Report

Raunak Kumar 23116082

Below is a detailed summary report for the Adult Census Income dataset assignment. This report encapsulates the data cleaning steps, exploratory analysis (univariate, bivariate, and multivariate), and key findings from the analysis.

Summary Report: Data Cleaning and Exploratory Data Analysis on the Adult Census Income Dataset

1. Introduction

The Adult Census Income dataset (sourced from the UCI Machine Learning Repository) comprises demographic and employment-related features. The main goal of this analysis was to prepare the dataset by addressing missing values, duplicate records, and outliers—and then to explore the data thoroughly using various visualization and statistical techniques. This report documents the steps taken during data cleaning, presents the exploratory data analysis, and summarizes key insights for further modeling.

2. Data Cleaning

2.1 Loading and Inspecting the Dataset

- **Dataset Source:** Loaded directly from the UCI repository.
- **Column Names:** Fifteen columns were assigned (including `age`, `workclass`, `fnlwgt`, `education`, `education_num`, `marital_status`, `occupation`, `relationship`, `race`, `sex`, `capital_gain`, `capital_loss`, `hours_per_week`, `native_country`, and `income`).

- **Initial Structure:** Basic inspection (using `.head()`, `.info()`, and `.describe()`) revealed that some columns were being treated as strings, which was corrected by converting key columns to numeric types.

2.2 Handling Missing Values

- **Numeric Imputation:** Missing values in numeric columns (e.g., `age`) were imputed using the median.
- **Categorical Imputation:** For categorical columns like `workclass`, `occupation`, and `native_country`, missing values were filled with the mode.

2.3 Removing Duplicates

- Duplicate records were identified and removed to ensure that subsequent analyses were not skewed by redundant data.

2.4 Outlier Detection and Treatment

- The Interquartile Range (IQR) method was applied to detect and remove outliers in columns such as `age` and `hours_per_week`, thus stabilizing the distributions for more robust analysis.

2.5 Standardizing Categorical Values

- All categorical values were standardized (converted to lowercase and stripped of extra spaces) to ensure consistency in subsequent grouping and visualization.

3. Exploratory Data Analysis (EDA)

3.1 Univariate Analysis

- **Summary Statistics:** Generated descriptive statistics (mean, median, variance, skewness) for numeric columns, providing insight into central tendencies and dispersion.
- **Frequency Distributions:** Analyzed categorical variables by displaying value counts, and visualized the most frequent categories (limiting to the top 10 where necessary for clarity).

- **Visualizations:**
  - **Histograms:** Displayed the distributions of numeric variables (e.g., `age`, `fnlwgt`, `capital_gain`).
  - **Box Plots:** Provided a visual check for outliers and distribution shapes across variables.

3.2 Bivariate Analysis

- **Correlation Matrix:** Computed the correlation matrix for numeric variables and visualized it with a heatmap, highlighting relationships (e.g., between `age` and `hours_per_week`).
- **Scatter Plots:** Explored relationships between continuous variables through scatter plots (e.g., `age` vs. `fnlwgt`).
- **Comparisons Across Groups:**
  - **Count Plots:** Compared the distribution of income categories across gender.
  - **Violin and Box Plots:** Examined the distribution of `hours_per_week` across income categories, and capital gain by workclass, providing insights into how different groups fare in these metrics.

3.3 Multivariate Analysis

- **Pair Plots:** Generated pair plots for key numeric features (including `age`, `education_num`, `hours_per_week`, and `capital_gain`) while using `income` as a grouping variable. This helped in identifying patterns and potential interactions between multiple variables.
- **Heatmap Revisited:** Reaffirmed the correlation findings through a detailed heatmap, emphasizing multivariate relationships.
- **Grouped Comparisons:** Analyzed the combined effect of income and sex on features such as `education_num` and `hours_per_week`, revealing nuanced group differences that could be further explored in predictive modeling.

4. Key Findings

- **Data Quality Improvements:** After addressing missing values, duplicates, and outliers, the dataset was significantly cleaner, ensuring that subsequent analysis was more reliable.
- **Variable Distributions:** Numeric variables generally exhibited a wide range with some skewness (e.g., `capital_gain` and `capital_loss`), indicating that transformations might be useful in future modeling.
- **Inter-variable Relationships:**
  - The correlation heatmap showed moderate correlations between several pairs of variables.
  - Grouped analysis suggested differences in working hours and education levels across income categories and gender, which could be valuable for income prediction models.
- **Categorical Consistency:** Standardizing categorical values improved the clarity of frequency distribution plots and aided in more accurate grouping for bivariate and multivariate analysis.

5. Conclusion and Next Steps

This comprehensive analysis of the Adult Census Income dataset provided the following:

- **Cleaned Data:** A robust, pre-processed dataset ready for further analysis or predictive modeling.
- **Insightful Visualizations:** A suite of univariate, bivariate, and multivariate plots that highlight key patterns and relationships in the data.
- **Future Directions:**
  - **Feature Engineering:** Additional features (e.g., combining certain categories or creating interaction terms) could enhance predictive performance.
  - **Predictive Modeling:** Building classification models (such as logistic regression or decision trees) to predict income levels, followed by rigorous model evaluation.

- **Advanced Analysis:** Exploring the impact of additional transformations or regularization techniques to better manage skewed variables.

This report serves as both documentation of the analytical process and a guide for subsequent steps in your data science pipeline.

End of Report.