# Clasificación de ingreso anual mayor a 50k

José Pablo Martínez Valdivia

7 de septiembre de 2024

**Abstract**

This paper presents a predictive model based on the Random Forest algorithm to classify income levels using the Census Income (50k) dataset. The objective of the study is to predict whether an individual's income exceeds $50,000 per year based on demographic and work-related attributes. Furthermore, I discuss the various preprocessing decisions made, such as handling encoding categorical variables, and the choice of hyperparameters to optimize model performance.

## 1. Introducción

El dataset "Census Income" fue proveido por el "UCI Machine Learning Repository". Contiene 14 variables y 48842 instancias de datos provenientes de un censo realizado en 1994 para determinar si un adulto genera un ingreso anual mayor a $50k. Emplearemos el modelo de bosque aleatorio (random forest) para emplear una predicción categórica y disminuir el overfit en el modelo.

## 2. ETL

### 2.1. Extracción

Los datos fueron descargados por medio de la librería de "ucimlrepo". Estos proveen datos demográficos como edad, estado civil, educación, ocupación, raza, sexo, pais de origen, etc.

Los datos se encuentran completos en su mayoría, pero hay datos faltantes en las columnas de tipo de trabajo, ocupación y pais de origen como se puede ver en 1. Se tomó la decision de deshacernos de las filas con datos faltantes las cuales representan 1221 instancias, lo cual es el 2.4 % de los datos; sobrándonos 47621 instancias. Esta decision se tomó considerando que algún tipo de inputación podría introducir errores para la taréa de clasificación.

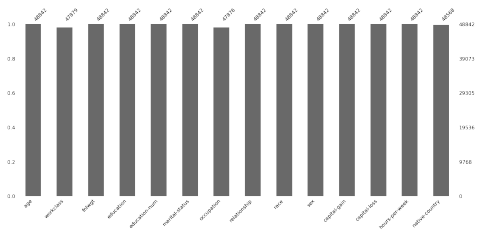

Figura 1: Valores faltantes.

### 2.2. Transformación

Para poder general el bosque aleatorio necesitamos convertir las columnas categóricas en valores numéricos. Para esto haremos uso de "one-hot encoding", esta técnica consta de tomar todas las categorías en una columna y crear una columna por categoría, dejando un uno en la categoría que prsenta la instancia y ceros en el resto.

| Colour |
|---|
| Green |
| Red |
| Blue |

| Green | Red | Blue |
|---|---|---|
| 0 | 1 | 1 |
| 1 | 1 | 1 |
| 1 | 0 | 1 |
| 0 | 0 | 0 |
| 0 | 1 | 0 |

Figura 2: Método de one-hot encoding.