# **Income Classification Model**

## **Introduction**

The income dataset was extracted from the 1994 U.S. Census database.

## **The importance of census statistics**

The census is a special, wide-range activity, which takes place once a decade in the entire country. The purpose is to gather information about the general population, in order to present a full and reliable picture of the population in the country - its housing conditions and demographic, social and economic characteristics. The information collected includes data on age, gender, country of origin, marital status, housing conditions, marriage, education, employment, etc.

This information makes it possible to plan better services, improve the quality of life and solve existing problems. Statistical information, which serves as the basis for constructing planning forecasts, is essential for the democratic process since it enables the citizens to examine the decisions made by the government and local authorities, and decide whether they serve the public they are meant to help.

## **Objective of the project**

The goal of this machine learning project is to predict whether a person makes over 50K a year or not given their demographic variation. To achieve this, several classification techniques are explored and the random forest model yields the best prediction result.

## **Features Description**

# **Attributes in the Dataset**

## **Categorical Attributes**

1. **Workclass**:
   * Categories: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
   * Description: Represents an individual’s work category.
2. **Education**:
   * Categories: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
   * Description: Indicates an individual’s highest education degree.
3. **Marital Status**:
   * Categories: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
   * Description: Represents an individual’s marital status.
4. **Occupation**:
   * Categories: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspect, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
   * Description: Indicates an individual’s occupation.
5. **Relationship**:
   * Categories: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
   * Description: Represents an individual’s relation within a family.
6. **Race**:
   * Categories: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
   * Description: Indicates an individual’s race.
7. **Sex**:
   * Categories: Female, Male
   * Description: Represents an individual’s gender.
8. **Native Country**:
   * Categories: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad Tobago, Peru, Hong, Holland-Netherlands
   * Description: Indicates an individual’s native country.

## **Continuous Attributes**

1. **Age**:
   * Type: Continuous
   * Description: Represents the age of an individual.
2. **Final Weight (fnlwgt)**:
   * Type: Continuous
   * Description: The weights on the CPS files are controlled to independent estimates of the civilian noninstitutional population of the US. These are prepared monthly by the Population Division at the Census Bureau.
3. **Capital Gain**:
   * Type: Continuous
   * Description: Represents capital gains (if applicable).
4. **Capital Loss**:
   * Type: Continuous
   * Description: Represents capital losses (if applicable).
5. **Hours per Week**:
   * Type: Continuous
   * Description: Indicates an individual’s working hours per week.