DATA MINING PROJECT REPORT Sl-515

Submitted by:
Sonu Gupta 22N0062

Supervised by:
Prof. Sanjeev Sabnis &
Prof. Radhendushka Srivastava
Associate Professor
Department of Mathematics
IIT Bombay

Census Income Dataset Report

Sonu Gupta, Roll No.: 22n0062

Data Information

Dataset Name:

Census Income Dataset

Overview:

This dataset is used to predict whether an individual's income exceeds $50K/year based on census data.

Creators:

Barry Becker Ronny Kohavi

Data Source:

https://archive.ics.uci.edu/dataset/2/adult

Variables

- **age:** This is a continuous variable representing the age of the individual.
- **workclass:** This categorical variable represents the type of workclass an individual belongs to. Possible values include:
  - Private
  - Self-emp-not-inc
  - Self-emp-inc
  - Federal-gov
  - Local-gov
  - State-gov
  - Without-pay
  - Never-worked
- **fnlwgt:** This is a continuous variable and stands for "final weight." It represents the number of people the census believes the entry represents.
- **education:** This categorical variable represents the highest level of education an individual has completed. Possible values include:
  - Bachelors

  - Some-college
  - 11th
  - HS-grad
  - Prof-school
  - Assoc-acdm
  - Assoc-voc
  - 9th
  - 7th-8th
  - 12th
  - Masters
  - 1st-4th
  - 10th
  - Doctorate
  - 5th-6th
  - Preschool
- **education-num:** This is a continuous variable and corresponds to the numerical representation of education levels.
- **marital-status:** This categorical variable represents the marital status of the individual. Possible values include:
  - Married-civ-spouse
  - Divorced
  - Never-married
  - Separated
  - Widowed
  - Married-spouse-absent
  - Married-AF-spouse
- **occupation:** This categorical variable represents the type of occupation the individual is engaged in. Possible values include:
  - Tech-support
  - Craft-repair
  - Other-service
  - Sales
  - Exec-managerial
  - Prof-specialty
  - Handlers-cleaners
  - Machine-op-inspct
  - Adm-clerical
  - Farming-fishing
  - Transport-moving

  - Priv-house-serv
  - Protective-serv
  - Armed-Forces

- **relationship:** This categorical variable represents the individual's relationship status. Possible values include:
  - Wife
  - Own-child
  - Husband
  - Not-in-family
  - Other-relative
  - Unmarried

- **race:** This categorical variable represents the individual's race. Possible values include:
  - White
  - Asian-Pac-Islander
  - Amer-Indian-Eskimo
  - Other
  - Black

- **sex:** This categorical variable represents the individual's gender. Possible values are Female and Male.
- **capital-gain:** This is a continuous variable representing capital gains.
- **capital-loss:** This is a continuous variable representing capital losses.
- **hours-per-week:** This is a continuous variable representing the number of hours an individual works per week.
- **native-country:** This categorical variable represents the native country of the individual. Possible values include a variety of countries such as the United-States, Cambodia, England, Puerto-Rico, and more.

Target Variable

income: This target variable is typically represented as greater than 50K or less than or equal to 50K.

Goal

The primary objective of this dataset is to make predictions about an individual's annual income. We want to determine whether a person's income exceeds $50K/year based on various factors captured in the census data. This task is framed as a classification problem where we categorize individuals into two income groups: greater than 50K or less than or equal to 50K. Our ultimate aim is to develop a predictive model that can accurately classify and predict income levels.

Data Cleaning

1. Missing Value Handling

In the dataset, there are three categorical columns with missing values. It's essential to address these missing values before building our predictive model.

Here are some common strategies we can use to handle missing categorical data:

1. **Imputation:** Replace missing values with the mode (most frequent category) of the respective column. This is a simple and effective approach, especially when the missing values are few.

2. **Create a New Category:** If it makes sense for the specific column, you can create a new category that represents missing values. This can be a meaningful option when missing values carry information.

3. **Deletion:** In cases where the missing values are extensive, you may consider removing rows or columns with missing values. However, this should be done cautiously, as it might result in a loss of valuable data.

The choice of how to handle missing values should be made based on the specific characteristics of the data and the goals of your analysis.

In this dataset, the **workclass**, **occupation**, and **native country** columns have missing values that were addressed using the strategies mentioned above.

We will use the imputation method to fix the missing value for **workclass**, **occupation**, and **native country** columns.

2. Handling Typos and Unexpected Characters

In the target variable, i.e., the income variable, there are some typos. Ideally, there should be two distinct entries representing income levels: greater than 50K or less than or equal to 50K. However, due to typos, there are four entries in the array: ['<=50K', '>50K', '<=50K.', '>50K.'], dtype=object. Therefore, it is necessary to correct these typos to ensure the accuracy of our data analysis.

To rectify this issue, we will standardize the entries to only two categories: '<=50K' and '>50K'. This correction will help maintain the consistency of the target variable for our analysis.

Exploratory Data Analysis

In this section, we delve into the results of our Exploratory Data Analysis (EDA) on the Census Income Dataset. EDA is a critical step in understanding the characteristics and patterns within the dataset. We have identified several key insights and trends that shed light on the socioeconomic factors affecting income levels.

Below, we present a series of 11 insights, each accompanied by a visualization. These insights provide a deeper understanding of the dataset and serve as a foundation for our subsequent analysis and modeling.

Let's dive into the insights obtained from our EDA.

Insight 1: Income count

The dataset reveals that a majority of individuals have an income of $50K or less, outnumbering those who earn more than $50K.

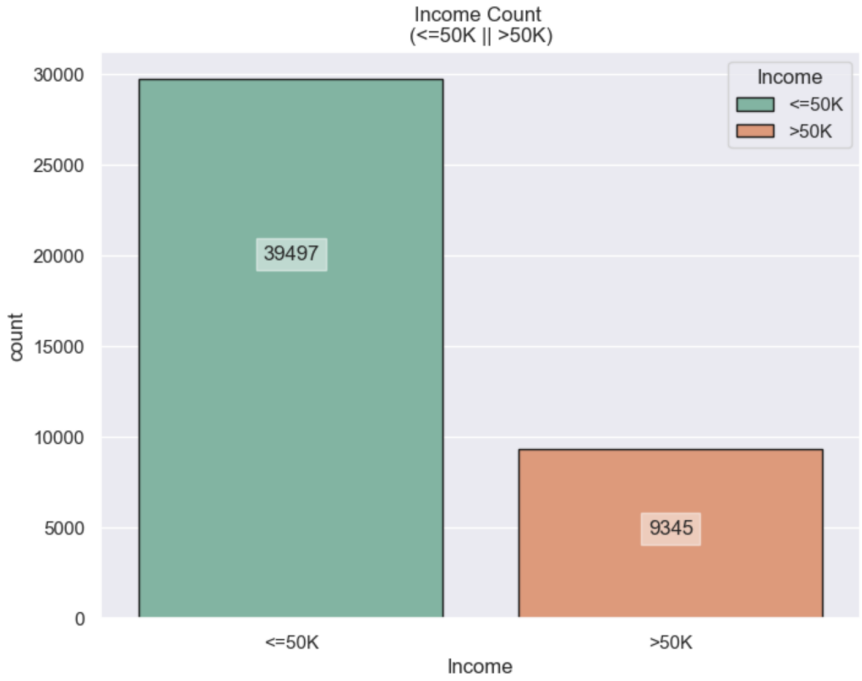

Figure 1: Income count ('<=50K' and '>50K')

Insight 2: Marital Status vs Income

The dataset includes various marital status categories, including "Married-civ-spouse," "Divorced," "Never-married," "Separated," "Widowed," and others. These categories capture the diverse marital situations of the individuals in the dataset. Marital status appears to be a significant factor influencing income levels, with "**Married-civ-spouse**" often associated with higher income. These findings provide valuable insights into the relationship between marital status and income in our dataset.

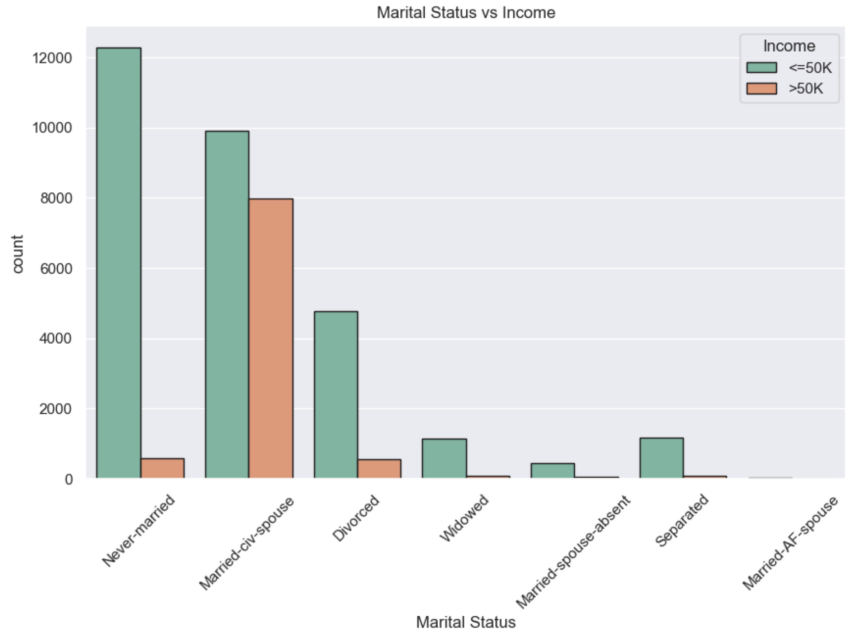

Figure 2: Marital Status vs Income

Insight 3: Sex vs Income

We explored the relationship between an individual's gender (sex) and their income level in the Census Income Dataset. Gender can be a significant factor influencing income disparities.

The plots illustrate the percentage of individuals with income greater than $50K and those with income less than or equal to 50K, segregated by gender.

Our analysis reveals that gender-based income disparities exist within the dataset. The percentage of males earning above $50K is higher than that of females. This finding suggests a gender-related income gap in the dataset.

Gender plays a significant role in influencing income levels, with males often having a higher proportion of income exceeding $50K compared to females. These findings shed light on the relationship between gender and income in our dataset.

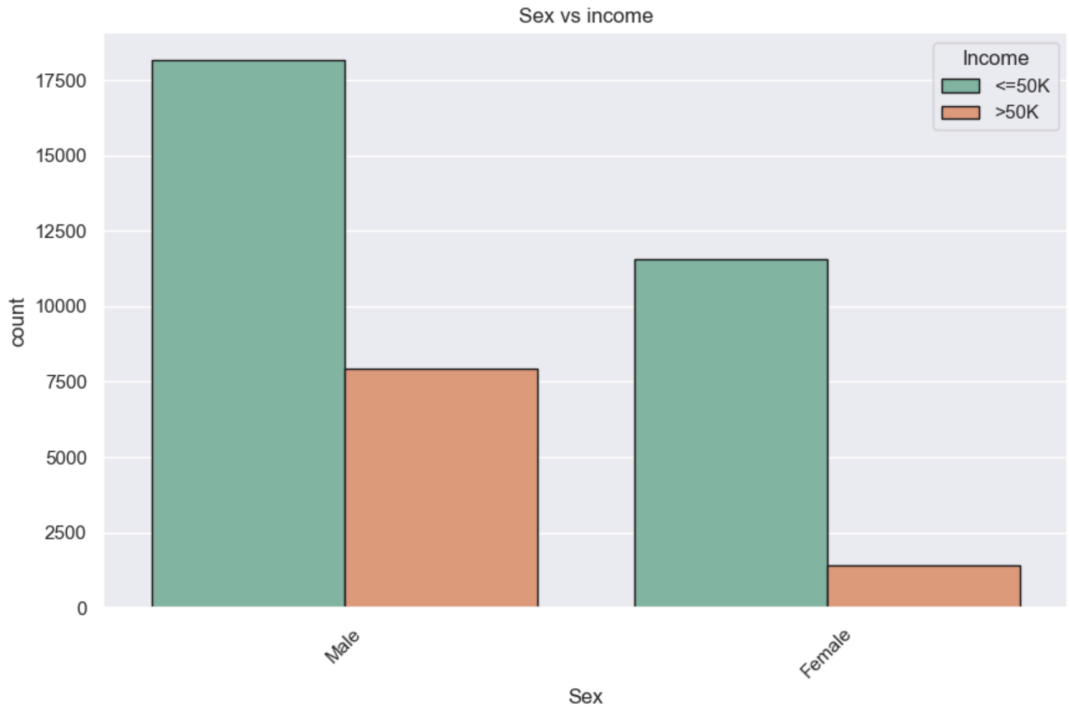

Figure 3: Sex vs Income

Insight 4: Race vs Income

In this section, we investigate the influence of an individual's race on their income level within the Census Income Dataset. It's important to understand the role that race plays in income disparities.

The data reveals significant income disparities among different racial groups. Notably, individuals who identify as White tend to have higher incomes compared to other racial groups. This finding underscores the presence of income inequality based on race.

Our analysis confirms that there are pronounced income disparities associated with race. White individuals have a higher percentage of individuals earning above $50K, while other racial groups exhibit lower percentages in this income bracket. This suggests that race can significantly impact an individual's income.

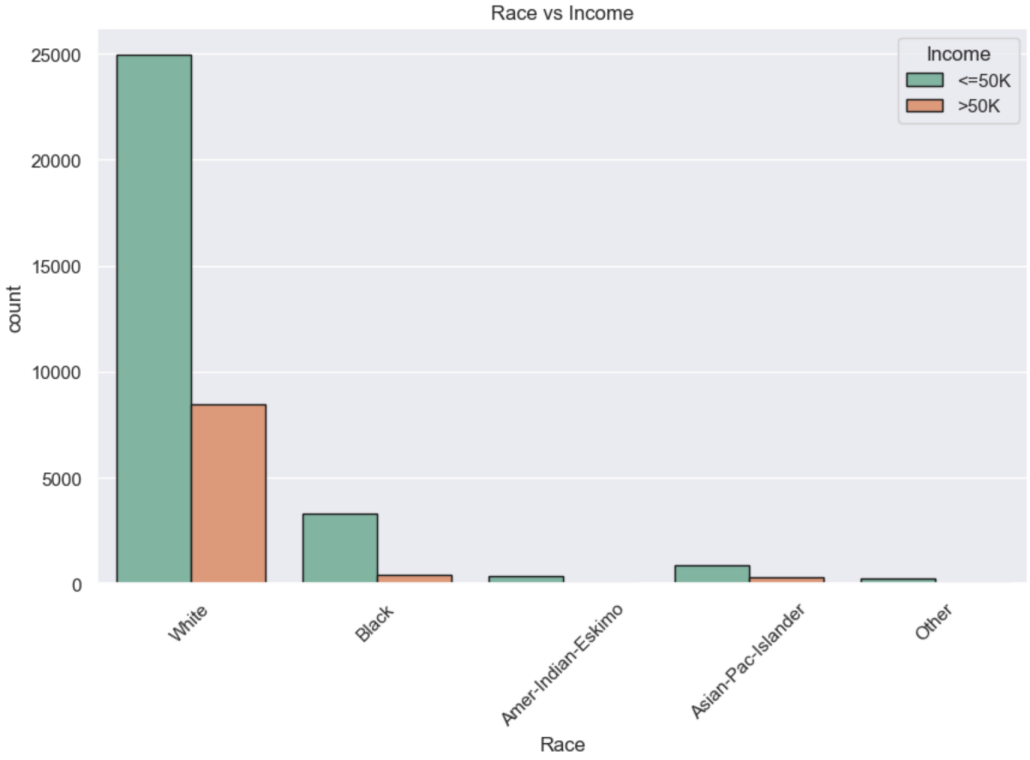

Figure 4: Race vs Income

Insight 5: Relationship Status vs Income

This section delves into the correlation between an individual's relationship status and their income level within the Census Income Dataset. The relationship status can be a significant factor in understanding income disparities.

The dataset contains various relationship statuses, including individuals who are husbands, wives, unmarried, not in a family, or not own children. We provide an overview of the distribution of individuals across these categories.

We specifically compare the income levels of husbands and wives to identify whether one group tends to have higher incomes than the other. This analysis helps us understand gender-based income disparities within married couples.

Our analysis reveals that there are income disparities based on relationship status. Interestingly, within married couples, husbands tend to have higher incomes compared to wives. Additionally, we observe income differences for unmarried individuals and those not in a family.

Relationship status is an important factor in understanding income disparities. The data highlights income variations among different relationship categories. Gender-based income disparities are noticeable within married couples, emphasizing the need to address such disparities.

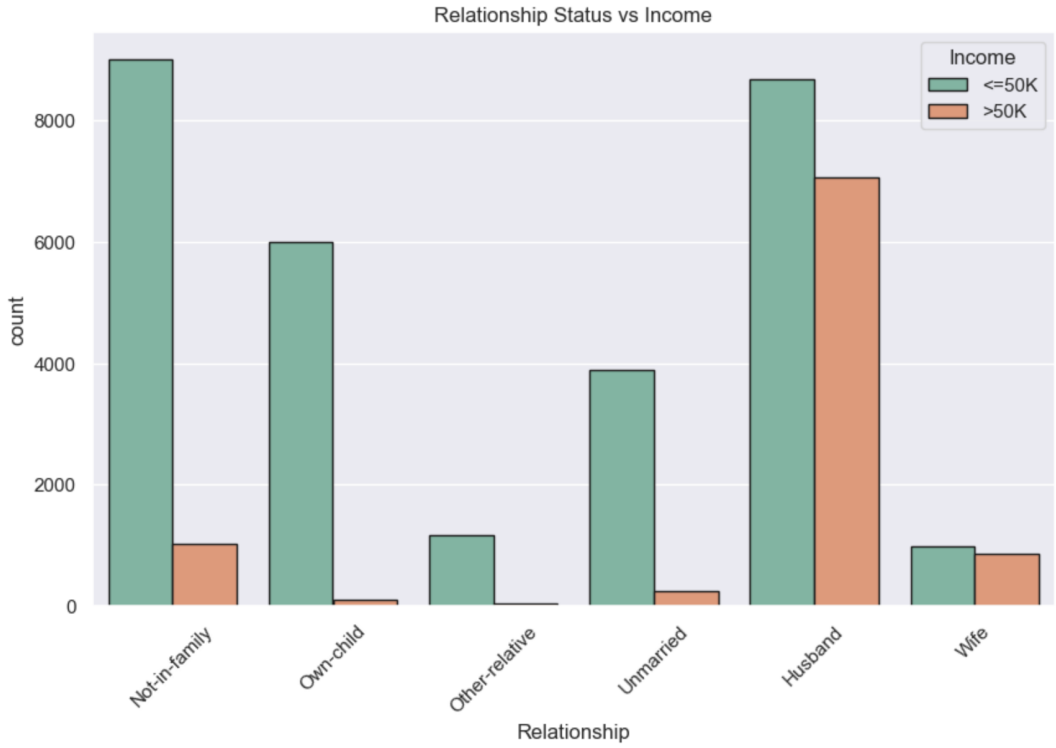

Figure 5: Relationship Status vs Income

Insight 6: Education vs Income

This section investigates the connection between an individual's level of education and their income level within the Census Income Dataset. Education is a pivotal factor in understanding income disparities.

The dataset encompasses a wide range of educational backgrounds, including individuals with various degrees, from high school graduates to doctorate holders. We observe that individuals with higher education tend to have higher incomes. We examine the income levels of highly educated individuals, such as those with high school degrees, college graduates, and postgraduate degrees. This analysis highlights the correlation between higher education and higher income.

Education plays a crucial role in determining income levels. The data underscores the trend that higher education is associated with higher income. This finding emphasizes the importance of investing in education for economic prosperity.

Insight 7: Occupation vs Income

This section delves into the influence of an individual's occupation on their income within the Census Income Dataset. Occupation is a key determinant of an individual's economic well-being.

The dataset encompasses a diverse range of occupations, spanning from technical roles to service-oriented positions.

We highlight the contrast in income levels between individuals working in the private sector and those in government jobs. This analysis underscores the income advantage of the private sector.

We identify specific occupations or job roles that are associated with higher income levels. This includes highlighting the income levels of professionals, executives, technical workers, and more.

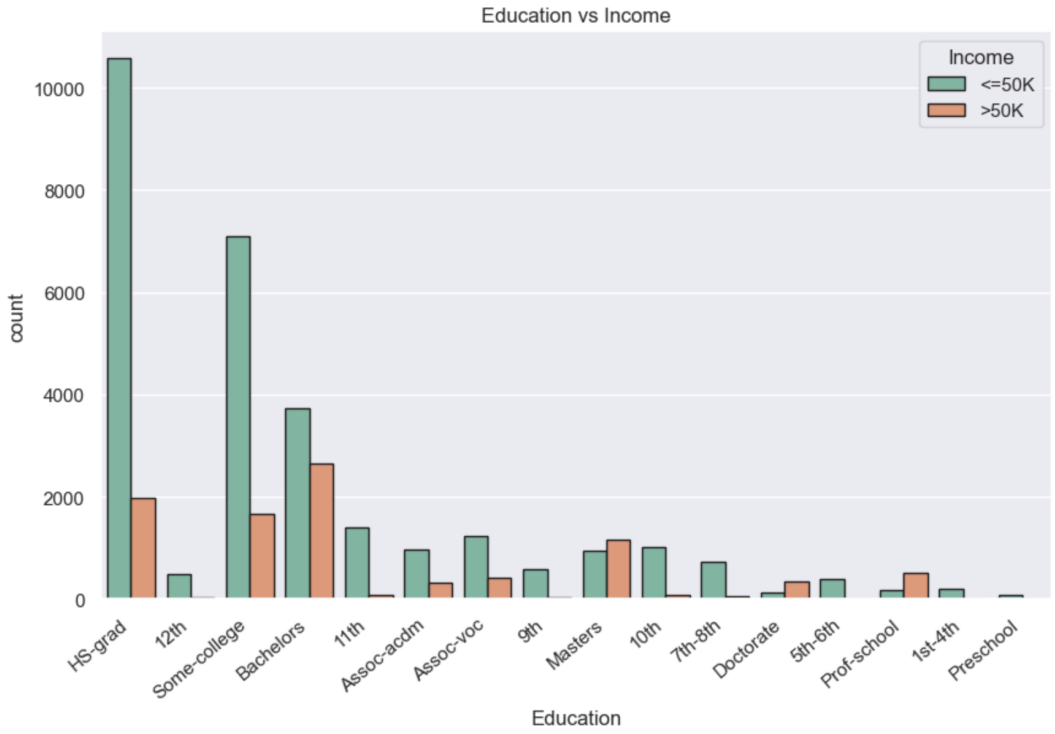

Figure 6: Education vs Income

Our analysis reveals a significant income difference between private sector employees and government workers. Private sector occupations generally yield higher incomes.

Occupation is a significant factor influencing income disparities within the dataset. Private sector jobs are associated with higher income levels, suggesting that the choice of occupation plays a crucial role in determining an individual's income.

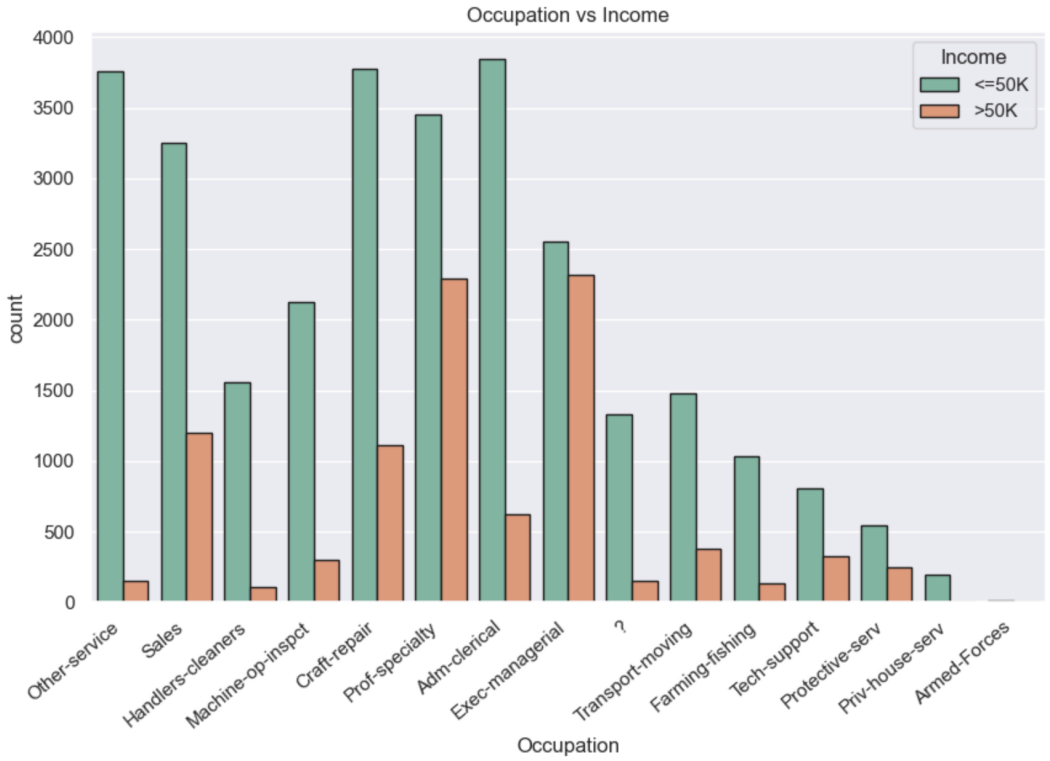

Figure 7: Occupation vs Income

Insight 8: Age Distribution

This section examines the age distribution of individuals within the Census Income Dataset. Understanding the age composition of the dataset is crucial in analyzing income trends across different age groups.

The dataset encompasses a wide range of ages, spanning from young adults to senior citizens. We provide an overview of the age distribution, highlighting the age groups with the highest representation.

Notably, we observe that the dataset's peak age group falls within the range of 30 to 40 years. This age range has the highest number of individuals, signifying the dataset's demographic composition.

We categorize the dataset into different age groups to examine income disparities. This segmentation allows us to evaluate whether income levels vary significantly based on age. Analyzing income trends within this age bracket can provide valuable insights into income disparities and economic factors affecting this age group. The age distribution within the dataset demonstrates the prominence of the 30 to 40 years age group, highlighting the importance of considering this segment when examining income levels and socioeconomic factors.

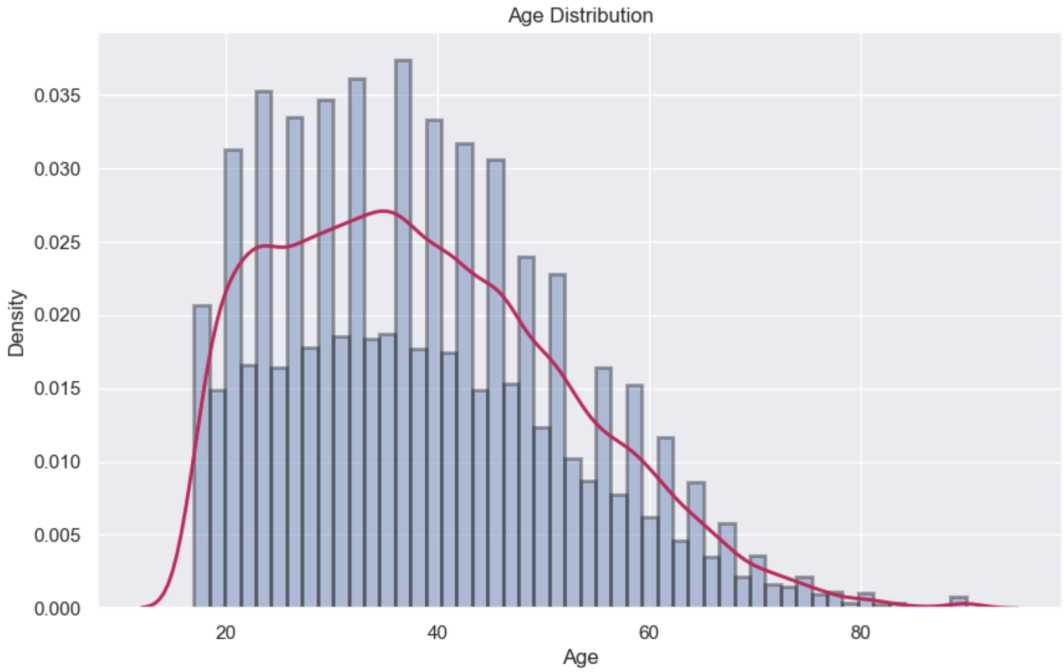

Figure 8: Age Distribution

Insight 9: Hours per week Distribution

This section explores the distribution of the number of hours individuals work per week, a significant factor in determining income levels and employment status.

We analyze the range of hours worked per week within the Census Income Dataset to gain insights into the working patterns of the individuals. This distribution plays a crucial role in understanding income disparities.

Notably, our analysis reveals that the majority of individuals in the dataset work around 40 hours per week. This specific number of hours stands out as the peak in the distribution, signifying that a significant portion of the workforce adheres to standard full-time employment.

The dataset's primary working pattern indicates that a substantial proportion of individuals maintain a standard full-time workweek of approximately 40 hours.

The hours per week distribution provides a clear understanding of the predominant working pattern within the dataset, with most individuals working around 40 hours per week. This insight forms a critical foundation for investigating the relationship between working hours and income levels.

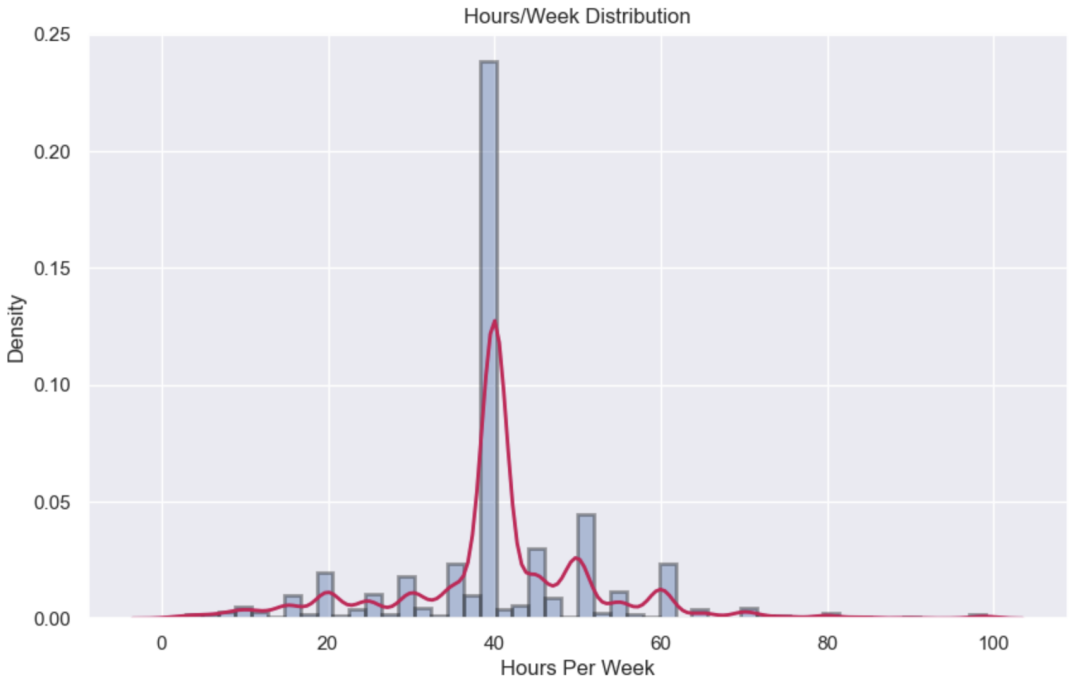

Figure 9: Hours per Week Distribution

Insight 10: Median Age vs Income

This section delves into the intriguing relationship between an individual's age and their income level. Understanding how age impacts income is crucial for gaining insights into the dataset's income distribution.

We scrutinize the dataset to uncover patterns in income levels with respect to an individual's age. This analysis is a fundamental step in exploring the impact of age on earning potential.

Our findings reveal a significant trend in the dataset. Individuals aged over 40 tend to earn more than $50,000 per year, while those under the age of 35 are more likely to earn less than $50,000 annually.

The dataset's median age distribution clearly delineates two distinct income thresholds. The demarcation at the age of 40 signifies that those older than this age are more likely to achieve a higher income status, while individuals below the age of 35 are predominantly in the lower income bracket.

This age-based income divide offers valuable insights for further exploring the demographic and economic dynamics captured in the data.

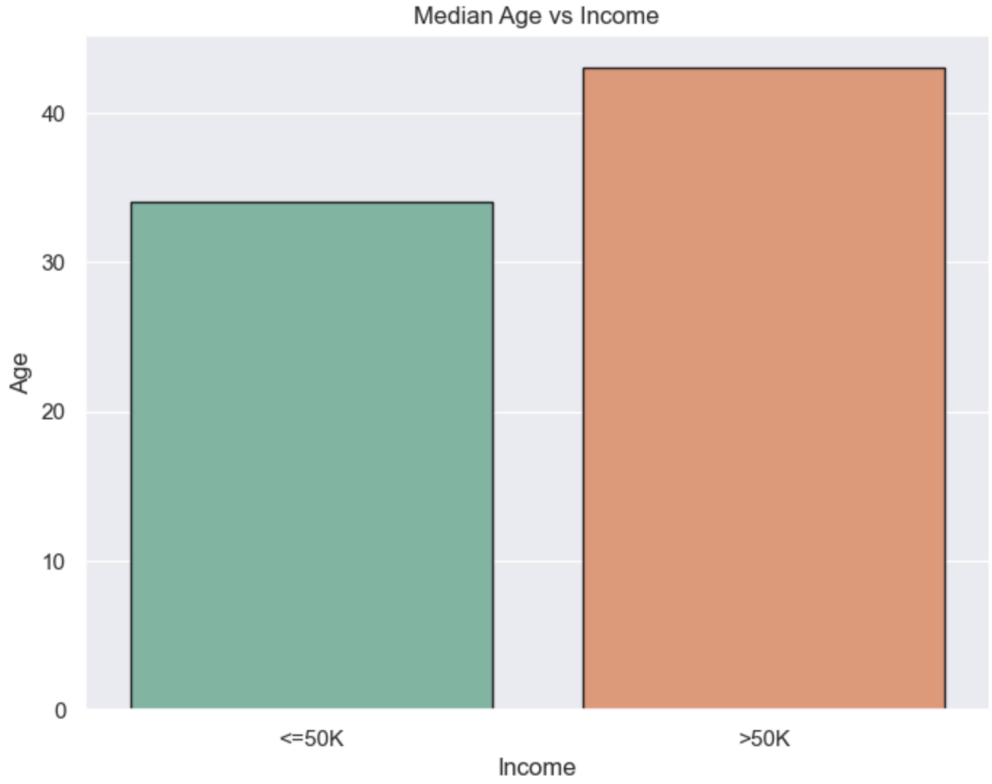

Figure 10: Median Age vs Income

Insight 11: Correlation Map

We explored the correlations between numeric variables within the dataset and investigated the presence of multicollinearity. Understanding the relationships between these variables is crucial for model building and feature selection.

Our analysis reveals that the majority of correlations fall within the range of -0.1 to 0.1. These correlations are considered weak and indicate that the numeric features are mostly uncorrelated. Weak correlations suggest that there is no significant linear relationship between these variables.

Multicollinearity, which occurs when two or more independent variables are highly correlated, can pose challenges for predictive modeling. In our dataset, the absence of strong correlations suggests that multicollinearity is not a significant concern. This is a positive outcome as it simplifies the model building process and reduces the risk of overfitting.

Understanding the correlations between numeric variables is essential for model selection and variable importance assessment. The absence of strong correlations indicates that each numeric feature contributes unique information to the predictive model.

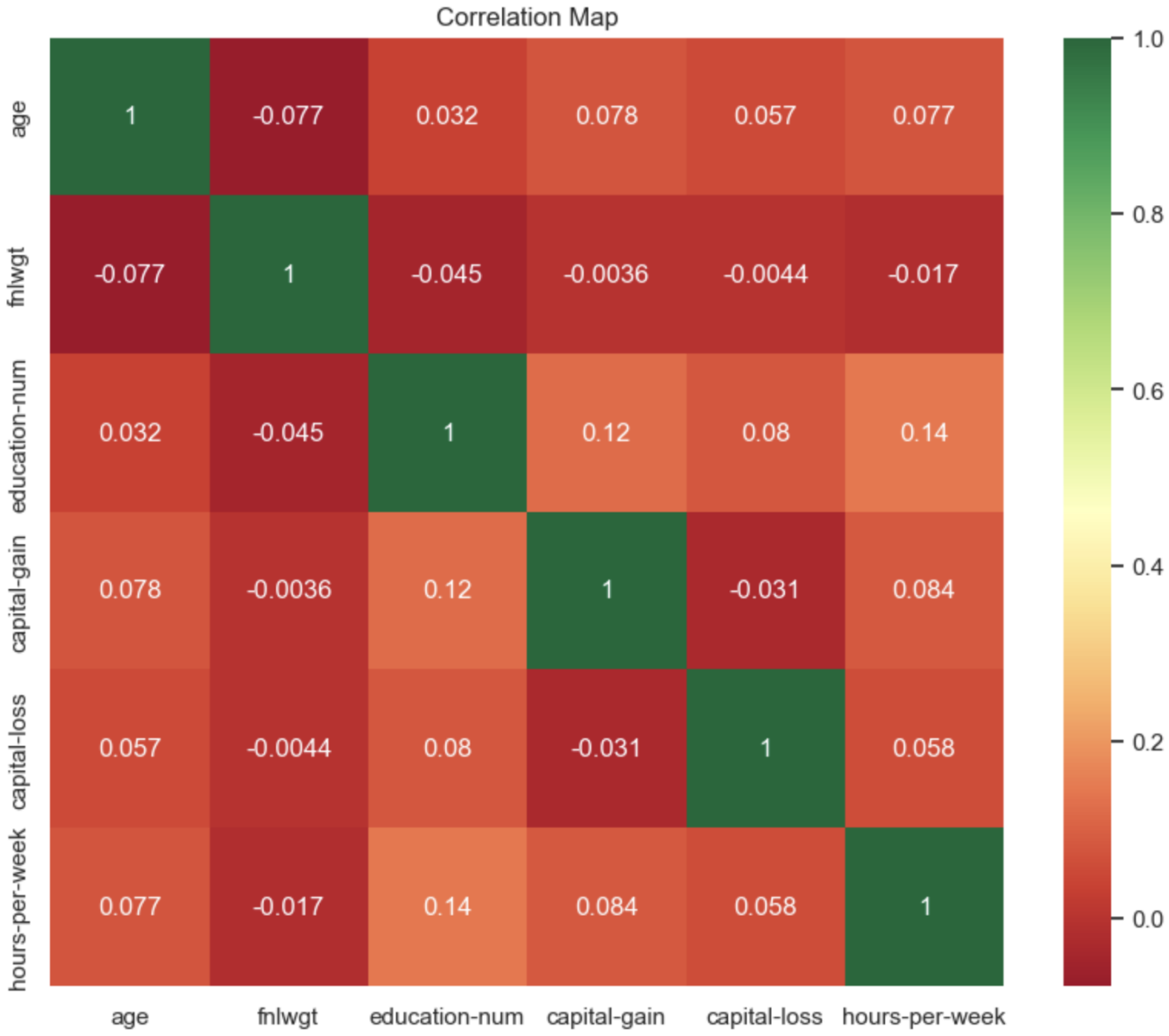

Figure 11: Correlation Map

Data Preprocessing

In the Data Preprocessing phase, we undertake essential tasks to prepare the dataset for model building. This section outlines two key steps in this process.

1. Handling Categorical Data

One of the initial challenges in working with the Census Income Dataset was the presence of categorical data. To address this, we applied Label Encoding, a technique that converts categorical variables into numerical representations. The result is a dataset where each category is assigned a unique integer value, making it suitable for machine learning algorithms.

Label Encoding simplifies the modeling process and ensures that our algorithms can work with categorical variables. It is an important step in transforming the data into a format that can be used for predictive modeling.

2. Data Splitting

To facilitate model training and evaluation, we divided the dataset into two subsets: a training set and a testing set. The data splitting was performed using an 80/20 ratio, with 80% of the data allocated to the training set and the remaining 20% to the testing set.

The training set is used to train machine learning models, while the testing set serves as an independent dataset for evaluating model performance. This approach helps us assess the model's ability to generalize to unseen data and make predictions effectively.

In the subsequent sections, we will apply various machine learning algorithms to the training set, evaluate their performance on the testing set, and present the results and insights.

The Data Preprocessing phase ensures that the dataset is ready for model building and testing, setting the stage for our predictive modeling and analysis.

3. Rescaling the Features

Some variables in the dataset have a wide range, which can lead to issues in model training and interpretation. Rescaling is essential to ensure that all features have a comparable scale. If features have disparate scales, it can result in coefficients that are either very large or very small, making the model harder to interpret and potentially leading to convergence issues.

There are two common methods for feature rescaling:

1. **Min-Max Scaling**: This technique scales features to a specified range, often between 0 and 1.
2. **Standardization (Mean-0, Sigma-1)**: Standardization transforms features to have a mean of 0 and a standard deviation of 1.

We have employed Standardization Scaling, which helps in achieving consistent feature scaling. By standardizing the features, we make sure that they have similar scales and are centered around a common mean and standard deviation. This facilitates gradient descent convergence and overall model stability.

After applying the Standard Scaler, the data exhibits a consistent scale across its features, ensuring that our machine learning algorithms can work effectively and produce reliable results.

Building a Linear Model

In this section, we will explore the application of various classification methods to predict income levels. We have employed the following classification algorithms:

1. Logistic Regression

Logistic Regression is a widely used classification algorithm. It models the probability of a binary outcome (in our case, income ¿50K or ¡=50K) using the logistic function. This method is efficient, interpretable, and provides probabilities of class membership. It's a good starting point for binary classification tasks.

2. K-Nearest Neighbors (KNN)

K-Nearest Neighbors is a non-parametric classification algorithm. It classifies data points based on the majority class of their k nearest neighbors. KNN is simple to understand and doesn't require model training. However, it can be sensitive to the choice of the parameter k.

3. Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It's robust, handles high-dimensional data, and is less prone to overfitting. Random Forest can capture complex relationships in the data and is a powerful classification tool.

4. Naive Bayes

Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem. It assumes that features are conditionally independent, given the class. Naive Bayes is simple, fast, and works well for text and categorical data. It's particularly useful for spam detection and document classification.

5. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis is a dimensionality reduction technique that aims to find a linear combination of features that best separates different classes. LDA is useful when dealing with multiple classes and can be a valuable tool for classification tasks.

6. Quadratic Discriminant Analysis (QDA)

Quadratic Discriminant Analysis is a variation of LDA that relaxes the assumption of equal covariance matrices for all classes. It's more flexible but requires more data and can handle situations where LDA is not appropriate.

In the subsequent sections, we will analyze the performance of these classification methods, assess their accuracy, and choose the most suitable model for predicting income levels based on the Census Income Dataset.

Let's apply each of the classification methods mentioned above one by one and learn how they work on my dataset.

Model Selection

I have used Logistic Regression, KNN, Random Forest, Naive Bayes, LDA, and CDA. Below is a brief explanation of each classification method, along with their performance metrics:

Logistic Regression

Logistic Regression is a linear model that predicts the probability of a binary outcome. It provides a good balance between simplicity and performance.

Method	Mean Score	Standard Deviation
Logistic Regression	0.81298	0.00305

K-Nearest Neighbors (KNN)

KNN is a non-parametric method that classifies data points based on the majority class among their k-nearest neighbors. It shows good performance.

Method	Mean Score	Standard Deviation
KNN	0.84773	0.00712

Linear Discriminant Analysis (LDA)

LDA is a method that finds linear combinations of features to separate classes. It's an effective technique for classification tasks.

Method	Mean Score	Standard Deviation
LDA	0.80417	0.00279

Random Forest

Random Forest is an ensemble method that combines multiple decision trees to improve classification accuracy. It shows the highest mean score but with a relatively high standard deviation, indicating some variability.

Method	Mean Score	Standard Deviation
Random Forest	0.88597	0.03300

Naive Bayes

Naive Bayes is a probabilistic method that assumes independence between features. It has the lowest mean score among the methods.

Method	Mean Score	Standard Deviation
Naive Bayes	0.69688	0.00757

Quadratic Discriminant Analysis (QDA)

QDA is a method similar to LDA but does not assume equal covariance matrices. It has a mean score of 0.5, indicating that it might not be suitable for this dataset.

Method	Mean Score	Standard Deviation
QDA	0.5	3.895e-05

Insight:

In ascending order of mean scores, the methods are ranked as follows: Random Forest, KNN, LDA, Logistic Regression, Naive Bayes, and QDA. Random Forest performs the best, while QDA has the lowest mean score and may not be suitable for this dataset.

Final Model Evaluation

Confusion Matrix

The confusion matrix is a critical tool in evaluating the performance of a classification model. It provides a clear summary of how well the model predicts the target variable. The matrix is divided into four key components:

- True Positives (TP): The number of correct positive predictions. These are cases where the model correctly predicted a positive outcome.
- True Negatives (TN): The number of correct negative predictions. These are cases where the model correctly predicted a negative outcome.
- False Positives (FP): The number of incorrect positive predictions. These are cases where the model predicted a positive outcome, but the actual outcome was negative.
- False Negatives (FN): The number of incorrect negative predictions. These are cases where the model predicted a negative outcome, but the actual outcome was positive.

The confusion matrix is a valuable tool to assess the model's ability to distinguish between the two classes ('<=50K' and '>50K') and understand where it may be making errors.

AUC and ROC Curve

The AUC (Area Under the Curve) and ROC (Receiver Operating Characteristic) curve are used to evaluate the performance of a binary classification model in terms of its ability to discriminate between the two classes.

The ROC curve is a graphical representation of a model's performance at different classification thresholds. It displays the True Positive Rate (Sensitivity) against the False Positive Rate (1-Specificity) at various threshold settings. A model with a higher AUC score and a curve that approaches the top-left corner of the plot is considered to have better discriminative power.

The AUC value, ranging from 0 to 1, quantifies the overall performance of the model. An AUC of 0.5 indicates random guessing, while an AUC of 1 indicates perfect classification. Higher AUC values suggest a better model.

Precision, Recall, and F1-Score

Precision, Recall, and F1-Score are important metrics for evaluating a classification model's performance.

- Precision: Precision measures the proportion of true positive predictions among all positive predictions. It helps assess the accuracy of positive predictions.

- Recall: Recall (Sensitivity) measures the proportion of true positive predictions among all actual positives. It indicates how well the model captures positive instances.

- F1-Score: The F1-Score is the harmonic mean of precision and recall. It provides a balanced measure of a model's accuracy and its ability to capture positive instances.

- Precision = $\frac{TP}{TP+FP}$
- Recall (Sensitivity) = $\frac{TP}{TP+FN}$
- F1-Score = $\frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$

In this section, we will analyze the confusion matrix, ROC/AUC metrics, precision, recall, and F1-Score to evaluate the performance of our final model.

Random Forest Model Validation

We present the results of the Random Forest classification model on the test data.

Classification Report

The Classification Report provides a comprehensive summary of the model's performance, including key metrics such as precision, recall, F1-score, and support for each class.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.87      0.89      7426
           1       0.63      0.71      0.67      2342

    accuracy                           0.83      9768
   macro avg       0.77      0.79      0.78      9768
weighted avg       0.84      0.83      0.83      9768
```

Figure 12: Random Forest Classification Report

Confusion Matrix

The Confusion Matrix allows us to examine the model's ability to correctly classify individuals' income levels.

ROC Curve

The ROC curve and AUC score are valuable for assessing the model's discriminative power in distinguishing between income levels.

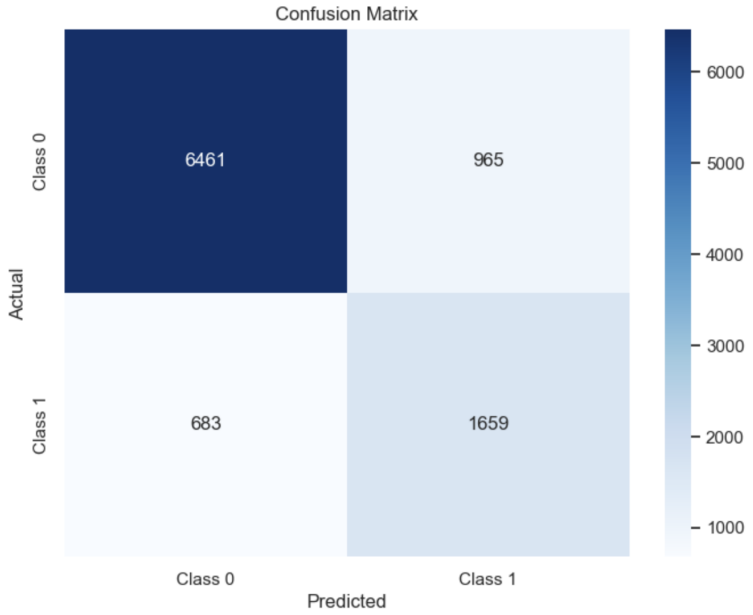

Figure 13: Random Forest Confusion Matrix

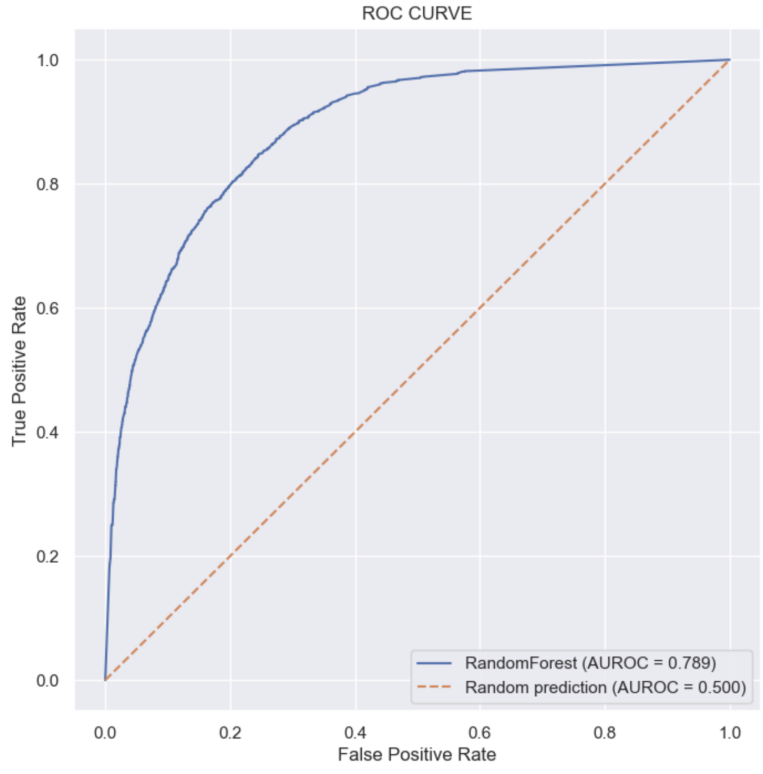

Figure 14: Random Forest ROC Curve and AUC

Logistic Regression Model Validation

We present the results of the Logistic Regression classification model on the test data.

Classification Report

The Classification Report for Logistic Regression provides a comprehensive summary of the model's performance, including key metrics such as precision, recall, F1-score, and support for each class.

Classification Report:

	precision	recall	f1-score	support
0	0.94	0.78	0.85	7426
1	0.55	0.83	0.66	2342
accuracy			0.79	9768
macro avg	0.74	0.81	0.76	9768
weighted avg	0.84	0.79	0.81	9768

Figure 15: Logistic Regression Classification Report

Confusion Matrix

The Confusion Matrix for Logistic Regression allows us to examine the model's ability to correctly classify individuals' income levels.

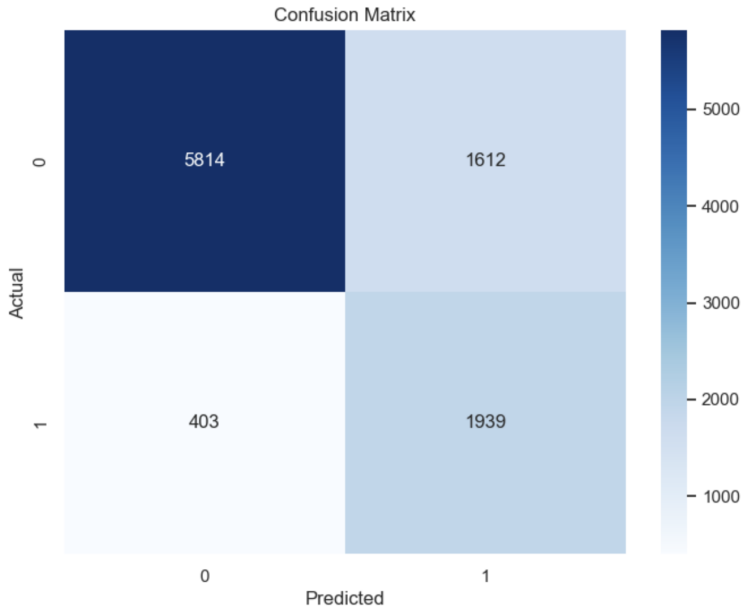

Figure 16: Logistic Regression Confusion Matrix

ROC Curve

The ROC curve and AUC score for Logistic Regression assess the model's discriminative power in distinguishing between income levels.

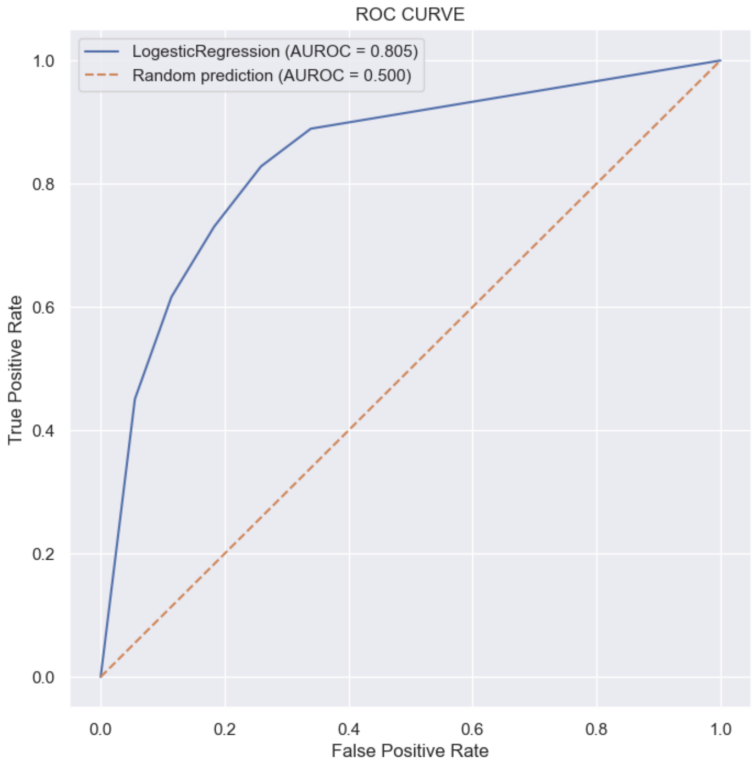

Figure 17: Logistic Regression ROC Curve and AUC

KNN Model Validation

We present the results of the K-Nearest Neighbors (KNN) classification model on the test data.

Classification Report

The Classification Report for KNN provides a comprehensive summary of the model's performance, including key metrics such as precision, recall, F1-score, and support for each class.

Classification Report:

	precision	recall	f1-score	support
0	0.91	0.82	0.86	7426
1	0.56	0.73	0.63	2342
accuracy			0.80	9768
macro avg	0.73	0.77	0.75	9768
weighted avg	0.82	0.80	0.80	9768

Figure 18: KNN Classification Report

Confusion Matrix

The Confusion Matrix for KNN allows us to examine the model's ability to correctly classify individuals' income levels.

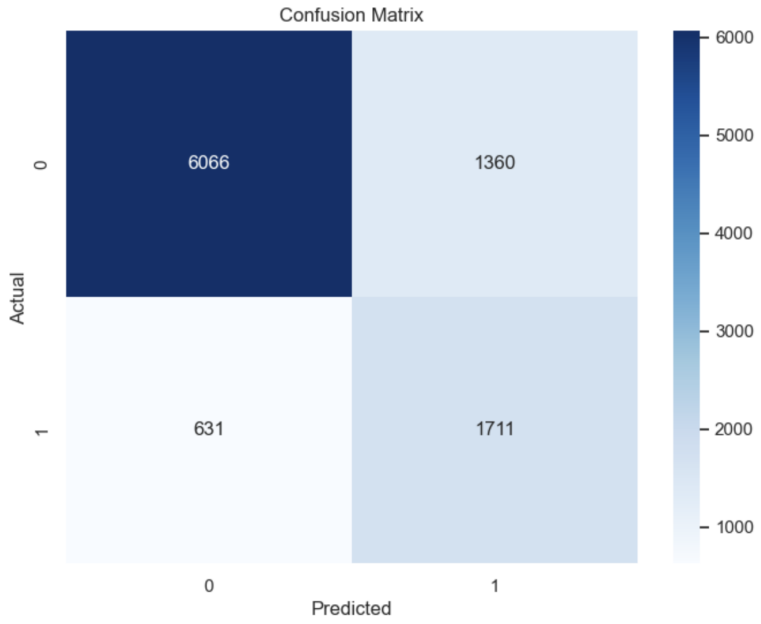

Figure 19: KNN Confusion Matrix

ROC Curve

The ROC curve and AUC score for KNN assess the model's discriminative power in distinguishing between income levels.

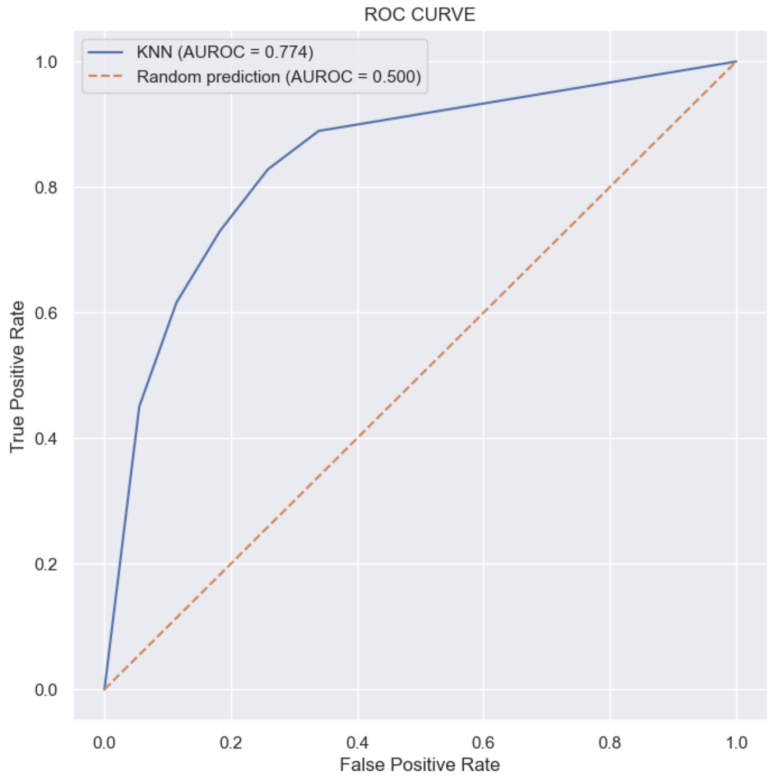

Figure 20: KNN ROC Curve and AUC

Conclusion

Based on the evaluation of the three classification methods, it is evident that Random Forest exhibits strong performance on the training data. However, Logistic Regression outperforms Random Forest on the test data, making it the preferred model for predicting income levels. KNN, while reasonable, falls slightly behind the other two methods.

In conclusion, Logistic Regression is the most suitable model for predicting income levels in this dataset. It offers a good balance between training and test performance, making it a reliable choice for real-world applications.