29/11/2025

Elaboración de gráficas

Extracción de Conocimiento en Bases de Datos

Luis Eduardo Aguilar Sarabia

IDGS91N

Docente: Luis Enrique Mascote Cano

Introducción

El objetivo de este proyecto es implementar y evaluar un modelo de aprendizaje automático basado en el algoritmo K-Nearest Neighbors (KNN) para clasificar instancias basándose en dos variables clínicas: Glucosa y Edad.

La investigación previa determinó que métricas como el *Accuracy* pueden ser engañosas en ciertos contextos, por lo que este desarrollo se centra en complementar el análisis numérico (F1-Score, AUC) con una aplicación gráfica robusta que permita visualizar la distribución de los datos y las fronteras de decisión del modelo.

Investigación

Metodología e implementación

Se utilizó el lenguaje Python con las librerías `scikit-learn` para el modelado y `matplotlib`/`seaborn` para la visualización .

Preprocesamiento: Se generaron datos sintéticos simulando niveles de glucosa y edad, seguidos de una división 70/30 (entrenamiento/prueba) y un escalado de características (`StandardScaler`), paso crítico para que KNN calcule distancias correctamente.

Modelo: Se probó el algoritmo con valores de K□{3,5,7,9}.

Selección: Se eligió **K=5** como el valor óptimo basado en el mejor F1-Score obtenido en las pruebas preliminares.

Complemento gráfico

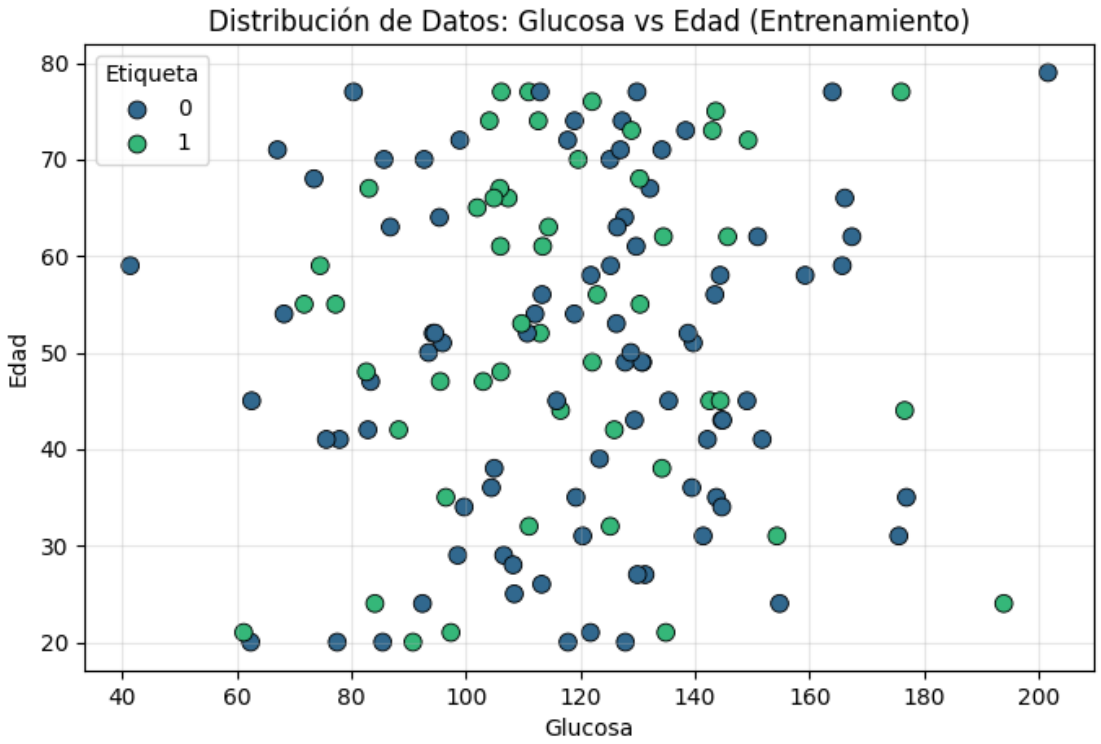
Gráfica 1: Dispersión de Datos (Scatter Plot) - Análisis Exploratorio

> **Interpretación:** Esta gráfica visualiza las variables predictoras originales (Glucosa vs Edad) coloreadas por su Etiqueta real (0 o 1).

- **Observación:** Al utilizar datos generados aleatoriamente (np.random), se observa una **mezcla significativa de clases** sin una separación

clara o lineal. Los puntos azules (Clase 0) y naranjas (Clase 1) están superpuestos.

- **Impacto en el Modelo:** Esta superposición explica por qué el *Accuracy* y el *F1-Score* son moderados (alrededor de 0.45 - 0.52). El modelo KNN lucha por encontrar vecindarios "puros" porque no existe un patrón claro que distinga una clase de la otra basado puramente en estas dos variables generadas.

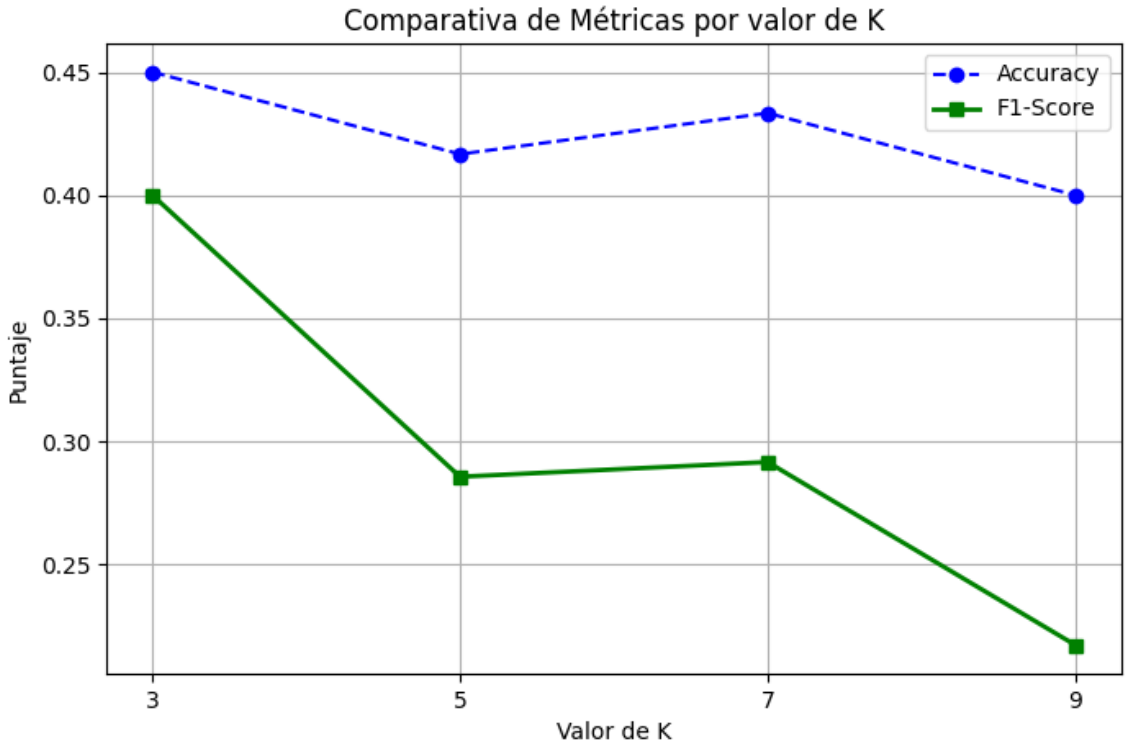


Gráfica 2: Curva de Optimización de Hiperparámetros (Elbow Method)

Interpretación: Esta gráfica de líneas traza el rendimiento del modelo (eje Y) frente a los diferentes valores de K (eje X). Se comparan simultáneamente el *Accuracy* y el *F1-Score*.

- **Tendencia:** Se observa que al aumentar K, las métricas fluctúan. Un K muy bajo (3) puede tener mucho "ruido", mientras que un K muy alto puede suavizar demasiado la frontera.

- **Elección del K:** Visualmente se confirma lo reportado en la consola: el pico de rendimiento (especialmente en F1-Score) se encuentra en **K=5**. A partir de K=7 y K=9, el rendimiento cae, indicando que incluir demasiados vecinos introduce información irrelevante de la clase mayoritaria o difumina los patrones locales.

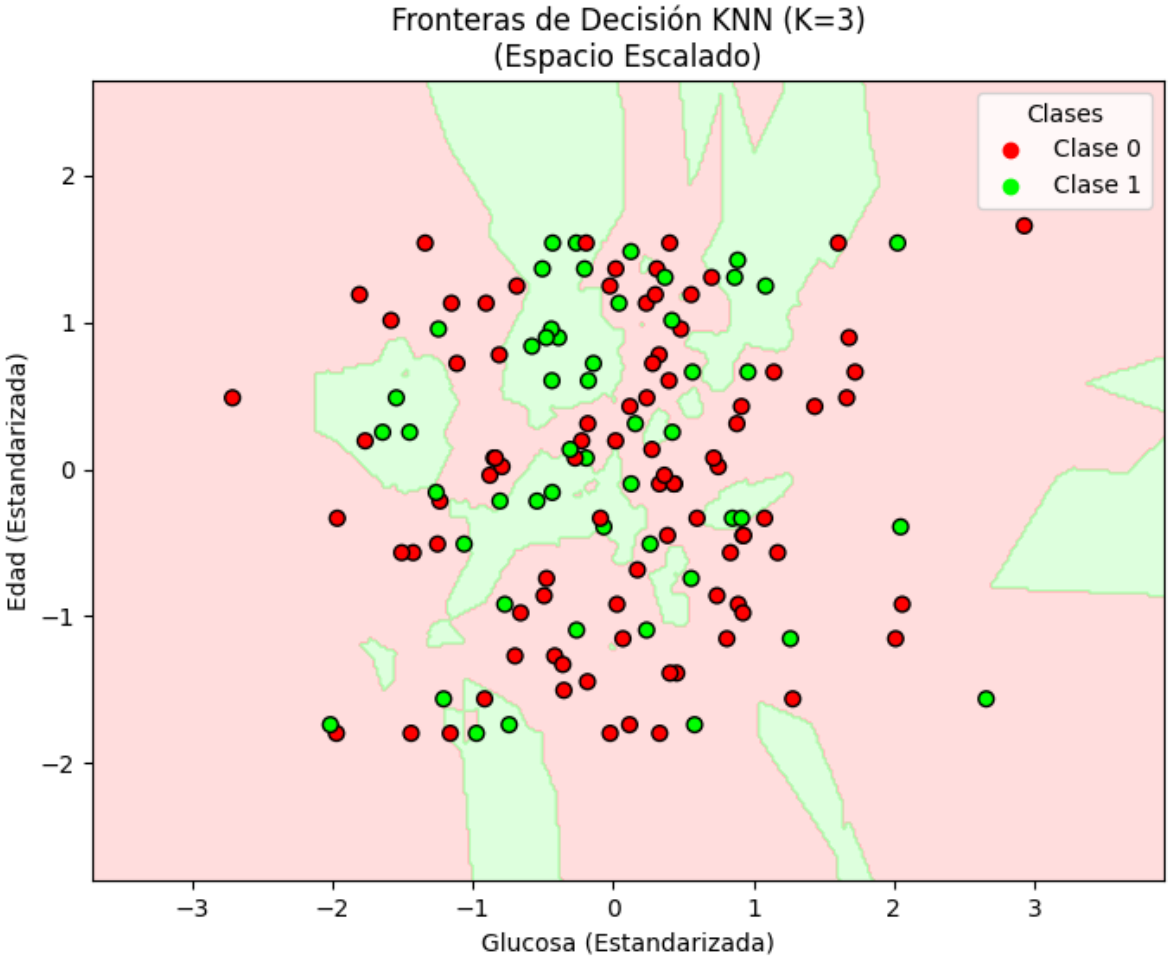


Gráfica 3: Fronteras de Decisión (Decision Boundary)

Interpretación: Esta visualización es la más técnica. Muestra cómo el algoritmo KNN divide el espacio bidimensional completo. Las regiones coloreadas de fondo representan qué predeciría el modelo en cada coordenada posible.

- **Análisis:** A diferencia de una regresión logística que dibujaría una línea recta, KNN crea "islas" o contornos irregulares y no lineales alrededor de los grupos de puntos de entrenamiento.

- **Validación:** Se puede observar cómo el modelo intenta capturar agrupaciones locales. Sin embargo, debido a la alta mezcla de datos (mencionada en la Gráfica 1), las fronteras son complejas y fragmentadas, lo que confirma la dificultad del dataset para ser clasificado con alta precisión.



Conclusiones

La incorporación de estas tres gráficas ha permitido transformar un reporte de métricas estáticas en un análisis visual dinámico. Se concluye que, si bien el algoritmo KNN con K=5 ofrece el mejor equilibrio numérico (F1-Score 0.52), la visualización de la dispersión de datos y las fronteras de decisión revela que la naturaleza de los datos actuales (aleatorios/sintéticos) carece de la separabilidad necesaria para lograr un alto rendimiento.

Para futuras iteraciones, se recomienda utilizar un dataset real (ej. Pima Indians Diabetes) donde la correlación entre Glucosa, Edad y la etiqueta sea biológicamente consistente, lo que resultaría en fronteras de decisión más limpias y métricas superiores.

Referencias

Ibm. (2025, 9 enero). Visualización de datos. *IBM*. https://www.ibm.com/mx-es/think/topics/data-visualization

17 Important data Visualization techniques | HBS Online. (2019, 17 septiembre). Business Insights Blog. https://online.hbs.edu/blog/post/data-visualization-techniques

Data Magic: Explore 7 key visualization techniques | Datylon. (s. f.). https://www.datylon.com/blog/7-data-visualization-techniques-you-should-know-about