College of Professional Studies
Northeastern University San Jose

MPS Analytics

Course: ALY6020: Predictive Analytics

Assignment:

Module 1 Project - Understanding Income Inequality

Submitted on:

Jan 16, 2024

Submitted to:

Prof: BEHZAD AHMADI

Submitted by:

NIKSHITA RANGANATHAN

INTRODUCTION

Understanding the dataset:

The Adult Income Dataset comprises 48,842 records with 15 variables, including 6 numerical and 9 categorical attributes.

This dataset, available on Kaggle (https://www.kaggle.com/datasets/wenruliu/adult-income-dataset), which contains 15 columns, with the target attribute being 'Salary,' categorized into two classes: <=50K and >50K. With the attributes capturing demographics and personal features, the dataset presents an opportunity to explore predicting income levels based on individual information.

Below are the data descriptions of each variable of the data that briefly describe the contents of the data set. The dataset's features are as follows:

1. Age: The age of the individual.
2. Work Class: The category of employment.
3. fnlwgt: A weighting factor indicating the estimated number of people represented by the entry according to the census.
4. Education: The highest level of education attained by the individual.
5. Education num: The numerical representation of the individual's education level.
6. Marital Status: The marital status of the individual.
7. Occupation: The specific occupation or job role of the individual.
8. Relationship: The person's relationship status, such as husband, wife, or own child.
9. Race: The individual's racial or ethnic background.
10. Sex: The gender of the individual.
11. Capital Gain: Profits gained by the individual through investments.
12. Capital Loss: Losses incurred by the individual through investments.
13. Hours per Week: The average number of hours the individual works per week.
14. Native Country: The country of origin for the individual.
15. Income: The yearly income of the individual categorized as <=50K or >50K.

Utilizing this census information, this research aims to create a model distinguishing citizens with low and high incomes. This in future could help in providing insights to organizations for effective policy formulation. The following sections will go into a detailed exploration, talking about the relationships between salaries and various features.

DATA CLEANING

- **Dropping the "fnlwgt" column**

  As I could see from the dataset, column – **fnlwgt** is not required or significant for our analysis, hence I dropped it.

- **Removing duplicate rows**

  Duplicate rows in the dataset can lead to inconsistencies and affect the accuracy of data analysis. It is essential to find and remove any duplicate rows from the dataset before starting the analysis. Duplicate records were removed from the dataset, resulting in a reduced dataset with 42,468 unique records.

- **Checking the number of missing values for each Attribute in the dataset**

  The variables do not seem to have any missing records.

- **Abnormal values in the dataset**

  During the initial analysis, abnormal values in the form of "?" were identified in the columns (Workclass, Occupation, and Native Country) were observed. These values were handled by replacing them with the term "Unknown".

DATA EXPLORATION AND EDA

The donut chart **(Figure 1)** illustrates that a significant majority, 75.4% (37,155 records) of the population represented earns $50,000 or less, while the minority, 24.6%(11,687 records), earns more than $50,000. This imbalance reveals that the dataset contains more than twice as many records for salaries under 50k compared to those above 50k and distribution skewed towards the lower income category.

The grouped bar chart **(Figure 2)** suggests that there are more males than females in both income categories. However, the discrepancy is larger in the higher income bracket (>$50K), where the count of males is significantly higher than that of females. This could indicate a gender income gap, with a higher proportion of males earning above $50,000 compared to females.

The visualization **(Figure 3)** reveals that majority of individuals earn over $50k during their mid-career (25-45) years. There is an overlap of the two distributions, particularly in the middle age ranges. The distribution indicates that earning potential increases with age until midlife and then declines as individuals approach retirement age.

From the multiple histograms **(Figure 4)**, the education-num displays a multi-modal distribution with peaks around 10 and 13 years, suggesting common stopping points in education, which may correspond to high school and bachelor's degree. The distribution of capital gain and capital loss is highly skewed to the right, with the vast majority of values at or near zero.
Lastly, a standard workweek of around 40 hours is the most common among the individuals represented in the data.

Next, I started working on the encoding process for analysis by converting text data into numerical form through label encoding. This is essential because most machine learning algorithms operate on numeric inputs and may not be capable of handling text data. Through this, I created a heatmap of the correlations between variables serves to identify how strongly the features are related to one another.

The level of education (education-num), age, and hours worked per week have the strongest relationships with income.

KNN MODEL

Initially, the dataset was divided into two segments: 80% dedicated to training the model, and the remaining 20% for testing purposes. This helps us build and check a model to predict income based on factors like age, job type, and education

We then applied standardization to both the training and testing datasets to adjust all feature values to a common scale, enhancing the accuracy and efficiency of our model.

The mean squared error is calculated for each k for the range 0 to 19, providing insights into the model's accuracy. The resulting plot shows how the mean squared error varies with changing k values, aiding in the selection of an optimal k for the KNN classifier.
Around k=7, the rate of decrease slows down significantly. This is referred to as the "elbow" (optimal value of k) of the curve and suggests that beyond this point, adding more neighbors does not significantly improve the model.

CONCLUSION

- **Feature Importance:**
  Education level (education-num), age, and hours worked per week were found to be the most important features in predicting income levels.

- **Explanation of Variables:**
  Education-num is a crucial factor as it represents an individual's educational, which often correlates with earning potential.
  Age plays a significant role, with income generally increasing until midlife and then declining.
  Hours worked per week also impact income, as more hours worked typically lead to higher earnings.

- **Determination of K:**
  The choice of K was determined through the Mean Squared Error curve. The optimal K value was found to be 7.

- **Accuracy and Model Evaluation:**
  The model achieves an accuracy of 80% and demonstrates better performance in correctly identifying class 0 (low income) instances with higher precision and recall, while its performance for class 1 (high income) is less satisfactory.

We can also understand that distinguishing between people with high and low incomes is quite challenging because of the more low income values in the data. To improve the classification, it is better to have a random subset of the dataset with equal rows of high and low-income individuals.

REFERENCES

Correlation matrix plot with coefficients on one side, scatterplots on another, and distributions on diagonal. (n.d.). Stack Overflow. https://stackoverflow.com/questions/48139899/correlation-matrix-plot-with-coefficients-on-one-side-scatterplots-on-another

Srivastava, T. (2024, January 4). *A complete guide to K-Nearest neighbors (Updated 2024)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

APPENDIX

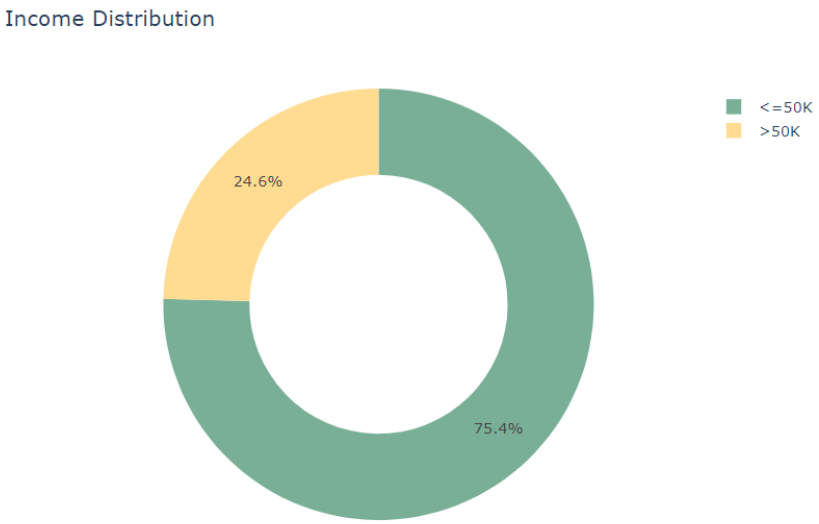

Figure 1 – Income distribution

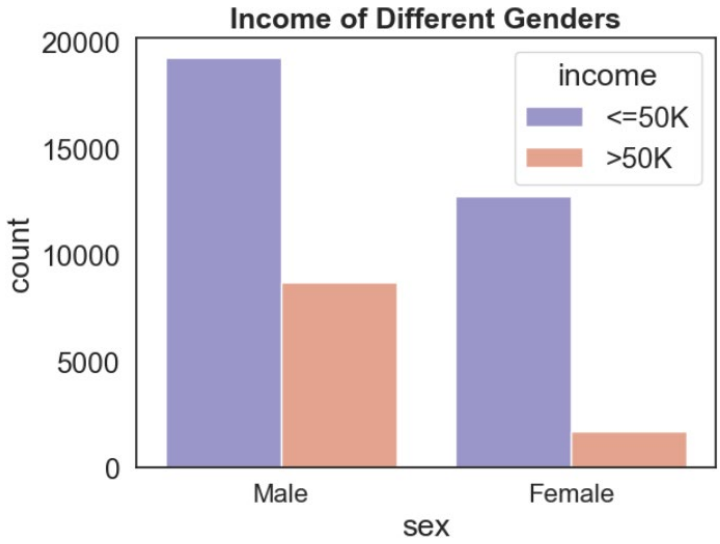

Figure 2 – Income – Male vs Female

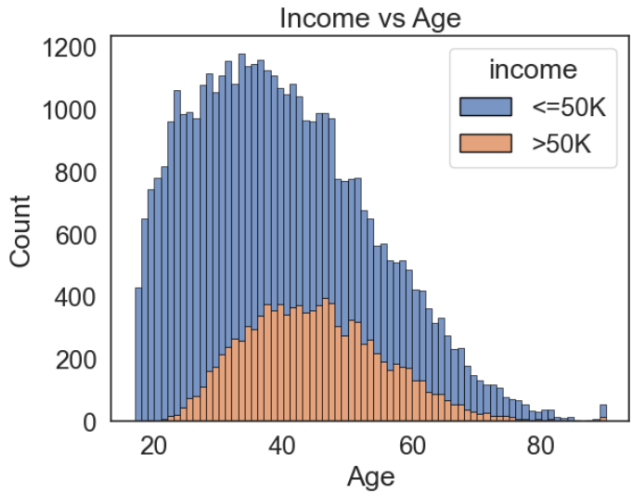

Figure 3 – Income vs Age

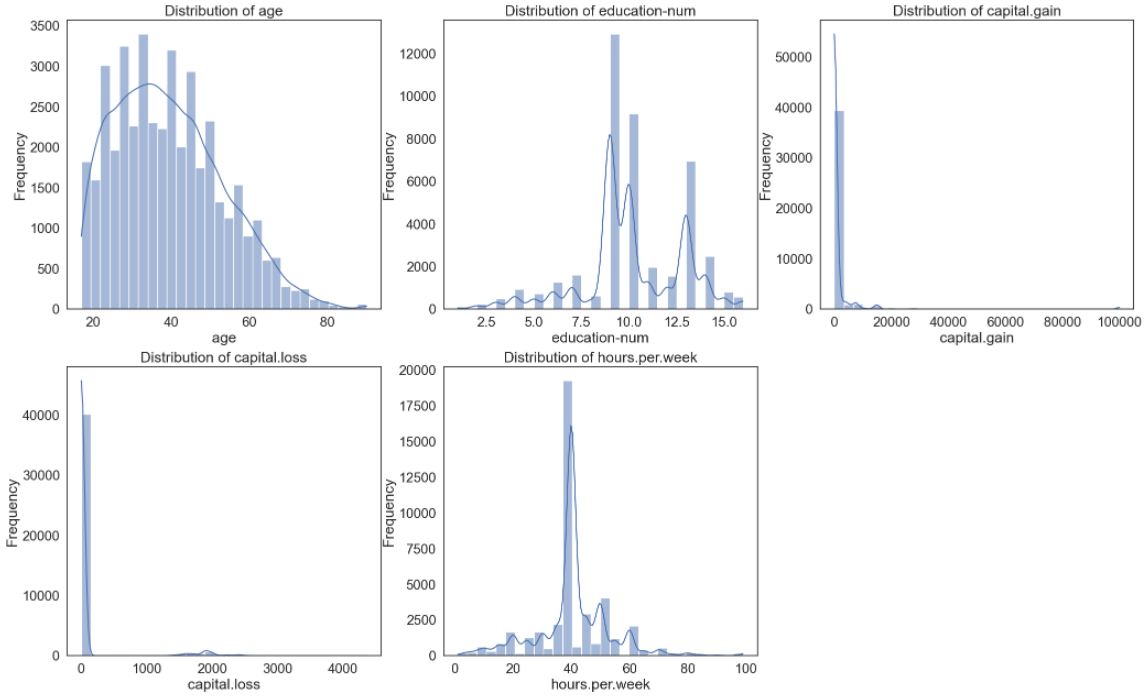

Figure 4 – Income vs Age

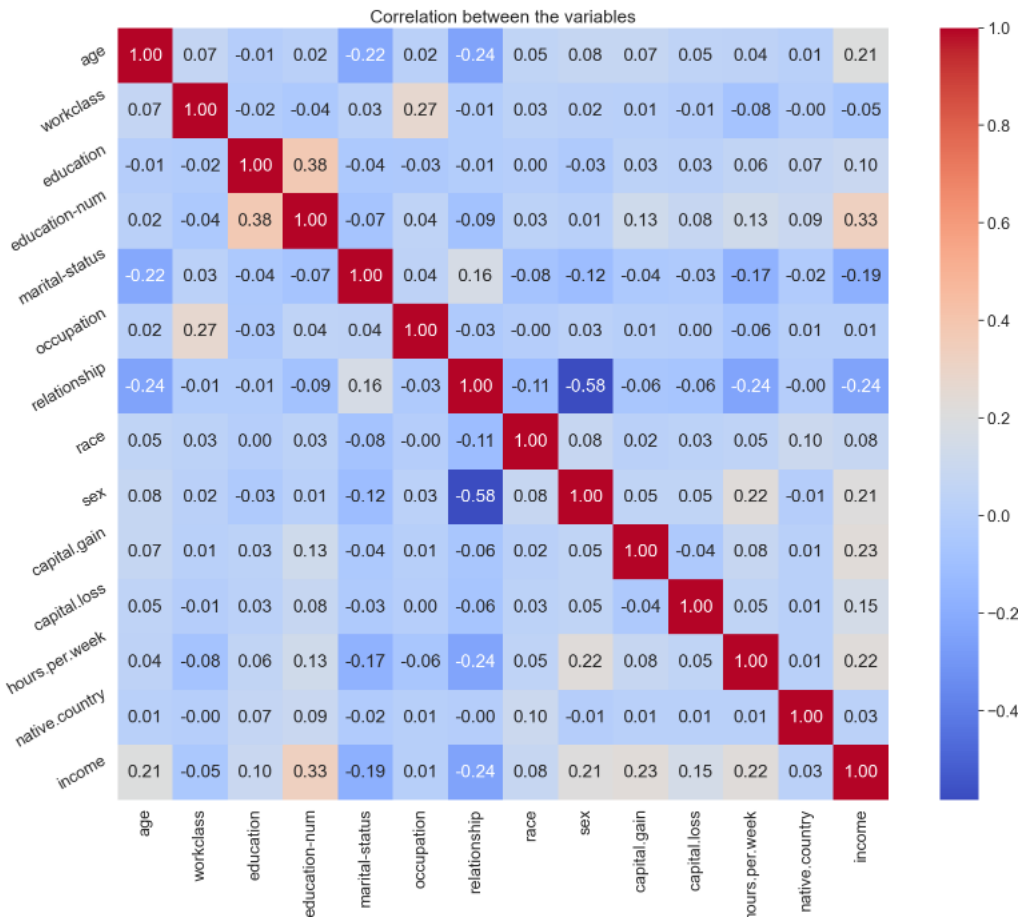

Figure 5 – Heatmap

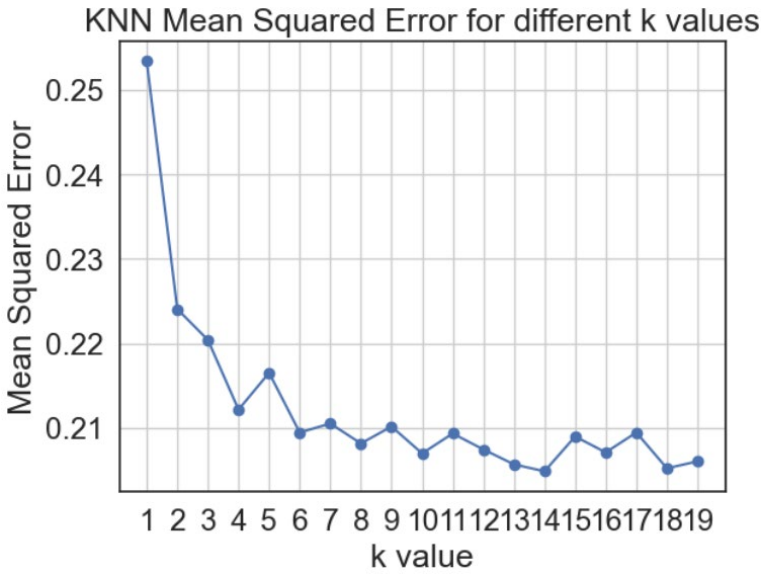

Figure 6 – Different k values

	precision	recall	f1-score
K=7			
0	0.86	0.88	0.87
1	0.59	0.55	0.57

Table 1: Classification Report

	Yes	No
Yes	5659	778
No	917	1140

Table 2: Confusion Matrix

K – value	Accuracy
7	80%

Table 3: Accuracy