MACHINE LEARNING - PROJECT 1

Income Prediction Analysis - An ML Approach Using the Adult Dataset

Bhavana Chinnamgari
MS in ES Data Science
University at Buffalo, New York

Introduction

The Adult Income dataset, with 32,561 training and 16,281 testing instances, aims to predict if individuals earn over $50,000 based on demographic features like age and education. This project includes data preprocessing, exploratory data analysis (EDA), and machine learning classification, providing insights for policy and economic research. The report summarizes the methodology, analysis, visualizations, and results.

Data Cleaning

Data cleaning ensures the dataset's quality for analysis and modeling. The following techniques were employed

1. **Handling Missing Values**:
   - Missing values indicated by "?" were replaced with pd.NA, and rows with any missing values were removed using dropna().
2. **Standardizing Data**:
   - The income column was cleaned by stripping spaces and removing periods (e.g., <=50K.) to ensure uniformity.
3. **Encoding Categorical Variables**:
   - Categorical columns like workclass, education, and sex were transformed using **LabelEncoder**, converting categories into unique integers for effective processing by machine learning algorithms.

Cleaned Train Data:

	age	workclass	fnlwgt	education	education_num	marital_status	occupation
0	39	5	77516	9	13	4	0
1	50	4	83311	9	13	2	3
2	38	2	215646	11	9	0	5
3	53	2	234721	1	7	2	5
4	28	2	338409	9	13	2	9

Cleaned Test Data:

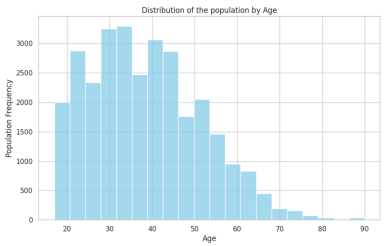
	age	workclass	fnlwgt	education	education_num	marital_status	occupation
0	25	2	226802	1	7	4	6
1	38	2	89814	11	9	2	4
2	28	1	336951	7	12	2	10
3	44	2	160323	15	10	2	6
5	34	2	198693	0	6	4	7

Data Analysis

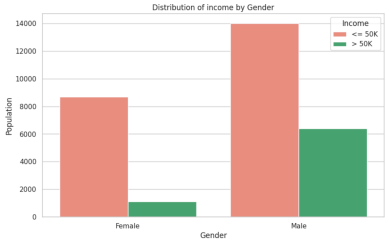
To gain insights into the dataset and prepare it for predictive modeling, the following analyses were conducted:

1. **Distribution of Age**: A histogram visualized the age distribution, highlighting the prevalence of different age groups across income categories.
2. **Income Distribution by Gender**: A count plot illustrated the income levels segmented by gender, allowing for the examination of income disparities between males and females.
3. **Box Plot Analysis**: Box plots were created to display the distribution of age by income levels while differentiating by gender. This analysis revealed outliers and the central tendency of age within each income category.

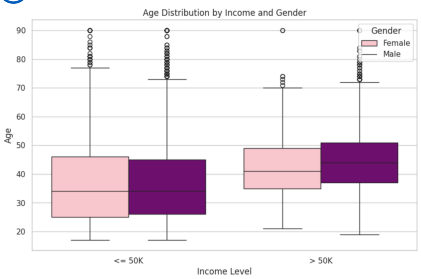
A



B



C



Predictive analysis

A **Random Forest Classifier** was implemented to predict income levels. The process involved:

- Splitting the dataset into features and target variables.
- Training the model with the training data.
- Making predictions on the test set.
- Evaluating the model's accuracy using the accuracy_score() function, achieving a result of **85.08%**.

Results

The model, trained using a Random Forest Classifier, achieved an accuracy of **85.08%**, surpassing the target of 85%.

Key Findings:

- **Age Distribution**: The histogram showed that older individuals are more likely to earn above $50K, indicating increased income potential with age.
- **Income by Gender**: The count plot revealed a significant disparity, with a higher percentage of males earning over $50K compared to females.
- **Box Plots**: These plots illustrated variations in age, hours worked, and capital gains among income groups, highlighting that individuals with higher capital gains are more likely to earn above $50K.

Conclusion

This presentation highlights the critical role of data preprocessing, exploratory data analysis, and visualization in predicting income levels on the Adult Income dataset. By employing structured data cleaning and robust machine learning methods, we gained meaningful insights into income disparities and their underlying factors. The Random Forest Classifier demonstrated high accuracy, validating the effectiveness of our approach.

References

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