FINAL PROJECT: INCOME LEVEL PREDICTION

Project timeline: 2 weeks (13th, June – 26th June)

Project deliverable: You are to submit a Jupyter notebook containing all your code and non-code tasks.

Project Overview

The aim of this project is to develop a machine learning model to predict whether a person's income exceeds $50K/yr based on census data. By leveraging various census data, you will explore the data, engineer relevant features, build and evaluate your predictive model, and provide insights from the data.

Feel free to be creative about the different ways you can make your model better.

1. Data Collection and Preparation

1. **Data Importation:**
    - Load the dataset containing the census data.
2. **Initial Data Inspection:**
    - Inspect the first few rows of the dataset to understand its structure.
    - Review data types and summary statistics to identify numerical and categorical variables and also convert variable to appropriate datatype.
    - Drop Irrelevant features
    - Check for missing values if any and handle them appropriately.

2. Exploratory Data Analysis (EDA)

- Generate at least 3 meaningful insights from the data with appropriate visualizations.

3. Data Preprocessing and Feature Engineering

1. **Handling Missing Values:**
    - Address missing values in the dataset if any, using appropriate imputation methods to ensure a complete dataset for analysis.

2. **Encoding Categorical Variables:**
    - Convert categorical variables into numerical format using label encoding or one-hot encoding to prepare them for machine learning algorithms.
3. **Feature Scaling:**
    - Standardize/normalize numerical features to ensure they are on a comparable scale, which can improve the performance of many machine learning algorithms.
4. You can also try out different feature engineering and preprocessing techniques like PCA feature selection etc.

4. Model Development

1. **Train-Test Split:**
    - Split the dataset into training and testing sets to evaluate the model's performance on unseen data.
2. **Model Selection and Training:**
    - Choose an appropriate machine learning algorithm (you can try out various algorithms and select the one the performs best)
    - Train the model on the training dataset and optimize hyperparameters for better performance.
3. **Model Evaluation:**
    - Evaluate the trained model on the testing dataset using metrics such as accuracy, precision, recall, F1 score, and ROC-AUC.
    - Generate a ROC curve and confusion matrix to visualize the performance of the model in predicting income level.

6. Summary and Recommendations

- Summarize key insights from the EDA, highlighting any patterns or trends observed in the data.
- Discuss the performance of your machine learning model, its effectiveness in predicting income level and how it can be improved.