**Ejercicio: Predicción de Diabetes con Machine Learning**

Un centro médico de investigación quiere desarrollar un modelo que **prediga si una paciente tiene diabetes** en base a 8 variables clínicas:

* Número de embarazos
* Glucosa en sangre
* Presión arterial
* Espesor de piel
* Insulina
* Índice de masa corporal (IMC)
* Función hereditaria de diabetes
* Edad

El objetivo es clasificar a cada paciente como:

* **1:** Tiene diabetes
* **0:** No tiene diabetes

Usaremos el **Pima Indians Diabetes Dataset**, disponible en Kaggle: <https://www.kaggle.com/datasets/mathchi/diabetes-data-set>

**Instrucciones**

1. **Análisis Exploratorio de Datos**

* Conocer el dataset antes de entrenar el modelo y eliminar cualquier dato atipico que no tenga sentido para la clasificación

1. **Preprocesamiento**

* Aplicar tecnicas como escalamiento, conversion de variables categoricas a variables numéricas, eliminar nulos y duplicados, etc.

---

1. **Separar datos en entrenamiento y prueba**

* Definir **X** (todas las columnas excepto la etiqueta) y **y** (la columna Outcome, que indica diabetes).
* Usar train\_test\_split de sklearn.model\_selection para dividir el dataset:

---

1. **Entrenamiento**

* Elegir un modelo de clasificación inicial. Opciones sugeridas:
* **Regresión Logística**
  + **Random Forest**
  + **K-Nearest Neighbors (KNN)**
* Entrenar el modelo con los datos de **entrenamiento** usando .fit().

---

1. **Predicción**

* Usar el modelo entrenado para predecir el conjunto de prueba (X\_test) con .predict().
* Guardar las predicciones en una variable y\_pred.

---

1. **Evaluación**

* Calcular las métricas principales usando sklearn.metrics

---

**Entrega**

* Un notebook de Python (Google Colab o Jupyter) con cada punto desarrollado, el cual se debe subir al repositorio de github trabajado durante el bootcamp.
* Deben incluir comentarios explicando brevemente lo que hacen en cada celda