Census Income Classification

Group Member:

Atharva Avinash Gosavi
Yishtavi Gedipudi
Hrithik Puri
Shubham Pandya
Raghavi Dube

1. Goal

The project's goal is to create a classification model that, using census data, forecasts if a person's annual income surpasses $50,000 or not. The UCI Machine Learning Repository provided the dataset, which has 32,561 observations and 14 characteristics and 1 target variable. As a binary variable, the goal variable indicates whether or not the income level exceeds $50,000 (1) or (0). The project's objective is to accurately classify people into various income categories by utilizing machine learning techniques.

2. Dataset Description

2.1 Dataset Overview

The dataset includes information on age, education, gender, marital status, occupation, and more—14 features total, with 1 target variable. "Income level," the goal variable, makes a distinction between people who make less than or equal to $50,000 (0) and those who make more than $50,000 (1).

2.2 Data Exploration and Cleaning

To find any anomalies or missing values, the first step is to explore the dataset. To learn more about feature distributions and correlations, a thorough exploratory data analysis (EDA) will be carried out. The dataset's integrity will be preserved by addressing any anomalies or inconsistencies and properly handling missing values.

3. Approach

3.1 Feature Engineering

Feature engineering will be used to increase the model's predictive power. To get pertinent data, this entails developing new features or altering current ones. Grouping education levels, total capital (capital gain - capital loss), and age groupings, for instance, could offer more information.

3.2 Handling Categorical Variables

Appropriate encoding techniques, like one-hot encoding, will be used because categorical data, like education and employment, are present. This guarantees the model's ability to use and interpret these features in an efficient manner.

3.3 Model Selection

We'll use a binary classification strategy, taking into account random forests, decision trees, and logistic regression. The model that performs best will be selected based on how well it can handle the unique properties of the dataset.

3.4 Model Training and Evaluation

A portion of the dataset will be used to train the chosen model, and a different testing set will be used to assess its performance. We'll use metrics like accuracy and F1 score to evaluate the model's efficacy. Metrics like precision and recall will be especially important for evaluating the model because of the possible class imbalance.