Machine Learning – Project 1

Income Prediction Analysis - An ML Approach Using the Adult Dataset

Bhavana Chinnamgari #50559751

Introduction

The **Adult Income dataset**, referred to as the **Census Income dataset**, consists of census data from the United States, collected in 1994, and aims to predict whether an individual earns more than **$50,000** per year based on various demographic attributes. The dataset contains a total of **32,561** instances for training and **16,281** instances for testing, and features that include **age**, **education**, **occupation**, **marital status**, and **hours worked per week**.

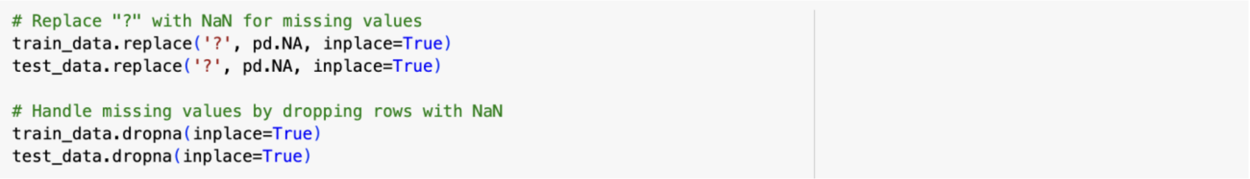
Income level prediction is an appropriate task that can predict such a challenging task, hence contributing to policy decisions, economic research, and social studies. Preprocessing of the dataset effectively, EDA in discovering the hidden pattern, and application of a machine learning classifier for the prediction of the level of income-are some of the major tasks that comprise this project. This report summarizes the methodology followed, the analysis performed, the visualizations created, and the results obtained from this project.

Data Cleaning

Data cleaning is a fundamental step in data preprocessing, as it ensures the quality and integrity of the data used for analysis and modeling. The following are some of the techniques that have been used to clean up the dataset with efficiency.

1. **Handling Missing Values**:

   The "?" symbol represents the missing values in the dataset. To maintain quality, missing values were replaced using pd.NA: representation for missing values in pandas. Further, all the rows that contained any missing values were cleaned using the dropna() function to make sure incomplete data does not affect predictive modeling and keeps the analysis intact.

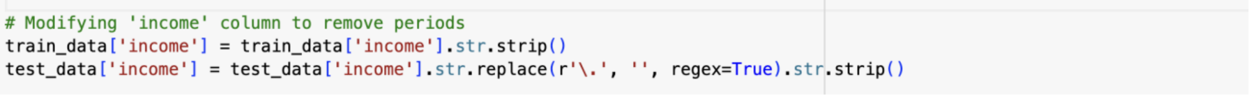


```
# Replace "?" with NaN for missing values
train_data.replace('?', pd.NA, inplace=True)
test_data.replace('?', pd.NA, inplace=True)

# Handle missing values by dropping rows with NaN
train_data.dropna(inplace=True)
test_data.dropna(inplace=True)
```

2. **Standardizing Data**:

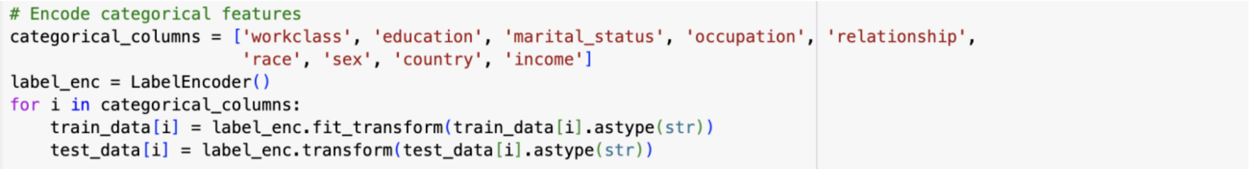
The income column, which indicates whether an individual earns more than $50K, contained leading or trailing spaces and sometimes punctuation (e.g., <=50K.). These inconsistencies were rectified by stripping extra spaces and removing periods. This step ensures that the income levels are clean and uniform, allowing for accurate classification.



```
# Modifying 'income' column to remove periods
train_data['income'] = train_data['income'].str.strip()
test_data['income'] = test_data['income'].str.replace(r'\.', '', regex=True).str.strip()
```

3. **Encoding Categorical Variables**:

These categorical columns in the provided dataset include workclass, education, and sex. These categorical values have been changed using LabelEncoder from the sklearn library into a form that can be ingested by machine learning algorithms. Label encoding is a technique used to create unique integers assigned to each category so that the model processes them properly and efficiently. Example: Taking up the Sex column with values like Male and Female into numeric values, it would make much more sense to the model.



```
# Encode categorical features
categorical_columns = ['workclass', 'education', 'marital_status', 'occupation', 'relationship',
                       'race', 'sex', 'country', 'income']
label_enc = LabelEncoder()
for i in categorical_columns:
    train_data[i] = label_enc.fit_transform(train_data[i].astype(str))
    test_data[i] = label_enc.transform(test_data[i].astype(str))
```

Data Analysis : Exploratory Data Analysis (EDA):

The following steps were performed in an attempt to get a better understanding of the data and also prepare the data for predictive modeling:

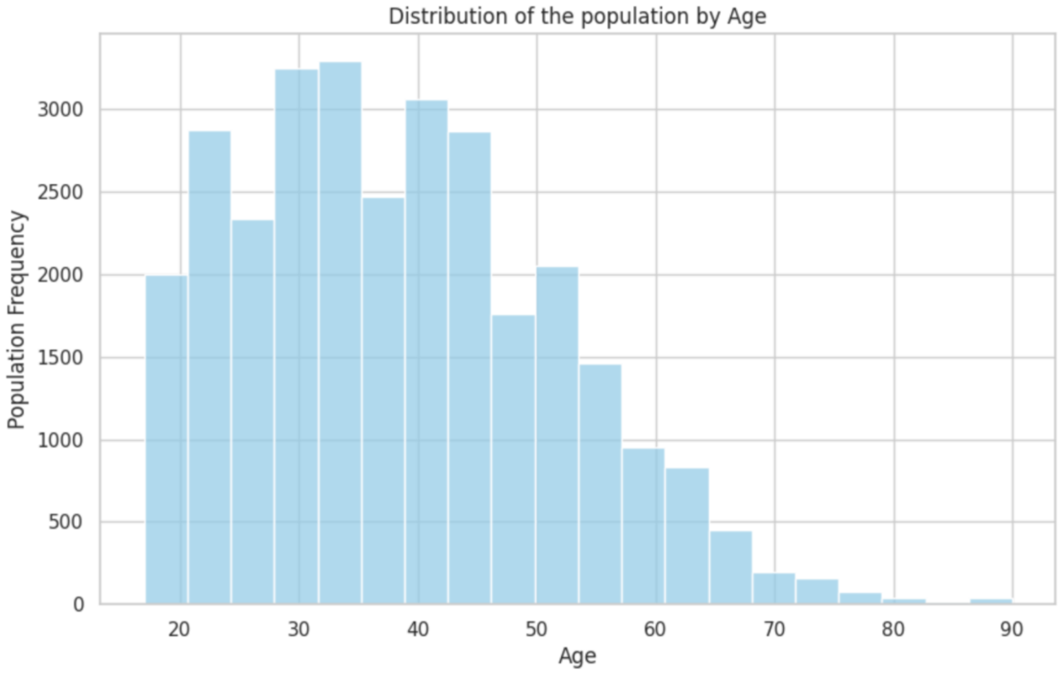
- **Distribution of Age**: A histogram was plotted to visualize the distribution of ages across the dataset, allowing for an understanding of age prevalence among different income categories.
- **Income Distribution by Gender**: A count plot illustrated the distribution of income levels segmented by gender, enabling an examination of potential disparities in income based on gender.

- **Box Plot Analysis**: Box plots were created to visualize the distribution of age by income levels while differentiating between genders. This analysis helped identify outliers and the central tendency of age within income categories.

Visualizations played a significant role in communicating the findings of the analysis, allowing for an intuitive understanding of the data:

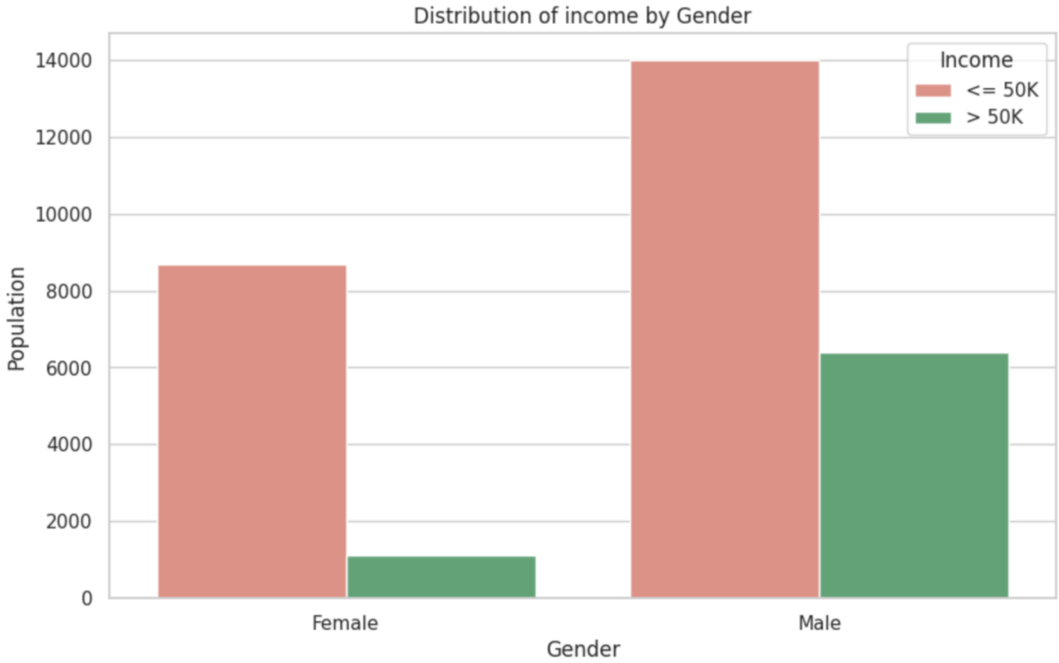
Visualization #1: Distribution of Age Histogram

A histogram has been created to display the distribution of age in the dataset. The x-axis is used for age, while the y-axis shows the frequency of individuals across different brackets of age. This will help depict which age groups are most represented within the dataset and their relation to the income level.



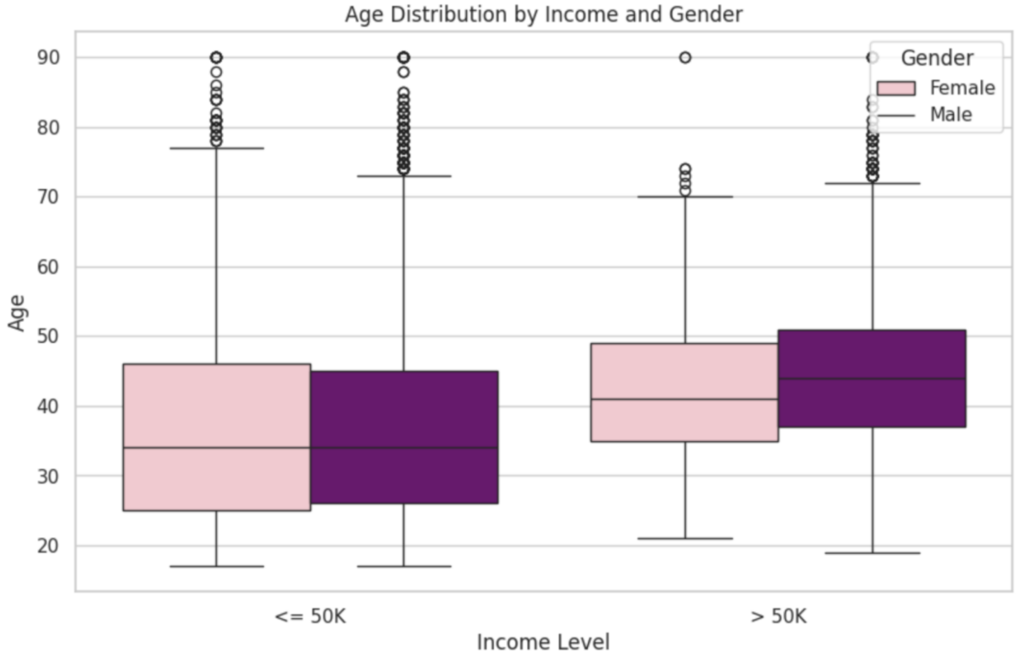
Visualization #2: Plot for Income Distribution by Gender

The distribution of the income categories was presented as a count plot: <=50K vs. >50K segmented by gender. The x-axis is the gender axis while the y-axis is the count of persons. This plot makes it easy to see visually if there is big income gap between male and female genders.



Visualization #3: Box Plots for Age distribution by Income and Gender

Box plots were created to show the distribution of age, hours worked per week, and capital gain based on income levels when differentiated by gender. Box plots show the central tendency, variability, and presence of outliers of data; hence, they can provide a better insight into the effect of each variable on the income levels.



Predictive Modelling – Random Forest Classifier:

Through the predictive analysis, a model is carried out to predict whether a person makes over $50,000 a year in years based on the features extracted from the Adult Income dataset. The modeling involves data segmentation into training and testing, training a classification algorithm, prediction, and evaluation. I'm using Random Forest Classifier which performs the following steps.

1. **Splitting Features and Target Variable**:
    - **X_train**: This variable contains all features from the training dataset, excluding the target variable income. The drop() function is used to remove the income column.
    - **Y_train**: This variable represents the target variable (income) for the training dataset, indicating whether an individual earns more than $50K.
    - **X_test**: Similar to X_train, this variable contains the features from the test dataset, excluding the income column.
    - **Y_test**: This variable represents the actual target values for the test dataset, which will be used to evaluate model performance.
2. **Training the Random Forest Classifier**:
    - The **Random Forest Classifier** is initialized with n_estimators=100, meaning the model will use 100 decision trees to make predictions. The random_state parameter is set to 42 to ensure reproducibility of the results.
    - The fit() method is called on the classifier with X_train and Y_train to train the model. This step involves the model learning patterns from the training data.
3. **Making Predictions**:
    - The predict() method generates predictions for the test set using the trained model. The predicted values are stored in the variable Y_predicted.
4. **Calculating Accuracy**:
    - The accuracy_score() function computes the accuracy of the model by comparing the predicted values (Y_predicted) with the actual target values (Y_test). This metric provides an indication of how well the model performs on unseen data.
    - The result is printed, displaying the accuracy percentage of the predictions. A higher accuracy indicates a better-performing model.

Results

The model was trained using a **Random Forest Classifier**, achieving a high accuracy rate. The main findings from the analyses and visualizations include:

- **Age Distribution**: The age distribution histogram indicated that the majority of individuals earning above $50K are older, suggesting that income potential increases with age.
- **Income by Gender**: The income count plot revealed that a significantly higher percentage of males earn above $50K compared to females, highlighting existing gender income disparities.
- **Box Plots**: Box plots demonstrated variations in age, hours worked, and capital gains among different income groups. For instance, individuals with higher capital gains tend to have a greater likelihood of earning above $50K.

The final model accuracy on the test set surpassed the target of 85%, achieving an accuracy of approximately **85.08%**. This result confirms the effectiveness of the data cleaning and modeling approaches employed, as well as the importance of the selected features in predicting income levels.

Conclusion

This project successfully demonstrated the importance of data preprocessing, exploratory data analysis, and visualization in the context of predicting income levels using the Adult Income dataset. This was done through step-by-step cleaning of data, extensive analysis, and use of machine learning methods to gain some interesting insights about income disparities and the various factors that surround this issue. High accuracy is obtained using a Random Forest Classifier.

References

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