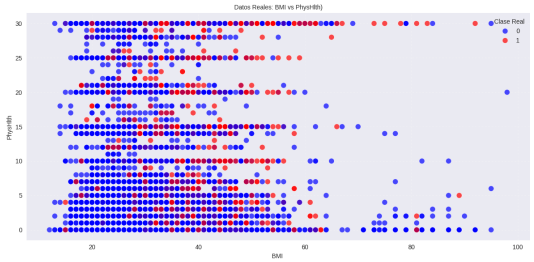

Figura 29: Datos reales

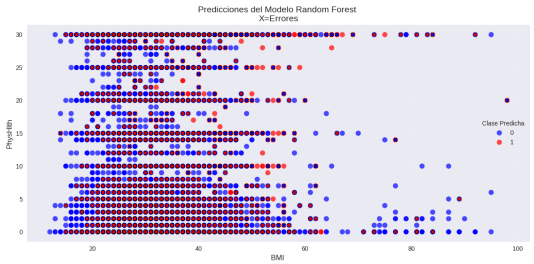

Figura 30: Predicción y errores

Estos gráficos son para Random Forest pero el resultado es similar en cualquiera de los modelos con los que se trabajó este dataset, notando que no se obtuvo una buena clasificación en las personas que tienen Diabetes.

2.3. Hipertensión

Para analizar esta enfermedad utilizamos el dataset llamado "Hipertensión Arterial México Data Set". Este conjunto de datos fue construido a partir de tres fuentes distintas provenientes de la Encuesta Nacional de Salud y Nutrición (ENSANUT) [9]. Incluye información biométrica y medidas asociadas con la salud de pacientes en México. Entre las variables disponibles se encuentran: sexo, edad, concentración de hemoglobina, temperatura ambiente, niveles de ácido úrico, albúmina, colesterol (HDL, LDL, total), creatinina, glucosa, insulina, proteína C reactiva, triglicéridos, glucosa promedio, hemoglobina glucosilada, ferritina, folato, homocisteína, transferrina, vitamina B12, vitamina D, peso, estatura, medida de cintura, entre otras.

La columna final, **riesgo_hipertensión**, indica si el paciente está en riesgo de desarrollar hipertensión arterial. Los valores posibles son:

- 1: El paciente está en riesgo de hipertensión.
- 0: El paciente no está en riesgo.

Este dataset es valioso para identificar factores de riesgo y patrones asociados a la hipertensión arterial en la población mexicana.

2.3.1. Preprocesamiento

El primer paso fue importar las librerías necesarias para la manipulación y visualización de datos (`pandas`, `numpy`, `matplotlib`, `seaborn`), así como herramientas de *machine learning* como escaladores, modelos de clasificación y funciones de evaluación.

Posteriormente, cargamos el dataset original y realizamos un análisis exploratorio de los datos. Este cuenta con 4363 registros con 36 columnas, de las cuales 34 son cuantitativas y 1 categórica. Además en esta etapa se revisaron aspectos como:

- **Datos faltantes o nulos**
- **Filas duplicadas**
- **Valores atípicos**
- **Distribución de los datos**

En este análisis se verificó que nuestro dataset no presenta valores faltantes o nulos, ni filas duplicadas. Para identificar valores atípicos (outliers) se empleó el método del rango intercuartílico (IQR), el cual evidenció la presencia de outliers en casi todas las características. Dado que se trata de datos bioquímicos y clínicos, la eliminación de estos outliers podría implicar la pérdida de información relevante y, en algunos casos, resultar riesgosa. Como podemos ver en la imagen, hay características en donde los outliers son mínimos, pero hay otros, donde son bastantes:

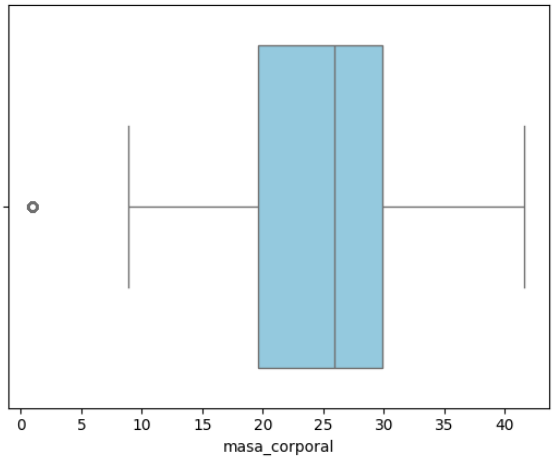

Figura 31: masa corporal

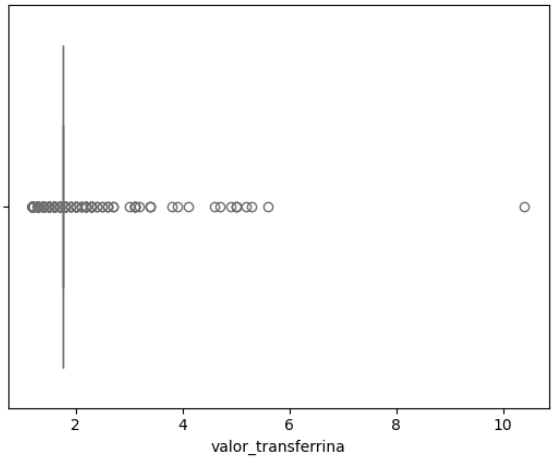

Figura 32: valor transferrina

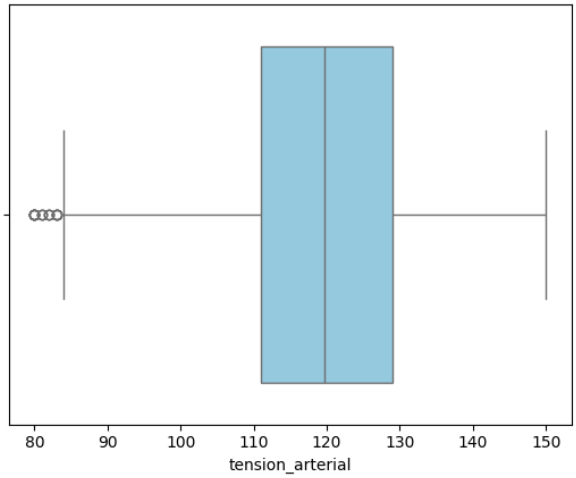

Figura 33: tensión arterial

Ante esta situación, se exploraron diversas técnicas para el tratamiento de los outliers. La estrategia que ofreció mejores resultados fue la imputación de los datos utilizando la media. No obstante, para aplicar esta técnica de forma adecuada y evitar descartar datos que pudieran ser reales, se consultaron diversas fuentes especializadas, entre ellas el National Institutes of Health (NIH). Esto permitió establecer límites adecuados para las distintas características, garantizando de este modo la integridad y validez de la información.

Una vez concluido el tratamiento de outliers, se procedió a evaluar el equilibrio de clases. Los resultados revelaron que las clases están desequilibradas, lo cual podría conducir a que el modelo se incline hacia la clase mayoritaria. Para contrarrestar este problema y reducir la posibilidad de falsos negativos—es decir, evitar clasificar erróneamente a un paciente de ”no riesgo” cuando en realidad sí lo tiene—se implementó la técnica de ajuste de ponderación de clases mediante el parámetro **class**

weight

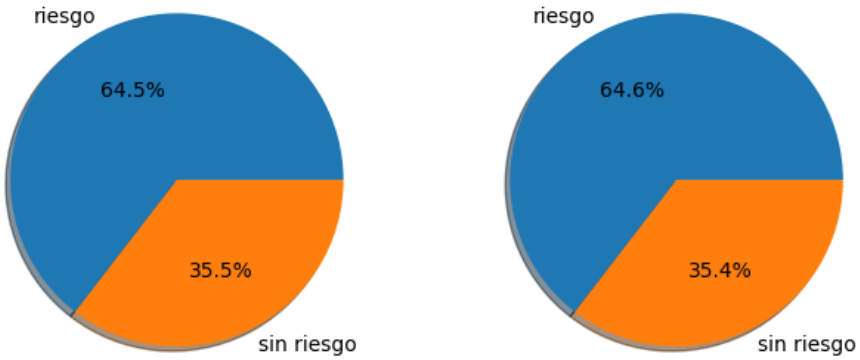

Figura 34: Equilibrio de clases

Esta técnica permite asignar un mayor peso a la clase minoritaria durante el proceso de entrenamiento. Esto significa que el algoritmo penaliza de manera más severa los errores de clasificación relacionados con la clase en menor cantidad, incentivando al modelo a prestar mayor atención a estos casos y mejorar su capacidad de detección. Este enfoque es especialmente valioso en contextos médicos, donde los errores de predicción pueden tener implicaciones significativas para la salud del paciente. Esta configuración garantiza que los datos se preparen de forma adecuada antes de entrenar el modelo, contribuyendo a una mejor detección del riesgo en pacientes y minimizando los falsos negativos.

2.3.2. Modelos de clasificación

Una vez completado el preprocesamiento, que incluyó el tratamiento de outliers y el ajuste de la ponderación de clases, se procede a dividir el conjunto de datos en entrenamiento y prueba. En este proceso, se utiliza una semilla (seed = 123) para garantizar la reproducibilidad y se asigna el 30 % de los datos al conjunto de prueba. Además, se emplea la estratificación para mantener la proporción de clases, lo cual es crucial considerando el desequilibrio presente en los datos. Cabe destacar que la variable objetivo es **riesgo hipertension**, la cual se utiliza para clasificar a las personas en dos grupos: aquellas que presentan un riesgo de desarrollar hipertensión en función de las demás características, y aquellas que no presentan dicho riesgo.

Posteriormente, se implementaron varios modelos de clasificación, entre los que se encuentran:

- RandomForestClassifier
- XGBClassifier
- LGBMClassifier
- StackingClassifier

De todos los modelos evaluados, el **XGBClassifier** fue el que arrojó mejores resultados, destacando por su alta precisión (accuracy: **0.9924**) y una excelente capacidad para detectar casos positivos (recall:**0.9941**). Esto lo convierte en una herramienta confiable para identificar pacientes en riesgo de hipertensión, minimizando especialmente los falsos negativos: La matriz de confusión muestra los siguientes resultados:

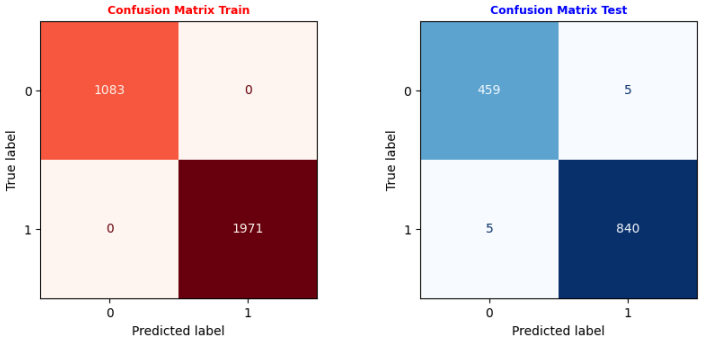

Figura 35: Matriz de Confusión

El recall cercano a 1.0 confirma su eficacia para captar casi todos los casos, lo que es crucial en entornos clínicos donde pasar por alto un diagnóstico puede tener consecuencias graves. Si bien presenta algunos falsos positivos, el equilibrio entre precision y recall asegura un rendimiento general sobresaliente.

A continuación podemos visualizar los resultados de cada modelo utilizado:

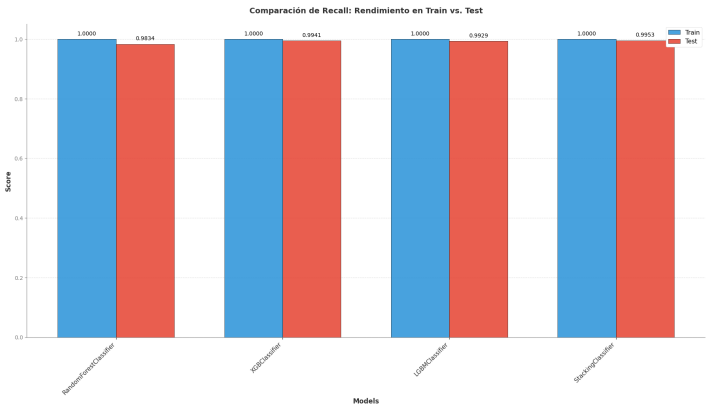

Figura 36: Comparación de rendimiento

Además en la siguiente gráfica, se muestran las características según su relevancia en la predicción del riesgo de hipertensión. Destaca especialmente **masa corporal**, que presenta el valor de importancia más importante, seguida de **actividad total, tensión arterial, valor de colesterol total, valor de hemoglobina y medida de cintura**. Estas variables tienen una influencia significativa en la clasificación, lo que sugiera que la actividad física, el peso corporal, la presión arterial, y el peso, son factores determinantes en la detección temprana del riesgo de hipertensión.

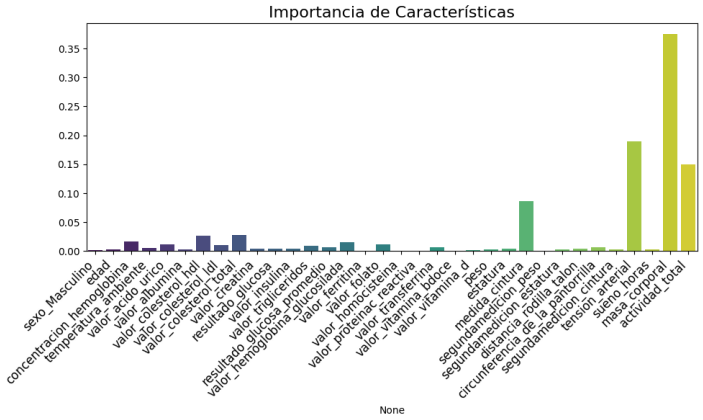

Figura 37: Caracteristicas significativas

Otra forma de evaluar la importancia de las características en nuestro modelo de **XGBoost** es a través de los valores **SHAP (SHapley Additive exPlanations)**. Esta metodología nos permite interpretar el impacto de cada variable en la predicción de hipertensión, analizando cómo los diferentes valores influyen en la salida del modelo.

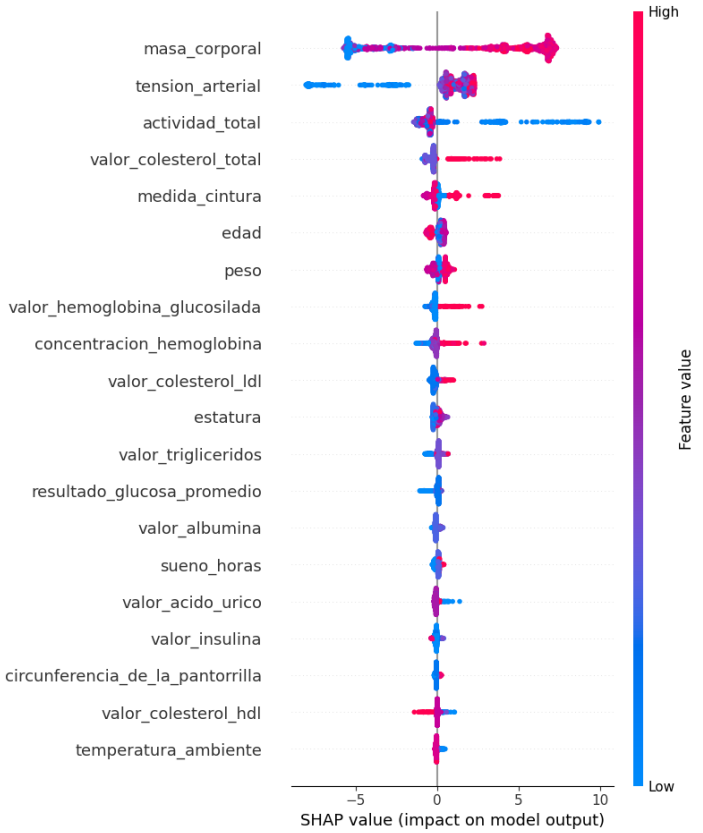

Figura 38: SHAP

De acuerdo con la gráfica SHAP, las características con mayor influencia en la predicción de hipertensión son:

- **Masa Corporal:** Es la característica con mayor impacto en la predicción. Se observa una clara tendencia donde valores bajos de masa corporal (azul) están asociados con una menor probabilidad de hipertensión, reflejada en valores SHAP negativos. En contraste, valores altos de masa corporal (rojo) se agrupan en la región positiva del eje X, indicando un aumento significativo en la probabilidad de hipertensión.
- **Tensión arterial:** Muestra un patrón similar. Valores bajos (azul) tienen poca influencia en la predicción, mientras que valores altos (rojo) están fuertemente correlacionados con un aumento en la probabilidad de hipertensión.
- **Colesterol:** Muestra un comportamiento un poco más diferenciado, en este se puede ver que en el rango positivo, se agrupan valores altos de colesterol (rojo), lo que indica que en ciertos escenarios, niveles elevados de colesterol, pueden aumentar la probabilidad de hiperensión, aunque el impacto es menor que en las anteriores.
- **Actividad total:** Esta tiene una relación inversa con la hipertensión, valores bajos de actividad física (azul) se asocian con una mayor probabilidad de hipertensión, lo que se refleja en valores de SHAP positivos, mientras que valores más altos de actividad tienen menor impacto en la predicción. Esto nos sugiera que la falta de actividad física es un factor de riesgo significativo en el modelo

3. Resultados

3.1. Obesidad

El mejor modelo en la clasificación fue el **XGBClassifier**, que alcanzó un **0.971** de *accuracy* y cuyas variables más importantes son el **género, el peso, la frecuencia de consumo de verduras, la frecuencia de consumo de alcohol y la altura**. El rendimiento promedio de los algoritmos probados se muestra en la siguiente tabla.

Performance promedio de los algoritmos

Modelo	Accuracy Media
SVM	0.9585
XGBClassifier	0.9682
KNN	0.83

Figura 39: Precisión promedio de los algoritmos

Los mejores resultados para la regresión los obtuvo la **red neuronal**, que alcanzó un **MSE de 0.094** y un R^2 **de 0.908**. Según *SHAP*, las variables más importantes del modelo fueron la **altura, el historial familiar, la edad, la frecuencia de consumo de verduras y el género**.

Las métricas promedio de los algoritmos y la red (se corrió varias veces y se promediaron sus métricas) se muestran en la siguiente tabla.

Performance promedio de los algoritmos

Modelo	r^2 media	MSE media
Linear Regression	0.551	0.448
SVR: Polinomial	0.725	0.272
SVR: RBF	0.87	0.128
Red neuronal	0.89	0.104

Figura 40: MSE medio y R^2 por modelo

En cuanto a la **clusterización**, obtuvimos **4 grupos**, los cuales nos ayudaron a comprender mejor la relación entre las variables **CAEC** y **Weight**.

3.2. Diabetes

Tras evaluar los 3 modelos de clasificación para el problema de la predicción de diabetes observamos los siguientes puntos claves: El accuracy es alto, pero para problemas de clases desbalanceadas como

en este caso , un modelo con alta precisión no necesariamente es el mejor, en este caso LGBMClassifier obtuvo un mejor accuracy seguido por XGBClassifier y por ultimo RandomnForest. Los 3 modelos mostraron un desempeño limitado, ya que el recall de la clase 1 fue baja en todos los casos, teniendo problemas para predecir correctamente a los pacientes con diabetes, lo que es un problema importante en el área medica.
La baja capacidad de predicción en la clase 1 resalta la necesidad de considerar técnicas adicionales para mejorar el desempeño.

Modelo	Accuracy	F1-Score
RandomForest Classifier	0.8600	0.26
XGBClassifier	0.8639	0.27
LGBM Classifier	0.8682	0.26

Figura 41: Precisión promedio de los algoritmos

Para el caso de **clusterización**, nuestro análisis muestra que la Diabetes y sus factores de riesgo están fuertemente influenciados por diversos aspectos como el estilo de vida, acceso a atención medica, indice de masa corporal, presión arterial y los hábitos de ejercicio,y de alimentación, los 3 grupos identificados reflejan diferentes niveles de riesgo y salud, siendo el Grupo 1 (Alto Riesgo de Diabetes y Enfermedades Cardiovasculares) con una alta prevalencia de diabetes y enfermedades cardiovasculares, en cambio el grupo 2 (Estilo de Vida Saludable y Bajo Riesgo) se caracteriza por un estilo de vida saludable, con bajo riesgo de diabetes y mejores indicadores de salud. Esto es importante ya que puede ayudarnos a prevenir ciertas enfermedades.

3.3. Hipertensión

El modelo que obtuvo los mejores resultados en la clasificación fue **XGBClassifier**, alcanzando una precisión (accuracy) de **0.9924**.

Modelo	Accuracy Media
RandomForest Classifier	0.9878
XGBClassifier	0.9924
LGBM Classifier	0.9829
Stacking Classifier	0.9924

Figura 42: Precisión promedio de los algoritmos

Las variables más influyentes en la predicción fueron masa corporal, tensión arterial, colesterol y peso (así como también medidas de cintura).

Además, gracias a la gráfica de **SHAP**, podemos observar que valores bajos de actividad física (representados en azul) incrementan la probabilidad de hipertensión. Es decir, las personas con menor actividad física están más asociadas con el diagnóstico de hipertensión según nuestro modelo. Estos resultados refuerzan la validez del modelo, ya que los factores identificados como más relevantes coinciden con aquellos comúnmente asociados a enfermedades cardiometabólicas en la literatura médica. La relación entre el peso corporal, la presión arterial y los niveles de colesterol con el desarrollo de hipertensión es ampliamente reconocida, lo que sugiere que nuestro modelo está capturando correctamente los patrones esperados en este tipo de afecciones.

4. Discusión y Conclusiones

A modo de conclusión, nuestro estudio evidencia el gran potencial de las técnicas de machine learning e inteligencia artificial para identificar y validar factores de riesgo clave en el desarrollo de enfermedades cardiometabólicas, permitiendo así un abordaje más preciso, personalizado y orientado a la prevención.

En el caso de la obesidad, los modelos aplicados han mostrado que, además de factores tradicionales como el exceso de peso, género y el historial familiar –elementos que reflejan la heredabilidad de los hábitos alimenticios–, se deben considerar también otros aspectos como el tipo de alimentación, el consumo frecuente de comidas calóricas y bebidas alcohólicas, así como la frecuencia en el consumo de verduras. Estos hallazgos coinciden con la literatura médica, donde se observa que las personas con hábitos sedentarios y una dieta rica en grasas saturadas tienen una mayor probabilidad de desarrollar obesidad.

Respecto a la hipertensión, se encontró que la masa corporal y la tensión arterial son variables críticas, mientras que la actividad física regular juega un rol protector. La inclusión de métricas adicionales como la medida de la cintura y los niveles de colesterol en nuestros modelos mejora la capacidad para la detección temprana, especialmente en contextos donde una parte significativa de la población (como se estima en México) desconoce su condición hipertensiva.

En el análisis de la diabetes, los resultados confirman que los factores de riesgo están estrechamente ligados al estilo de vida, el acceso a la atención médica y los hábitos de ejercicio y alimentación. La identificación de diferentes grupos de riesgo –por ejemplo, un grupo con alto riesgo de diabetes y enfermedades cardiovasculares versus otro de estilo de vida saludable y bajo riesgo– ofrece un marco de referencia útil para la implementación de intervenciones preventivas y estrategias de salud pública.

Cabe resaltar que la alta precisión obtenida en los modelos utilizados respalda su utilidad como herramientas en la detección y prevención de estas enfermedades. Sin embargo, se identifican oportunidades de mejora, tales como la necesidad de ampliar y enriquecer la base de datos con información adicional y de mayor fiabilidad, lo que permitiría un análisis más profundo y la exploración de relaciones complejas entre obesidad, diabetes e hipertensión.

En síntesis, la integración de modelos de machine learning en el análisis de factores de riesgo cardiometabólicos no solo permite identificar con precisión las variables involucradas, sino también validar su coherencia con el conocimiento médico actual. Estos hallazgos pueden orientar intervenciones personalizadas y estrategias de salud pública que prioricen cambios en el estilo de vida y mejoras en la atención médica, contribuyendo de manera significativa a la reducción de la carga de estas enfermedades en la población.

Referencias

[1] Centro de Atención Integral del Paciente con Diabetes CAIPaDi. *Resultados de la Encuesta Nacional de Salud y Nutrición 2022*. Accedido: 2025-03-26. 2022. URL: `https://www.incmnsz.mx/opencms/contenido/departamentos/CAIPaDi/boletines/BoletinJULIO2023.html?utm_source=chatgpt.com`.

[2] M. C. Escamilla-Núñez et al. «Detección, diagnóstico previo y tratamiento de enfermedades crónicas no transmisibles en adultos mexicanos. Ensanut 2022». En: *Salud Pública de México* 65.supl 1 (2023), S153-S162.

[3] Medscape en Español. *La diabetes en México: cifras preocupantes y retos por superar*. Accedido: 2025-03-26. 2022. URL: https://espanol.medscape.com/verarticulo/5911153.

[4] Faria Ferdowsy et al. «A machine learning approach for obesity risk prediction». En: *Current Research in Behavioral Sciences* 2 (2021), pág. 100053. ISSN: 2666-5182. DOI: https://doi.org/10.1016/j.crbeha.2021.100053. URL: https://www.sciencedirect.com/science/article/pii/S2666518221000401.

[5] S. M. S. Islam et al. «Machine Learning Approaches for Predicting Hypertension and Its Associated Factors Using Population-Level Data From Three South Asian Countries». En: *Frontiers in Cardiovascular Medicine* 9 (2022), pág. 839379. DOI: 10.3389/fcvm.2022.839379.

[6] El País. *México se enfrenta a un futuro de obesidad y diabetes: el 56 % de la niñez sufrirá sobrepeso en 2035*. Accedido: 2025-03-26. 2025. URL: https://elpais.com/mexico/2025-03-20/mexico-se-enfrenta-a-un-futuro-de-obesidad-y-diabetes-el-56-de-la-ninez-sufrira-sobrepeso-en-2035.html.

[7] Fabio Mendoza Palechor y Alexis De la Hoz Manotas. «Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico». En: *Data in Brief* 25 (2019). URL: https://api.semanticscholar.org/CorpusID:201195793.

[8] Instituto Nacional de Salud Pública. *29.9 % de las personas adultas en México vive con hipertensión*. Accedido: 2025-03-26. 2022. URL: https://www.insp.mx/avisos/299-de-las-personas-adultas-en-mexico-vive-con-hipertension.

[9] Instituto Nacional de Salud Pública. *La salud de los mexicanos en cifras: resultados de la ENSANUT 2022*. Accedido: 2025-03-26. 2022. URL: https://www.insp.mx/informacion-relevante/la-salud-de-los-mexicanos-en-cifras-resultados-de-la-ensanut-2022.

[10] S. A. Thamrin et al. «Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research 2018». En: *Frontiers in Nutrition* 8 (2021), pág. 669155. DOI: 10.3389/fnut.2021.669155.

[11] Álvaro Torres-Martos et al. «Multiomics and eXplainable artificial intelligence for decision support in insulin resistance early diagnosis: A pediatric population-based longitudinal study». En: *Artificial Intelligence in Medicine* 156 (2024), pág. 102962. ISSN: 0933-3657. DOI: https://doi.org/10.1016/j.artmed.2024.102962. URL: https://www.sciencedirect.com/science/article/pii/S0933365724002045.