Predictive Modeling for Income Analysis 50K

Python tasks using dataset to analyze income inequality (CSE 574: Introduction to Machine Learning - Assignment 1)

Sahithya Arveti Nagaraju (Person number – 50559752)

INTORUCTION

This project will talk about the analysis of the Census Income dataset, which allows for predictions of individual earnings above USD 50,000 per year. It first preprocess the data and then apply a various classifiers and provides an accuracy of 86.6%. The work is intended to illustrate some basic concepts of the treatment of data within machine learning.

DATA CLEANING

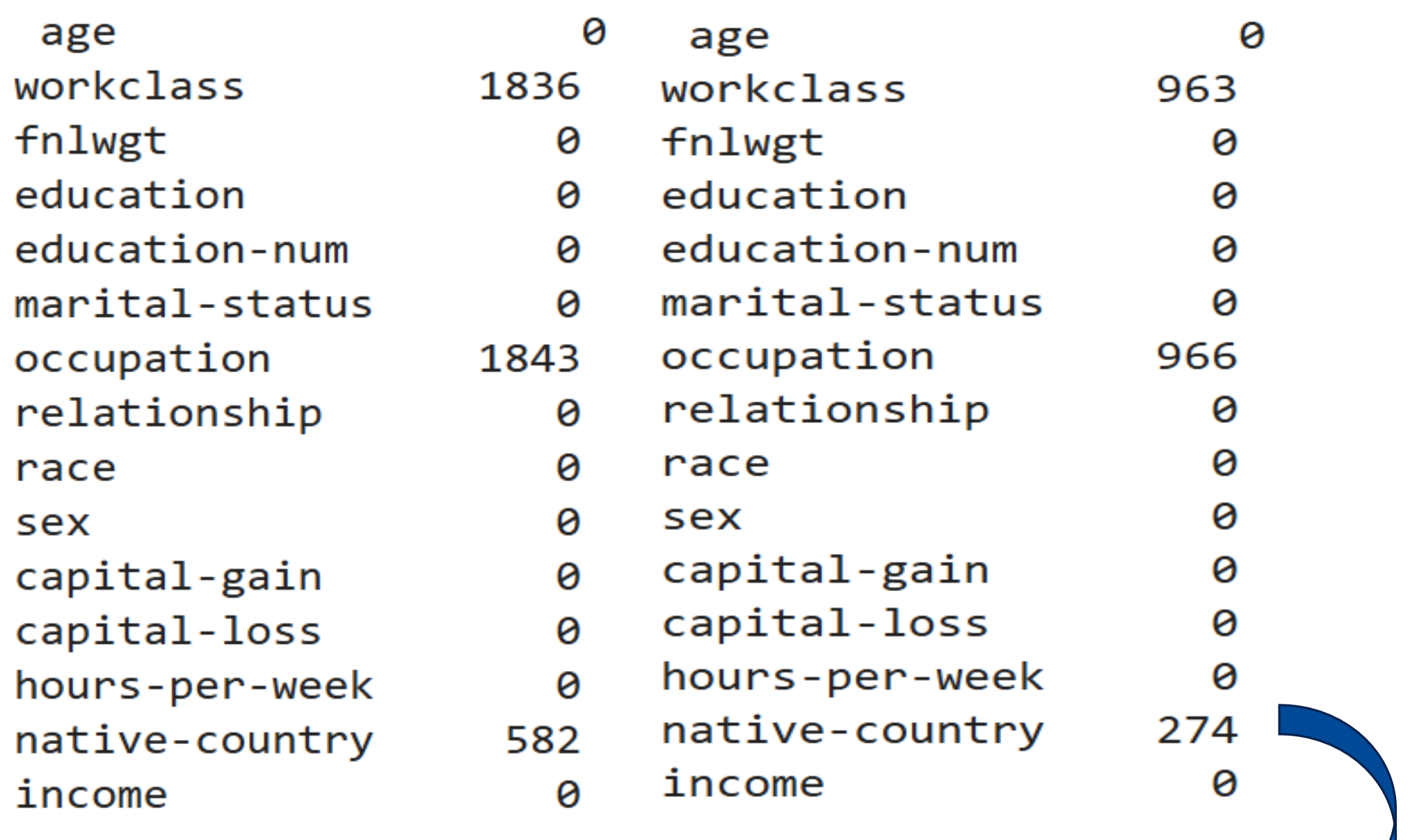
A. Duplicate Removal: Detected and managed 24 duplicate rows to maintain data integrity.

B. Handling Missing Values: Identified and replaced '?' values with NaN, then imputed missing entries in workclass and occupation using the mode, and categorized native-country missing values as 'Unknown'.

C. Data Encoding and Conversion: Converted float columns to integers for consistency and applied LabelEncoder to transform categorical variables into numerical format, preparing the dataset for machine learning.

```
Number of duplicate rows: 24
```

? Missing value Counts



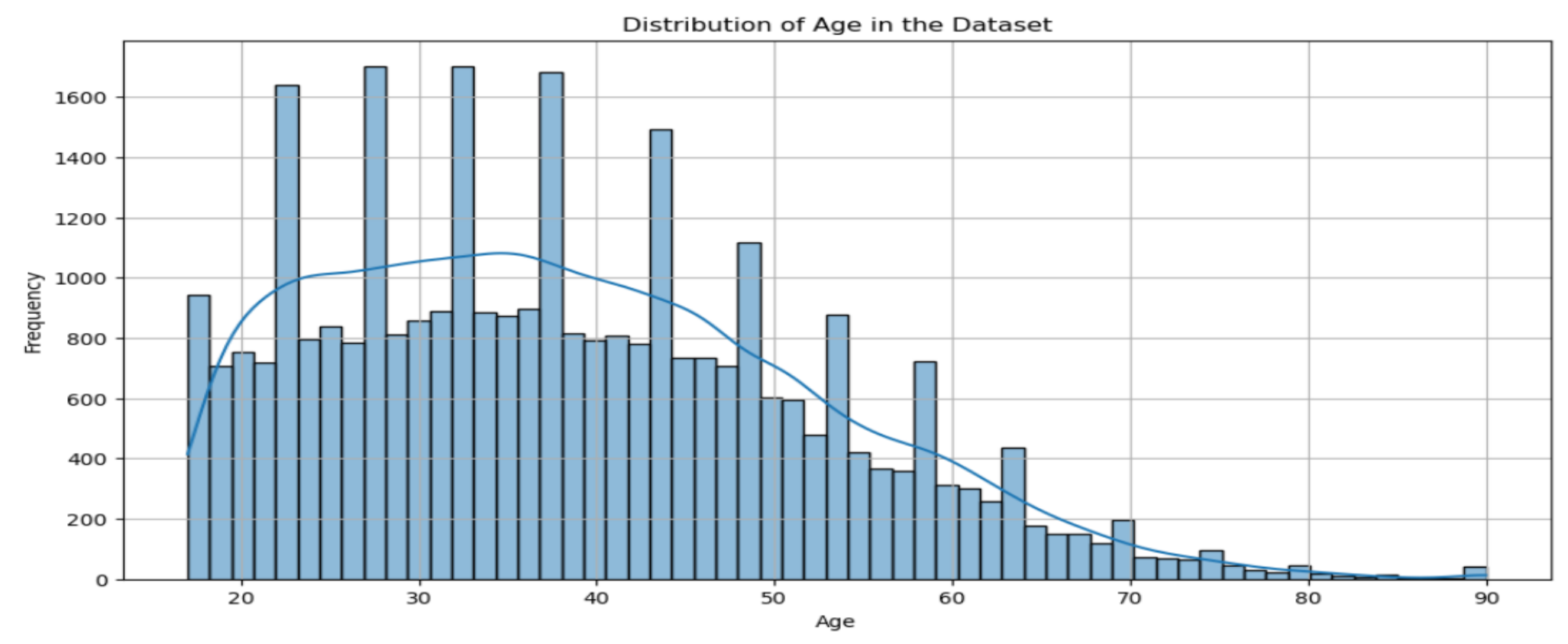
```
 age                  0    age                  0
workclass          1836   workclass          963
fnlwgt                0   fnlwgt               0
education             0   education            0
education-num         0   education-num        0
marital-status        0   marital-status       0
occupation         1843   occupation         966
relationship          0   relationship         0
race                  0   race                 0
sex                   0   sex                  0
capital-gain          0   capital-gain         0
capital-loss          0   capital-loss         0
hours-per-week        0   hours-per-week       0
native-country      582   native-country     274
income                0   income               0
```

Training data   Testing data

EXPLORATORY DATA ANALYSIS

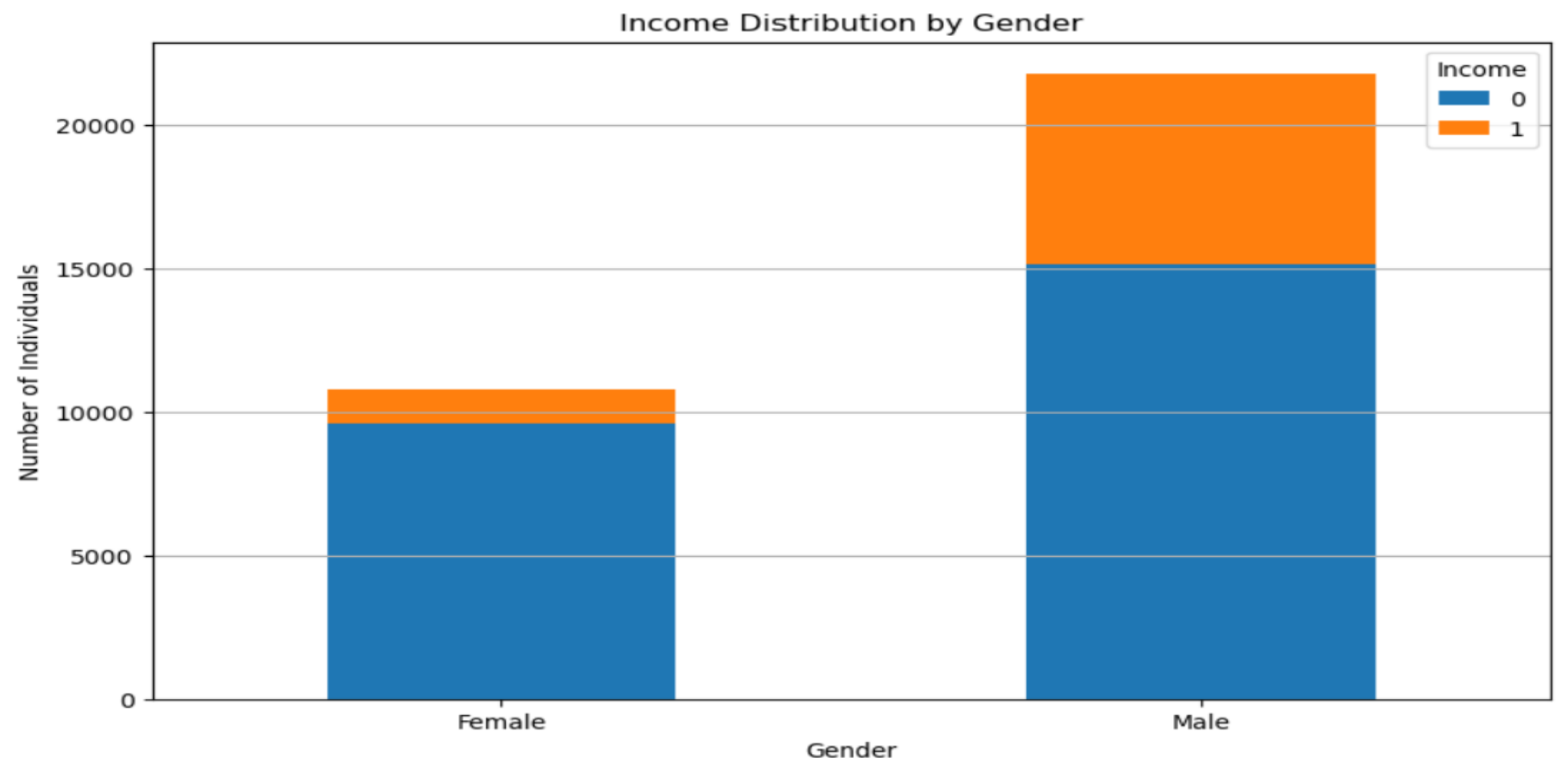
❑ Histogram of distribution of age:

The distribution is skewed to the right, indicating a larger proportion of individuals in the adult age group. The peak around the age of 30-35 suggests that this age range is the most common in the dataset.



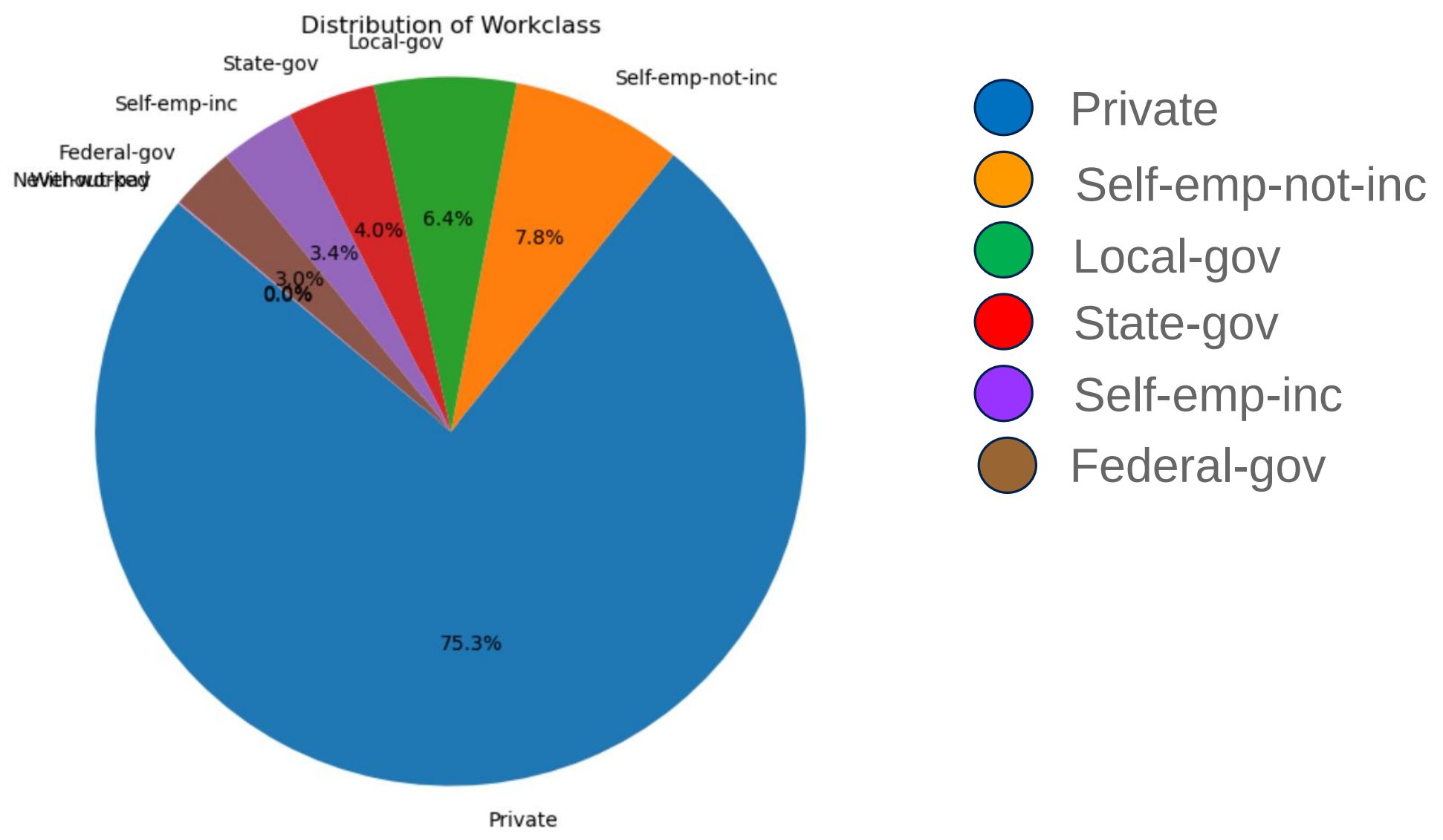
❑ Stacked bar chart for Income v/s Gender:

Males have a higher proportion in the higher income bracket. More females are in the lower income bracket.



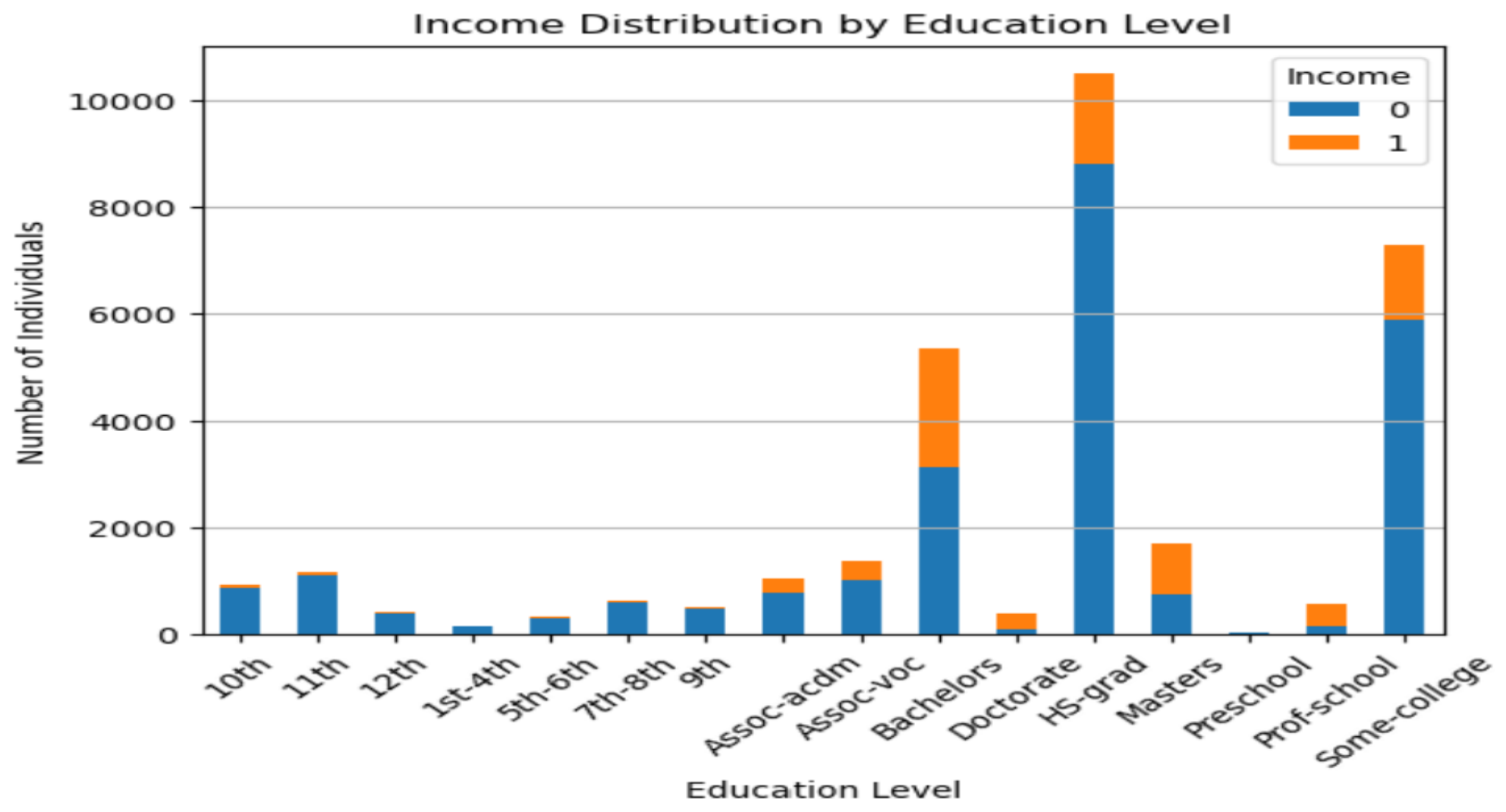
❑ Pie Chart for Workclass Distribution:

Most individuals work in the private sector. Other sectors have smaller proportions. Few individuals are unemployed or have never worked.



❑ Income Distribution by Education Level:

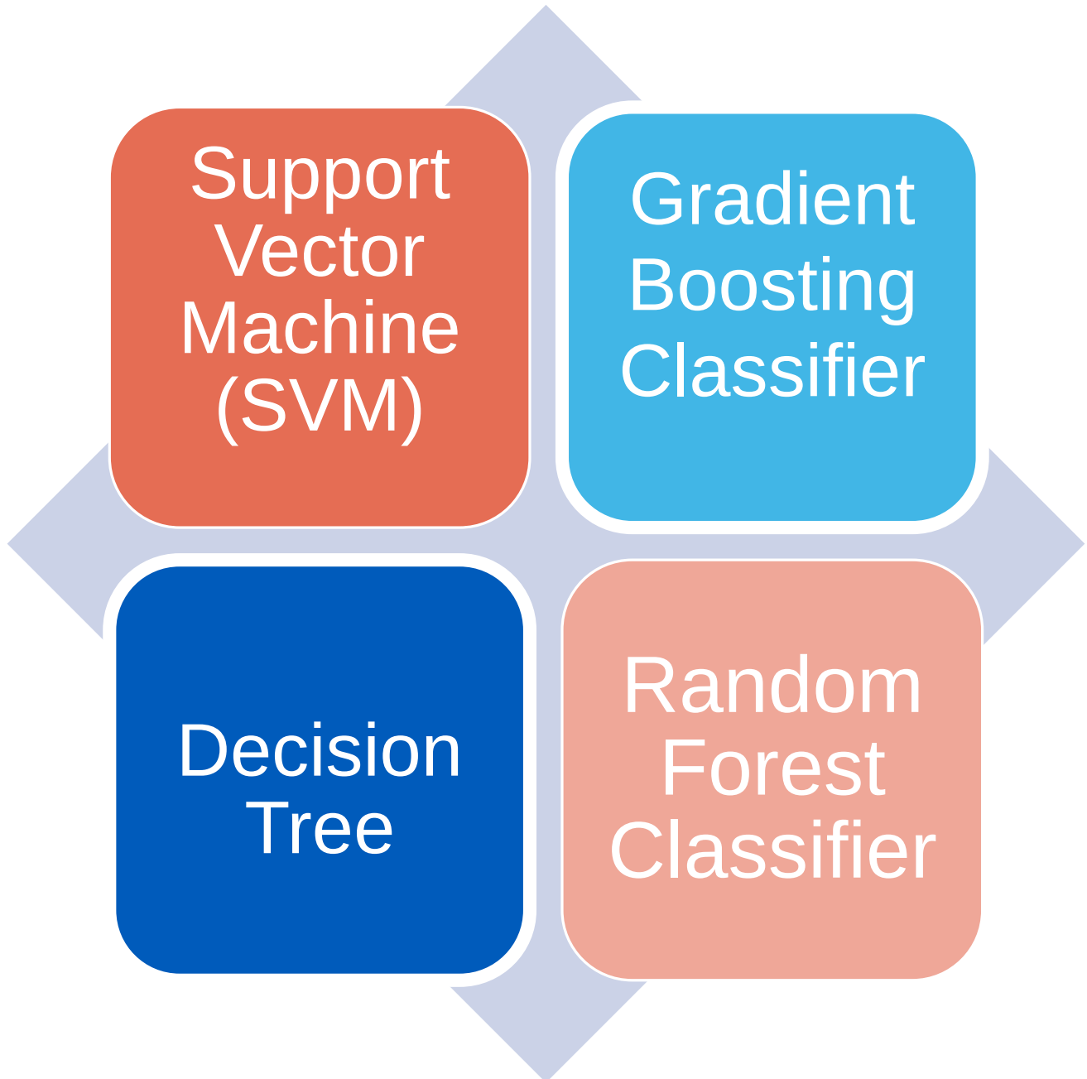
Education level is positively correlated with income. Lower education levels are associated with lower income. Factors beyond education can influence income.



TRAINING MODELS

Started by dividing the dataset into features (X) and target variable (y) for both train and test datasets. The last column, which indicates income, has been set as the target variable.

Considered four different classifiers and evaluated their performance on the test data.



Support Vector Machine (SVM)

Gradient Boosting Classifier

Decision Tree

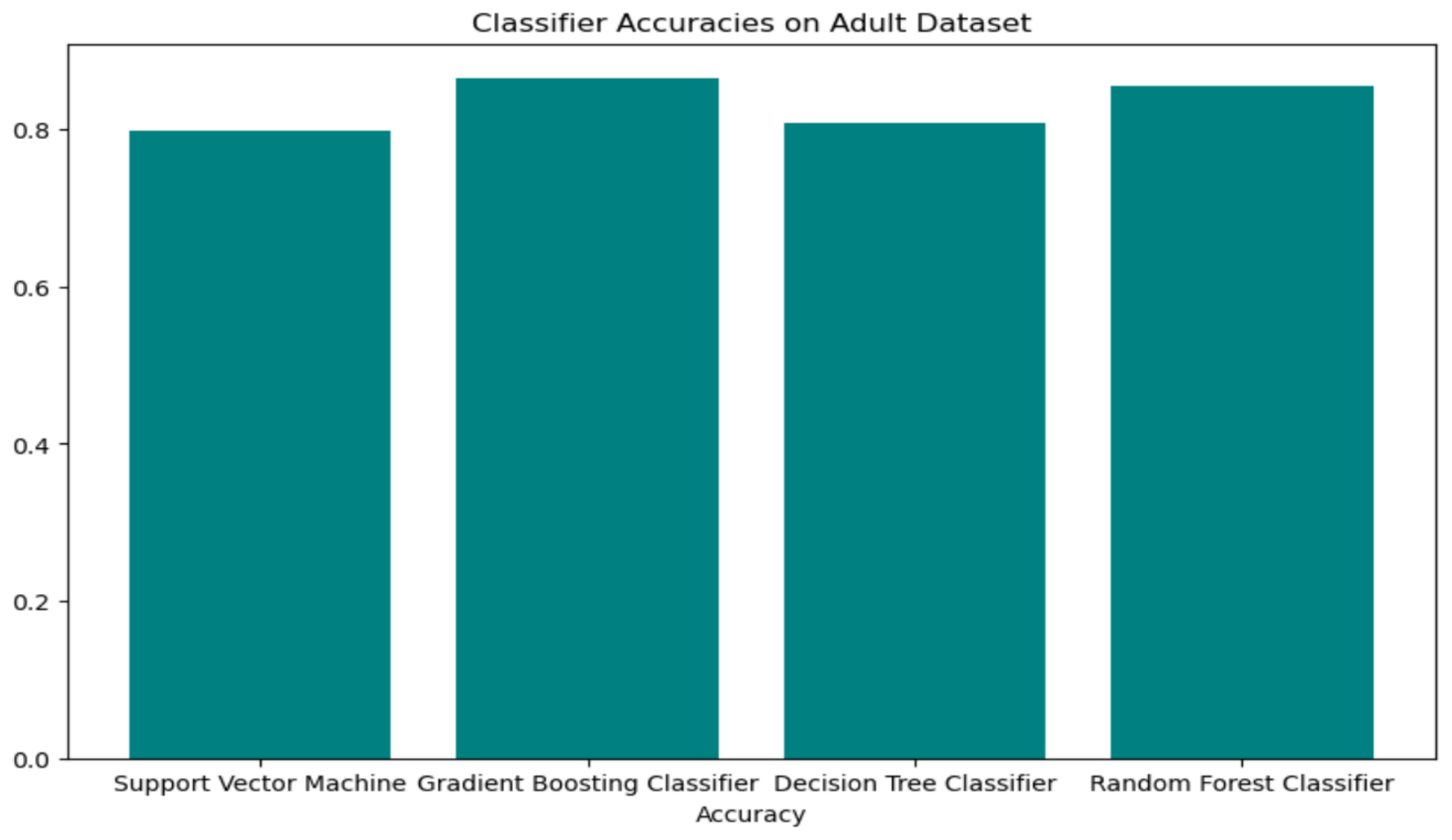
Random Forest Classifier

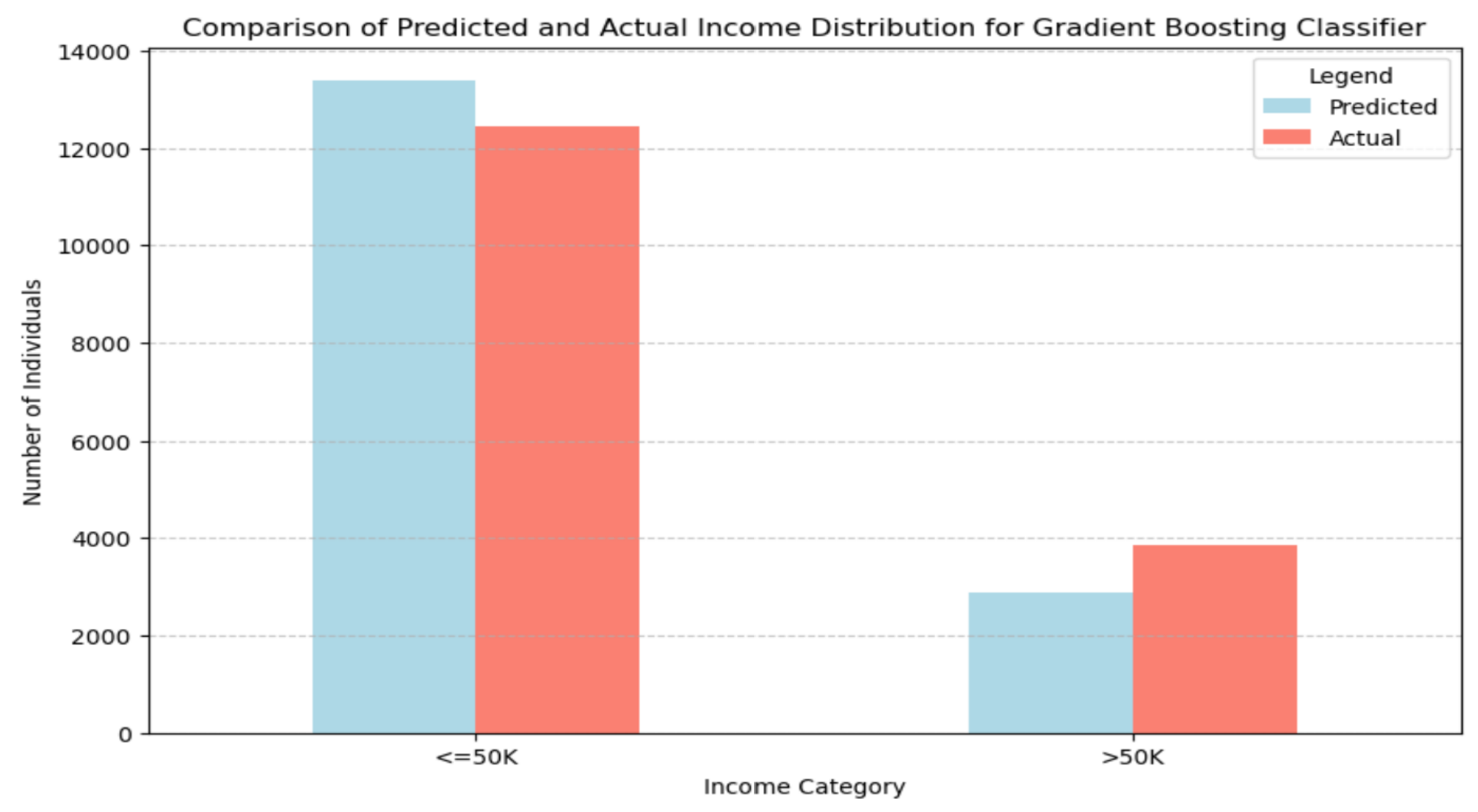
1. **Support Vector Machine**: A powerful classifier used for classification.
2. **Gradient Boosting Classifier**: Builds models sequentially to improve accuracy.
3. **Decision Tree**: Interpretable model that splits data based on feature values
4. **Random Forest**: Combines multiple decision trees to enhance prediction stability.

MODEL EVALUATION

The performance of each model was evaluated using accuracy scores and confusion matrices:

- **Accuracy Score:** Measures the proportion of correct predictions made by the model.
- **Confusion Matrix:** Provides a detailed breakdown of correct and incorrect predictions, allowing insights into model performance for each class.





CONCLUSION

This project successfully analyzed the Census Income dataset to predict whether individuals earn over USD 50,000 per year. Through data cleaning, preprocessing, and the application of various machine learning models, achieved a satisfactory level of accuracy of 86.67% in predictions (through Gradient Boost Classifier).

References

1. Becker, B. & Kohavi, R. (1996). Adult [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5XW20.