**US Census Income Prediction (1994 – 1995)**

**PROJECT OVERVIEW:**

This data set comprises weighted census data from the US Census Bureau's current population surveys conducted between 1994 and 1995.

Our task is to predict if a person makes over $50,000 a year or not based on several employment and demographic-related attributes like age, marital status, citizenship, stock dividends, family status, Migration status, income, etc.

**DATA DESCRIPTION:**

1. The dataset was collected from the *UCI Machine Learning Repository* - <https://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29>
2. There are approximately *1,99,523* instances in the data file and *99,762* in the test file (*2,99,285* instances), with *42* employment and demographic-related attributes.
3. The dataset contains both Numeric and Categorical variables.

|  |  |  |
| --- | --- | --- |
| COLUMN NAME | DESCRIPTION | DATATYPE |
| age | Age of the worker | Numeric |
| class\_worker | Class of worker | Categorical |
| det\_ind\_code | Industry code | Numeric |
| det\_occ\_code | Occupation code | Numeric |
| hs\_college | Enrolled in educational institution | Categorial |
| education | Level of education | Categorial |
| wage\_per\_hour | Wage per hour | Numeric |
| major\_ind\_code | Major industry code | Categorial |
| major\_occ\_code | Major occupation code | Categorial |
| hisp\_origin | Hispanic origin | Categorial |
| sex | Sex | Categorial |
| region\_prev\_res | Region of previous residence | Categorial |
| stock\_dividends | Dividends from stocks | Numeric |
| det\_hh\_fam\_stat | Detailed household and family status | Categorial |
| det\_hh\_summ | Detailed household summary in household | Categorial |
| union\_member | Member of a labor union | Categorial |
| mig\_chg\_msa | Migration code - change in MSA | Categorial |
| unemp\_reason | Reason for unemployment | Categorial |
| mig\_chg\_reg | Migration code - change in region | Categorial |
| full\_or\_part\_emp | Full- or part-time employment status | Categorial |
| capital\_losses | Capital losses | Numeric |
| state\_prev\_res | State of previous residence | Categorial |
| mig\_move\_reg | Migration code - move within region | Categorial |
|  | Tax filer status | Categorial |
| tax\_filer\_stat | Live in this house one year ago | Categorial |
| mig\_same | Migration - previous residence in sunbelt | Categorial |

**APPROACH:**

To predict if a person makes over $50,000 a year or not, we will follow the following approach:

1. **Understand the Problem:**

In this approach, we will try to understand the business problem or do a sort of pre-analysis where-in we will understand which variables are essential and ask questions to perform the analysis.

1. **Data Preparation and Cleaning:**

This step will clean the data by handling missing values, performing string manipulations, removing duplicates, and converting them into appropriate data types. In short, we will improve the data quality for our analysis.

1. **Exploratory Data Analysis:**

This step will discover patterns and anomalies present in our dataset. We also understand if there is any correlation between any variables.

1. **Feature Scaling and Feature Engineering:**

In this step, we will transform the raw data into suitable features for machine learning. We will perform “*one-hot*” and “*dummy encoding”* to convert the categorical features into numeric attributes.

Then, we will scale the data using “*Normalization*” and “*Standardization*” techniques.

Next, we will perform “*Principal Component Analysis*” (PCA) to remove the set of highly correlated variables and keep only those variables that convey most of the information.

1. **Applying Machine Learning Models:**

We will divide the dataset into training, testing, and validation dataset.

Since it is a classification problem, we will apply *Logistic Regression*, *Decision trees*, *Random Forest*, *Naïve Bayes*, *K Nearest Neighbors*, and *Support Vector Machine* algorithms.

1. **Model Performance and Evaluation:**

For evaluating the model performance, we will use *Confusion Matrix*, *AUC* and *ROC* curve, *Precision*, and *Recall.*