UNIVERSIDAD DE GRANADA

MÁSTER UNIVERSITARIO EN INGENIERÍA INFORMÁTICA

SISTEMAS INTELIGENTES PARA LA GESTIÓN EN LA EMPRESA

CURSO 2024 / 2025

Práctica 1: Pre-procesamiento de datos y clasificación binaria

---

Trabajo realizado por:
Mario Martínez Sánchez
martinezmario@correo.ugr.es

Índice

1. Exploración

En esta sección, se ha realizado un análisis exploratorio de datos con el objetivo de comprender mejor la estructura del dataset y detectar posibles patrones, desequilibrios y relaciones entre variables.

En primer lugar, se llevó a cabo una **distribución de la variable objetivo**. Para ello, se generó un gráfico de barras para visualizar la distribución de esta, que indica la presencia (1) o ausencia (0) de diabetes en los individuos del dataset.

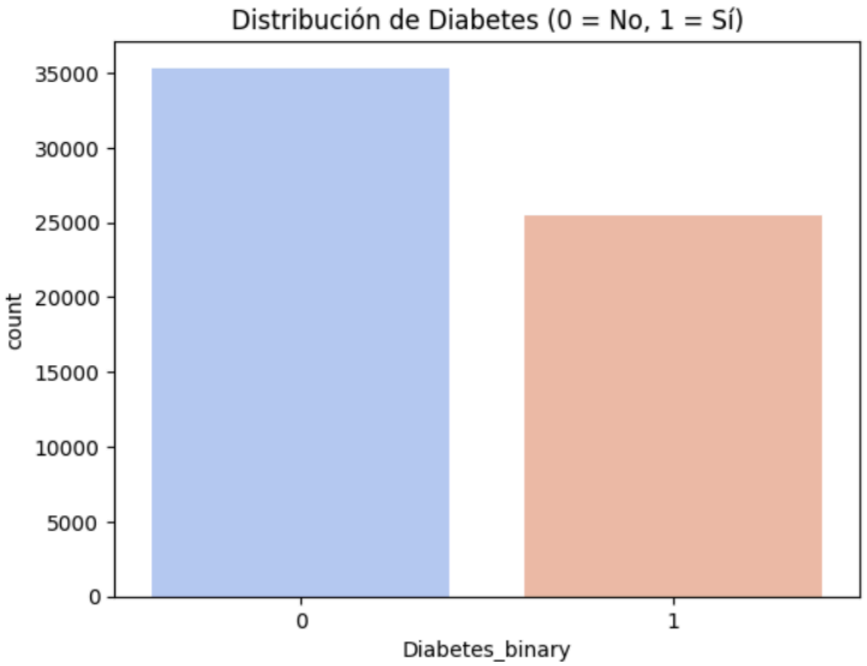

Figura 1. Distribución de diabetes

Como se puede observar, el gráfico muestra un desbalance de clases, con una mayor cantidad de muestras en la clase "No Diabetes" (0) en comparación con la clase "Diabetes" (1). Esta información es relevante, ya que un desbalance significativo podría afectar el rendimiento de los modelos de clasificación, requiriendo técnicas de balanceo como SMOTE.

Para analizar la distribución de las variables continuas, se han generado **histogramas individuales**. Estos permiten identificar tendencias, sesgos en los datos y la posible presencia de valores extremos (outliers).

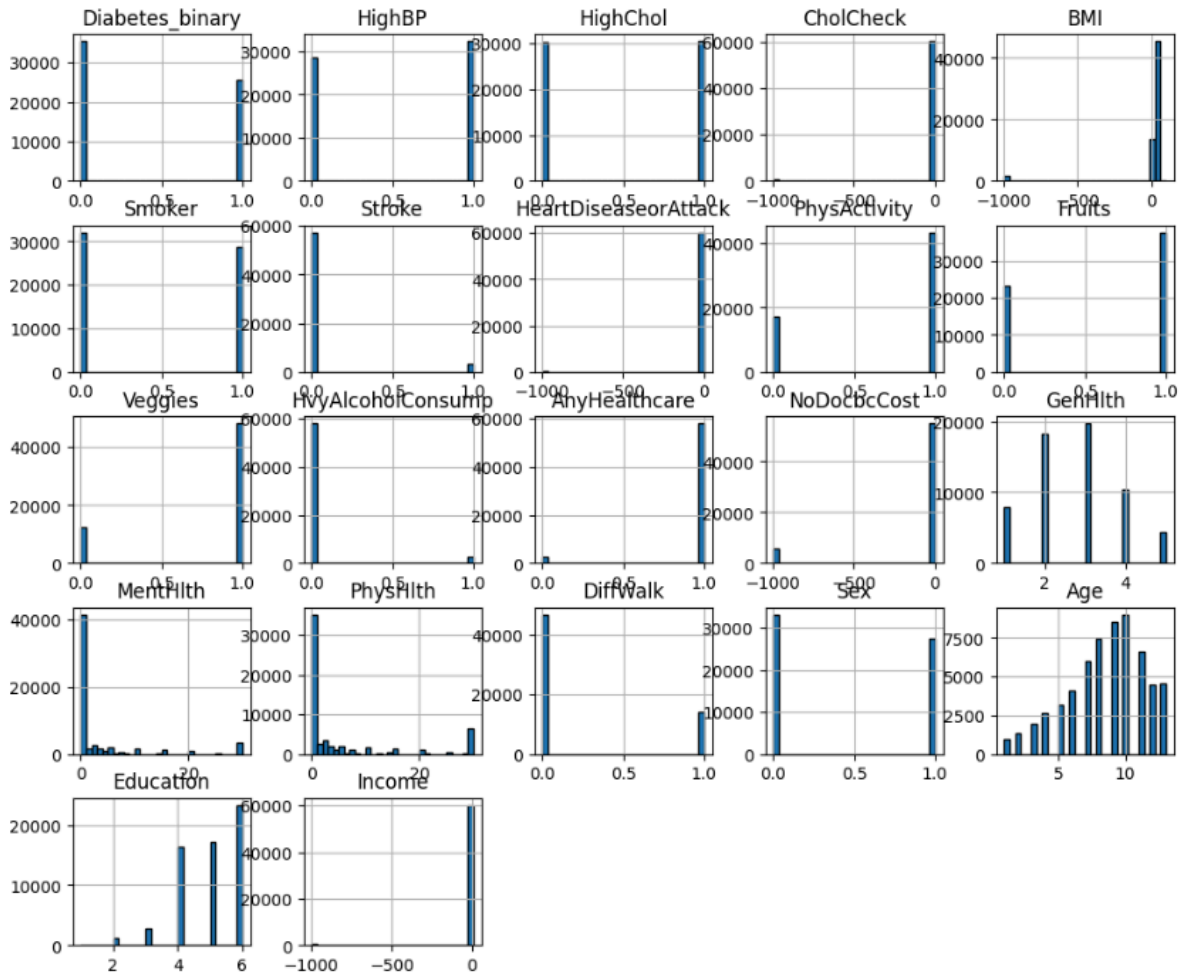

Figura 2. Histogramas de variables numéricas

Al observar los histogramas, se detecta que gran parte de las variables presentan distribuciones asimétricas, lo que sugiere que puede ser necesario aplicar técnicas de normalización para mejorar el desempeño de los modelos. Además, ciertas variables presentan concentraciones de valores en rangos específicos, lo que podría influir en su capacidad predictiva.

Finalmente, se llevó a cabo la **matriz de correlación**, es decir, un mapa de calor de correlaciones entre las variables numéricas para detectar relaciones significativas entre ellas. La correlación se mide mediante el coeficiente de Pearson, donde valores cercanos a +1 o -1 indican una fuerte relación lineal y valores cercanos a 0 indican ausencia de correlación.

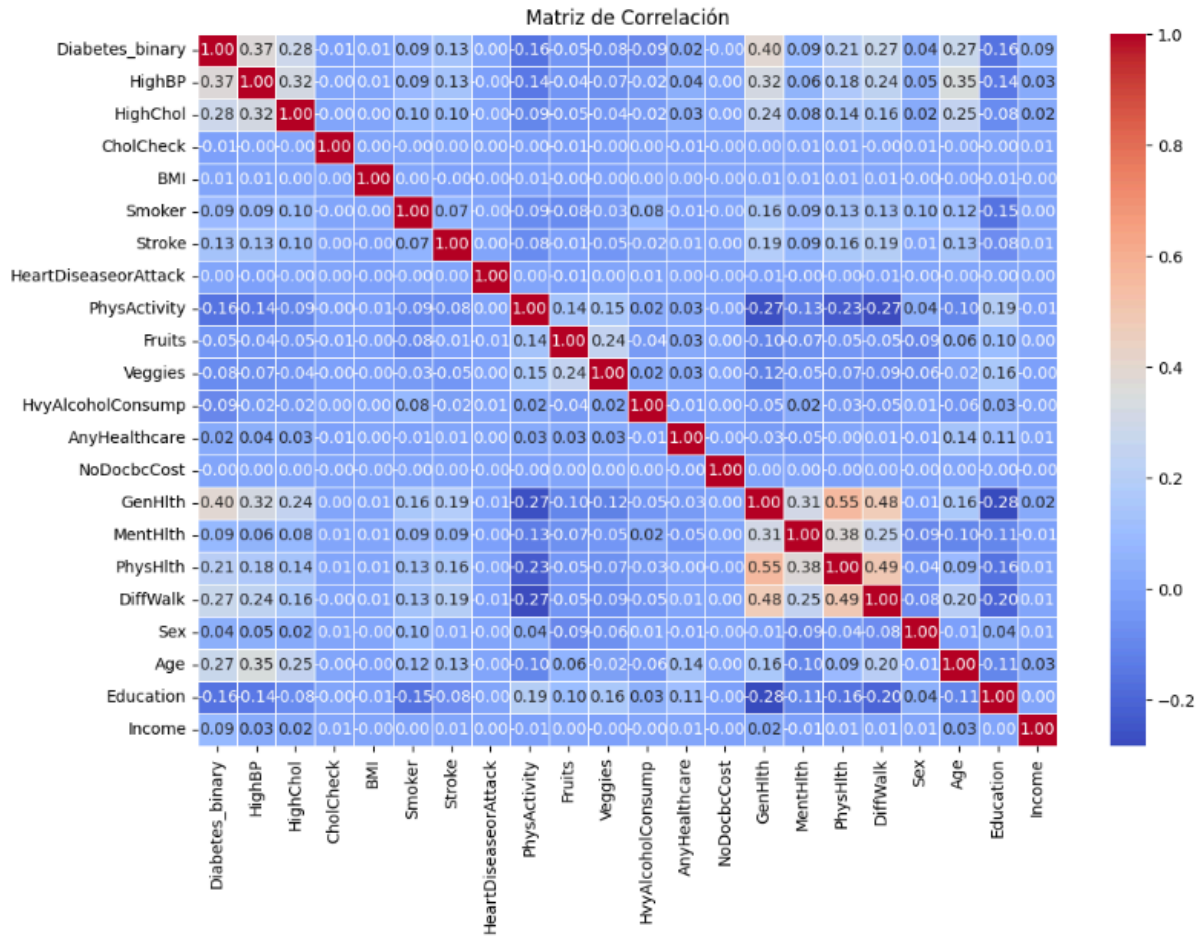

Figura 3. Matriz de correlación

Se puede observar que no existen relaciones excesivamente fuertes entre las variables. La mayor correlación encontrada es de aproximadamente 0.55, lo que indica que ninguna variable está altamente redundante con otra.

Algunas relaciones a destacar serían las siguientes:

- "Diabetes_binary" presenta correlaciones moderadas con variables como "HighBP" (-0.37) y "GenHlth" (-0.40), lo que sugiere que la hipertensión y la percepción de la salud general pueden ser factores asociados a la diabetes.
- Correlaciones moderadas entre algunas variables de salud como "PhysHlth" y "GenHlth" (correlación de 0.55) es lógico, ya que la percepción general de la salud está influenciada por problemas físicos.
- La mayoría de las correlaciones son bajas (cercanas a 0), lo que indica que las variables aportan información diversa y no hay una gran redundancia en los datos.

De este modo, dado que no se han encontrado correlaciones extremadamente altas (por encima de 0.8), no parece necesario eliminar variables por colinealidad en esta etapa.

2. Preprocesamiento

El preprocesamiento de los datos es una etapa fundamental para mejorar la calidad del conjunto de datos y garantizar un mejor desempeño de los modelos de clasificación. A continuación, se detallan los pasos llevados a cabo en este proceso.

2.1. Manejo de valores perdidos

Se identificó que los valores perdidos en el conjunto de datos estaban representados por el valor -999. Para corregir esto, se reemplazaron por valores NaN y posteriormente se imputaron con la media de cada variable numérica utilizando SimpleImputer. Esta estrategia permite conservar la mayor cantidad de datos posible sin introducir sesgos significativos.

2.2. Selección de características

Para mejorar la eficiencia del modelo y reducir la dimensionalidad, se utilizó SelectKBest para seleccionar las 10 variables más relevantes en la predicción de la diabetes. Tras aplicar esta técnica, se identificaron las siguientes características como las más importantes: HighBP (Presión arterial alta), HighChol (Colesterol alto), BMI (Índice de Masa Corporal), Smoker (Tabaquismo), Fruits (Consumo de frutas), GenHlth (Percepción general de salud), Sex (Sexo), Age (Edad), Education (Nivel educativo) e Income (Nivel de ingresos).

2.3. Identificación y tratamiento de outliers

Se implementó el método del rango intercuartílico (IQR) para detectar y eliminar valores atípicos en las variables numéricas. Se establecieron límites basados en 1.5 veces el IQR, descartando aquellos datos que se encontraban fuera de este rango. Esta eliminación ayuda a reducir el ruido en el modelo y mejorar su estabilidad.

2.4. Discretización

Para convertir variables continuas (numéricas) en variables discretas (categóricas), se ha recurrido a la discretización. En este proyecto, se aplicó la discretización a las variables Income (Ingresos) y Education (Educación) con el objetivo de simplificar el análisis y facilitar la interpretación de los modelos.

2.5. Normalización de los datos

Dado que las variables numéricas presentan escalas diferentes, se aplicó una normalización mediante StandardScaler, que estandariza los valores con una media de 0 y una desviación estándar de 1. Esto evita que variables con valores en rangos muy distintos dominen el proceso de modelado y facilita la convergencia en los algoritmos de aprendizaje automático.

2.6. Tratamiento del desbalance de clases

Tal y como se indicó en la exploración, se observó un desbalance en la variable objetivo (Diabetes_binary), lo que podría afectar negativamente el rendimiento del modelo. Para corregirlo, se aplicó SMOTE, una técnica de sobremuestreo que genera ejemplos sintéticos de la clase minoritaria para equilibrar la distribución de clases. Esto ayuda a evitar que el modelo aprenda un sesgo hacia la clase mayoritaria.

3. Clasificación

Para abordar la tarea de clasificación de la diabetes, se han seleccionado tres modelos distintos con el objetivo de comparar su desempeño y determinar cuál ofrece la mejor capacidad predictiva.

3.1. Regresión Logística

La Regresión Logística es un modelo lineal ampliamente utilizado en problemas de clasificación binaria. Se basa en la función sigmoide para predecir la probabilidad de que una observación pertenezca a una clase determinada.

La regresión logística es un modelo interpretable y eficiente para datos con relaciones lineales, lo que permite obtener información sobre la importancia de cada variable en la predicción de la diabetes.

3.2. Random Forest

El Random Forest es un modelo de ensamblado basado en múltiples árboles de decisión. Su principal ventaja es la reducción del sobreajuste gracias a la aleatorización en la selección de características y muestras en cada árbol.

Este modelo es robusto ante datos ruidosos y puede capturar relaciones no lineales entre las variables, lo que lo hace una buena opción para este problema.

3.3. Gradient Boosting

El Gradient Boosting Classifier es un modelo de ensamblado que construye árboles de decisión de manera secuencial, corrigiendo los errores cometidos en iteraciones previas.

Este modelo suele ofrecer un alto rendimiento en tareas de clasificación, ya que optimiza continuamente los errores de predicción, aunque su costo computacional es mayor en comparación con los otros modelos.

4. Discusión de resultados

Para la visualización de los resultados, se ha realizado una **curva ROC**, es decir, una evaluación del rendimiento de modelos de clasificación binaria. Respecto al AUC medido, podemos observar que el más alto es el de Gradient Boosting, aunque con una diferencia mínima. Esto significa que los tres modelos tienen una buena capacidad para distinguir entre las clases positivas (diabéticos) y negativas (no diabéticos).

Por otro lado, la forma de la curva ROC también proporciona información valiosa. Cuanto más rápido la curva se eleve hacia la esquina superior izquierda, mejor será el rendimiento del modelo. Se aprecia cómo la curva de Gradient Boosting se eleva más rápidamente, lo que sugiere una mejor capacidad para identificar verdaderos positivos con una baja tasa de falsos positivos. No obstante, las curvas de Random Forest y Regresión Logística también son buenas, pero muestran una menor capacidad inicial para discriminar entre clases.

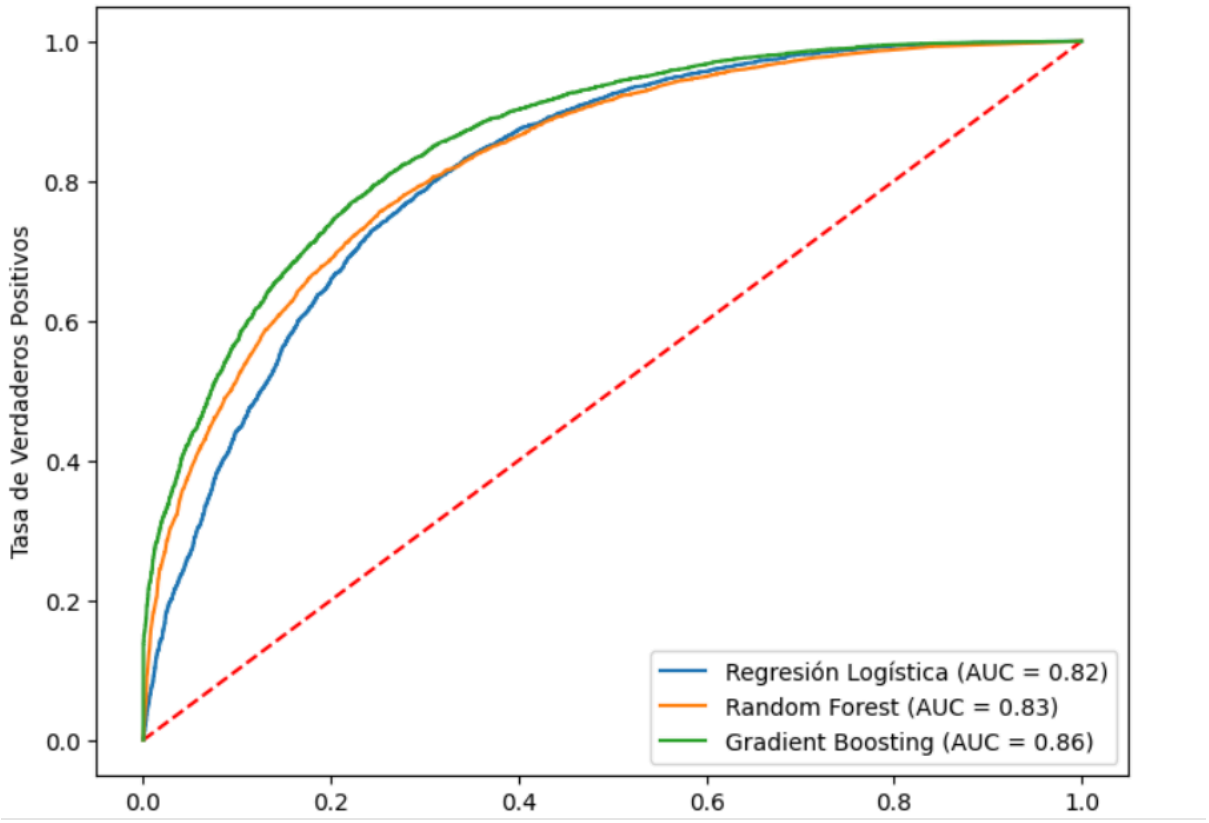

Figura 4. Curvas ROC

Tras esto, se llevó a cabo un análisis de las **matrices de confusión**, es decir, se representaron tablas que describen el rendimiento del modelo de clasificación al mostrar el número de verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN). Aquí está el análisis:

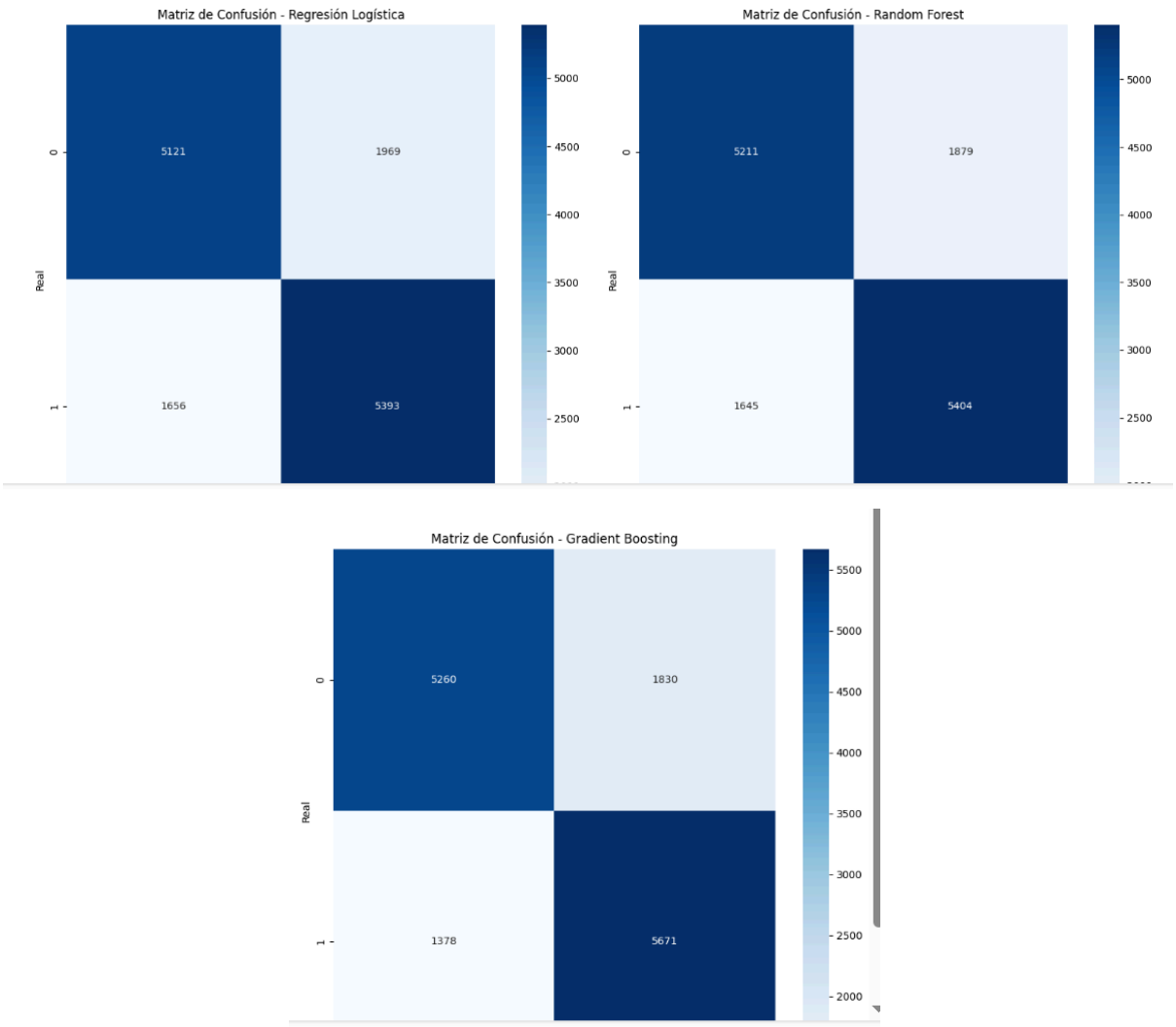

Figura 5. Matrices de confusión

Al analizar las matrices de confusión, vemos que Gradient Boosting es el modelo más equilibrado y preciso, minimizando tanto los falsos positivos como los falsos negativos, y maximizando los verdaderos positivos y negativos. Esto lo convierte en la mejor opción para la detección de diabetes.

5. Conclusiones

Tras un proceso de exploración y preprocesamiento de los datos, se evaluaron tres modelos de clasificación binaria: Gradient Boosting, Random Forest y Regresión Logística, utilizando la curva ROC y las matrices de confusión.

El modelo de Gradient Boosting mostró el mejor desempeño, con el AUC más alto y una curva ROC que asciende rápidamente hacia la esquina superior izquierda, lo que indica una alta capacidad para identificar correctamente los diabéticos y minimizar los falsos positivos. Las matrices de confusión también reflejan su precisión y equilibrio, maximizando los verdaderos positivos y negativos mientras minimiza los errores.

Aunque Random Forest y Regresión Logística también ofrecieron buenos resultados, sus curvas ROC y matrices de confusión sugieren que tienen una ligera desventaja en comparación con Gradient Boosting, que se destaca por su mayor capacidad de discriminación y balance en la clasificación.

En conclusión, Gradient Boosting es el modelo más eficaz para la detección de diabetes, superando a los otros modelos en precisión y equilibrio.

6. Bibliografía

Teboul, A. (s.f.). Diabetes health indicators dataset. Kaggle. Recuperado de https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

Preprocesamiento de datos. (s.f.). Recuperado de https://pradoposgrado2425.ugr.es/pluginfile.php/371208/mod_resource/content/1/Preprocesamiento.pdf