INFORME FINAL DEL PROYECTO

Curso: Programación PRG-2025-1
Título: Predicción de ingresos con datos del censo de EE.UU.

1. ¿De qué trata este proyecto?

En este proyecto usamos programación y ciencia de datos para **predecir si una persona gana más o menos de $50,000 dólares al año**, basándonos en datos reales del censo de Estados Unidos.

Este ejercicio es parte del trabajo final del curso y sirve para aplicar todo lo que aprendimos sobre manejo de datos, preprocesamiento, y entrenamiento de modelos.

2. ¿Qué datos usamos?

Utilizamos un dataset llamado **Adult Dataset**, que está disponible en internet en la página de UCI Machine Learning Repository.

Este archivo tiene información de casi **49 mil personas**, con datos como:

- Edad
- Nivel educativo
- Estado civil
- Ocupación
- Horas trabajadas por semana
- País de origen
- Si gana más de $50K o no

Este último dato es lo que queríamos predecir (la **variable objetivo**).

3. ¿Qué hicimos con los datos?

Primero hicimos un análisis general para conocer los datos: ver los tipos de columnas, si había valores faltantes, y cómo se distribuían las edades, horas trabajadas, etc.

Después, limpiamos el dataset:

- Quitamos las filas que tenían datos faltantes (como ocupación vacía)
- Convertimos los textos (como “Private” o “Male”) en números para que los modelos pudieran entenderlos
- Normalizamos las columnas numéricas (edad, horas, ingresos, etc.)

Guardamos todo eso en un archivo final llamado **06 - dataset.csv**, que usamos en los modelos.

4. ¿Qué modelos usamos?

Probamos dos modelos para hacer predicciones:

🧠 Modelo 1: Random Forest

- Es un modelo basado en varios árboles de decisión
- Precisión: **85.2%**
- Se entrenó con: 200 árboles, profundidad máxima de 20

🤖 Modelo 2: Support Vector Machine (SVM)

- Es un modelo que separa los datos con líneas (o curvas)
- Precisión: **84.4%**
- Usamos un kernel RBF y valor C=10

5. ¿Cuál funcionó mejor?

Ambos funcionaron bien, pero **Random Forest fue un poco mejor** en precisión y velocidad.

También medimos qué tanto se parecían las predicciones entre ambos modelos usando una métrica llamada **Cohen's Kappa**, que nos dio:

- **Kappa ≈ 0.71**, lo que indica buena coincidencia entre los dos modelos

6. Conclusiones

- El modelo puede predecir con buena precisión si alguien gana más de $50K o no.
- Limpiar y preparar los datos fue fundamental para que todo funcionara.
- Random Forest fue más estable y rápido que SVM.
- Aprendimos a trabajar con datos reales y a entrenar modelos en Python.

7. Integrantes

- SANTIAGO OSORIO CAICEDO