ITAI 1371

Professor Viswanatha Rao

October 28, 2025

Group 4. Maria Jose Viveros

Data Preprocessing and Cleaning for the Adult Income Dataset

The goal of this project is to predict which factors influence whether a person earns more or less than $50,000 per year. The dataset includes 32,561 participants and 15 variables representing demographic and socioeconomic characteristics. These variables are both numerical (e.g., age and hours worked per week) and categorical (e.g., education, occupation, and income level).

In the overall data quality analysis, the dataset initially contained 32,561 rows, with 24 duplicates accounting for just 0.07 percent of the data. After removing these duplicates, the final dataset had 32,537 unique rows. Missing values were found in the occupation (5.66 percent), workclass (5.63 percent), and native-country (1.79 percent) columns. Using the command df.replace('?', pd.NA, inplace=True) was crucial because it replaced question marks with recognized missing values, enabling proper handling of null data during cleaning and analysis. To fill in missing data, mode imputation was used for categorical variables with nulls. The occupation variable was filled with the mode “Prof-specialty,” workclass with “Private,” and native-country with “United-States.” Mode imputation was suitable here because it maintained the original structure of the categorical variables without changing the distribution of the dominant categories. Since the distribution after imputation remained nearly the same as initially, this approach helps ensure the model will not be biased by over-replacements. As a result, the dataset quality stays stable and ready for further analysis or modeling.

Certain variables were removed from the dataset to enhance data quality and model performance. The fnlwgt variable was excluded because it is undefined and likely acts as a combined indicator from a previous calculation. The string form of the education variable was removed since it duplicates the information found in education.num. Additionally, capital.gain and capital.loss were eliminated due to their extreme imbalance, with zeros in 96 and 98 percent of cases, respectively, resulting in highly skewed distributions that contribute little to predictive analysis. The correlations among numerical variables (e.g., education.num, hours-per-week, age, etc.) were generally low (below 0.5), with the exception of a strong correlation between education and education.num, since both represent the same data—one as a string and the other as a number. To avoid redundancy, one of these variables was removed.

Because categorical variables are stored as text strings, they need to be converted to numerical format for machine learning algorithms. Initially, One-Hot Encoding was tested, but it created too many new variables, which is effective for logistic or linear regression models but inefficient for decision tree algorithms. Therefore, label encoding was used for categorical variables, helping to produce a more balanced distribution, preserve semantic meaning, and reduce bias related to gender, occupation, and country of origin. For example, the marital.status variable was recoded because the original data showed a highly uneven distribution. “Married-civ-spouse” made up 46.01 percent of the sample, while other categories like "Married-AF-spouse" had only 0.07 percent, creating imbalance. To create a more uniform distribution and reduce sparsity, categories were grouped into two broader classes: "Married," which includes Married-civ-spouse, Married-spouse-absent, and Married-AF-spouse, and "Not Married," which includes Never-married, Divorced, Separated, and Widowed. This recoding resulted in a balanced split of 47.36 percent and 52.64 percent, simplifying the variable and improving model efficiency.

To prepare numerical features for modeling, both age and hours-per-week were scaled using two methods: Standard Scaling (Z-score) and Min-Max Scaling (0-1 normalization). Standard Scaling adjusted the features to have a mean of 0 and a standard deviation of 1, while Min-Max Scaling normalized them between 0 and 1. Histograms comparing original and scaled data showed that scaling preserved the data shape while adjusting the range.

Outliers in these variables were treated using the Interquartile Range (IQR) method. Values below Q1 - 1.5 IQR or above Q3 + 1.5 IQR were capped to these bounds. This approach reduced extreme values while maintaining the overall distribution. Boxplots before and after capping confirmed that the influence of outliers was minimized, ensuring more stable input for machine learning models.

In summary, the Adult Income dataset was carefully preprocessed for machine learning. Missing values in categorical variables were handled using mode imputation, ensuring no null data remained. Redundant and highly skewed columns such as fnlwgt, education (string), capital.gain, and capital.loss were removed to improve data quality. Categorical variables were encoded with label encoding, and categories with high imbalance, such as marital.status, were recoded for better distribution. Numerical features like age and hours-per-week were scaled with both Standard Scaling and Min-Max normalization, and outliers were capped using the IQR method to reduce their impact.