# 

# Memoria de Capstone Project

# Análisis del Impacto Nutricional de Alimentos Ultraprocesados con Machine Learning

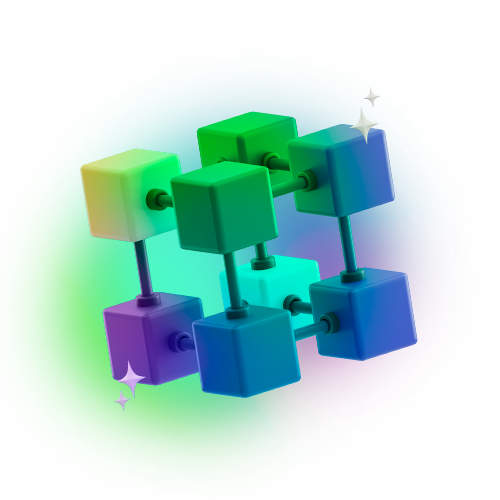
BootCamp Data Analytics OCT24

Esteban Miller

Sergio Martín

Lidia Martin Sanchez

Sergio Caballero Garcia



## ÍNDICE

1. **INTRODUCCIÓN**
2. **ESTADO DEL ARTE**
3. **DESARROLLO**
4. **PRUEBAS Y RESULTADOS**
5. **VISUALIZACIONES**
6. **CONCLUSIONES Y TRABAJOS FUTUROS**

## 

1. INTRODUCCIÓN

**Motivación**

**La obesidad como problema de salud pública**

La obesidad se ha convertido en uno de los problemas de salud pública más graves del siglo XXI. Según la Organización Mundial de la Salud, la prevalencia mundial de la obesidad se ha triplicado desde 1975, afectando a más de 650 millones de adultos globalmente. Este problema de salud está asociado con múltiples comorbilidades como diabetes tipo 2, enfermedades cardiovasculares y algunos tipos de cáncer.

Una clasificación precisa de los diferentes niveles de obesidad puede ayudar a los profesionales de la salud a implementar intervenciones personalizadas y preventivas. El machine learning ofrece herramientas potentes para analizar patrones complejos en datos médicos y antropométricos, potencialmente mejorando la precisión de los diagnósticos y facilitando la identificación temprana de pacientes en riesgo.

En este proyecto vamos a aplicar múltiples modelos de clasificación con el objetivo de predecir el nivel de obesidad en función de hábitos y características personales. Se han elegido modelos de diferentes familias de algoritmos para asegurar una comparación robusta y comprender mejor qué tipo de modelo se ajusta mejor a los datos.

**Objetivos**

**> OBJETIVO 1:**

Desarrollar un sistema de clasificación capaz de predecir con alta precisión los diferentes niveles de obesidad basado en datos antropométricos y hábitos de vida.

**> OBJETIVO 2:**

Identificar los factores más relevantes que contribuyen a los diferentes niveles de obesidad, proporcionando insights valiosos para los profesionales de la salud.

**> OBJETIVO 3:**

Comparar y evaluar el rendimiento de nuestros modelos optimizado con los métodos tradicionales de clasificación de obesidad

**Organización del proyecto**

1. **Exploración y Preprocesamiento de Datos:** Análisis inicial del conjunto de datos, tratamiento de valores faltantes, normalización y codificación de variables.
2. Análisis descriptivo y exploratorio: **EDA**
3. **Modelado:** Implementación y optimización de diferentes algoritmos para la clasificación multiclase.
4. **Resultados y validación:** Análisis exhaustivo del rendimiento del modelo mediante diversas métricas y técnicas de validación cruzada.
5. **Interpretación y visualización:** Extracción de insights y visualización de la importancia de las características en el modelo final.

**Contratiempos y cambio de rumbo**

1. **Contratiempos y solución** en la búsqueda de nuestro dataset dada la ausencia de un recurso completo: Ningún repositorio público combinaba de forma detallada hábitos alimentarios y datos clínicos (obesidad, diabetes, enfermedades crónicas) en un mismo archivo.
2. Muchos conjuntos tenían menos de 400 registros o provenían de estudios locales con sesgos por edad, región o nivel socioeconómico, además de presentar valores faltantes e inconsistencias en unidades. Sumado a la incompatibilidad de formatos y licencias: Los pocos datasets disponibles venían en CSV, JSON, XLSX o PDF, con convenciones de nombres y codificaciones dispares, y cada fuente aplicaba licencias distintas (CC0, CC BY, datos abiertos gubernamentales), lo que dificultaba su fusión. -Muestras pequeñas y sesgadas:
3. Exploración de crear un dataset propio: Se recopilaron datos reales de diversas fuentes oficiales (ENSE, INE, ALADINO, EUROSTAT, ENRICA) y se empleó IA para homogeneizar formatos, limpiar registros y documentar orígenes y licencias. Sin embargo, implicaba un gran esfuerzo de limpieza manual y recopilación de datos reales, por lo que decidimos variar ligeramente el enfoque del proyecto.
4. **Decisión final:** Se optó por un variar el enfoque del proyecto y centrarse en los hábitos de las personas a la hora de comer alimentos calóricos, bebidas azucaradas, frecuencia de consumo de verduras, calorías entre comidas, frecuencia de la actividad física etc. y como todos esto factores influye en su nivel de obesidad. A través del repositorio de Kaggle con datos ya estructurados y unas variables acordes a nuestro objetivo.

2. ESTADO DEL ARTE:

El estado actual de la investigación sobre la obesidad destaca que esta condición es una enfermedad crónica y multifactorial, influenciada por una compleja interacción de factores genéticos, biológicos, socioeconómicos, culturales y ambientales. Este enfoque integral ha sido respaldado por diversos estudios recientes, que subrayan la necesidad de estrategias multidisciplinarias para su prevención y tratamiento.

**Fuente: El País**

Los factores más influyentes en los análisis anteriores son los siguientes:

* **Factores Genéticos y Biológicos**

La predisposición genética a la obesidad es un factor importante que determina parte de las causas del aumento de peso, aunque está muy influenciado por muchos factores externos

* **Estilo de vida**

El estilo de vida moderno, el alto consumo de alimentos ultraprocesados, la variación en los horarios de sueño, son factores que aumentan considerablemente el riesgo de obesidad. Estudios como el ANIBES en España han demostrado que una mayor actividad física y un nivel educativo más alto se asocian con un menor riesgo de obesidad general y abdominal. Asimismo, el tiempo excesivo frente a pantallas y la falta de sueño de calidad se han vinculado con un mayor riesgo de obesidad.

* **Factores socioeconómicos y ambientales**

La obesidad afecta en un porcentaje mucho más alto a las personas con un nivel socioeconómico más bajo, personas que tienen manos acceso a alimentos frescos, saludables y con una posibilidad mucho más reducida de realizar actividad física. Como indica el estudio ALADINO de 2023, el entorno físico influye drásticamente, como el acceso a restaurantes de comida rápida, espacios verdes donde realizar actividad física etc …

* **Salud Mental**

Finalmente el estigma social asociado a la obesidad en muchas ocasiones causa mayores dificultades a la hora de realizar una perdida de peso, por lo que los estudios recomiendan abarcar estos problemas con empatía y comprensión, para evitar que puedan derivar en enfermedades asociadas con la comida como la anorexia o la bulimia.

>>ESTUDIOS PREVIOS

Existen estudios recientes que analizan la obesidad en relación con diversos factores del estilo de vida en España. Uno de los más destacados es el Estudio ALADINO 2023, llevado a cabo por la Agencia Española de Seguridad Alimentaria y Nutrición (AESAN). Este estudio se centra en la alimentación, actividad física, desarrollo infantil y obesidad en niños de 6 a 9 años.

Las principales conclusiones de estos estudios son:

* **Prevalencia de exceso de peso:** Más de un tercio de los escolares españoles entre 6 y 9 años presenta exceso de peso.
* **Diferencias socioeconómicas entre las familias encuestadas:** La obesidad es más prevalente en niños de familias con ingresos más bajos, evidenciando una asociación inversa entre el nivel socioeconómico y el exceso de peso.
* **Hábitos alimentarios y actividad física:** Los niños de familias con menores ingresos tienden a consumir menos frutas y verduras y a realizar menos actividad física, factores que contribuyen al aumento de la obesidad.

Otros estudios realizados por el Instituto Salud Carlos III (ISCIII) y AESAN, basado en datos de la encuesta ENE-COVID, revela que:

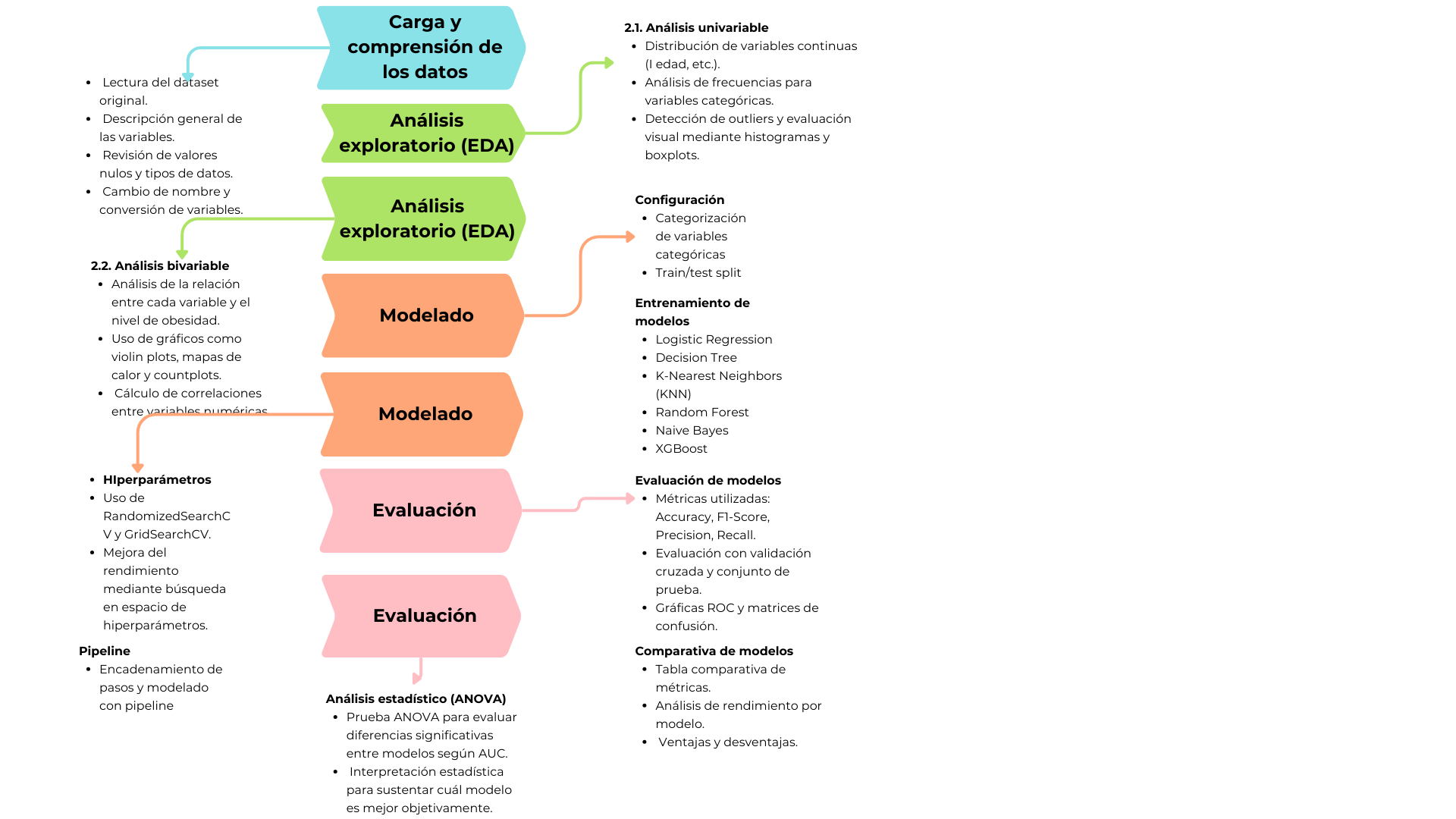
* **Una conclusión general:** El 55,8% de los españoles presentan exceso de peso, siendo un 18,7% que tiene obesidad.
* **Otros factores a tener en cuenta:** En familias con menor poder adquisitivo, niveles educativos más bajos, es más común que existan niveles de obesidad más elevados.

>>APORTACIÓN AL ESTADO DEL ARTE

En la parte de EDA nuestro proyecto **aporta conclusiones sólidas** que confirman las aportadas por los estudios anteriores, en cuanto a la importancia de los hábitos de alimentación, actividad física, consumo de bebidas azucaradas y demás factores antes mencionados.

Además de **aportar un modelo de ML capaz de predecir el nivel de obesidad**, si es que existe en una persona basándonos en sus hábitos y estilo de vida, aportando una herramienta más para poder detectar con antelación, cuáles son los factores que afectan en mayor medida al aumento de peso de los sujetos.

3. DESARROLLO

****

3.1 Carga y Exploración del Dataset

>>CARACTERÍSTICAS DEL DATASET

El conjunto de datos utilizado contiene **2111 registros** con **17 variables** que incluyen;

* **Datos demográficos:** Género, Edad
* **Medidas antropométricas:** Altura, Peso
* **Historia familiar:** Antecedentes familiares de sobrepeso
* **Hábitos alimenticios:** Consumo de alimentos hipercalóricos (FAVC), frecuencia de consumo de verduras (FCVC), número de comidas principales (NCP), consumo entre comidas (CAEC)
* **Estilo de vida:** Consumo de tabaco (SMOKE), consumo de agua (CH20), monitoreo de calorías (SCC), frecuencia de actividad física (FAF), tiempo de uso de dispositivos tecnológicos (TUE).
* **Movilidad:** Medio de transporte utilizado (MTRANS)
* **Variable objetivo:** Nivel de obesidad (NObeyesdad) con 7 categorías desde peso insuficiente hasta obesidad tipo III

>>ANÁLISIS INICIAL DE LOS DATOS

El conjunto de datos seleccionado no necesita mucha limpieza de datos, verificamos los siguientes puntos:

* Convertimos todo a minúsculas
* Quitamos todos los caracteres diacríticos
* Renombramos algunas variables
* Comprobación de datos nulos: No había ningún dato nulo

Esta fase nos sirve para estandarizar los datos que vamos a utilizar en el proyecto y poder trabajar de una manera más sencilla y fácil con los datos.

3.2 Análisis descriptivo y exploratorio (EDA)

>>ANÁLISIS UNIVARIABLE

Se analizaron individualmente todas las variables para comprender su distribución y características generales:

* **Variables continuas:**
  + Usando histogramas y boxplots para explorar la distribución, simetría y valores atípicos.
  + Por ejemplo, se observan valores extremos en la edad.
* **Variables categóricas:**
  + Uso de gráficos de barras para estudiar la frecuencia de categorías como género, tipo de transporte o frecuencia de comida rápida.
  + Se identificaron desequilibrios en algunas categorías, lo que ayudó a tomar decisiones de categorización.
  + La variable objetivo (nivel de obesidad) mostró clases más frecuentes como “Obesity Type I” y “Overweight Level II”, lo cual planteó un reto de balanceo para el modelado.
* **Análisis de valores atípicos:** 
  + Identificamos algunos valores extremos en peso (> 150kg) y altura (< 1.50 m y 1.95 m)
  + Verificamos que estos valores extremos son biológicamente plausibles y corresponden a casos reales. Decidimos mantenerlos en el análisis ya que representan la variabilidad natural de la población.

>>ANÁLISIS BIVARIABLE

Examinamos las relaciones entre variables para identificar patrones y correlaciones significativas:

* Para **variables categóricas** vs. objetivo:
  + Se utilizaron gráficos de barras agrupadas para observar cómo varía el nivel de obesidad según hábitos de vida (por ejemplo, consumo de verduras o tipo de transporte).
* Para **variables continuas** vs. objetivo:
  + A través de los boxplots para comparar distribuciones de edad, tiempo de uso de dispositivos tecnológicos entre los distintos niveles de obesidad.
* También calculamos una **matriz de correlación** para variables continuas:
  + Correlación positiva moderada entre el peso y la altura (0,46)
  + Correlación positiva débil entre actividad física (FAF) y altura (0,29)
  + Correlación negativa débil entre edad y tue (-0,30)

Estos análisis sirven para detectar variables relevantes para la predicción.

3.3 MODELADO

>>ALGORITMOS UTILIZADOS

Se entrenan varios algoritmos de clasificación con distintos enfoques: Modelos lineales (Regresión logística), no lineales (KNN, Árboles de decisión) y ensambles (Random Forest, XGBoost) para evaluar cómo distintos enfoques capturan las relaciones entre variables y asegurar la robustez en los resultados mediante técnicas complementarias.

* **Modelos lineales:** Regresión logística, ampliamente utilizado para problemas de clasificación. Su simplicidad lo convierte en un buen modelo base, permitiendo identificar relaciones lineales entre variables independientes y la clase objetivo.
* **Modelos basados en vecinos:** K-Nearest Neighbors. Modelo no paramétrico basado en la similitud entre observaciones. Asigna una clase a una muestra en función de la mayoría de sus vecinos más cercanos.
* **Árboles de decisión:** Decision Tree. Modelo interpretable que segmenta el espacio de decisiones mediante reglas basadas en variables predictoras. Permite comprender fácilmente qué características son más relevantes.
* **Random forest:** Modelo de conjunto de árboles de decisión entrenados de manera aleatoria. Mejora la capacidad de generalización respecto a un solo árbol, reduce el sobreajuste y captura interacciones complejas entre variables.
* **XGBoost:** Modelo basado en boosting que combina múltiples árboles de decisión de forma secuencial, corrigiendo errores del anterior. Altamente preciso y eficiente, especialmente en datasets estructurados.

>>CONFIGURACIÓN DEL ENTRENAMIENTO

Para todos los modelos:

* Se seleccionaron variables continuas y la variables objetivo a utilizar
* Se utilizó una división 80% train y 20% test
* Se ajustaron hiper parámetros mediante GridSearchCV o RandomizedSearch cuando fue posible.

>CONFIGURACIÓN DE LAS MÉTRICAS

Aplicamos funciones específicas para evaluar el rendimiento de los modelos entrenamos a través de las siguientes métricas:

* + - **Accuracy** (precisión general, representa la proporción de predicciones correctas respecto al total).
    - **F1 Scores** (La usaremos como métrica principal ya que equilibra la precisión y la sensibilidad y es especialmente útil en problemas multiclase donde no todas las clases tiene la misma importancia)
    - **Precisión** (Indica cuántas de las predicciones positivas fueron realmente positivas).
    - **Recall** (sensibilidad, refleja cuántos de los casos positivos fueron correctamente identificados).
    - **Matrices de confusión**

Estas métricas nos permiten evaluar los resultados de una manera completa en todos los modelos utilizados.

>>REGRESIÓN LOGÍSTICA SIMPLE Y MEJORADA

Comenzamos con un modelo base de regresión logística utilizando una configuración estándar:

****

Este modelo proporciona **una línea base de rendimiento** para comparar con implementaciones más complejas.

Implementamos una versión mejorada de r**egresión logísticas con parámetros optimizados para problemas multiclase**:

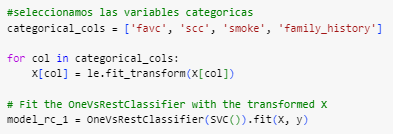
****

Las mejoras incluyen:

* **Configuración multi\_class=’multinomial’** para manejar mejor las múltiples categorías de obesidad
* **Solver newton-cg** que funciona mejor para problemas multiclase
* **Mayor número de iteraciones** para garantizar la convergencia.

>>ONE VS REST

Para abordar la naturaleza multiclase de nuestro problema, implementamos un clasificador One-vs-Rest utilizando SVC:

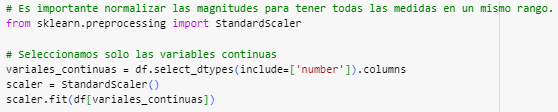
****

En este caso:

* **Transformamos variables categóricas** a continuas mediante LabelEncoder
* Se utiliza la estrategia **OneVsRest** para manejar **múltiples clases**

>>MODELO KNN (K-Nearest Neighbors)

Implementamos un modelo de KNN con normalización de variables para tener todas las medidas en un mismo rango:

****

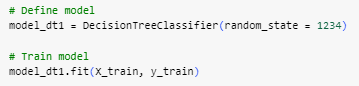
****

En este caso:

* **Normalización de variables** mediante StandardScaler para mejorar el rendimiento de KNN
* Configuración base de **5 vecinos** para la clasificación

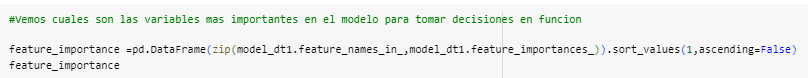
>>MODELO DE ÁRBOL DE DECISIÓN

Implementamos un modelo de árbol de decisión para explorar su rendimiento y analizar las características importantes (feature importance):

****

Con este modelo:

* Permitimos que el árbol **crezca al máximo** para evaluar su rendimiento sin restricciones
* Realizamos un **análisis de características importantes** para identificar variables clave
* Evaluamos la **profundidad del árbol y el número de hojas**

****

**| Índice | Variable | Valor |**

**|--------|----------|-----------|**

**| 0 | age | 0.054852 |**

**| 1 | height | 0.275331 |**

**| 2 | weight | 0.470960 |**

**| 4 | fcvc | 0.142853 |**

**| 5 | ncp | 0.024265 |**

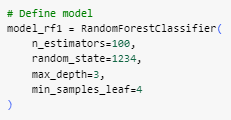
**| 8 | ch2o | 0.006701 |**

**| 10 | faf | 0.009398 |**

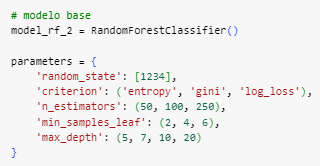
**| 11 | tue | 0.015640 |**

>> RANDOM FOREST CON OPTIMIZACIÓN DE HIPER PARÁMETROS

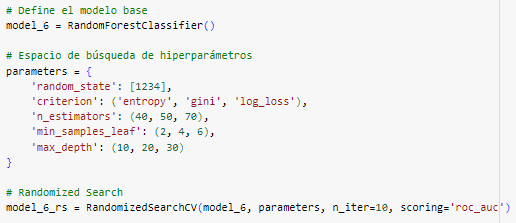
Implementamos varias variantes del modelo Random Forest con diferentes técnicas de optimización:

****

**Random Forest** con la técnica **GridSearchCV**

****

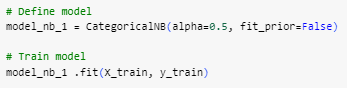
**Random Forest** con la técnica **RandomizedSearchCV**

****

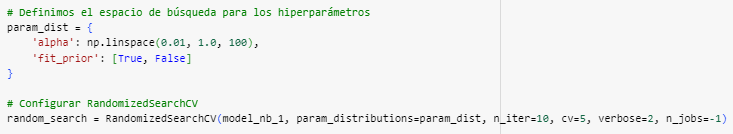
Para cada modelo, realizamos un análisis de las **características importantes: usamos Feature importances**

>>MODELO NAIVE BAYES

Implementamos un clasificador **Naive Bayes** para problemas con variables categóricas:

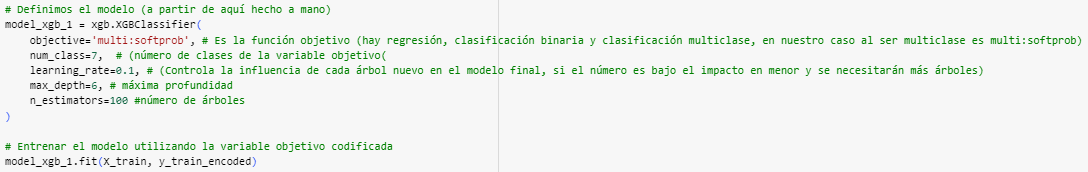
****

Y optimizamos sus hiper parámetros con **RandomizedSearchCV:**

****

>>MODELO XGBOOST

Implementos el modelo XGBoost, un algoritmos de gradient boosting de alto rendimiento para problemas de clasificación multiclase:

****

Con este modelo:

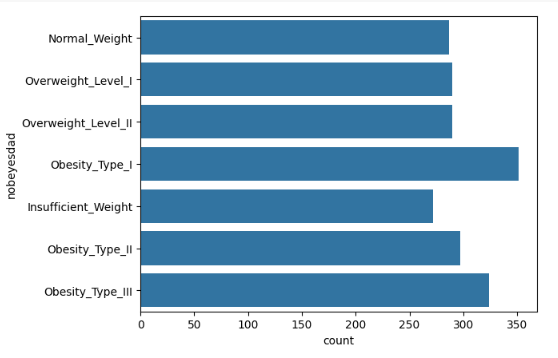
* Realizamos una **codificación específica de variables categóricas** para XGBoost
* Configuramos **parámetros clave como la función objetivo** para clasificación **multiclase**
* Aplicamos una **configuración balanceada entre profundidad y número de árboles**

4. PRUEBAS Y RESULTADOS

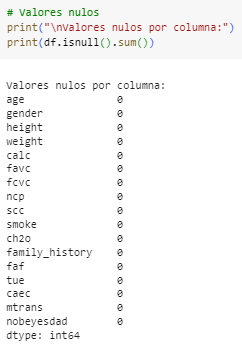
Antes de aplicar algoritmos de Machine Learning, se realizó un análisis exhaustivo de la calidad del dato y un estudio exploratorio de las variables incluidas en el dataset, lo que permitió detectar anomalías, validar consistencia y comprender relaciones y patrones clave entre variables.

4.1 CALIDAD Y PREPROCESAMIENTO DE DATOS

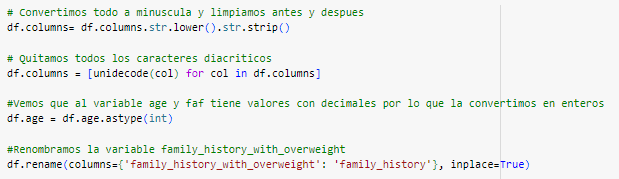
En el dataset utilizado en el proyecto se describen 17 variables relacionadas con la salud y estilo de vida. Estas variables incluyen edad, peso, altura, frecuencia de consumo de verduras, antecedentes familiares, entre otras, y el objetivo final es clasificar a cada persona según su nivel de obesidad (NObeyesdad). Se identificó que el dataset era sencillo, bien estructurado y aborda múltiples dimensiones relevantes para un problema de salud como la obesidad. La variable objetivo tiene varias clases, lo que hace del problema una clasificación multiclase.



En cuanto al procesamiento de datos se han realizado diferentes transformaciones aplicadas para un mejor manejo de los datos asegurando consistencia en los nombres de las variables y tipos de datos. El no tener datos nulos simplifica el tratamiento del dataset.



* Limpieza y estandarización de nombres de columnas.
* Conversión de variables numéricas a tipos adecuados.
* Eliminación de caracteres especiales (como acentos) para facilitar la manipulación.



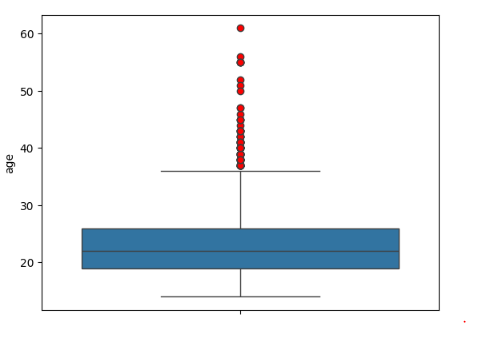
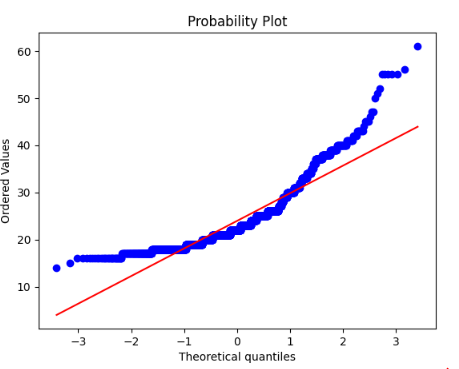
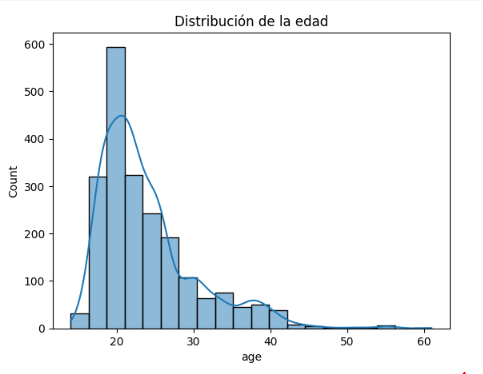
4.2 ANÁLISIS DESCRIPTIVO Y EXPLORATORIO (EDA)

Aquí se analiza la forma y distribución del dataset, se realizan análisis univariables y bivariables para comprobar las relaciones entre variables y algunas pruebas estadísticas.

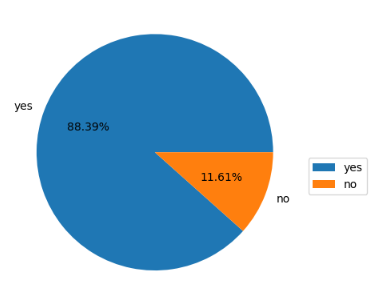
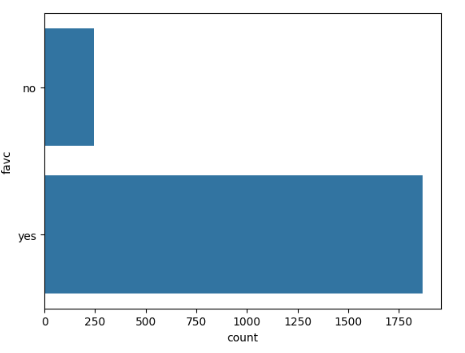
>>Análisis univariable

* **Test de normalidad:** Se aplicó la prueba de **Shapiro-Wilk** para evaluar la distribución de las variables continuas. Al tratarse de datos de salud, **ninguna de las variables sigue una distribución normal.**
* **Comprobación de outliers:** Se aplicó el método **zscore** y **iqr** para detectar valores atípicos. Aunque **se detectaron outliers en variables como la edad, la altura y el peso**, estos se mantuvieron por tratarse de datos reales representativos de condiciones clínicas atípicas pero válidas.
* **Visualizaciones:** Se usaron histogramas, boxplots, gráficos de barras, gráficos de tarta

📌 Ejemplo de visualizaciones de la variable ‘age’



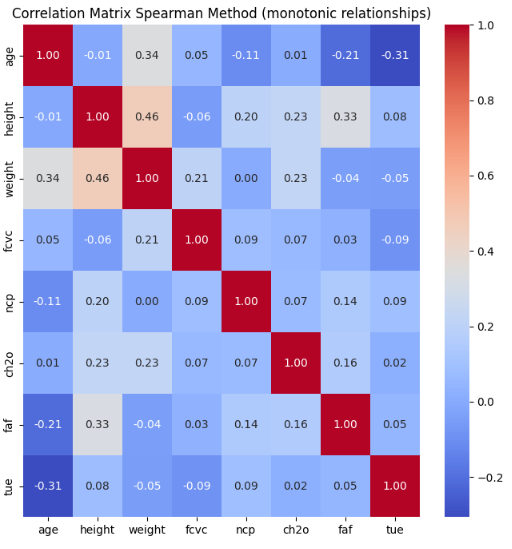
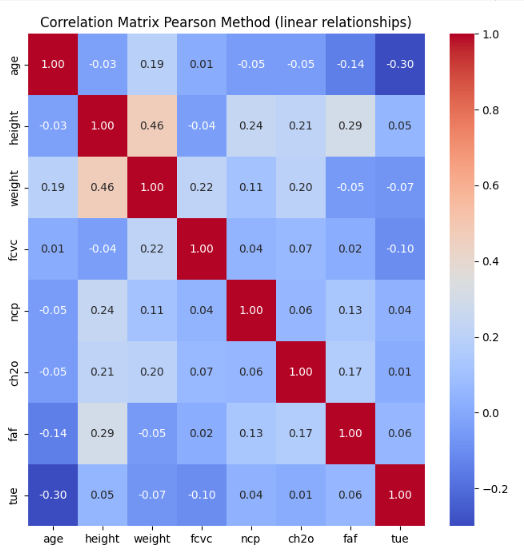
📌 Ejemplo de visualizaciones de la variable ‘favc’ (Consumo frecuente de alimentos ricos en calorías)



>>Análisis bivariable

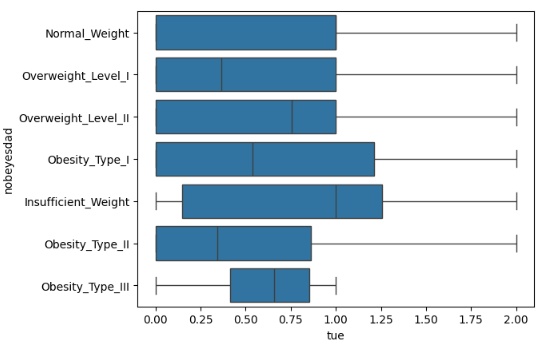
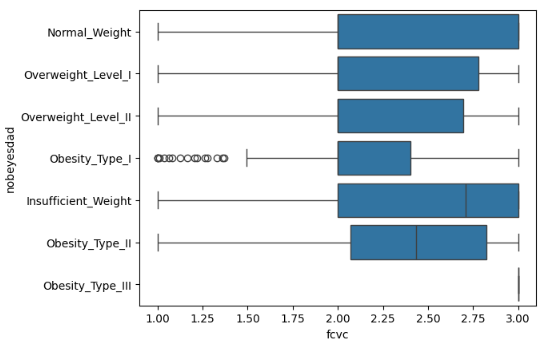
Se exploran las relaciones entre las variables y la variable objetivo para identificar patrones y correlaciones significativas

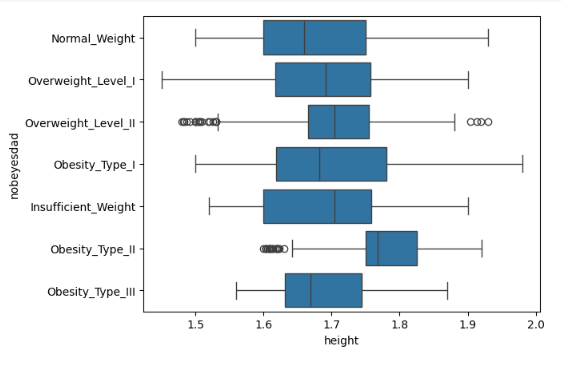
* **Correlaciones**: Se realizan matrices de correlaciones con el sistema **“pearson”** y el sistema **“spearman”** y en ninguna se observan correlaciones muy fuertes o determinantes
  + Correlación positiva moderada entre el peso y la altura (0,46)
  + Correlación positiva débil entre actividad física (FAF) y altura (0,29)
  + Correlación negativa débil entre edad y tue (-0,30)



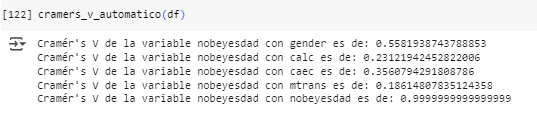
* **Kruskal-wallis:** Utilizaremos esta prueba para comparar variables continuas frente a variables categóricas, con este análisis observamos que en todas las variables continuas hay variaciones significativas con respecto al nivel de obesidad, por lo que se descarta la hipótesis nula en todos los casos. Sin embargo, al analizar las variables continuas con la variable de peso, vemos que no se rechaza la hipótesis para las siguientes variables:
  + FCVC: Frequency of vegetable consumption.
  + NCP: Number of main meals
  + CH2O: Daily water intake.
  + FAF: Physical activity frequency.
  + TUE: Time using technology devices.

Por lo que no hay diferencias entre los grupos según su peso, pero si en según su nivel de obesidad

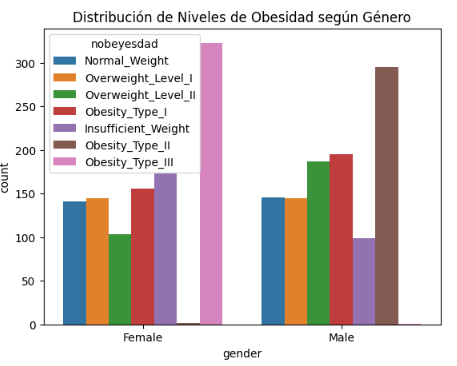




* **Cramer’s V:** Utilizaremos esta prueba para comparar variables categóricas frente a variables categóricas.



Obteniendo como resultado que **no hay una correlación significativa** entre variables categóricas ni con la variable de salida.



4.3 MODELADO Y RESULTADOS

Después de entrenar diferentes modelos y tras evaluar los resultados obtenidos estos son algunos de los principales hallazgos:

El modelo **XGBoost** obtuvo el **mejor rendimiento general**, seguido por **Random Forest optimizado**, mostraron un rendimiento superior en comparación con modelos más simples

La o**ptimización de hiper parámetros** mejoró significativamente el rendimiento de los modelos base

Las técnicas de búsqueda (**RandomizedSearchCV y GridSearchCV**) fueron fundamentales para encontrar **configuraciones óptimas**

Se observó que variables como frecuencia de consumo de vegetales (FCVC), consumo calórico (CALC), actividad física (FAF), consumo de agua (CH2O), y antecedentes familiares de sobrepeso son más predictivas para determinar niveles de obesidad

Además del análisis de rendimiento estándar, se realizó una **comparación estadística mediante ANOVA** y **pruebas post-hoc**, lo que confirmó que el modelo **XGBoost supera significativamente a los demás** en términos de **AUC**. Esto refuerza su selección como modelo final para apoyar decisiones clínicas relacionadas con los niveles de obesidad.

TABLA COMPARATIVA RENDIMIENTO ESTÁNDAR

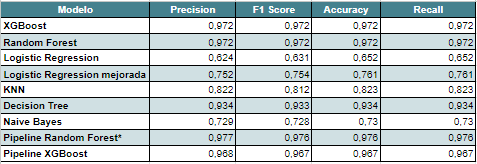
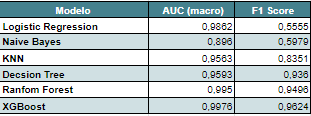


TABLA COMPARATIVA ANOVA



A continuación dejamos un breve análisis de cada modelo entrenado:

>>MODELO DE REGRESIÓN LOGÍSTICA (MODELO BASE)

El modelo de regresión logística muestra un rendimiento aceptable pero con margen de mejora.

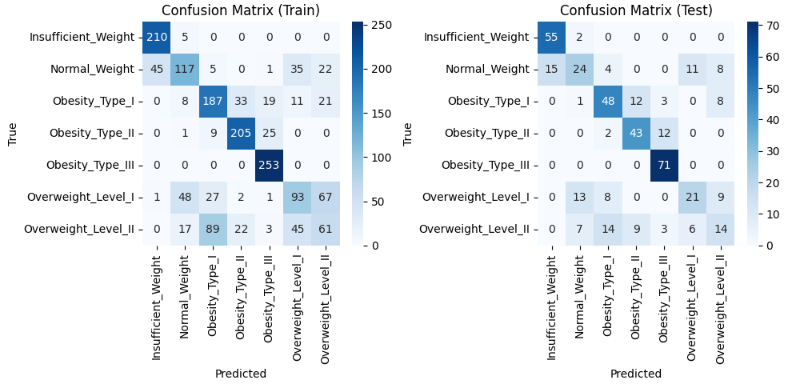
Se observa un ligero sobreajuste, con mejor rendimiento en el conjunto de train.

La matriz de confusión indica confusión entre algunas clases de obesidad.



Los mejores resultados de este modelo:

LogisticRegression accuracy (train) is 0.771  
LogisticRegression accuracy (test) is 0.761  
LogisticRegression F1 score (train) is 0.766  
LogisticRegression F1 score (test) is 0.754  
LogisticRegression Precision (test) is 0.752  
LogisticRegression Recall/Sensitividad (test) is 0.761



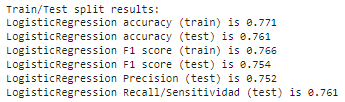
>>MODELO DE REGRESIÓN LOGÍSTICA MULTICLASE

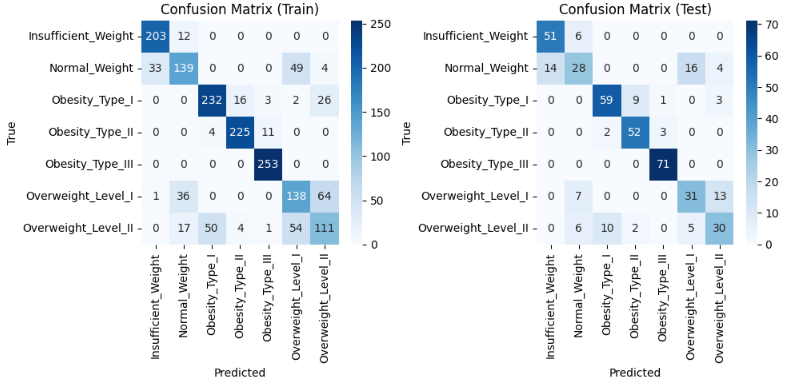
Al utilizar un modelo de regresión logística mejorada como parámetro de multiclase observamos cómo obtuvo un incremento notable en todas las métricas respecto al modelo base.

Hay un mejor equilibrio entre train y test, indicando un menor sobreajuste.

Y por último una mayor capacidad para distinguir entre categorías similares de obesidad.







>>ESTRATEGIA CLASIFICACIÓN MULTICLASE - ONEVSREST

Con la estrategia de clasificación OneVsRest entrenamos varios modelos binarios, cada uno aprendiendo a distinguir una clase frente al resto con el resultado de un rendimiento menor comparado con los modelos de regresión logística.

Los mejores resultados de este método de clasificación:

**OneVsRestClassifier accuracy (train) is 0.538**

**OneVsRestClassifier accuracy (test) is 0.501**

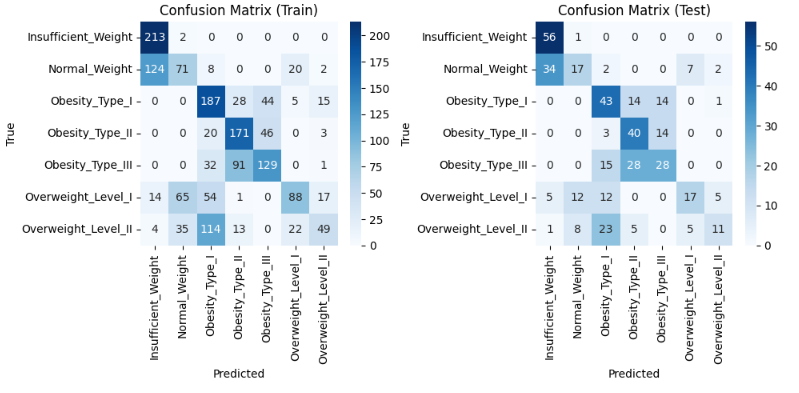
**OneVsRestClassifier F1 score (train) is 0.513**

**OneVsRestClassifier F1 score (test) is 0.472**

**OneVsRestClassifier Precision (test) is 0.513**

**OneVsRestClassifier Recall/Sensitividad (test) is 0.5391196343755448**

**OneVsRestClassifier Recall/Sensitividad (test) is 0.49869614246469446**



>>MODELO KNN (K-Nearest Neighbors)

Este modelo muestra un mejor rendimiento que los modelos de regresión logística.

La normalización de las medidas mejoró el rendimiento.



Los mejores resultados de este modelo:

KNeighborsClassifier accuracy (train) is 0.899

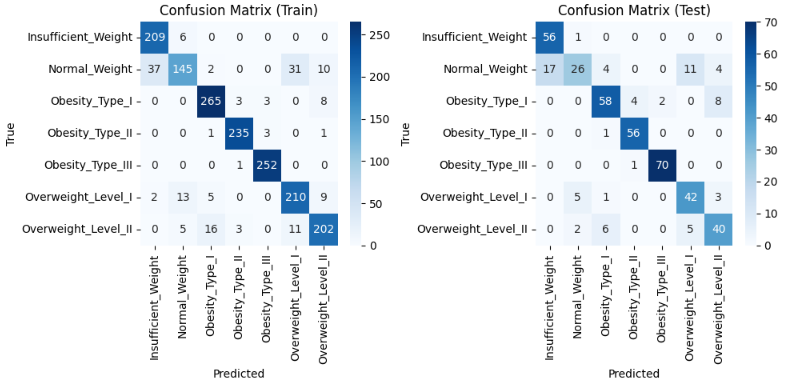
KNeighborsClassifier accuracy (test) is 0.823

KNeighborsClassifier F1 score (train) is 0.896

KNeighborsClassifier F1 score (test) is 0.812

KNeighborsClassifier Precision (test) is 0.822

KNeighborsClassifier Recall/Sensitividad (test) is 0.823



>>MODELO DECISION TREE

Este modelo muestra un **mejor rendimiento** que el modelo de **KNN**.

Muestra un alto rendimiento en el conjunto de train y algo menor en test indicando un mínimo sobreajuste.

**La profundidad del árbol llegó a 12 niveles con 111 hojas**.

El análisis de **características importantes** identificó **weight** con un **0.470960**, **height** con un **0.275331** y **fcvc** con un **0.142853** como los factores más determinantes para la clasificación de obesidad



Los mejores resultados de este modelo(sobreajustando):

DecisionTreeClassifier accuracy (train) is 1.000

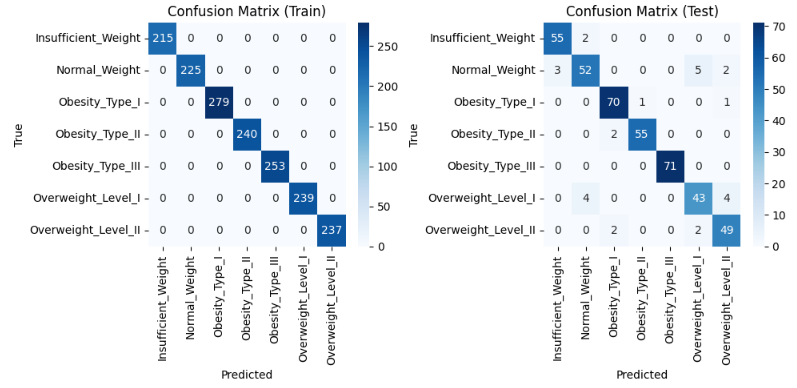
DecisionTreeClassifier accuracy (test) is 0.934

DecisionTreeClassifier F1 score (train) is 1.000

DecisionTreeClassifier F1 score (test) is 0.933

DecisionTreeClassifier Precision (test) is 0.934

DecisionTreeClassifier Recall/Sensitividad (test) is 0.934



>>MODELO RANDOM FOREST

Es uno de los modelos entrenados con mejor rendimiento.

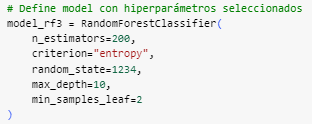
**Número de estimadores: 200, criterio:”entropy”, profundidad máxima:10, mínimo de muestras por hoja: 2**

Los modelos Random Forest muestran mejor capacidad de generalización que los árboles de decisión.

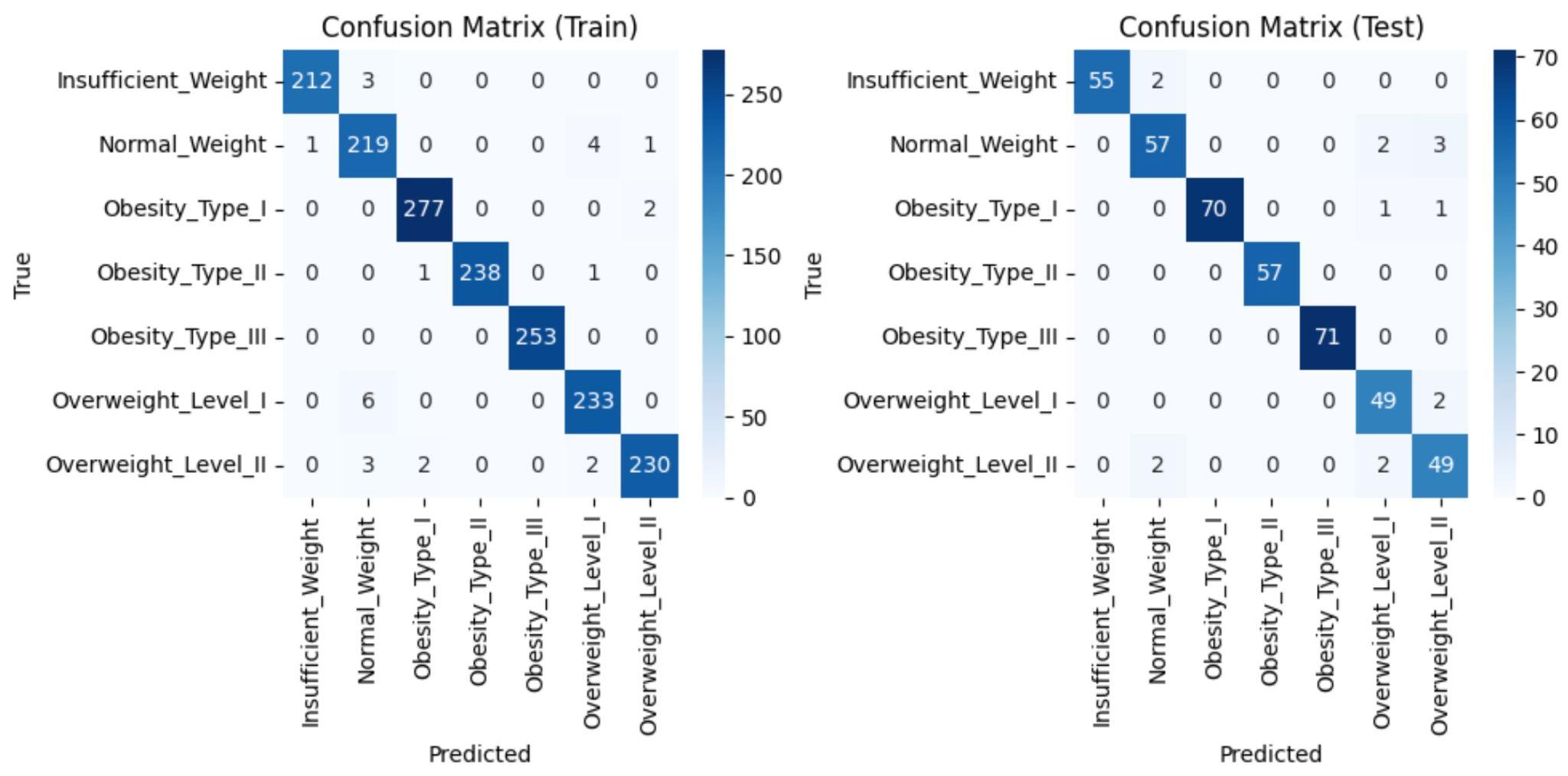
La herramienta **GridSearchCV** proporcionó los mejores hiper parámetros.

**Mejor balance entre train y test**, indicando menos sobreajuste

El análisis de las características importantes mantuvo coherencia con los resultados del árbol de decisión individual.



RandomForestClassifier accuracy (train) is 0.993  
RandomForestClassifier accuracy (test) is 0.972  
RandomForestClassifier F1 score (train) is 0.993  
RandomForestClassifier F1 score (test) is 0.972  
RandomForestClassifier Precision (test) is 0.972  
RandomForestClassifier Recall/Sensitividad (test) is 0.972



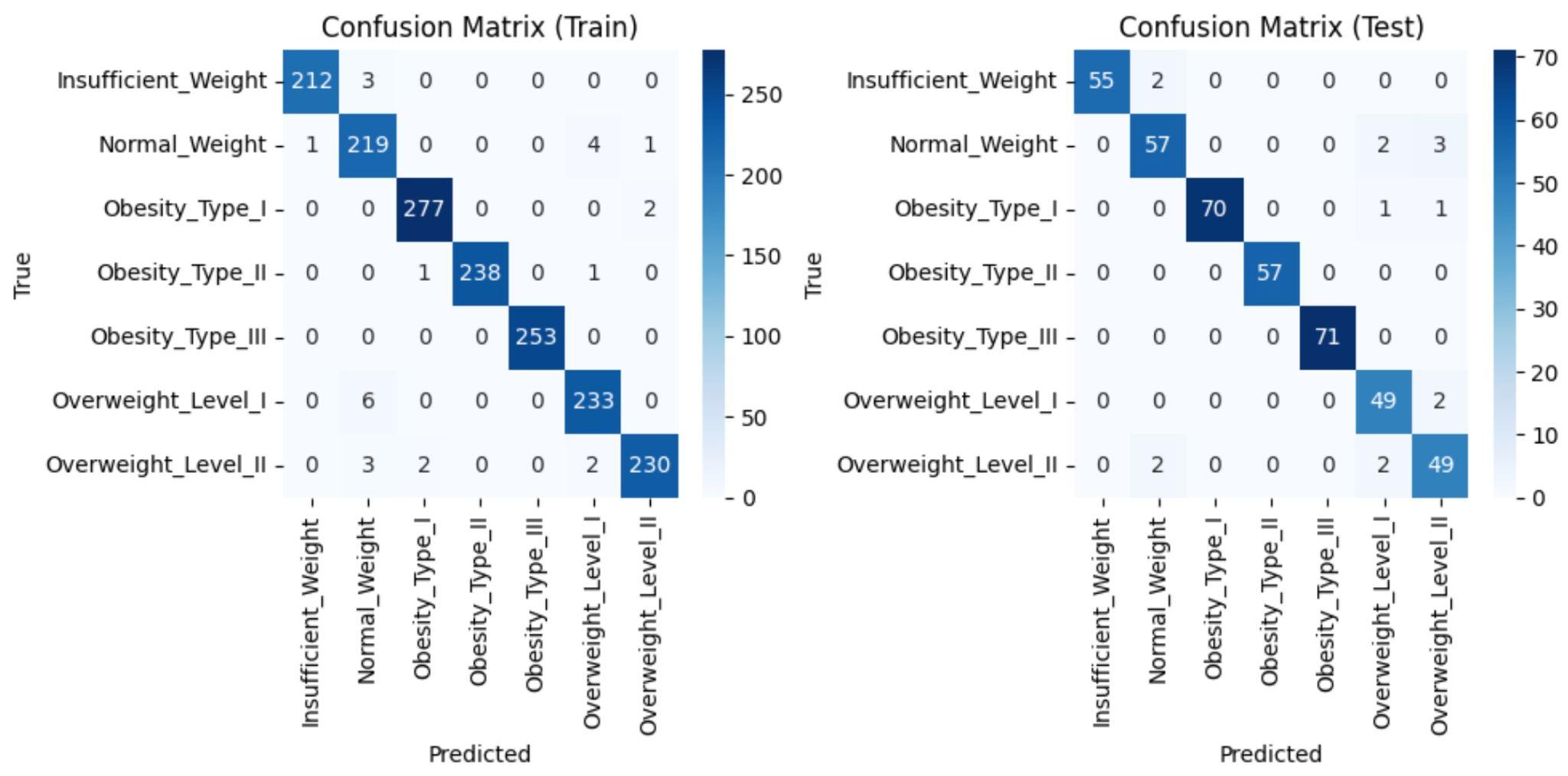
>>MODELO NAIVE BAYES

Este modelo tiene un rendimiento más bajo comparado con modelos más complejos, pero con train significativamente más rápido.

Mejor rendimiento en clases bien separadas



CategoricalNB accuracy (train) is 0.770  
CategoricalNB accuracy (test) is 0.730  
CategoricalNB F1 score (train) is 0.767  
CategoricalNB F1 score (test) is 0.728  
CategoricalNB Precision (test) is 0.729  
CategoricalNB Recall/Sensitividad (test) is 0.730



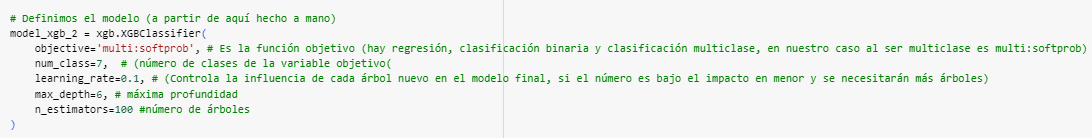
>>MODELO XGBOOST

Este modelo muestra **el rendimiento más alto** entre todos los modelos entrenados

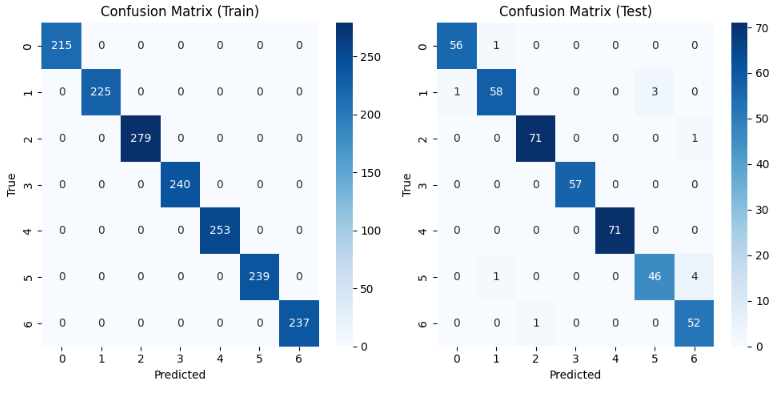
Buen equilibrio entre precisión y capacidad de generalización

El análisis de **características importantes** identificadas fueron **fcvc** con **0.329448**, **weight** con **0.254517** y **height** con **0.106170**.

Se realizó una codificación para variables categóricas, pero esto se vio compensado por la mejora en el rendimiento.



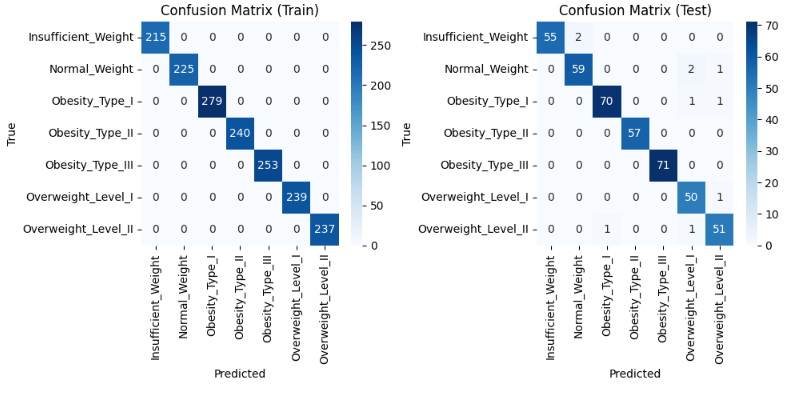
XGBClassifier accuracy (train) is 1.000  
XGBClassifier accuracy (test) is 0.972  
XGBClassifier F1 score (train) is 1.000  
XGBClassifier F1 score (test) is 0.972  
XGBClassifier Precision (test) is 0.972  
XGBClassifier Recall/Sensitividad (test) is 0.972



>>PIPELINE

El metodo presenta un rendimiento sobresaliente con precisión y F1 score cercanos a 1 en entrenamiento y 0.976 en prueba. Mantiene un buen equilibrio entre precisión y generalización. Las variables más relevantes fueron fcvc, weight y height. La codificación de variables categóricas mejoró notablemente el desempeño.

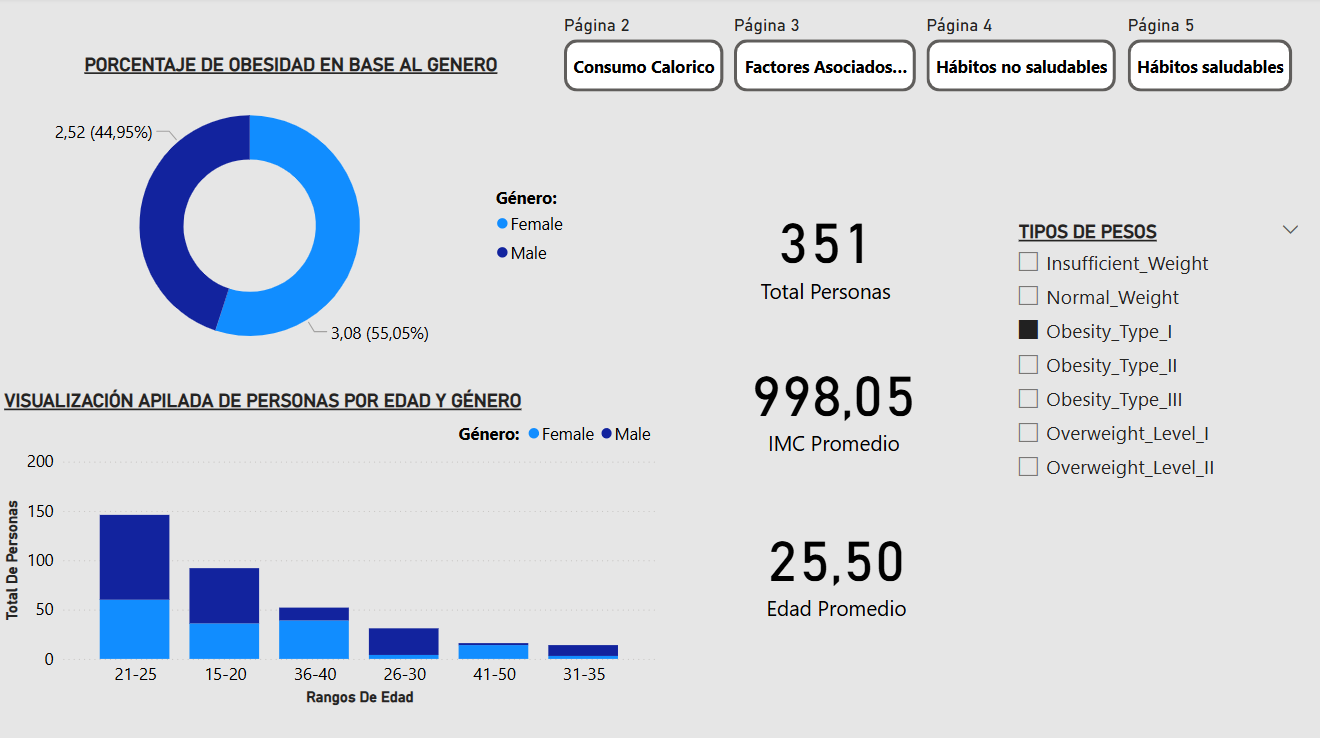
Pipeline accuracy (train) is 1.000  
Pipeline accuracy (test) is 0.976  
Pipeline F1 score (train) is 1.000  
Pipeline F1 score (test) is 0.976  
Pipeline Precision (test) is 0.977  
Pipeline Recall/Sensitividad (test) is 0.976



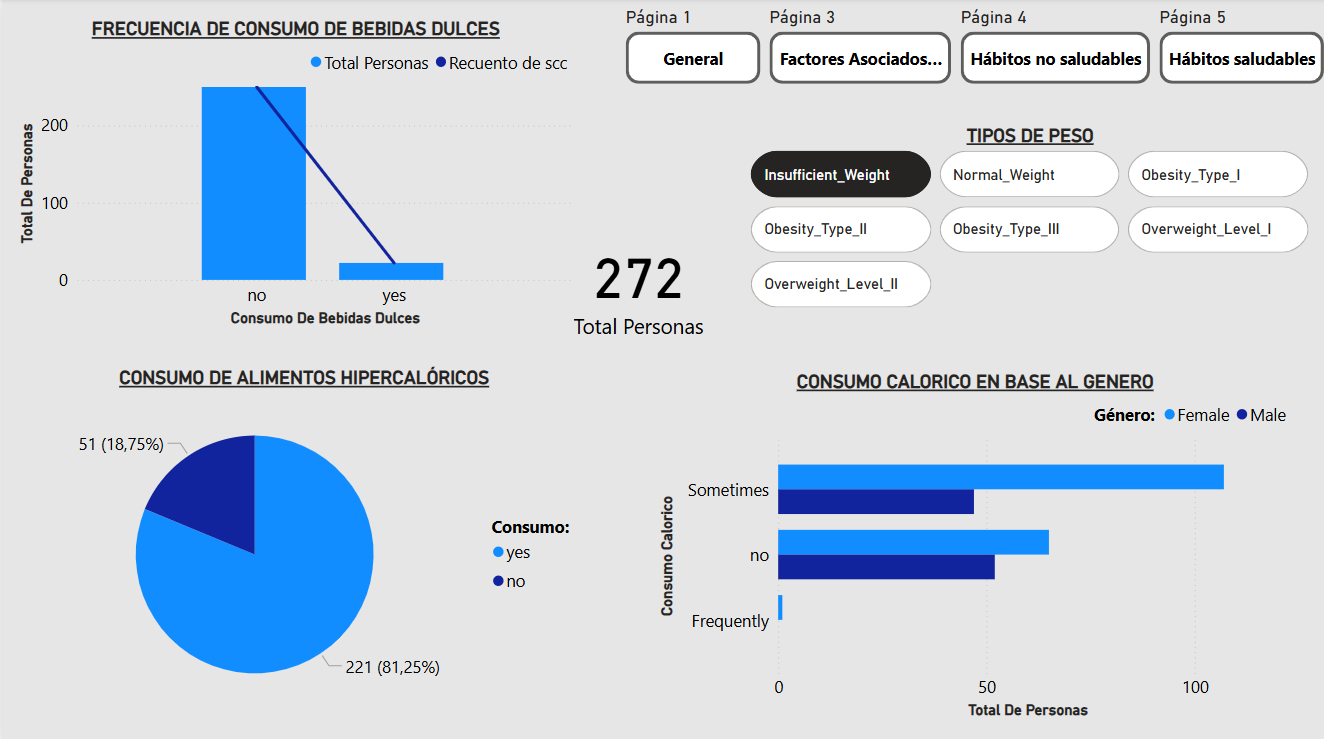
El método de optimización Pipeline nos da los mejores resultados para el modelo de Random Forest contra el mismo método tratando de optimizar XGBoost. Por lo tanto, en cuanto a modelo el XGBoost es el que mejores resultados F1 Score ha obtenido, pero en cuanto a la optimización que aporta de Pipeline a ambos modelos, Random Forest obtiene mejores resultados.

5. VISUALIZACIONES

>>INTRODUCCIÓN AL DASHBOARD

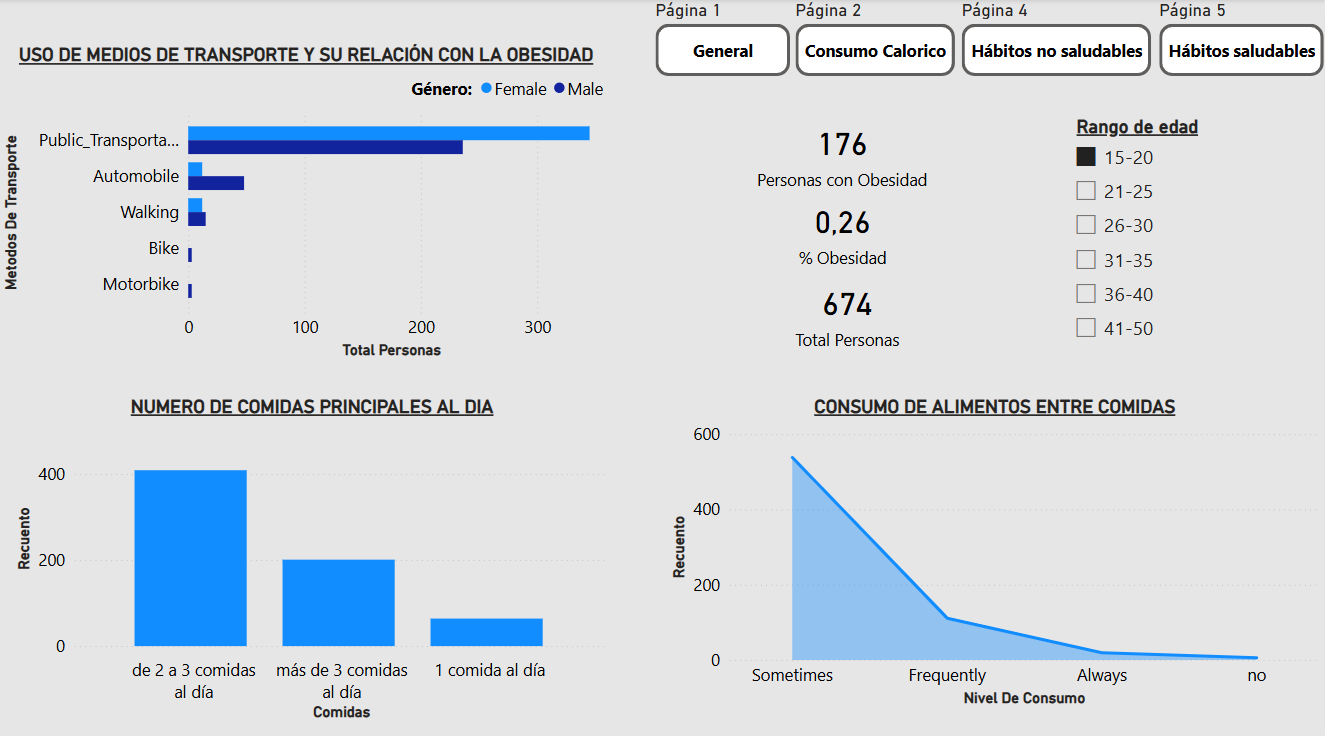
* Este dashboard interactivo tiene como objetivo presentar un análisis detallado sobre la distribución de la obesidad en base al género, la edad y otros factores asociados. En la página general, se destacan métricas clave como el porcentaje de obesidad por género, la cantidad total de personas analizadas, el índice de masa corporal (IMC) promedio y la edad promedio.
* Además, se ofrece una visualización que permite observar cómo se distribuyen los rangos de edad y género entre las personas evaluadas. Las páginas adicionales del informe profundizan en aspectos específicos como el consumo calórico, los factores asociados a la obesidad, y los hábitos saludables y no saludables, permitiendo explorar con más detalle las variables que impactan en el peso y la salud de la población analizada.
* Este informe busca proporcionar una herramienta visual clara para comprender mejor los patrones relacionados con la obesidad, apoyar la toma de decisiones y diseñar intervenciones dirigidas a mejorar la salud pública.

5.1 CONSUMO CALORICO  
En esta página se analiza el impacto del consumo calórico en la población evaluada. Se muestran gráficos sobre la frecuencia de consumo de bebidas dulces y alimentos hipercalóricos, permitiendo identificar qué proporción de personas mantiene estos hábitos. Además, se detalla el consumo calórico segmentado por género, lo que ayuda a entender si existen diferencias relevantes entre hombres y mujeres en términos de ingesta calórica. Esta sección es clave para visualizar cómo los hábitos alimentarios pueden estar relacionados con los tipos de peso identificados.



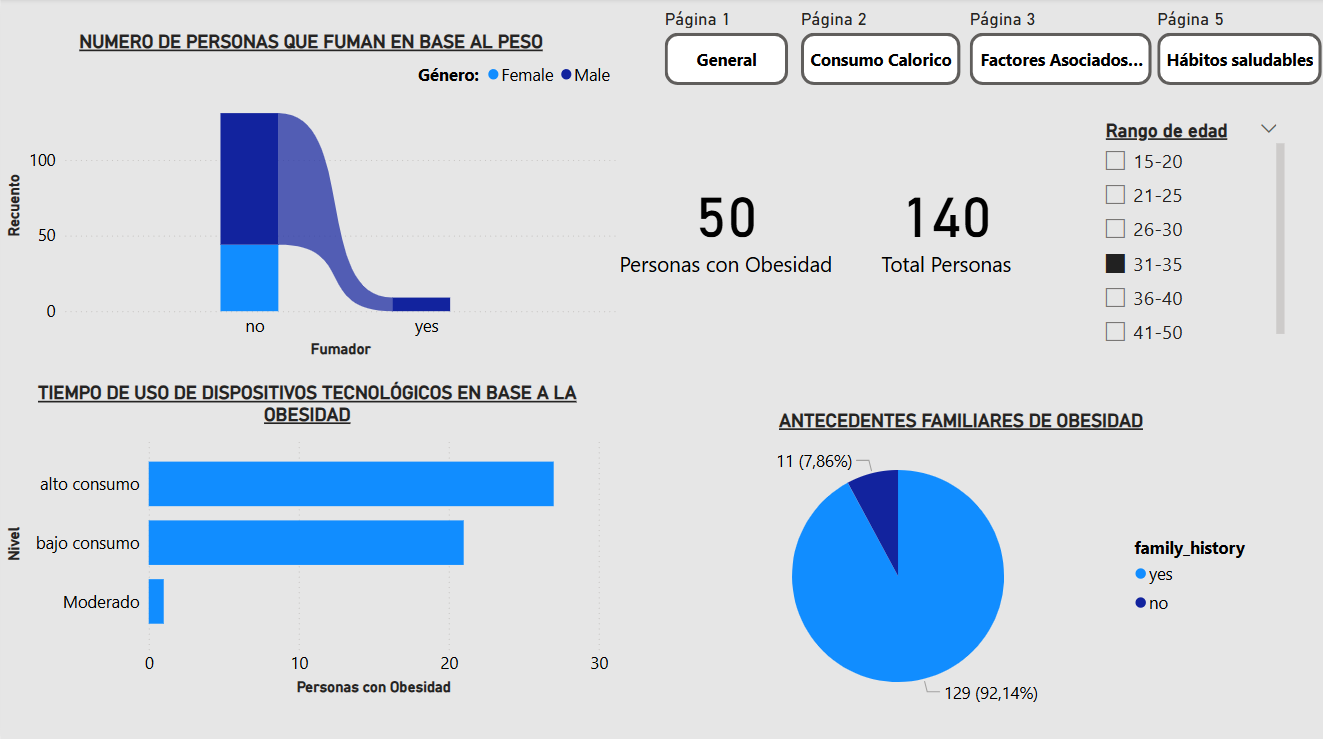
5.2 FACTORES ASOCIADOS A LA OBESIDAD

Aquí se exploran factores contextuales y de estilo de vida que podrían relacionarse con la obesidad. Se analiza, por ejemplo, el uso de medios de transporte y cómo se distribuyen entre las personas con obesidad. También se examina el número de comidas principales al día y el consumo de alimentos entre comidas, proporcionando una visión integral de los comportamientos cotidianos que podrían influir en el estado nutricional y el peso corporal.



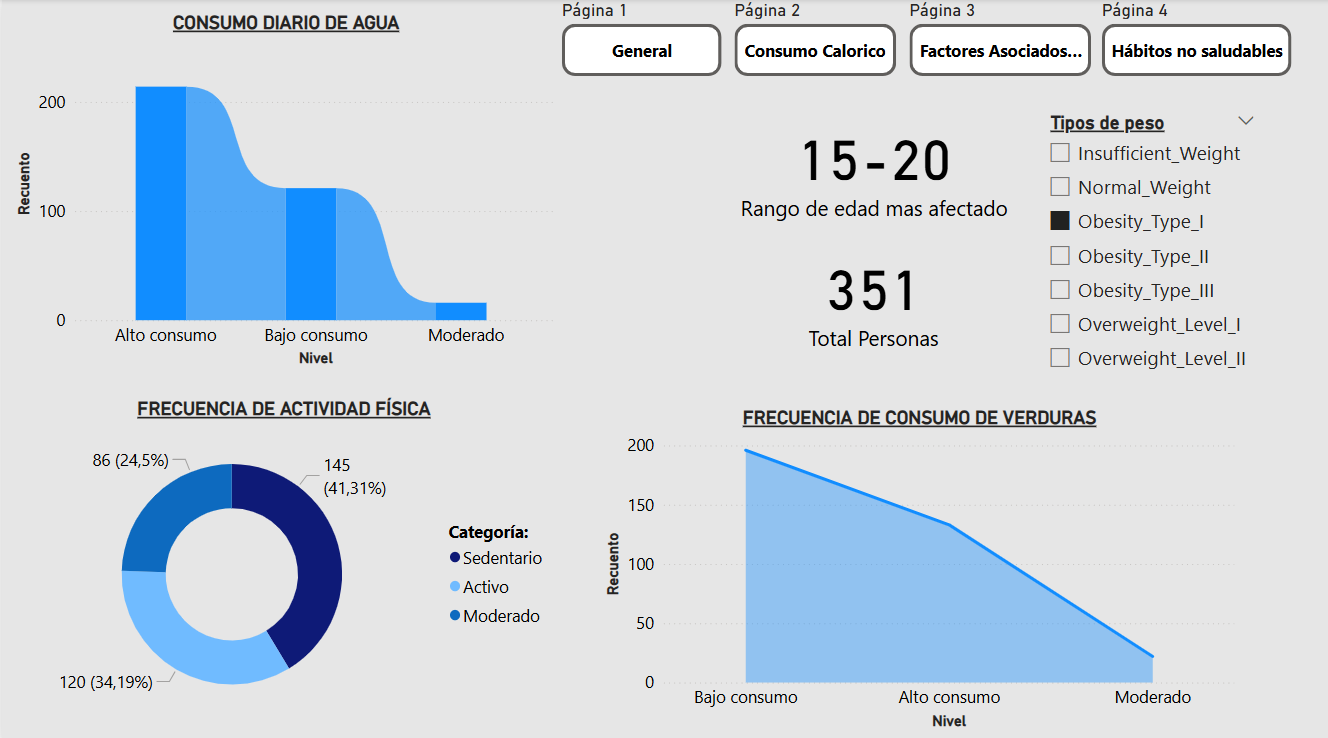
5.3 HÁBITOS NO SALUDABLES

En esta sección se examinan conductas que pueden contribuir negativamente al desarrollo de la obesidad. Se incluye un análisis del tabaquismo, segmentado por género y peso, permitiendo observar cuántas personas fuman y si existe alguna relación con su condición de peso. También se analiza el tiempo de uso de dispositivos tecnológicos, revelando que un alto consumo está significativamente asociado con la obesidad. Además, se presenta información sobre antecedentes familiares de obesidad, donde se observa que una gran mayoría de las personas evaluadas tienen historia familiar de esta condición. Estos factores ofrecen una perspectiva clara sobre cómo ciertos hábitos cotidianos y antecedentes genéticos pueden incidir en el desarrollo de la obesidad.



5.4 HÁBITOS SALUDABLES

Esta página se centra en las prácticas que favorecen un estilo de vida saludable. Se visualiza el consumo diario de agua, con una mayoría reportando un alto consumo, lo cual es un indicador positivo. También se analiza la frecuencia de actividad física, mostrando que el grupo más numeroso es el de personas sedentarias, seguido por quienes realizan actividad moderada y finalmente los activos. Asimismo, se incluye un gráfico sobre la frecuencia de consumo de verduras, en el que predomina el bajo consumo. La sección permite identificar oportunidades para promover mejores prácticas en hidratación, alimentación y ejercicio físico, factores clave para la prevención de la obesidad y la mejora de la salud general.



5.5 CONCLUSIÓN GENERAL DEL INFORME

* El análisis presentado en este dashboard ofrece una visión integral sobre los diversos factores que inciden en la prevalencia de la obesidad en la población evaluada. A través de las diferentes secciones, se evidencia cómo los hábitos alimentarios, el nivel de actividad física, el entorno familiar y el estilo de vida influyen de manera significativa en el estado de salud y el peso corporal.
* En particular, se identifican patrones preocupantes como el bajo consumo de verduras, el alto porcentaje de sedentarismo y el uso excesivo de dispositivos tecnológicos, que se asocian con una mayor prevalencia de obesidad, especialmente en grupos etarios específicos como los jóvenes entre 15 y 20 años. Al mismo tiempo, el consumo adecuado de agua se destaca como un hábito positivo que podría potenciarse junto con otras prácticas saludables.
* Estos hallazgos refuerzan la necesidad de implementar estrategias de prevención y promoción de la salud que se enfoquen en modificar comportamientos de riesgo y fomentar estilos de vida más activos y equilibrados. El dashboard se propone así como una herramienta útil para la toma de decisiones informadas, tanto a nivel institucional como comunitario, con el objetivo de reducir los índices de obesidad y mejorar la calidad de vida de la población.

6. CONCLUSIONES

Este proyecto se centra en el desarrollo de un modelo predictivo multiclase para clasificar el nivel de obesidad en individuos, utilizando datos fisiológicos y hábitos de vida. Se trabajó con un conjunto de datos que incluye 17 variables relacionadas con la salud, como edad, peso, altura y hábitos alimenticios, con el objetivo de categorizar a los participantes según su nivel de obesidad (NObeyesdad).

El proceso comenzó con la carga y comprensión del dataset, que se caracterizó por su estructura clara y la ausencia de valores nulos, lo que facilitó su manejo. Posteriormente, se realizó un preprocesamiento de datos que incluyó la limpieza y estandarización de nombres de columnas, así como la conversión de variables a tipos adecuados.

En el análisis exploratorio de datos (EDA), se examinaron las distribuciones y relaciones entre variables, confirmando que factores como el consumo de alimentos hipercalóricos y la frecuencia de ejercicio están relacionados con el nivel de obesidad. Se aplicaron pruebas estadísticas, como Kruskal-Wallis, para evaluar diferencias significativas entre las clases de obesidad.

Para la fase de modelado, se evaluaron diversos algoritmos de clasificación, incluyendo Regresión Logística, Naive Bayes, K-Nearest Neighbors, Árboles de Decisión, Random Forest y XGBoost. Los modelos más avanzados, como Random Forest y XGBoost, mostraron una mayor capacidad predictiva, especialmente en la identificación de niveles extremos de obesidad, mientras que los modelos más simples tuvieron un rendimiento inferior.

En resumen, se logró construir un modelo efectivo para la clasificación de niveles de obesidad, destacando la importancia de utilizar algoritmos que integren múltiples predicciones.

Para trabajos futuros, se sugiere explorar la incorporación de más variables, así como la implementación de técnicas de optimización de modelos para mejorar aún más la precisión de las predicciones. Además de:

> Incluir nuevas variables relacionadas con hábitos de sueño, estrés, enfermedades previas o entorno ambiental.

> Recoger datos longitudinales (a lo largo del tiempo) para predecir la evolución del estado de obesidad.

> Validación en entornos clínicos: Comparar las predicciones del modelo con diagnósticos clínicos reales para evaluar precisión y utilidad.

7. REFERENCIAS BIBLIOGRÁFICAS

**Estudio ALADINO 2023:**

Agencia Española de Seguridad Alimentaria y Nutrición (AESAN).(2023). Estudio ALADINO 2023 sobre alimentación, actividad física, desarrollo infantil y obesidad en niños de 6 a 9 años en España. Ministerio de Consumo, Gobierno de España.

Disponible en <https://www.aesan.gob.es/AECOSAN/web/nutricion/detalle/aladino_2023.htm>

**Obesidad en adultos - ISCIII y AESAN (basado en ENE-COVID):**

Agencia Española de Seguridad Alimentaria y Nutrición (AESAN), & Instituto de Salud Carlos III (ISCIII). (2022). Informe sobre la prevalencia de la obesidad en la población adulta en España y sus determinantes

según la Encuesta ENE-COVID. Gobierno de España.

Disponible en:

<https://www.aesan.gob.es/AECOSAN/docs/documentos/nutricion/Informe_OBESIDAD_ENECOVID_AESAN_ISCIll.pdf>