# Informe de Análisis de Datos con Orange - Adult Income Dataset

## 1. Exploración de Datos

El conjunto de datos utilizado es Adult Income Dataset, que contiene información socioeconómica de individuos, incluyendo edad, nivel educativo, ocupación, género, horas trabajadas por semana, entre otros.  
  
Objetivo: Predecir si una persona gana más o menos de 50.000 dólares anuales.

Insights iniciales:

* - El dataset incluye tanto variables categóricas como continuas.
* - Algunas columnas relevantes: age, education, occupation, race, sex, hours-per-week, income.
* - Se detectaron posibles valores faltantes o categorías ambiguas, por ejemplo: '?' en workclass o native-country.

## 2. Preprocesamiento de Datos

Para preparar los datos antes del modelado, se aplicaron los siguientes pasos:

1. 1. Selección de columnas relevantes mediante el nodo Select Columns.
2. 2. Edición del dominio con Edit Domain para renombrar o recodificar categorías.
3. 3. Preprocesamiento automático con Preprocess: imputación, codificación y normalización.
4. 4. Muestreo de datos con Data Sampler para dividir en entrenamiento y prueba.

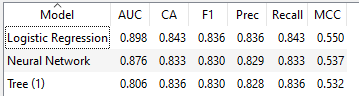
Justificación del preprocesamiento:

Estos pasos son fundamentales para mejorar la calidad de los datos y permitir que los algoritmos de aprendizaje automático funcionen correctamente. La codificación y normalización permiten comparar escalas y trabajar con modelos sensibles a la estructura de los datos.

## 3. Comparación entre Modelos

Métricas consideradas:

* - Accuracy (Precisión): porcentaje de predicciones correctas.
* - AUC (Área bajo la curva ROC): capacidad del modelo para distinguir entre clases.
* - F1 Score: equilibrio entre precisión y recall, útil en clases desbalanceadas.
* - Matriz de confusión: para interpretar falsos positivos y negativos.



Todos los modelos paracen ser bastante precisos ya que presentan valores de las métricas AUC, CA, F1, Prec, Recall y MCC mayores a 0,8. El mejor modelo es el logístico ya que tiene el valor mas alto en todas las métricas mencionadas.

AUC:

Regresión Logística tiene el AUC más alto (0.898), lo que indica excelente discriminación. La Red Neuronal también tiene buen AUC (0.876), pero inferior. El Árbol de Decisión está por debajo (0.806), lo que indica menor capacidad de separación entre clases

CA: Todas tienen valores parecidos e indican un porcentaje entre 83,6% y 84,3% predicciones correctas.

F1: En este caso es útil mirarlo ya que hay un gran desbalance entre los datos. Los tres modelos tienen valores de F1 altos.

MCC: Las tres matrices indican valores entre 5,3 y 5,56. Esto nos indica que que los modelos son robustos a pesar de los desbalances.