Data Analysis Report: U.S. Adult Census Dataset

Introduction

The purpose of this report is to provide a comprehensive analysis of the U.S. Adult Census dataset, which contains information about various demographic and socioeconomic factors of individuals. The dataset was loaded and analyzed using Python, with the assistance of libraries such as Pandas, NumPy, Matplotlib, and Seaborn.

Data Overview

The dataset, named 'adult.csv,' was imported and loaded into a Pandas DataFrame. Initially, we explored the dataset to gain an understanding of its contents. This included displaying the entire dataset, examining the first and last 10 rows, and determining its dimensions, which revealed 45,175 rows and multiple columns.
Data source: https://www.kaggle.com/datasets/wenruliu/adult-income-dataset

Data Cleaning

The dataset underwent a rigorous data cleaning process to ensure its reliability for analysis. Several steps were taken:

Missing Value Handling: We identified and handled missing values, which were represented as '?' in the dataset. The affected columns were 'workclass,' 'occupation,' and 'native-country.' The '?' values were replaced with NaN to facilitate further analysis. Heatmaps were used to visualize the locations of missing values.
Removal of Missing Data: Rows with any missing values were removed from the dataset, resulting in a reduction in the number of rows from 48,842 to 45,222. This step was crucial to ensure data integrity.
Duplicate Data Removal: Duplicate rows were identified and removed from the dataset, further refining the dataset to improve the quality of the analysis.
Exploratory Data Analysis (EDA)

The following exploratory data analysis tasks were conducted to gain insights into the dataset:

Age Distribution: The distribution of ages in the dataset was examined. The histogram revealed that the majority of individuals are approximately between 17 and 48 years old.
Income and Workclass: The distribution of workclasses in the dataset was explored. Most individuals are employed in the private sector. Additionally, we identified that 'Self-emp-inc' workclass appears to have the highest income on average.
Education Level: We investigated the number of individuals with Bachelors or Masters degrees, identifying that 6,966 individuals in the dataset possess either of these higher education qualifications.

Income and Gender: The relationship between gender and income was analyzed. It was observed that males have a higher chance of earning an income greater than $50,000 compared to females.

Data Transformation

Some data transformations were performed to enhance the dataset for analysis:

Income Encoding: The 'income' column was encoded to have binary values (0 and 1) instead of the original labels ('<=50K' and '>50K') for easier analysis.

Data Type Conversion: The 'workclass' column's data type was converted to the 'category' data type to optimize memory usage.

Conclusion

In summary, this data analysis report provides an in-depth exploration of the U.S. Adult Census dataset. Through data cleaning, EDA, and data transformations, we gained valuable insights into various aspects of the dataset, including age distribution, workclass, education levels, and the relationship between gender and income.