

Venkatesh Guri

[Module 1 Project -](https://northeastern.instructure.com/courses/159846/assignments/1927626) Understanding Income Inequality

Northeastern University- College of Professional Studies

ALY6020: Predictive Analytics

Summer 2 2024 -Quarter

Professor: [Justin Grosz](https://northeastern.instructure.com/courses/182858/users/24150)

**Introduction and Dataset Overview:**

In contemporary data analysis, the classification of demographic and economic attributes serves as a pivotal tool for understanding societal trends and individual behaviors. This report focuses on a comprehensive analysis of a dataset that encompasses diverse demographic and economic attributes, including age, education, occupation, and more. These attributes are critical in delineating patterns and trends that aid in decision-making processes across various domains.

The dataset encompasses a diverse range of attributes, each offering unique perspectives on the individuals under study. Key variables include demographic details such as age, race, and sex, educational attainment, employment status, and financial metrics like capital gains and losses. These attributes not only paint a portrait of the individuals but also serve as critical factors in understanding broader social and economic dynamics.

The primary objective of this report is to conduct exploratory data analysis (EDA) and prepare the dataset for classification modeling tasks. The methodology includes rigorous data preprocessing to handle missing values, remove duplicates, and transform variables as necessary for analysis. Exploratory techniques such as descriptive statistics, visualizations, and correlation analyses will be employed to uncover insights and patterns within the dataset. By leveraging classification modeling, we aim to predict income levels based on demographic and socio-economic attributes, providing valuable insights into factors influencing economic outcomes.

**Data Cleaning:**

**1.Handling Missing Values:**

Upon initial inspection, the dataset exhibited various missing values across multiple columns. Each column with missing data was addressed using specific imputation strategies to ensure data completeness and integrity.

**Age Column:**

The 'age' column initially contained 4,134 missing values (approximately 8.46% of the data). To address this:

* Missing values were filled using group-specific medians based on 'relationship' and 'marital-status' to maintain demographic integrity.
* Remaining null values were imputed with the overall median age of 37 to ensure consistency across the dataset.

A graph of a number of people

Description automatically generated

**Workclass Column:**

The 'workclass' column had 7,645 missing values (approximately 15.65% of the data). The strategy employed:

* Missing values were replaced with the mode of the 'workclass' column to reflect the most frequent employment classification.

**Education and Education-Num Columns:**

The 'education' column had 3,223 missing values (approximately 6.60% of the data), while the 'education-num' column had corresponding missing values related to 'education'. These were addressed as follows:

* Missing values in 'education' were filled with the mode of the column to maintain educational distribution.
* 'Education-num' values were imputed based on the mode of each respective 'education' group to align numeric education levels with their categorical counterparts.

**Other Columns:**

Missing values in the dataset were handled with a targeted approach to maintain data integrity and model accuracy. Imputation methods included using modes for categorical variables like marital status, occupation, relationship, race, and sex to ensure consistency within related groups. Numerical features such as capital gain, capital loss, and hours per week were imputed with their respective medians to preserve central tendencies. For native country, the most frequent value, 'United-States', was used. Rows lacking salary information, critical for classification modeling, were omitted to maintain the dataset's robustness and reliability in predictive analyses**.**

**2.Removed Column:**

The 'fnlwgt' column was identified as unnecessary for the analysis and was subsequently removed from the dataset to streamline further processing.

**3.Removing Duplicates:**

After identifying and assessing the dataset for duplicate entries, a total of 379 duplicate rows were identified and subsequently removed. This step was crucial to maintain data integrity and avoid bias in subsequent analyses.

**Exploratory Data Analysis (EDA) and Variable Selection:**

Exploratory Data Analysis (EDA) is crucial for understanding the dataset's structure, relationships between variables, and identifying patterns that may influence our target variable, salary.

During EDA, we inspected the dataset to identify missing values and apply appropriate imputation techniques to ensure data integrity. We examined the distribution of variables such as age, education level (education-num), work hours per week (hours-per-week), and demographic factors like sex and race.

Visualizations such as box plots and heatmaps were employed to gain insights:

* A graph of different colored boxes

  Description automatically generated with medium confidenceBoxplot of Age by Salary: This visualization revealed how age relates to salary, showing potential trends such as higher salaries among older individuals.

**Encoding Categorical Variables**

Categorical variables such as workclass, marital-status, occupation, relationship, race, sex, and native-country were encoded using Ordinal Encoding. This transformation converts categorical values into numeric form, allowing machine learning algorithms to process them effectively.

**Variable Selection:**

Variable selection is crucial for building predictive models and understanding which features most significantly influence the target variable, salary.

From our correlation analysis, three key variables were identified based on their strong correlations with salary:

* Education Number (education-num) has the highest positive correlation with salary (0.3188), indicating that higher education levels are linked to higher incomes, reflecting skill and qualification impacts on job opportunities and salary negotiations.
* Age shows a moderate positive correlation with salary (0.2271), suggesting older individuals typically earn more. This correlation reflects experience and seniority translating into higher wages.
* Hours per Week (hours-per-week) also positively correlates with salary (0.1958), highlighting the role of work commitment and productivity in income levels. Longer hours often mean higher earnings due to increased output and dedication to work.

A screenshot of a graph

Description automatically generated**A screenshot of a cell phone

Description automatically generated**

**Justification for variable Selection:**

* Education Number (education-num): Education significantly impacts earning potential by providing specialized skills and qualifications valued in the job market. Higher education-num levels correlate with higher salaries due to increased expertise and the ability to tackle complex challenges, which are highly sought after by employers.
* Age: Age correlates with higher earnings as professionals accumulate experience, leadership qualities, and industry-specific knowledge over their careers. This experience enhances their value to employers, leading to roles with greater responsibility and higher compensation.
* Hours per Week (hours-per-week): The number of hours worked per week directly affects income through hourly wages, overtime pay, and productivity levels. Longer hours-per-week can demonstrate dedication and may lead to higher earnings in roles where extra effort is rewarded. However, maintaining a healthy work-life balance is crucial for sustained career satisfaction and productivity.

**K Value selection:**

1. **K = 5**:
   * This represents a relatively low K value, which focuses on capturing local patterns and details within the data.
   * **Advantages**: It can capture fine-grained nuances and subtle variations in the dataset, potentially leading to a model that closely fits the training data.
   * **Challenges**: It is more susceptible to overfitting, where the model learns noise and specific quirks of the training data that may not generalize well to new, unseen data.
2. **K = 15**:
   * This value is around the point where the accuracy stabilizes or reaches an optimal point in the plot, balancing between bias and variance.
   * **Advantages**: It strikes a balance between capturing local patterns and generalizing broader trends in the data.
   * **Expected Performance**: Models with K = 15 are expected to exhibit robust performance on both the training and test datasets, providing a reliable compromise between underfitting and overfitting.
3. **K = 25**:
   * This represents a higher K value, which emphasizes broader generalization over local details.
   * **Advantages**: It tends to smooth out noise and reduce variance, making the model more robust to outliers and small fluctuations in the data.
   * **Challenges**: It may oversmooth the data, potentially missing important local patterns or variations that could be significant for accurate predictions.
   * **Expected Performance**: Models with K = 25 might exhibit lower variance but could introduce higher bias, resulting in less precise predictions that might not capture all nuances of the data.

In summary, the choice of K value in K-nearest neighbors (KNN) algorithms directly influences how the model interprets and predicts based on the data. Lower K values focus on local details but risk overfitting, while higher K values generalize more broadly but may overlook important local patterns. A moderate K value, such as K = 15, often strikes a balance that yields stable and effective performance across different datasets.

**Model building and Evaluation**:

To assess the performance of the K-Nearest Neighbors (KNN) algorithm in predicting salary categories, three models were built using varying values of K: 5, 15, and 25. The models were trained on a dataset consisting of three key features—age, education-num, and hours-per-week—which were scaled using StandardScaler to standardize their distributions.

Each KNN model was trained on the scaled training data and evaluated on the scaled test data. The performance metrics were assessed using confusion matrices, highlighting the model's ability to correctly classify instances into their respective salary categories.

**K = 5:**

A blue squares with white text

Description automatically generatedThe model with K = 5 achieved an accuracy of 0.75. The confusion matrix revealed 5224 true negatives and 745 true positives, along with 764 false positives and 1183 false negatives. This K value appears to lean towards a lower bias but may be susceptible to overfitting due to its closer proximity consideration.

**K = 15:**

With K = 15, the model's accuracy improved to 0.79. The confusion matrix showed 5580 true negatives and 648 true positives, with 408 false positives and 1280 false negatives. This indicates a balanced approach between bias and variance, capturing a more generalized view of the dataset.

A blue squares with white text

Description automatically generated

**K = 25:**

Similarly, the model with K = 25 maintained an accuracy of 0.79. It exhibited 5610 true negatives and 617 true positives, accompanied by 378 false positives and 1311 false negatives. This K value also demonstrates stable performance, reinforcing the model's ability to generalize effectively.

A blue squares with white text

Description automatically generated

**Comparative Analysis**

Comparing these models, increasing the value of K from 5 to 15 led to improved accuracy and a reduction in false positives and false negatives. This trend suggests that a larger K value helps mitigate overfitting by considering a broader neighborhood of data points for classification.

However, while increasing K beyond 15 to 25 did not significantly impact accuracy, it maintained a consistent level of performance. The choice of K should be tailored to balance between capturing sufficient neighborhood information for classification and avoiding underfitting or overfitting the model to the training data.

In conclusion, the selection of K in KNN models significantly influences performance metrics such as accuracy and the trade-off between bias and variance. For this dataset predicting salary categories, K = 15 appears to strike a balanced compromise, achieving high accuracy while generalizing well to unseen data. Further adjustments in K should consider specific application requirements and dataset characteristics to optimize model performance effectively.

**Conclusion and Real-world Applicability:**

Based on the KNN models developed using demographic features to predict income categories, the findings suggest practical applicability in real-world scenarios, particularly in understanding economic disparities. These models, configured with optimal K values, demonstrate a capacity to classify individuals into income groups based on age, education level, and work hours. This capability is crucial for analyzing and addressing inequalities in income distribution across diverse populations.

In real-life settings, the robustness of these models lies in their ability to leverage local patterns within data to make classifications. By not assuming underlying data distributions, KNN models can adapt to varying conditions and dataset complexities, enhancing their reliability. However, ensuring their effectiveness requires ongoing validation against updated data and careful consideration of data quality and model assumptions to maintain accuracy and relevance over time.

References:

1. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Science & Business Media.
2. Raschka, S., & Mirjalili, V. (2019). *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2.* Packt Publishing.
3. Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists.* O'Reilly Media.