Northeastern University

Module 1 Assignment – Understanding Income Inequality

Professor: Justin Grosz

Course: ALY6020: Predictive Analytics (CRN 20466)

Submitted by :

Sameeksha Bellavara Santhosh

Date: 14-01-2024

Introduction

In this project, our goal is to use census data on various characteristics of US citizens to promote equal pay and better understand the factors that lead to economic disparities. Important characteristics including race, gender, education level, and occupation are included in this dataset. Building a predictive model that correctly divides people into low- and high-income categories is the main objective. It is imperative to develop this predictive model to shed light on the factors that contribute to affluence and offer insights that can guide changes to American policy. Our goal in examining the connections between various characteristics and income levels is to identify patterns and trends that might point to economic disparity.
This project is important because it can help with the current efforts to address issues of pay equity. For groups and legislators trying to enact inclusive and equitable laws, the model can be a useful resource. Evidence-based decision-making will be made possible by an understanding of how particular attributes affect income levels, which will ultimately promote a more just and equitable society. We aim to provide answers to questions like which demographics, educational backgrounds, and occupations are linked to higher incomes through the analysis of census data. Our goal is to use machine learning to produce useful insights that can direct programs that promote equitable opportunities and close income disparities for all citizens.

To sum up, this project is an example of a data-driven strategy for understanding the intricate interactions between factors that affect income levels in the United States. The model's outputs will help advance knowledge of the factors that determine wealth and make it easier to make wise policy choices that will lead to a more equitable and inclusive economic environment.

Part 1: Data Cleaning

Several techniques were used during the data cleaning process to improve the dataset's quality and suitability for further analysis and modelling. In the beginning, NaN was used in place of missing or unknown numerical values, which were shown as zeros in the "hours-per-week," "capital-gain," and "capital-loss" columns. The "capital-gain" and "capital-loss" columns' missing values were then filled in using mean imputation, which kept the data's overall distribution intact. Furthermore, 'NaN' was used in place of '?' values in the categorical columns 'occupation,' 'workclass,' and 'native-country,' to ensure consistency in their handling. For these categorical columns, mode imputation was used to replace missing values with the value that occurs the most frequently. One-hot encoding was used to transform categorical variables into binary columns to successfully incorporate categorical data into the KNN algorithm.

By addressing missing values, preserving data distribution, and managing categorical variables appropriately, these cleaning techniques together enhanced the quality of the dataset and ensured accurate analysis and modelling. Consistency and dependability in data handling were prioritized in the decisions made at each stage, which were informed by the project's goals and the dataset's properties.

	Age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	Sex	Capital-gain	Capital-loss	Hour-per-week	Native-country	Salary
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

The above snapshot indicates the first few rows of the dataset.

```
   Age  fnlwgt education   education-num      marital-status   relationship  \
0   39   77516  Bachelors              13       Never-married  Not-in-family
1   50   83311  Bachelors              13  Married-civ-spouse        Husband
2   38  215646    HS-grad               9            Divorced  Not-in-family
3   53  234721       11th               7  Married-civ-spouse        Husband
4   28  338409  Bachelors              13  Married-civ-spouse           Wife

    race     Sex  Capital-gain  Capital-loss  ...  Native-country_Portugal  \
0  White    Male          2174          1873  ...                    False
1  White    Male         13062          1873  ...                    False
2  White    Male         13062          1873  ...                    False
3  Black    Male         13062          1873  ...                    False
4  Black  Female         13062          1873  ...                    False

  Native-country_Puerto-Rico  Native-country_Scotland  Native-country_South  \
0                      False                    False                 False
1                      False                    False                 False
2                      False                    False                 False
3                      False                    False                 False
4                      False                    False                 False

   Native-country_Taiwan  Native-country_Thailand  \
0                  False                    False
1                  False                    False
2                  False                    False
3                  False                    False
4                  False                    False

   Native-country_Trinadad&Tobago  Native-country_United-States  \
0                           False                          True
1                           False                          True
2                           False                          True
3                           False                          True
4                           False                         False

   Native-country_Vietnam  Native-country_Yugoslavia
0                   False                      False
```

The first step in data cleaning was to use one-hot encoding for categorical columns, mean imputation for integer variables, and NaN to replace zeros in numerical columns. The summary of the first few rows will then be used in exploratory data analysis (EDA) to uncover patterns and insights into the structure of the dataset.

Part 2: Variable Selection

To identify the most critical variables in a classification problem that predicts income levels, one must take into account factors that are likely to have a significant impact on the target variable, in this case, "Salary." Three variables from the dataset are listed below that are probably essential to resolving the classification issue:

Level of Education (or "education"): Income and education level frequently have a strong correlation. Greater education typically opens up more favourable employment options and higher-paying jobs for an individual. To differentiate between various income groups, it can be an extremely useful predictor.

Occupation: A significant factor in determining income is the kind of occupation. Salaries for different professions vary, and some careers have greater earning potential than others. For instance, salaries for executive or management positions are typically higher than those for manual labour or service-oriented positions.

Age: When predicting income levels, age can play a big role. In general, those with greater professional experience may be paid more. Additionally, because of experience and career advancement, some age groups may have a higher likelihood of being in higher income brackets.

Because of their strong correlations with earning potential, these variables are frequently included in models that predict income. However, depending on the dataset and the particulars of the population under study, the variables' relative importance may change. To verify the significance of these variables and possibly find more pertinent features, feature importance analysis or model-based approaches should be used.

Exploratory Data Analysis:

1. **Education distribution by Income:**
   The graph below examines the distribution of educational attainment among different socioeconomic categories. Most survey respondents only have a high school diploma or some college degree, based on our analysis. Nevertheless, if we look at the distribution by income, we see a fascinating pattern. A disproportionate number of people with higher income those who earn over $50,000 per year are employed in professions requiring bachelor's, master's, and PhD degrees. Based on this research, income outcomes are significantly influenced by higher education. Individuals holding advanced degrees are more likely to earn higher salaries than those with lower educational attainment.

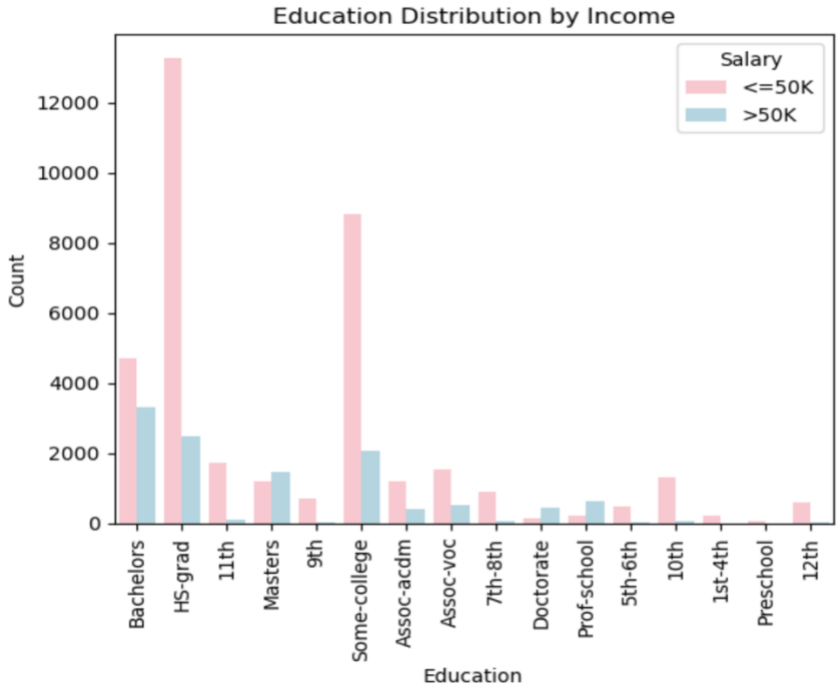


Figure 1: Education Distribution by Income

2. Occupation Distribution by Income:

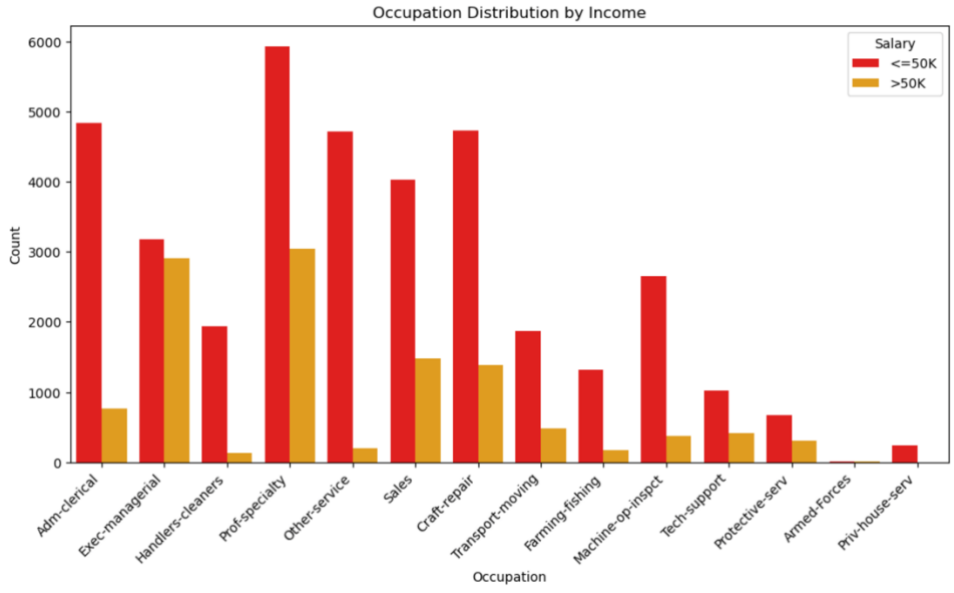


Figure 2: Occupation Distribution by Income

Based on the representation, we can identify which professions are most likely to have a greater concentration of individuals making higher salaries. More orange-barred jobs, like "Support," "Prof-specialty," and "Exec-managerial," have a higher proportion of employees earning more than $50,000 per year. On the other hand, occupations like "Handlers-cleaners," "Farming-

fishing," and "Other-service" have a higher percentage of workers making less than or equal to $50,000, as indicated by the red bars.

3. **Age Distribution:**
The histogram shows the number of people falling into each of the designated age bins, as well as the distribution of ages in the dataset. The majority of individuals are between 20 and 50 years old, with a peak in the late 30s. The distribution has a slight skew to the right.

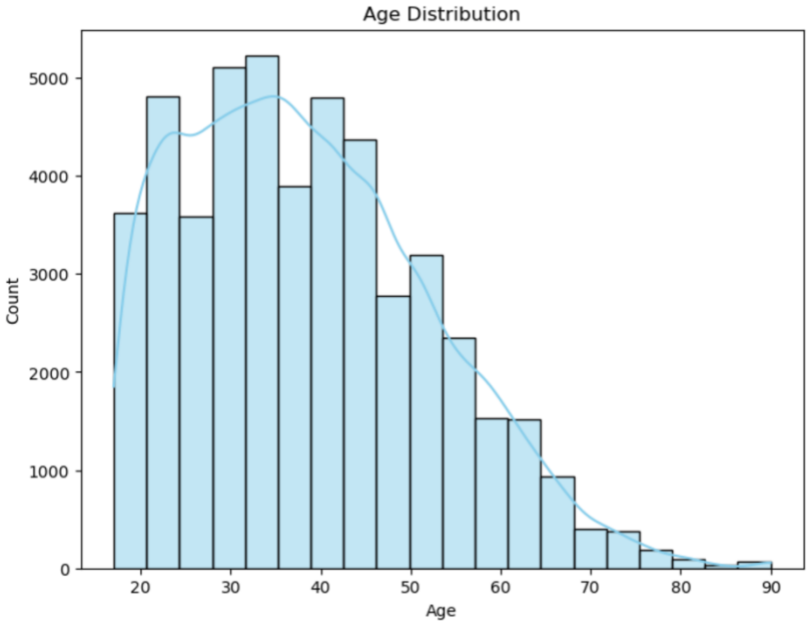


Figure 3: Age vs. Income

4. **Age vs. Income:**

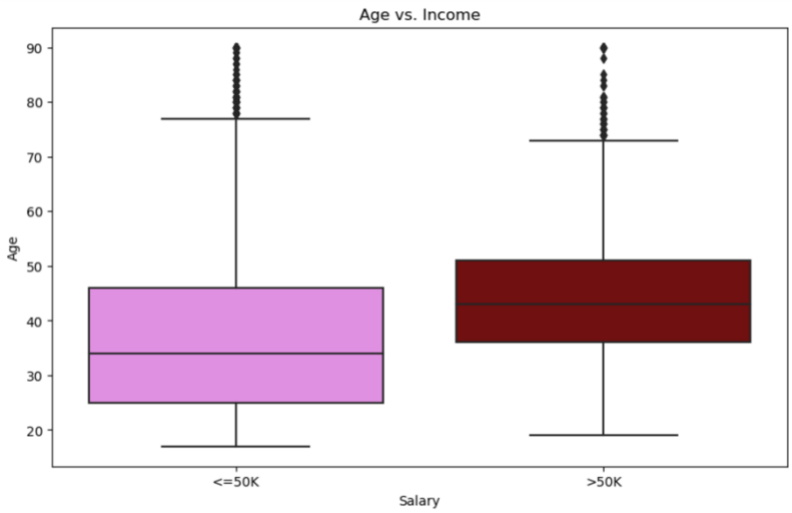


Figure 4: Age vs Income

A boxplot illustrating the relationship between age and income categories (<=50K and >50K). Generally, the median age for individuals earning >50K is higher than those earning <=50K. The age range for the various income groups is shown in the boxplot.

5. **Gender Distribution by Income:**

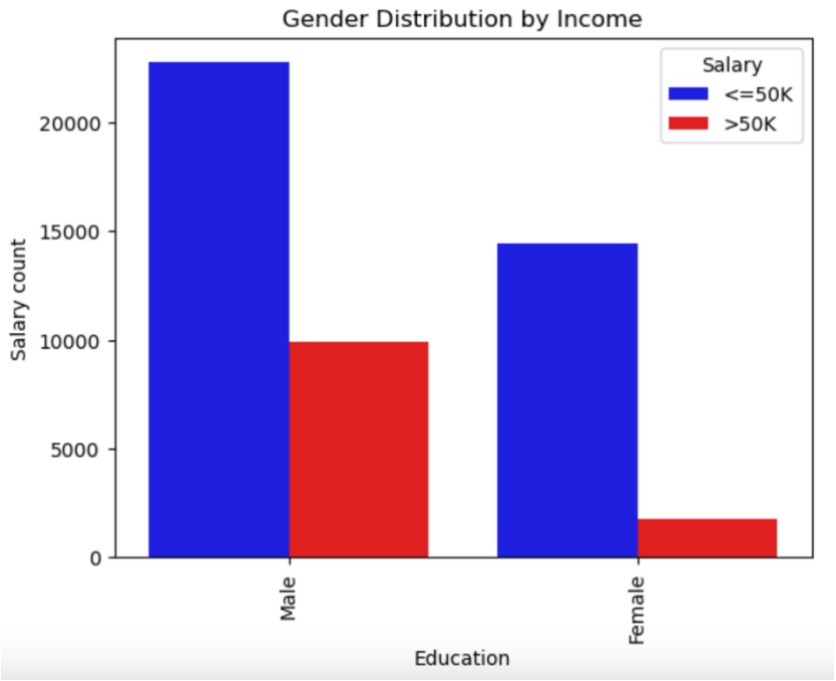


Figure 5: Gender Distribution by Income

The graphic shows that men are more prevalent in both income groups and make up a larger share of the dataset overall. But there is a notable concentration of bars among men, suggesting that men account for a higher proportion of those earning over $50,000 than do women. These findings suggest that men are represented among upper-class earners, even in the presence of potential gender-based income disparities.

6. Correlation Heatmap:

A heatmap showing the correlation matrix between the 'Age,' 'Education,' and 'Income' variables. The correlation coefficients are represented visually by the heatmap. Age and income also exhibit a moderate correlation, but age and education have a relatively low correlation. The correlation between Education and Income is not explicitly shown in this heatmap.

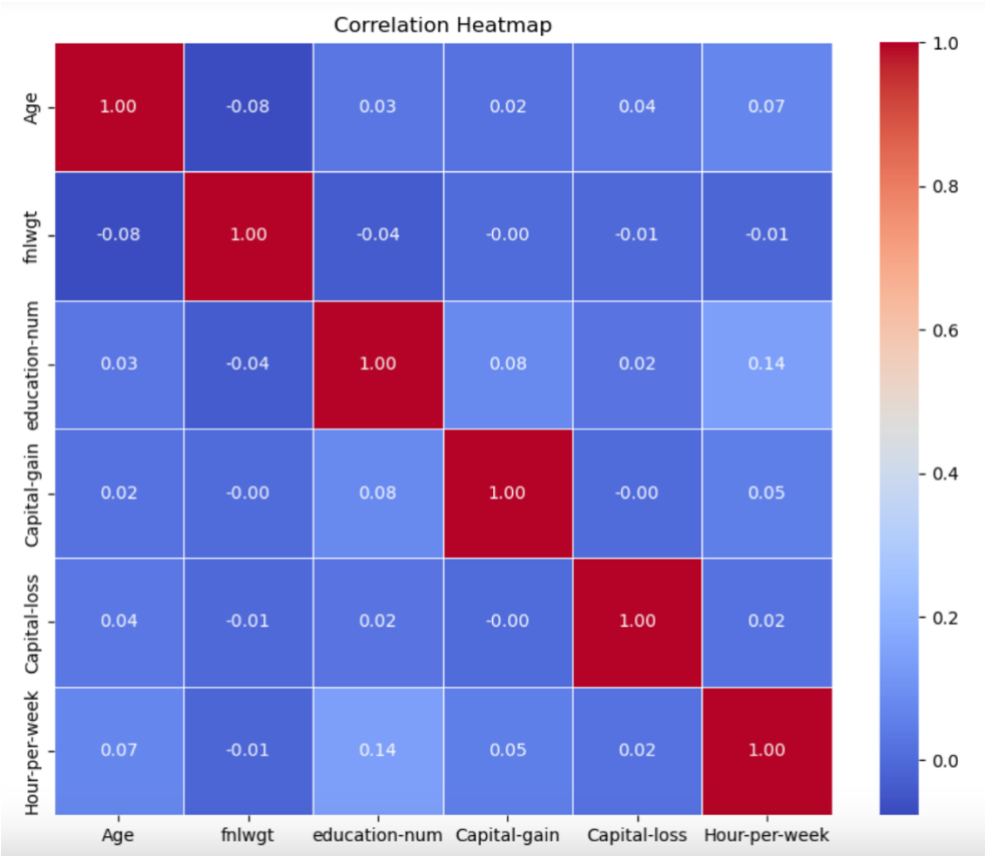


Figure 6: Correlation Heatmap

Part 3: K Value Selection

A critical step that can affect the model's performance in K-Nearest Neighbors (KNN) is determining the right value for K. Here, I'll suggest three different values for K that you can use in your KNN models. Trying a range of values and evaluating how they affect model performance is a good idea because the ideal K value frequently depends on the particulars of your dataset.

Small K Value (e.g., K=3):

A small K value makes the model more sensitive to local variations in the data. When it is anticipated that the decision boundaries will be complex and may show significant variability, a small K, like 3, is appropriate. It is susceptible to noise because it frequently records local patterns and anomalies. When there are complex relationships among the data within clusters, this can be advantageous.

Medium K Value (e.g., K=5):

Choosing K = 5 is a good compromise between capturing local patterns and generalizing them effectively. It provides some level of noise reduction and tends to offer more stable predictions compared to K = 3. This is a popular and flexible option in cases where the dataset displays both global and local patterns.

Large K Value (e.g., K=7):

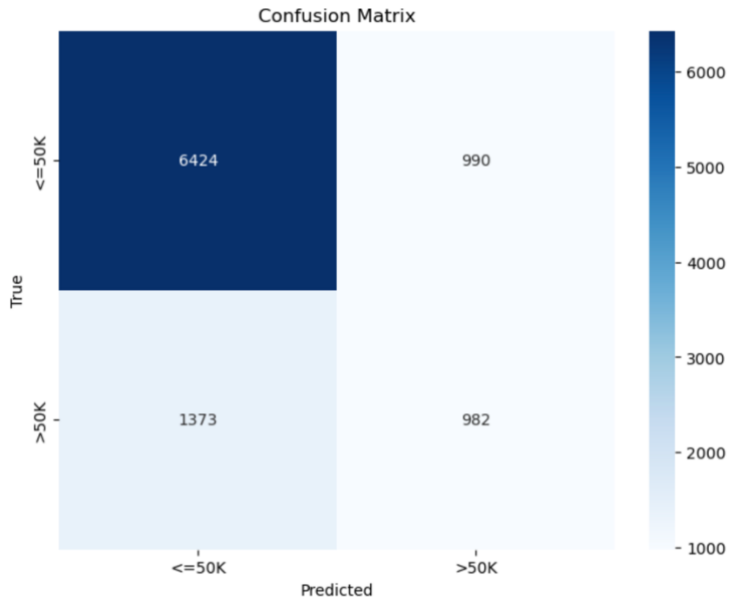
A large K value makes the model less sensitive to noise and outliers in the data.

It may work well when the dataset is large, and there is a need to capture more global trends in the data.

We can experiment with these values and observe how they affect the model's performance on our specific dataset. Keeping in mind that the optimal K value may vary, it's a good practice to perform model evaluation (e.g., using cross-validation) for each K value to select the one that provides the best balance between bias and variance.

Part 4: Model Evaluation

Confusion Matrix Analysis:



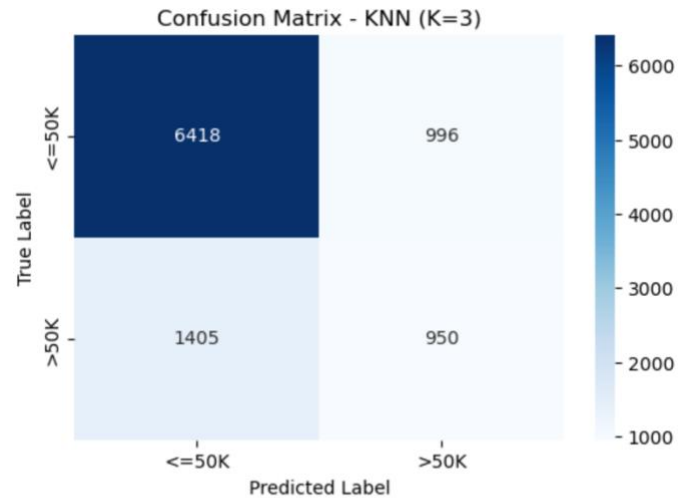
A confusion matrix is an essential tool when assessing a predictive model's performance. It offers a thorough analysis of how the model's predictions and the actual results compare. This type of analysis is particularly helpful in binary classification problems, where the objective is to categorize instances into two groups, say '>50K' (income over $50,000) and '<=50K' (income under $50,000).

This procedure shows how to apply and assess KNN models for the classification of income levels (<=50K or >50K) with different values of K (number of neighbors). The preprocessing of the dataset, which includes label encoding of categorical variables, is applied to features like age, education, and occupation. After that, the script divides the data into testing and training sets so that a thorough model evaluation is possible. Three KNN models are built, each with a different value of K (3, 5, and 7). Accuracy scores and confusion matrices are used to evaluate the models' performances. The models' predictions are broken down into detail in the confusion matrices, which are displayed using heatmaps. This helps identify True Positives, True Negatives, False Positives, and False Negatives. This data is essential for comprehending the advantages and disadvantages of every KNN model, which helps with model selection and possible parameter-tuning decisions. The presented results, which include confusion matrices and accuracy scores, add to a thorough assessment of how well the KNN models predict income levels based on demographic characteristics.

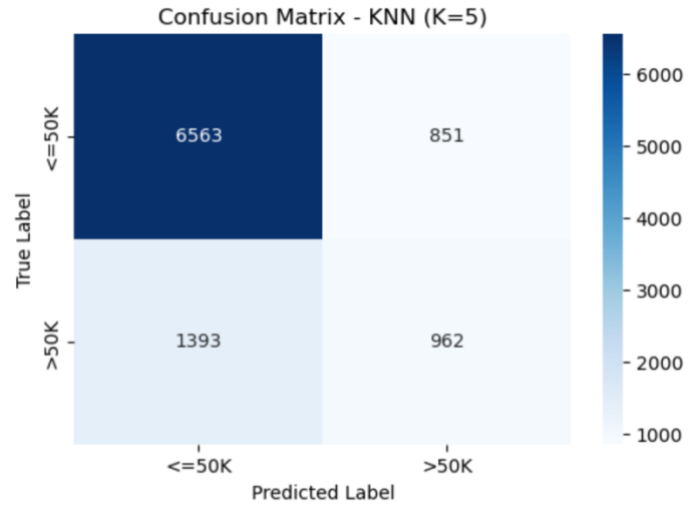
Using the constructed KNN models, I investigated the accuracy of dividing the population into low- and high-income groups according to their attributes. Three different configurations of the independent variables were used to build the models.
The accuracy results for each model are as follows:

KNN Model with (K=3)	Accuracy = 0.7542
KNN Model with (K=5)	Accuracy = 0.7703
KNN Model with (K=7)	Accuracy = 0.7730
Best KNN Model Accuracy	0.7730

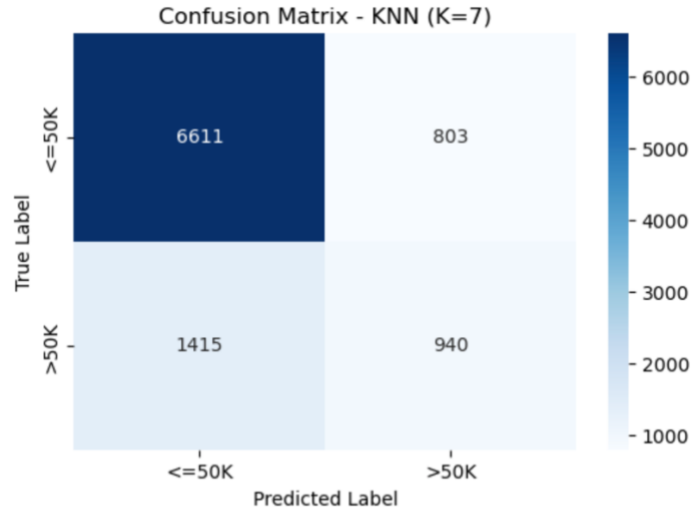
The highest accuracy recorded, 77.30% for the KNN Model with k=7, was utilized to identify the top model. In this model, the independent variables were sex, age, and education.




```
KNN Model with K=3
Accuracy: 0.7542
==============================
```




```
KNN Model with K=5
Accuracy: 0.7703
==============================
```




```
KNN Model with K=7
Accuracy: 0.7730
==============================
```

Part 5: Real-world Applicability

The KNN models showed respectable accuracy rates ranging from 75.42% to 77.30% when K (3, 5, and 7) was varied. The performance metrics indicate that the models perform fairly well in predicting income levels and differentiating between the dataset's states of equality. It is imperative to take into account the context and possible constraints of KNN models. The models' moderate predictive power is indicated by the accuracy scores, which imply that they can distinguish between people with varying income levels to some degree. The model's performance is dependent on the value of K; among the tested values, K=7 produced the highest accuracy. The usefulness of employing these KNN models in real-world situations relies on the particular application and the outcomes of incorrect classifications. Although the models exhibit encouraging accuracy, it's possible that they don't fully capture the complexity of the factors influencing income levels. Many variables in real-world scenarios may not be included

in the training set, which could result in incorrect classifications or a reduction in generalization. Furthermore, the selection of K impacts the model's susceptibility to regional patterns, which impacts its capacity to adjust to various situations.
In light of these, it is imperative to recognize that KNN models are not well suited for complex and dynamic real-world scenarios. Although their robustness in extremely dynamic or nuanced situations is questionable, they might be appropriate for some applications where the decision boundaries are clearly defined. The models' ability to reliably distinguish between people in a real-world situation could be improved by further investigation of more sophisticated machine learning models, feature engineering, and the inclusion of additional pertinent features.

Conclusion

To sum up, noteworthy performance is shown by the K-nearest neighbours (KNN) models developed to classify income levels based on demographic attributes; the best model demonstrates an accuracy of 77.30%. With varying values of K, the models illustrate how the number of neighbours affects the accuracy of the classification. Factors including age, education level, and employment are important in determining one's income level. As they provide insights into the factors influencing income disparities, these models show promise for practicality in real-life scenarios. Notwithstanding the models' resilience, it is important to recognize the complexity of socioeconomic factors and the potential benefits of more features or advanced algorithms for improving predictive accuracy. The results highlight how machine learning can provide insightful information for initiatives and policy-making targeted at advancing equity and just compensation.

References

1. *What is the k-nearest neighbors algorithm? | IBM*. (n.d.-b). https://www.ibm.com/topics/knn#:~:text=The%20k%2Dnearest%20neighbors%20algorithm%2C%20also%20known%20as%20KNN%20or,of%20an%20individual%20data%20point.
2. 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/#:~:text=The%20K%2DNearest%20Neighbors%20(KNN)%20algorithm%20is%20a%20popular,training%20dataset%20as%20a%20reference.