UNIVERSIDAD DE GRANADA

PRÁCTICA 1

Pre-procesamiento de datos y clasificación binaria

Sistemas Inteligentes para la Gestión en la Empresa

Alumna: Cristina del Águila Martín, cristinadam@correo.ugr.es

ÍNDICE

Exploración de Datos (EDA)

Para entender la estructura del conjunto de datos y detectar posibles problemas antes del preprocesamiento, he hecho un Análisis Exploratorio de Datos (EDA) para identificar valores atípicos, patrones en las variables y posibles correlaciones con la variable objetivo.

Carga de los datos y conversión de tipos

He cargado el conjunto de datos *diabetes.csv* y he convertido todas las columnas a tipo numérico para evitar problemas con valores incorrectamente interpretados como texto.

Distribución de la variable objetivo

Para evaluar el balanceo de clases, he generado un gráfico de barras de la variable *Diabetes_binary*. En el que se observa que la clase 0 (personas sin diabetes) es más frecuente que la clase 1 (personas con diabetes). Esto parece tener lógica ya que en la población general hay más personas sin diabetes que con ella, lo que podría indicar un desbalance de clases que será necesario abordar en el preprocesamiento para mejorar el rendimiento de los modelos de clasificación.

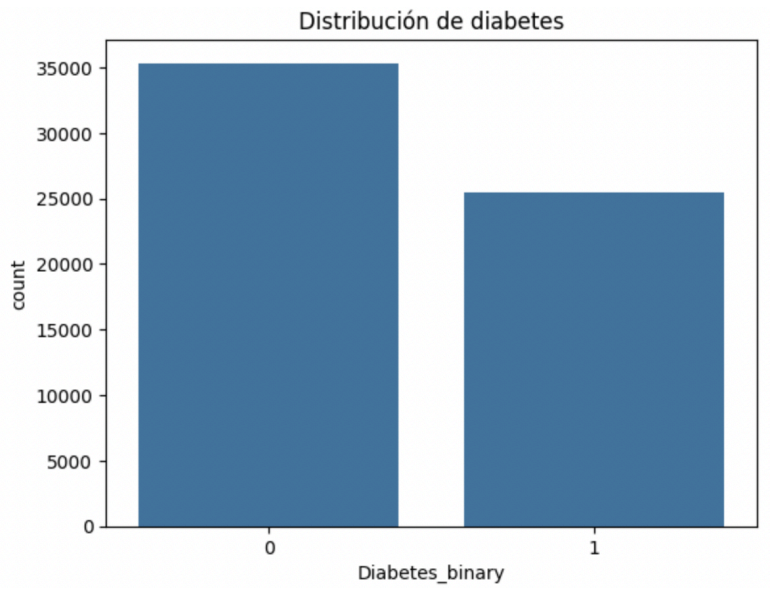

Imagen 1: Gráfico de barras de la variable Diabetes

Análisis estadístico descriptivo

He analizado el dataset y he observado lo siguiente:

En cuanto a las condiciones de salud, el 41.4% de los individuos están clasificados como diabéticos. Además, el 52.8% presenta presión arterial alta y el 49.9% tiene colesterol elevado. El 5.6% de los encuestados ha sufrido un accidente cerebrovascular.
Respecto a los hábitos de vida, el 46.9% de las personas son fumadoras. Sin embargo, un 72% realiza actividad física con regularidad. El consumo de frutas y verduras es relativamente elevado, con promedios de

62.2% y 79.5%, respectivamente. Por otro lado, el consumo excesivo de alcohol es bajo, presente solo en el 4.6% de los casos.

En términos de acceso a servicios de salud, el 95.5% ha dicho que tiene acceso a algún tipo de atención médica. La percepción general de salud, medida en una escala de 1 a 5, tiene una mediana de 3, lo que sugiere una autopercepción mayormente positiva. Sin embargo, tanto la salud mental como la física mostraron alta variabilidad, con casos extremos de hasta 30 días con problemas en el último mes.

A nivel demográfico, el género está relativamente equilibrado, con una ligera mayoría femenina. La edad promedio está en una categoría media-alta (8.48 en una escala de 1 a 13), correspondiente a adultos mayores. El nivel educativo medio es de 5, lo que representa estudios secundarios completos o superiores no universitarios.

Por último, un 22.2% de los usuarios ha dicho tener dificultades para caminar. En cuanto a la variable outlier, el 96% de los registros no presentan valores atípicos, pero voy a tratar los que si tienen en el preprocesamiento.

Matriz de correlación

Para analizar la relación entre las variables he generado una matriz de correlación.

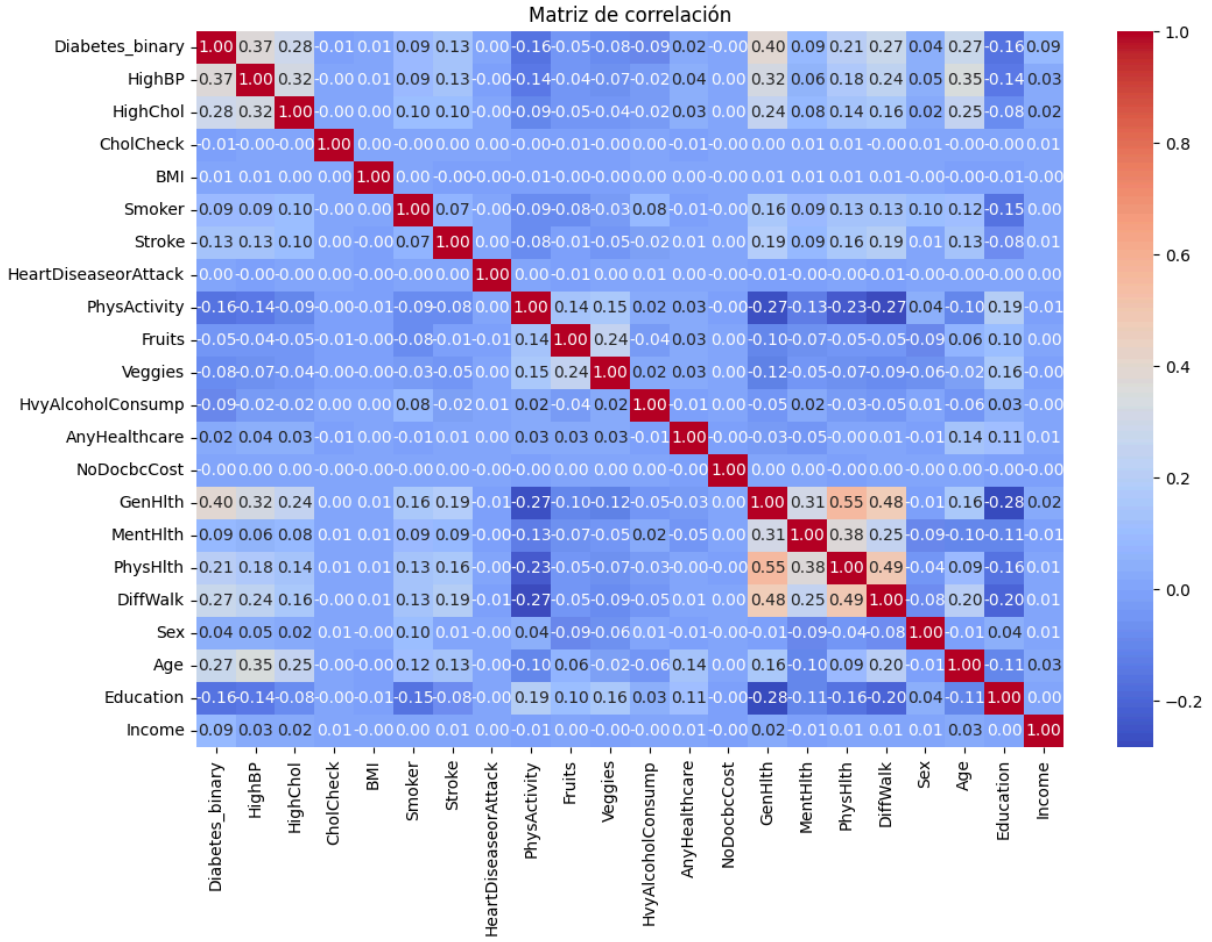

Imagen 2: Matriz de correlación

Correlaciones positivas (aumentan la probabilidad de diabetes):

- La presión arterial alta (HighBP) presenta una correlación positiva moderada con la diabetes, con un valor de 0.37. Esto indica que las personas que sufren de hipertensión tienen una mayor probabilidad de desarrollar diabetes. Ambas condiciones comparten factores de riesgo comunes, como el sobrepeso, el sedentarismo y una dieta poco saludable, por lo que no me parece raro que se presenten juntas.
- El colesterol alto (HighChol) también muestra una correlación positiva moderada con la diabetes, con un valor de 0.28. Las personas con niveles elevados de colesterol en sangre tienden a tener más riesgo de padecer diabetes tipo 2. Esto se debe, en parte, a que el colesterol alto suele ser consecuencia de hábitos poco saludables, los cuales también influyen en el desarrollo de la diabetes.
- La edad (Age) tiene una correlación positiva moderada con la diabetes, igualmente con un valor de 0.28. Esto significa que a medida que las personas envejecen, aumenta la probabilidad de que desarrollen diabetes. El envejecimiento suele venir acompañado de cambios metabólicos, menor actividad física y acumulación de factores de riesgo, lo que incrementa la vulnerabilidad frente a esta enfermedad.

Correlaciones negativas (disminuyen la probabilidad de diabetes):

- La salud general auto-percibida (GenHlth) tiene una correlación negativa moderada con la diabetes, con un valor de -0.39. Esto quiere decir que las personas que consideran que tienen buena salud en general tienen menos probabilidades de padecer diabetes.
- El nivel educativo (Education) tiene una correlación negativa débil con la diabetes, con un valor de -0.16. Esto sugiere que las personas con mayor nivel educativo tienen una probabilidad ligeramente menor de padecer esta enfermedad. Un mayor nivel de educación puede estar asociado con un mayor acceso a información sobre salud, mejor comprensión de los riesgos y una mayor tendencia a adoptar conductas preventivas.
- La actividad física regular (PhysActivity) también presenta una correlación negativa débil con la diabetes, con un valor de -0.15. Las personas que realizan ejercicio físico con frecuencia tienen menos probabilidades de desarrollar diabetes. La actividad física ayuda a controlar el peso, mejora la sensibilidad a la insulina y contribuye al buen funcionamiento del sistema metabólico, todos factores clave en la prevención de la diabetes.

Preprocesamiento de Datos

Detección y tratamiento de valores atípicos

Como primer paso, he realizado una exploración visual de las variables numéricas más relevantes para identificar posibles valores atípicos. En concreto, he generado boxplots de las variables BMI, Income, PhysHlth y MentHlth.

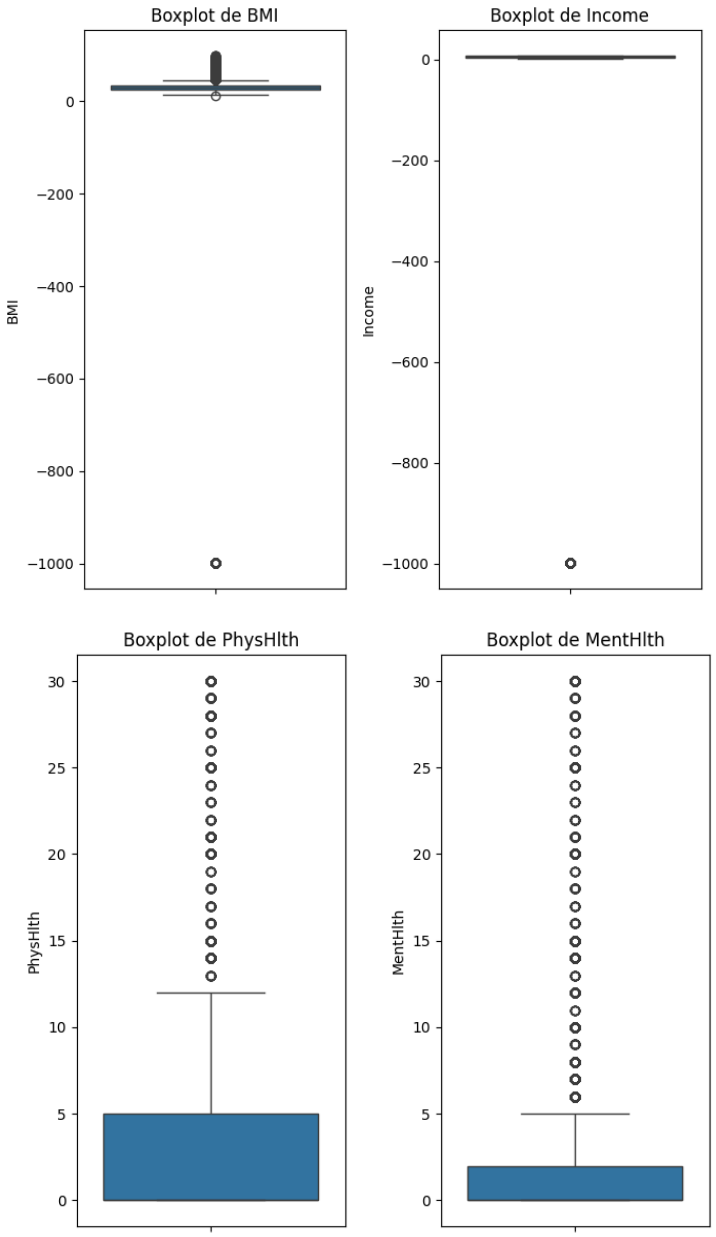

Imagen 3: Boxplots de variables relevantes

Como se puede observar en la imagen, en los dos primeros gráficos (BMI e Income), se observa claramente un valor atípico extremo con un valor cercano a -1000. Este tipo de valor es el valor -999 que se usa para representar datos ausentes, lo cual voy a corregir en la siguiente parte del preprocesamiento. La mayoría de los datos válidos de ambas variables se encuentran dentro de un rango estrecho y positivo, lo que también indica una fuerte concentración de valores cerca de la mediana.

En cuanto a PhysHlth y MentHlth, los boxplots muestran distribuciones fuertemente asimétricas a la derecha, con una gran cantidad de valores atípicos por encima del tercer cuartil. En ambas variables, muchos individuos reportaron entre 15 y 30 días de mala salud física o mental en el último mes, lo que podría tener una relación importante con la variable objetivo. Sin embargo, la mayor parte de los datos se concentra en

valores bajos (cerca de 0), por lo que la mayoría de los usuarios a los que se ha encuestado, no reportaron problemas de salud prolongados.

Sustitución de -999 por NaN

Hay algunas columnas donde se han usado valores como -999 para representar datos ausentes. He reemplazado esos valores por NaN usando la función replace, y posteriormente veo el patrón de valores faltantes mediante un mapa de calor.

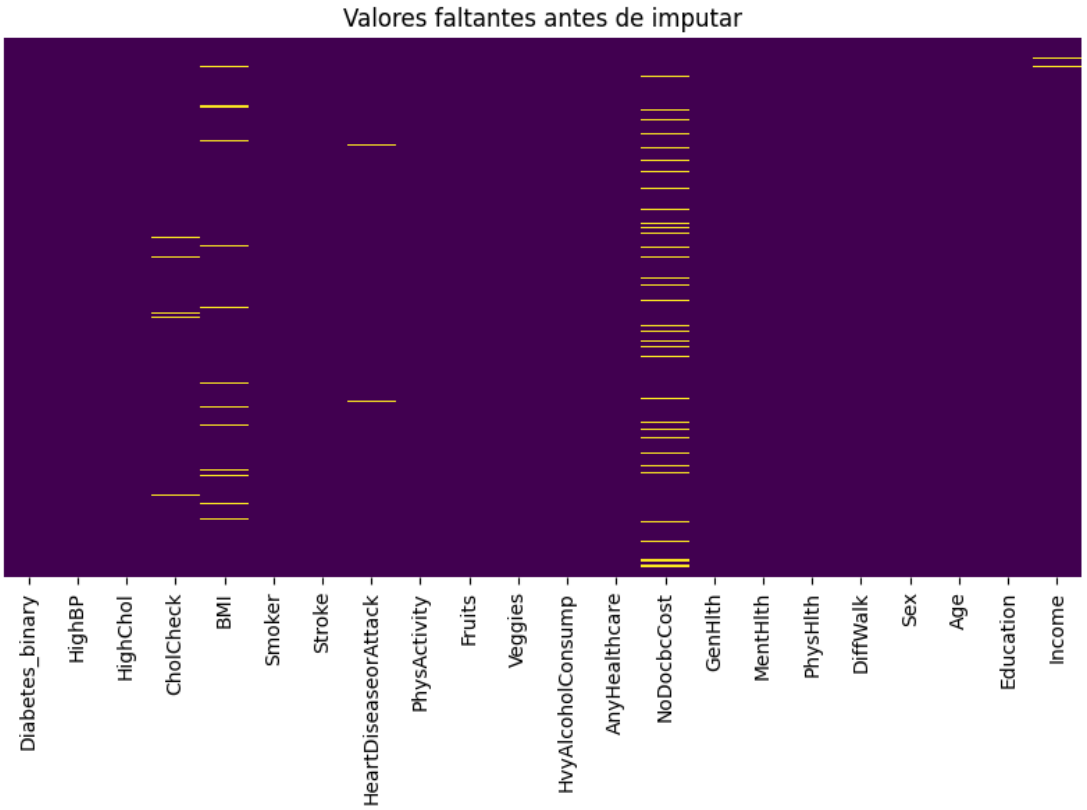

Imagen 4: Heatmap con valores ausentes

Para el tratamiento de los valores faltantes, he utilizado diferentes estrategias dependiendo del tipo de variable:

- Para variables numéricas como BMI e Income, he imputado la mediana para evitar el efecto de los valores extremos
- Para variables categóricas binarias como NoDocbcCost, HeartDiseaseorAttack y CholCheck, he imputado la moda (el valor más frecuente)

Después de la imputación, compruebo otra vez que no quedan valores faltantes con otro mapa de calor.

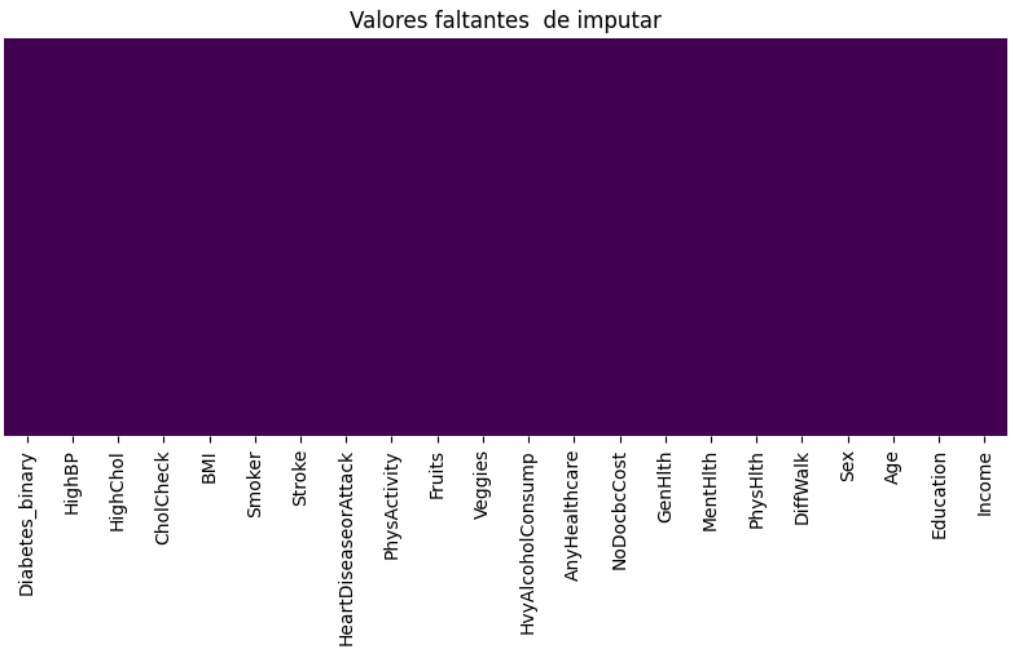

Imagen 5: Heatmap sin valores ausentes

Eliminación de registros inconsistentes

He identificado y eliminado registros con combinaciones de valores que he considerado que eran clínicamente inconsistentes. Por ejemplo:

- He eliminado los registros donde Income es negativo, ya que no tiene sentido
- Quito los casos en los que una persona reporta más de 25 días de mal estado físico (PhysHlth > 25) o mental (MentHlth > 25) en un mes, pero aun así califica su salud general (GenHlth) como "Buena" o "Excelente" (valor ≤ 2). Ya que que esta combinación es contradictoria
- Había pensado en eliminar registros donde AnyHealthcare = 0 (sin seguro médico) pero CholCheck = 1 (se ha hecho chequeos de colesterol), ya que en países como EE.UU. (donde la sanidad es privada y cara), esta situación sería muy improbable.
  Sin embargo, al no confirmar el origen del dataset, he decidido no aplicar este filtro, porque en países con sistemas de salud públicos o subvencionados (como España o Canadá), esta combinación sí podría darse.

Estas inconsistencias reflejan errores en la recogida de datos o respuestas poco fiables, por lo que he optado por eliminarlos. En total, se han eliminado 47 registros por este motivo.

Detección y tratamiento de outliers

Para una detección más robusta de valores atípicos en el conjunto de datos numéricos, he utilizado el algoritmo Isolation Forest con un umbral de contaminación del 2%.
Este método etiqueta 1215 registros como outliers, que he eliminado para evitar que distorsionen el aprendizaje de los modelos.

Selección de variables relevantes

Para reducir la dimensionalidad y centrar el análisis en las variables más relevantes, he aplicado tres métodos complementarios de selección de características:

- Mutual Information (MI): permite identificar relaciones no lineales entre variables y la variable objetivo.
- Importancia de características en Random Forest
- Importancia en XGBoost

He comparado las puntuaciones obtenidas por cada método y calculado una suma total de importancia. De este análisis me quedo con un conjunto de 10 variables más relevantes, que son: HighBP, GenHlth, BMI, Age, HighChol, Income, PhysHlth, Education, DiffWalk y CholCheck.

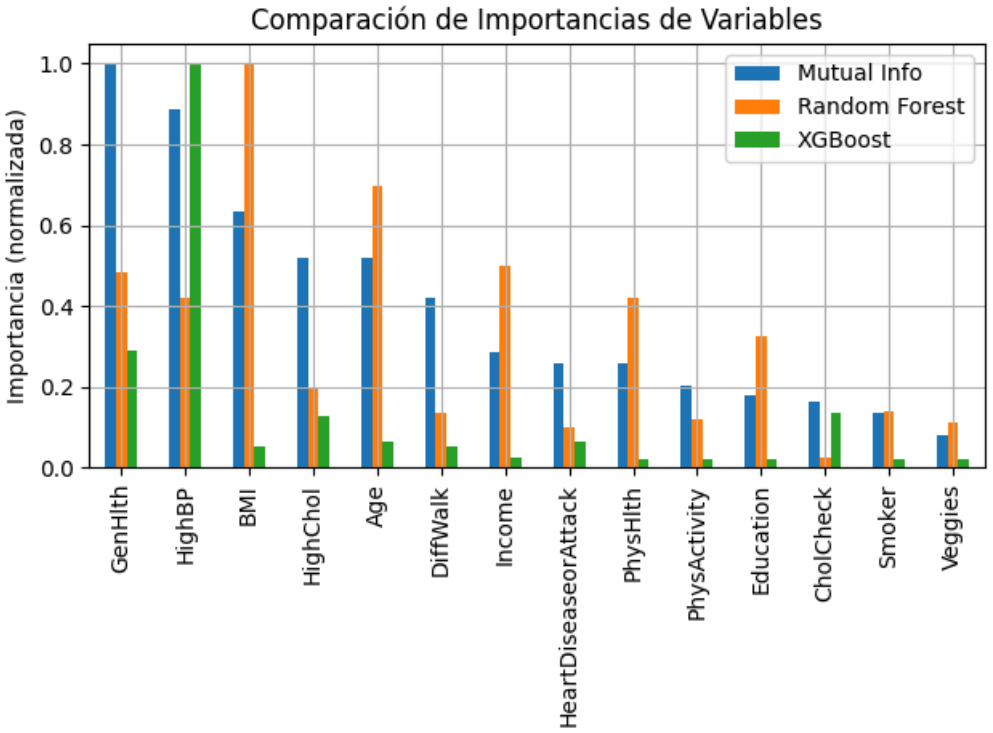

Imagen 6: Comparación de importancias de variables

Normalización

Para evitar que las diferencias de escala entre variables numéricas influyeran en los modelos, especialmente aquellos basados en distancia como KNN, he normalizado los datos utilizando StandardScaler, que transforma cada variable a una distribución con media 0 y desviación típica 1.

A modo de ejemplo, he visualizado la distribución de la variable BMI antes y después del escalado:

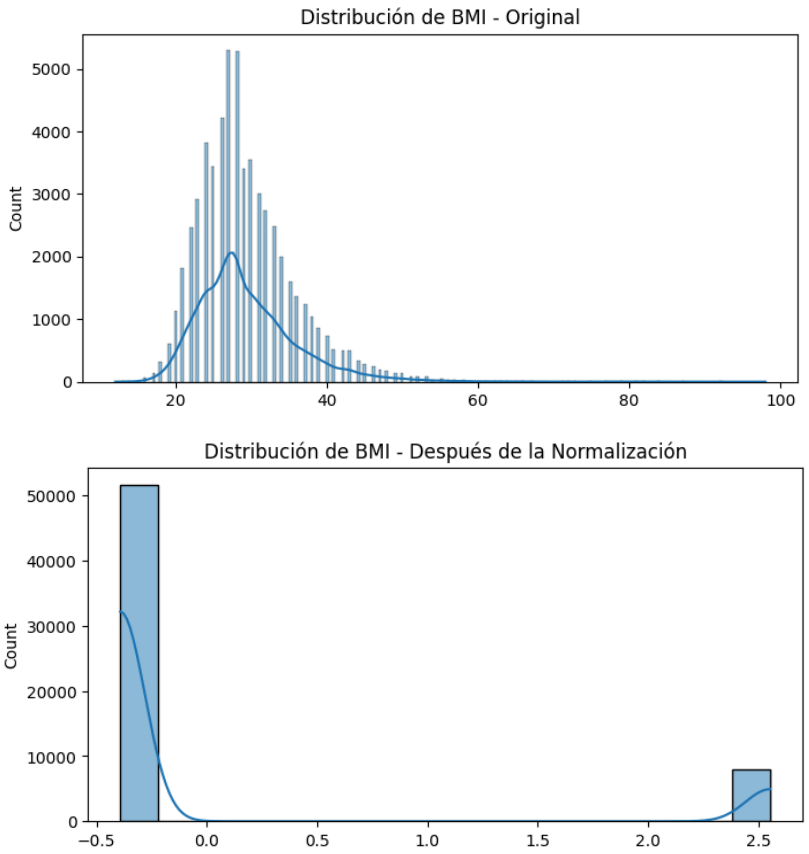

Imagen 7: Original vs Normalizado

Discretización

Por último, he discretizado la variable BMI en 4 categorías (Bajo, Moderado, Alto, Muy Alto) utilizando la técnica KBinsDiscretizer con estrategia de cuantiles. Esto lo he hecho para ver cómo se distribuye la variable objetivo en función del rango de índice de masa corporal.

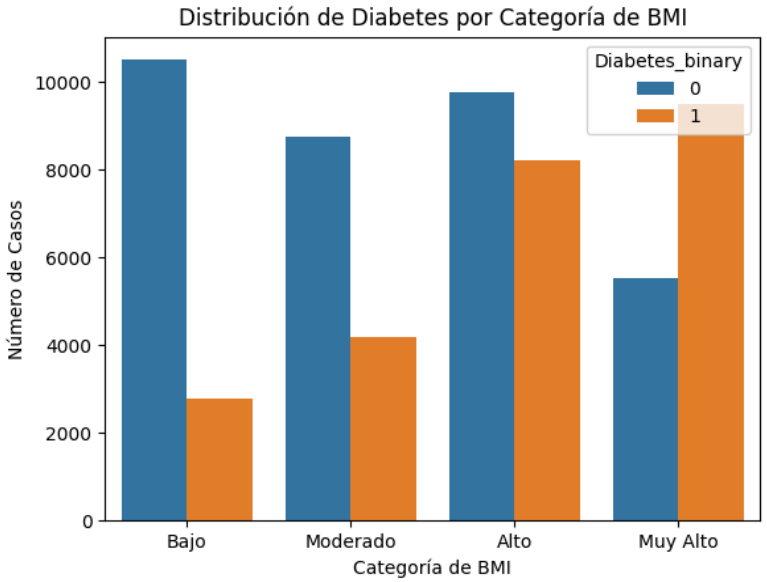

Imagen 8: Distribución de diabetes (ejemplo con BMI)

En el gráfico, podemos observar claramente que, a medida que aumenta el BMI, también lo hace la proporción de personas con diabetes. En la categoría 'Muy Alto', el número de casos con diabetes supera al de los casos sin diabetes, lo que sugiere una fuerte asociación entre un alto índice de masa corporal y la presencia de diabetes.(ya lo habíamos visto en la matriz de correlación).

Clasificación

Ahora voy a entrenar y evaluar distintos modelos de clasificación para predecir la variable objetivo Diabetes_binary. Antes de entrenar los modelos, hago una división del conjunto de datos
Primero, divido los datos escalados (X_scaled) y la variable objetivo en tres subconjuntos: entrenamiento, validación y prueba. Para ello, utilizo la función train_test_split

- Divido en un 80% para entrenamiento y validación
- Un 20% para prueba
- Luego, ese 80% lo vuelvo a dividir en 75% para entrenamiento y 25% para validación.

Después, aplico SMOTE sobre el conjunto de entrenamiento para generar ejemplos sintéticos de la clase minoritaria (personas con diabetes) para equilibrar las clases (ya que como he mostrado al principio, las clases estaban desbalanceadas). Esto es importante porque los modelos tienden a rendir peor con clases desbalanceadas.
Antes de SMOTE, el conjunto de entrenamiento contenía 20.944 casos negativos y 14.776 positivos. Tras aplicar SMOTE, ambos grupos tienen 20.944 ejemplos, es decir, quedan balanceados.

Una vez hecho esto, defino y entreno cuatro modelos distintos:

- Regresión Logística
- Random Forest
- Gradient Boosting

Para evaluar el rendimiento de los modelos, utilizo varias métricas: accuracy, AUC, recall, precision, f1 score, specificidad (calculada a partir de la matriz de confusión) y balanced accuracy, para entender mejor cómo se comporta cada modelo, sobre todo teniendo en cuenta el desbalanceo original del dataset.

Resultados obtenidos:

- Regresión Logística muestra un buen equilibrio entre sensibilidad y especificidad, con una AUC de 0.8221 y una balanced accuracy de 0.7446. Esto indica que el modelo distingue razonablemente bien entre clases.
- Random Forest tiene una precisión parecida, aunque con una menor sensibilidad (recall = 0.6648). Esto quiere decir que identifica menos casos positivos correctamente que otros modelos. Su AUC es 0.7824.

- Gradient Boosting logra los mejores resultados generales, con una accuracy de 0.7503 y una AUC de 0.8274. Tiene un buen equilibrio entre precisión y recall, y además su especificidad es alta (0.7793).

Por último, represento visualmente las curvas ROC para cada modelo, para comparar gráficamente la tasa de verdaderos positivos frente a falsos positivos

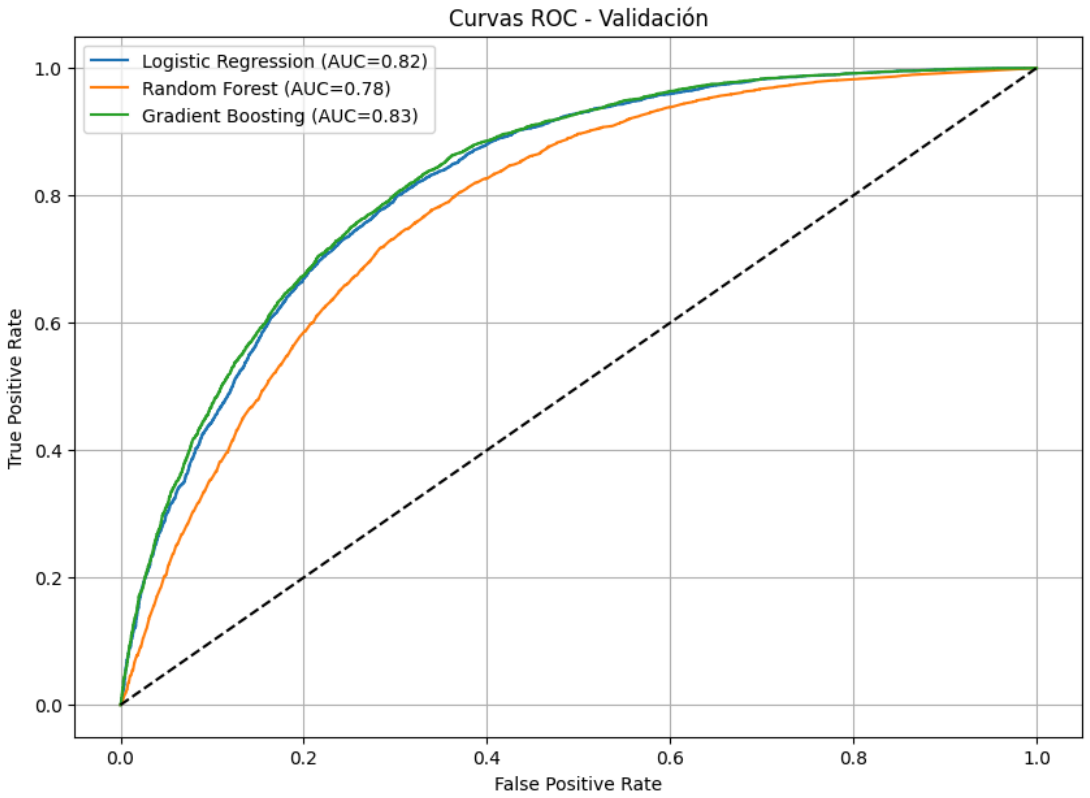

Imagen 9: Curvas ROC

Conclusión

Gradient Boosting ha sido el que ha obtenido el mejor rendimiento global, con buenos valores en métricas como AUC y balanced accuracy. Después, Regresión Logística ha tenido también buenos resultados, sobre todo en sensibilidad. Por último, Random Forest ha tenido menor recall, lo que puede afectar en contextos donde es crítico detectar positivos, por lo que lo clasificaría como el que ha dado peores resultados.
Para acabar, mencionar que el uso de SMOTE creo que ha sido bastante importante para mejorar la detección de la clase minoritaria en todos los modelos.

URL al código fuente ejecutable y ejecutado

https://colab.research.google.com/drive/1WCBIhIfpT5MAHJSkIFYeTW0K6ZEcH44K?usp=sharing

Bibliografía

- EDA: https://www.aprendemachinelearning.com/analisis-exploratorio-de-datos-pandas-python/
- Pre-Procesamiento de Datos en Python: https://www.youtube.com/watch?app=desktop&v=bY7OIJvTMrE
- Preprocesamiento de datos: Una guía completa con ejemplos en Python: https://www.datacamp.com/es/blog/data-preprocessing
- SMOTE para Clasificación Desbalanceada en Python: https://iartificial.blog/aprendizaje/smote-clasificacion-desbalanceada-python/
- Información de los modelos: https://medium.com/@nischitasadananda/the-battle-between-logistic-regression-random-forest-classifier-xg-boost-and-support-vector-46d773c70f41