Primer entregable Trabajo Final (TB1)

Alumno: Elmer augusto Riva Rodriguez

1. Tema elegido

Desarrollo de un modelo predictivo de Machine Learning para la detección temprana de Diabetes Mellitus.

Este proyecto se enmarca en el área de la salud predictiva y utilizará la metodología CRISP-DM para desarrollar una solución de Data Science. El enfoque es aplicar técnicas de machine learning para construir un modelo capaz de identificar patrones en datos médicos y de estilo de vida que indiquen la probabilidad de que una persona padezca diabetes. El fin último es crear una herramienta que pueda servir de apoyo para la detección temprana de esta enfermedad crónica.

2. Objetivo del proyecto

Construir un modelo de machine learning de clasificación supervisada utilizando el "Diabetes Prediction Dataset" para predecir la presencia de diabetes, logrando una precisión superior al 90% en el conjunto de datos de prueba y entregando el informe final asociado al desarrollo del modelo antes del 22 de noviembre de 2025

3. Dataset seleccionado

He seleccionado el dataset titulado “Diabetes prediction dataset” del repositorio Kaggle, y a continuación presento los detalles del mismo:

Fuente: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

Descripción: Este dataset contiene información médica, demográfica y de estilo de vida de 100,000 pacientes, recopilada con el propósito de predecir la presencia de diabetes. Asimismo, el publicador del dataset, indica que es un conjunto de datos estructurado, ideal para un problema de clasificación binaria.

Tamaño: 100,000 registros y 9 columnas.

Variables Principales:

- gender: Género del paciente.
- age: Edad del paciente.
- hypertension: 0 si no tiene hipertensión, 1 si la tiene.
- heart_disease: 0 si no tiene enfermedades cardíacas, 1 si las tiene.
- smoking_history: Historial de tabaquismo del paciente.
- bmi: Índice de Masa Corporal.
- HbA1c_level: Nivel de hemoglobina A1c, un indicador clave para el monitoreo de la glucosa.

- blood_glucose_level: Nivel de glucosa en la sangre.
- diabetes (Variable Objetivo): 0 si el paciente no tiene diabetes, 1 si tiene diabetes.

4. Enlace al Repositorio

Enlace: https://github.com/elmer-riva/Deteccion_Temprana_Diabetes

5. Asignación de Roles y Tareas

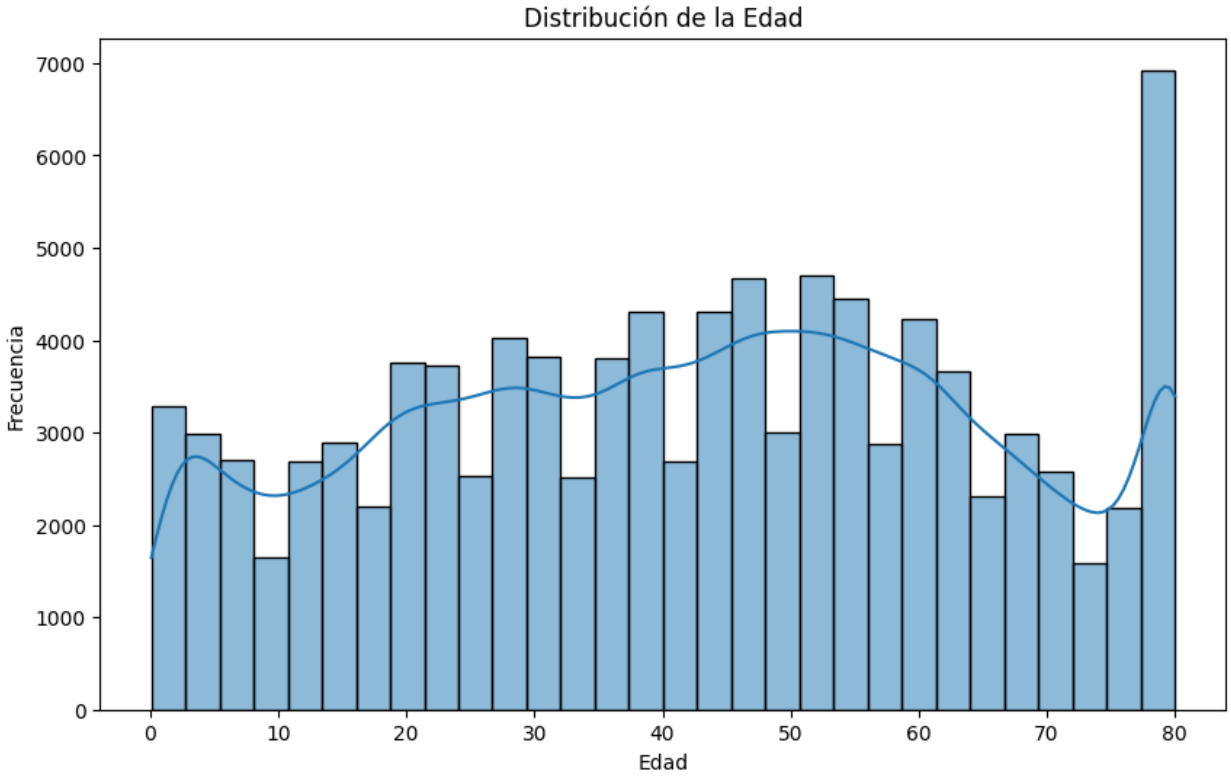
Integrante: Elmer Augusto Riva Rodrioguez

Roles Asumidos:

- Project Lead: Definición del alcance del problema, planificación de tareas y elaboración del informe final.
- Data Engineer: Identificación y propuesta de tratamiento para datos duplicados y outliers.
- Data Analyst: Ejecución del análisis exploratorio, generación de estadísticas descriptivas y creación de las visualizaciones clave.
- Data Scientist: Interpretación de las relaciones entre variables y evaluación de la idoneidad del dataset para un futuro modelo predictivo.

6. Tres visualizaciones significativas

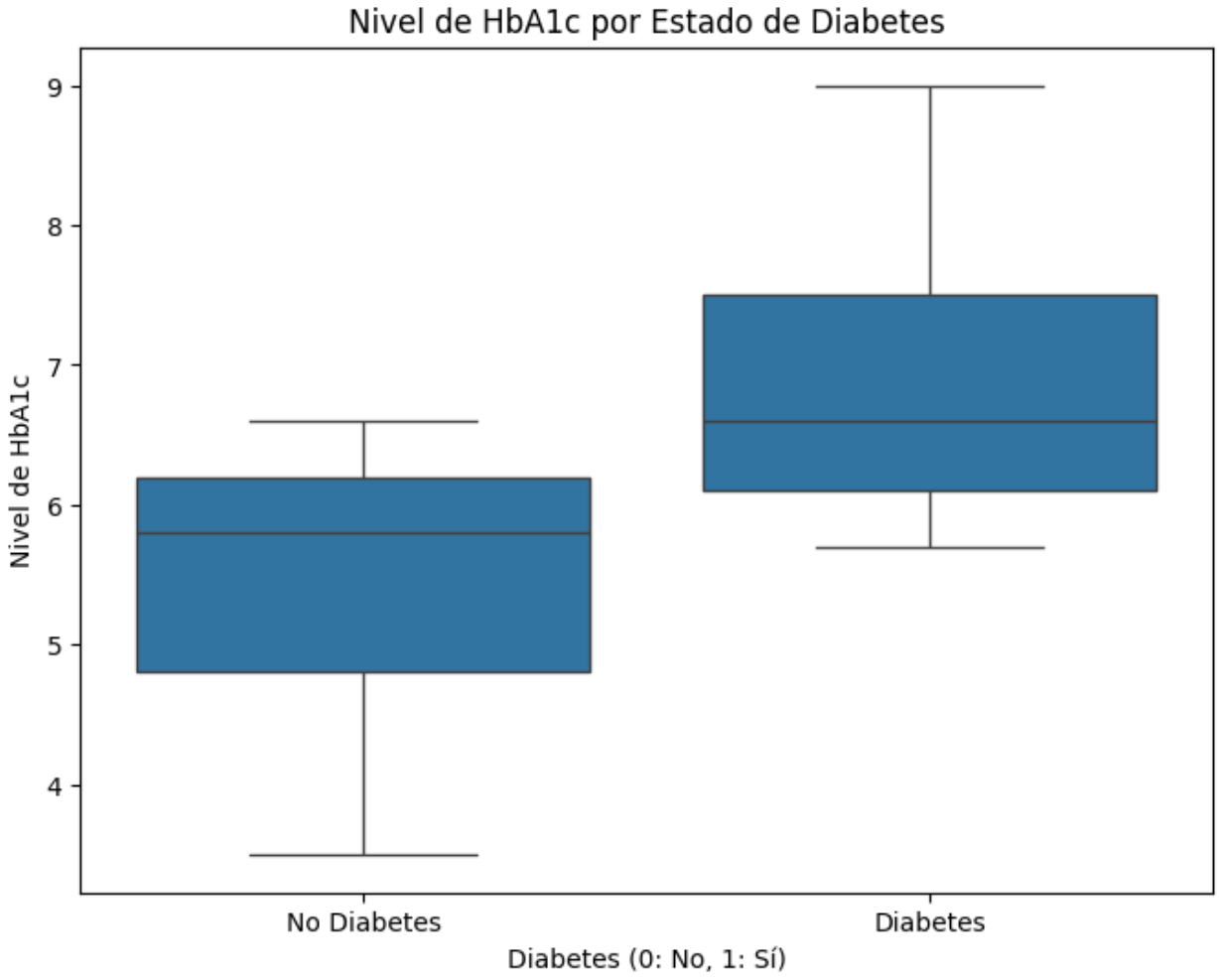
Distribución de la Edad (Histograma)



Este gráfico nos muestra cómo se distribuye la edad en el dataset. Podemos ver que hay una amplia variedad de edades, con un pico notable en las edades más jóvenes y otra

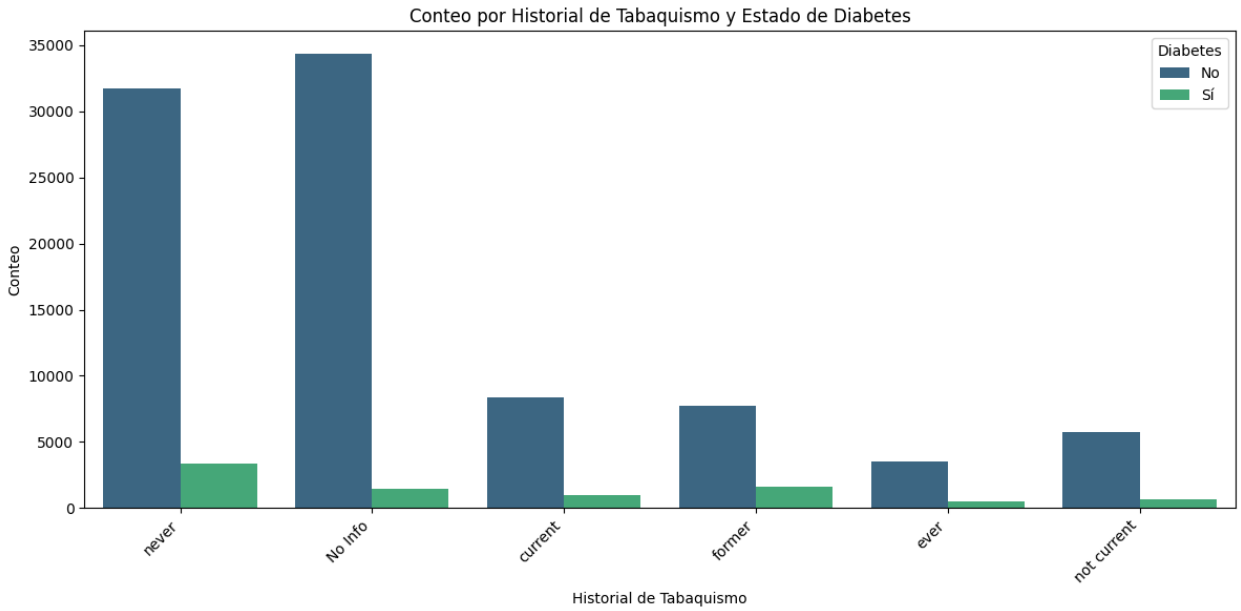
concentración en las edades adultas mayores. La forma general nos da una idea de la composición por edad de la población estudiada.

Nivel de HbA1c por Estado de Diabetes (Box Plot)



Este diagrama de caja es muy revelador. Muestra claramente que las personas con diabetes (caja de la derecha, marcada como 'Sí' o 1) tienen niveles de HbA1c significativamente más altos en comparación con las personas sin diabetes (caja de la izquierda, marcada como 'No' o 0). Los niveles de HbA1c son un indicador clave del control del azúcar en sangre a largo plazo, por lo que esta diferencia es esperada y refuerza la importancia de esta variable para predecir la diabetes.

Conteo por Historial de Tabaquismo y Estado de Diabetes (Countplot)



Este gráfico nos permite comparar la cantidad de personas con y sin diabetes dentro de cada categoría de historial de tabaquismo. A simple vista, parece haber una proporción mayor de casos de diabetes entre las personas que fuman actualmente ('current') o que fumaron en el pasado ('former'). Sin embargo, también es muy importante notar la gran cantidad de personas en la categoría 'No Info'. Esto significa que una parte considerable de la información sobre el historial de tabaquismo no está disponible, lo cual limita la solidez de las conclusiones que podemos sacar sobre la relación entre tabaquismo y diabetes basándonos solo en este gráfico. Se necesitaría una estrategia para manejar estos datos faltantes ('No Info') para un análisis más preciso.

7. Identificación y manejo de calidad de datos

En esta sección, realicé un análisis para identificar posibles problemas en la calidad de los datos que podrían afectar mi análisis predictivo.

7.1. Valores Faltantes:

Procedí a calcular la suma de valores nulos para cada columna del dataset (df.isnull().sum()). Mis resultados mostraron que no hay valores faltantes en ninguna de las columnas. Este es un hallazgo positivo ya que no requeriré técnicas de imputación para datos nulos en este dataset.

7.2. Duplicados:

Para detectar filas exactamente duplicadas, utilicé el método duplicated().sum(). Encontré que hay 3,854 filas duplicadas en el dataset. Si bien la presencia de duplicados puede indicar errores en la recolección de datos o ser intencional dependiendo del contexto, para un modelo predictivo, las filas duplicadas pueden sesgar el entrenamiento. Una estrategia común que consideraría sería eliminar estas filas duplicadas para asegurar que cada

observación represente un individuo único, a menos que haya una justificación explícita para mantenerlos (lo cual no parece ser el caso aquí).

7.3. Outliers:

Visualicé la distribución de las variables numéricas (age, bmi, HbA1c_level, blood_glucose_level) utilizando box plots para identificar posibles outliers.

- En la variable age, no observé outliers significativos.
- En bmi (Índice de Masa Corporal) y blood_glucose_level (Nivel de Glucosa en Sangre), identifiqué un número considerable de puntos que se extienden más allá de los "bigotes" de los box plots, indicando la presencia de outliers. Estos valores representan individuos con IMC muy alto (obesidad extrema) y niveles elevados de glucosa en sangre, lo cual es plausible y relevante en un contexto de estudio de diabetes.

Dado que estos outliers parecen ser valores reales y no errores de entrada, eliminarlos por completo podría resultar en la pérdida de información valiosa, especialmente porque los valores altos de bmi y blood_glucose_level son predictores importantes de diabetes.

7.4. Propuestas para el manejo de Outliers:

Aunque no ejecuté una manipulación directa de los outliers en esta fase exploratoria, propongo las siguientes estrategias para considerar en etapas posteriores del preprocesamiento de datos, antes de construir un modelo predictivo:

a. Truncamiento: En lugar de eliminar filas, podría "limitar" los valores extremos. Esto implica establecer un umbral (por ejemplo, basado en un percentil, como el 95% o 99%) y reemplazar cualquier valor por encima de ese umbral con el valor del umbral mismo. Esta técnica reduce la influencia de los valores extremos sin descartar los datos por completo.

b. Transformación de variables: Aplicar transformaciones matemáticas a las variables con outliers (como la transformación logarítmica) puede ayudar a que su distribución sea más simétrica y reducir el impacto de los valores extremos en algunos modelos.

c. Reemplazo por media/mediana/moda: Como una técnica muy simple, podría considerar reemplazar los valores atípicos detectados con la media la mediana de la variable correspondiente. Sin embargo, debo tener en cuenta que esto puede distorsionar la distribución original y reducir la variabilidad, por lo que debe usarse con cuidado, especialmente si hay muchos outliers.

8. Análisis del dataset en relación al problema y limitaciones

El dataset de predicción de diabetes que he analizado parece responder en gran medida al problema planteado, que es predecir la probabilidad de que un individuo tenga diabetes basándose en diversas características de salud. La inclusión de variables como la edad es

crucial, ya que el riesgo de diabetes tipo 2 aumenta con la edad. El IMC (índice de masa corporal) es otro factor de riesgo bien establecido, donde un IMC elevado se asocia con una mayor probabilidad de desarrollar resistencia a la insulina y diabetes. Los niveles de HbA1c y glucosa en sangre son indicadores directos del control del azúcar en el cuerpo; niveles elevados son marcadores clave de prediabetes o diabetes ya establecida. Además, la presencia de hipertensión y enfermedad cardíaca** son comorbilidades comunes que a menudo se presentan junto con la diabetes o son factores de riesgo relacionados. La variable objetivo, diabetes, está claramente definida (0 para no diabetes, 1 para diabetes), lo que permite el entrenamiento de modelos de clasificación para predecir esta condición. Por lo tanto, el dataset contiene los atributos fundamentales necesarios para abordar el objetivo principal de la predicción de diabetes.

Sin embargo, a pesar de ser un buen punto de partida, el dataset presenta algunas limitaciones importantes en este contexto que observé durante mi análisis:

a. Información incompleta sobre el historial de Tabaquismo: Una limitación significativa que encontré es la gran cantidad de registros con el valor 'No Info' en la columna `smoking_history`. Aproximadamente el 35.8% del dataset carece de información sobre si la persona fuma o ha fumado. El historial de tabaquismo es un factor de riesgo conocido para muchas condiciones de salud, incluida la diabetes. La falta de esta información en una porción tan grande del dataset podría limitar la capacidad de un modelo predictivo para capturar completamente la influencia del tabaquismo en el riesgo de diabetes, potencialmente llevando a predicciones menos precisas para los individuos en esta categoría o sesgando los resultados si decido ignorar esta variable o imputarla de forma inadecuada.
b. Ausencia de información longitudinal o historial completo: El dataset proporciona una instantánea de las métricas de salud en un momento dado. No incluye información sobre el historial médico completo del individuo, como diagnósticos previos de prediabetes, la duración de la hipertensión o enfermedades cardíacas, cambios en el estilo de vida a lo largo del tiempo, o el progreso de los niveles de glucosa/HbA1c. Para una predicción más robusta y personalizada de la diabetes, especialmente para identificar el riesgo de desarrollar la enfermedad en el futuro (no solo el estado actual), sería muy valioso contar con datos longitudinales o un historial más detallado. La falta de esta información temporal limita mi capacidad de analizar tendencias y el impacto acumulado de los factores de riesgo a lo largo del tiempo.