Income Factors: A Data Mining Analysis of Demographic and Socioeconomic Influences

CIS 9660

Group 3:

Salena Huynh
Cathy Luy
Ngoc Khanh Ha Nguyen
Phuong Trang Tran
Brandon Yan
Calvin Zhang

1. Introduction

There is a constant pursuit of a higher income in current society. With economic changes like the rising cost of living and increasing inflation rates, many strive for higher-paying jobs. Companies are hiring individuals based on career progression, with negotiations being an important determinant of salary. It is important to understand how different factors impact one's income. This report will aim to accomplish this by exploring the "Adult Income" dataset available on Kaggle [1]. The dataset provided two CSV files, and we chose to use the train set since there were more data points. This dataset is collected from a Census database that details an individual's income and information like their demographic group. We used income as our independent variable, with the remaining columns as our dependent variable. This contains various predictors including age, work class, education, marital-status, occupation, relationship, race, sex, hours-per-week, and native country.

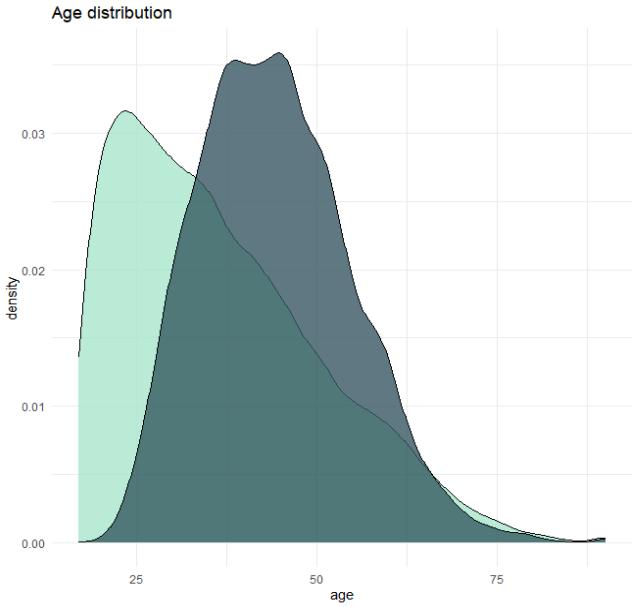
Using data mining methods to analyze the dataset, we wanted to investigate how different demographic and socioeconomic factors influence an individual's income. We hoped to uncover patterns that can guide us to understand fairness and equality across the workforce.

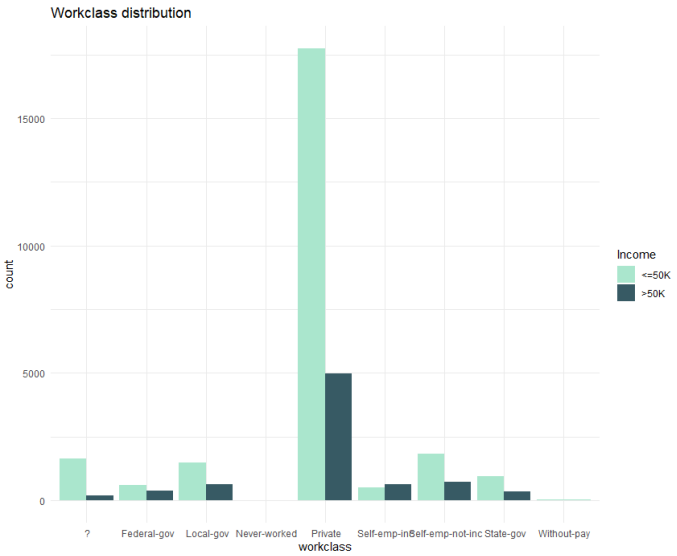
2. Data cleaning

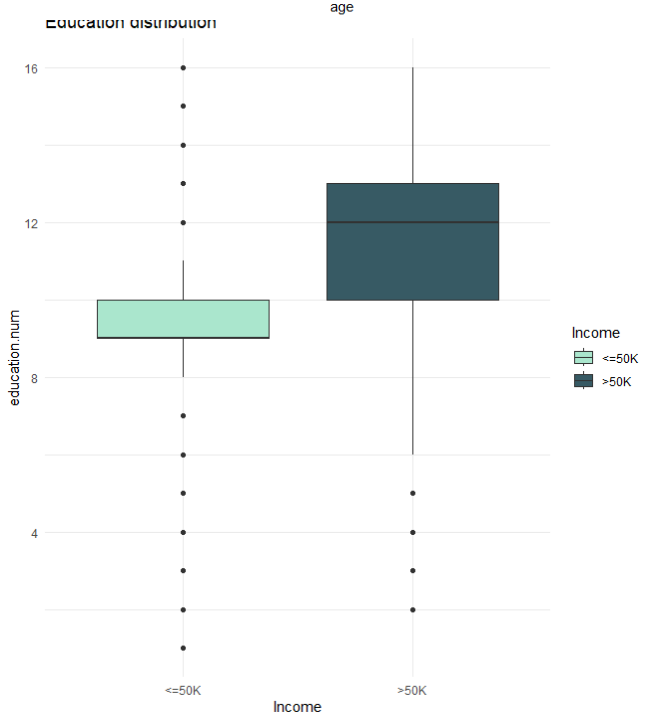
The first step of the analysis process was to clean the data. We examined the column names and removed columns we did not need such as capital.gain, capital.loss, and X. Looking through the dataset, we noticed "?" was used for cells with missing information. To remove these null cells, we replaced all the "?" with NA and omitted the entire row. Since the majority of the variables are categorical, we created new columns to transform qualitative into quantitative data by using as.factor(). This included our independent variable; we encoded it into the binary numeric variable Income.num, where we receive individuals earning more than $50K labeled as 0 and 1. With the cleaned dataset, we moved on to the summary statistics.

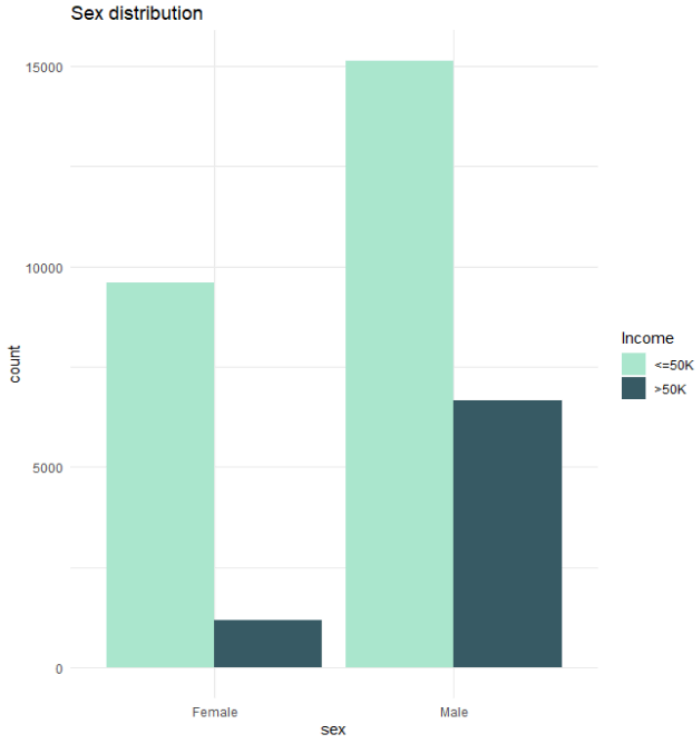
3. Summary Statistics

Before diving deeper into the different data mining techniques, it was essential to understand the key attributes of this dataset. We did this by looking into the summary statistics. This data consists of 30,162 data points that detail ten different demographic and socioeconomic factors that could influence an individual's income. Looking at the statistical summary and visualizations, the dataset shows a wide range of individuals.









The dataset covers individuals aged between 17 and 90, with an average of 38.44 years and a median age of 37. The older generation tends to have a higher income than the young. Looking into work classes, the most common is private, with a disparity within private. This is followed by self-emp-not-inc and local-gov. Most of the individuals in this dataset only finished education as a high school graduate, followed by some college and Bachelor's degrees. Individuals with a higher education level typically earn above 50 thousand a year. Also, understanding the familial and social dynamics of individuals is important for comprehending their socioeconomic contexts. Most of their relationship status are husbands, not-in-family, and own-child. The reason individuals are mostly husbands could be because of the 30,162 data points in this set: 20,380 are males and 9,782 are females. When it comes to racial diversity, it shows White, Black, and Asian-Pacific-Islander as prominently represented. Most individuals in this dataset work on average 40.93 hours per week with a minimum of 1 and a maximum of 99 hours per week. Based on the summary statistics, individuals who have a higher education level typically earn above $50,000 a year. To dive deeper into these findings and develop a comprehensive understanding of the demographics and socioeconomic landscape, we will analyze our dataset with data mining methods.

4. Data Mining Methods

Data mining methods are important to gain insight and make decisions based on datasets to uncover hidden patterns and structures of complex datasets. We will explore techniques like logistic regression, KNN, decision trees, clustering, and unsupervised learning to uncover hidden patterns of adult income.

4.1. Logistic Regression

For the dataset, we chose to run a logistic regression analysis. The reason is our independent variable is binary, which means that we had an adult's income as a variable and wanted to find out if it exceeded $50,000 per year or not. Logistic regression is effective for large datasets and could train more complex algorithms. The first model analyzed if factors, such as age, race, and gender affect income greater than $50,000.

	Untrained Model	Trained Model
(Intercept)	-4.499 (0.195) ***	-4.510 (0.227) ***
age	0.042 (0.001) ***	0.043 (0.001) ***
race Asian-Pac-Islander	0.996 (0.203) ***	0.985 (0.237) ***
race Black	0.178 (0.196)	0.068 (0.229)
race Other	-0.206 (0.300)	-0.296 (0.354)
race White	0.863 (0.188) ***	0.825 (0.219) ***
sex Male	1.212 (0.036) ***	1.218 (0.042) ***
AIC	30498.903	22782.672
BIC	30557.103	22838.858
Log Likelihood	-15242.451	-11384.336
Deviance	30484.903	22768.672
Num. obs.	30162	22621

*** p < 0.001; ** p < 0.01; * p < 0.05

By running the regression, age seemed to affect income. The coefficient was at 0.0419, which meant that for each one-unit increase in age, the log odds of having an income higher than $50,000 increased by 0.0419. Calculating our coefficient of age, $e^{0.042}$, we got 1.042 which indicates that with each one-year increase in age, the odds of having an income greater than $50,000 increased by 4.2%. Age had a very small p-value of less than 0.05, which is statistically significant. The older the individual is, the higher the chances of earning more than $50,000. For race, the different categories included Asian-Pacific-Islander, Black, Other, and White. Looking at it more specifically, we saw that race Asian-Pacific-Islander had the highest coefficient, which meant that Asians-Pacific-Islander had 2.71 times higher odds of having income greater than $50,000 compared to the other categories. Race White seemed to have 2.37 odds of earning more than $50K. Also, the p-values of both were very small, which indicates race Asian-Pacific-Islander and White are statistically significant in predicting income. However, race Other and Black p-values were greater than 0.05, demonstrating that those race categories are not statistically significant and do not impact income. The next predictor was gender and based on our model, we saw that males had 3.36 higher odds of earning income greater than $50K compared to females.

We chose to train the dataset with a random sample of 75% of the dataset to see a more accurate model. Training the dataset helps build a predictive model because algorithms can learn patterns, relationships, and the structure of the data. In addition, the trained dataset allows us to evaluate the model performance better. The training dataset was a subset of our trained logistic regression with the same predictors. After running the logistic regression with the trained subset, the coefficients and statistical significance of the dependent variables (age, race, and sex) remained consistent between the untrained and trained datasets. With this observation of the trained and untrained, we concluded that there was a consistency between the predictors and target variables. However, the difference was in the model statistics, such as null deviance, residual deviance, degree of freedom, and AIC. The null deviance represents the model's fit when no predictors are included. The untrained dataset had 33,851 on 30,161 degrees of freedom while the trained dataset decreased to 25,386 on 22,620 degrees of freedom. The reduction indicates that the trained model provided a better model fit compared to the untrained one. The residual deviance measures the model's fit including the predictors. The untrained data had a residual deviance of 30,485 on 30,155 degrees of freedom compared to the trained dataset with 33,769 on 22.614 degrees of freedom. This decrease in residual deviance shows that the trained model was a better fit for the data. Overall, it explains the variability in the outcome variable is greater. The AIC measures the relative quality of the statistical model. In the untrained dataset, the AIC was 30,499 while it decreased in the trained set to 22.783. In conclusion, the model statistics of the trained dataset were better than the untrained because all the numbers decreased, which indicates a better model fit for our data.

We created a contingency table (confusion matrix) to compare the predicted model with the actual for our target variable Income. With the confusion matrix, we calculated the accuracy rate of the proportion of the correct predictions, true positives, and true negatives to the total number of observations. The accuracy rate for the untrained dataset was approximately 77.55%, which meant that the model correctly predicted an income greater than $50,000 for 77.55% of individuals in the dataset. The precision rate was 62.62% and represented the proportion of correctly predicted true positives out of true positives and false

positives. It shows the proportion of correctly predicted high-income individuals out of all individuals with a predictive high income. The recall rate was 24.32% and represents all the true positives out of all true positives and false negatives. It means the proportion of correctly predicted high-income individuals out of all actual high-income individuals in the dataset. Compared to the untrained rates, the trained dataset rates are 72.01% for accuracy rate, 25.24% for precision rate, and a recall rate of 6.34%. All the rates decreased, which suggests that the model's performance worsened when evaluating new and unseen data. The reason for the decrease in performance could be overfitting, generalization, missing values, and outliers. It could also be due to the income proportion prior to training the dataset. The dataset contained mostly individuals with under $50,000 income, and splitting the data to 75% can increase the ratio of under 50k.

Overall, the trained model performed better for our model because the coefficients and their p-values showed the same outcome of statistical significance. It is consistent between the trained and untrained models. The model statistics are lower in the trained model and lower deviance values and AIC indicates a better fit of the model to the data. Even though the accuracy, precision, and recall rates were higher in the untrained model, it could be due to the model's ability to generalize to new and unseen data. The most important factor is that the trained model can generalize unseen data. The lower performance in the trained model reflects a more realistic performance in predicting income greater than $50,000.

Logistic regression is a powerful tool for binary classification; however, it could not capture the outcome effectively. This is a limitation of logistic regression and to explore alternative classification methods, we considered the k-nearest neighbors (KNN) method. It will classify data points based on the k nearest neighbors.

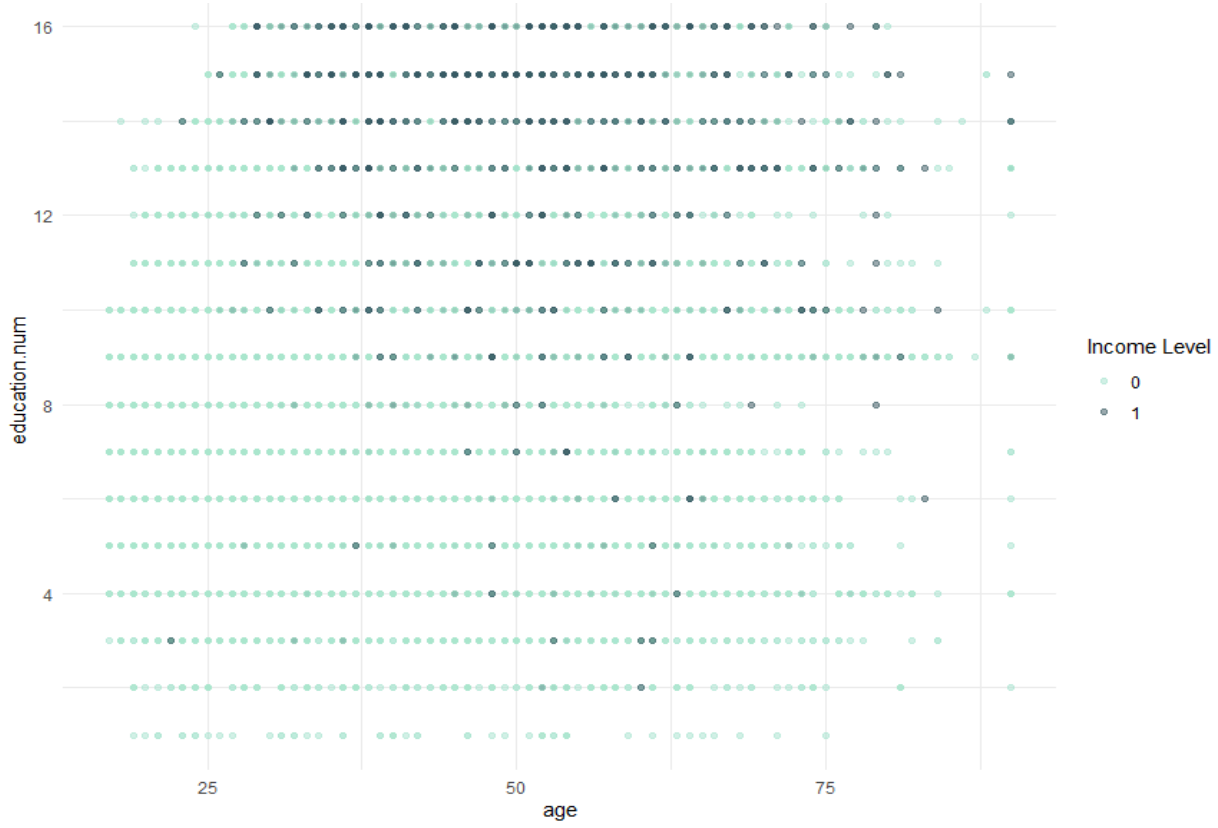
4.2. KNN Model

We created a KNN model to predict income levels using certain features such as age, hours worked per week, and education levels. We included age as a predictor under the assumption that career progression

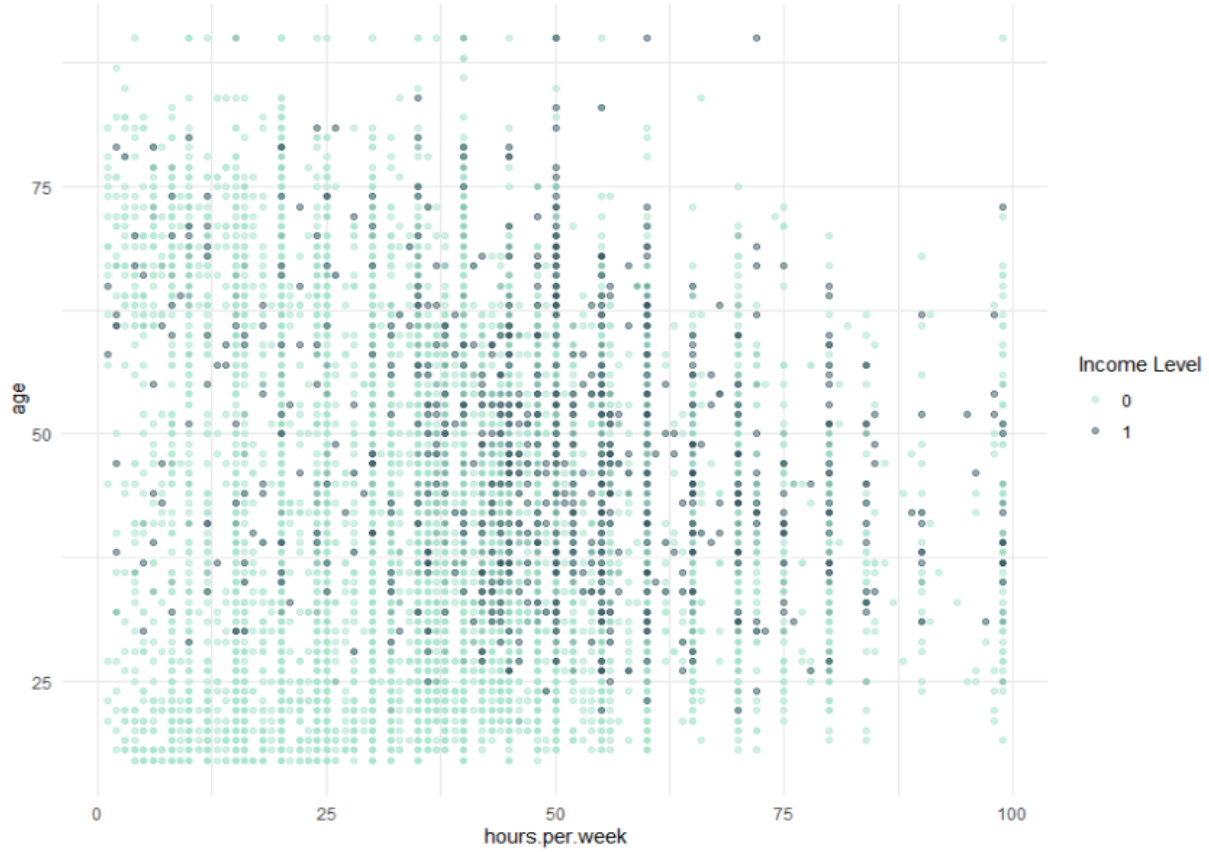
and experience might lead to higher income based on a research study by Erica York “Average Income Tends to Rise with Age,” where she discusses how income changes throughout an individual's life [2]. We picked hours per week to explore the hypothesis that higher work investment translates to higher income, based on the U.S. Census Bureau’s May 2022 current population survey which shows that America’s top 10% income percentile works 4.4 hours more each week than those in the bottom 10% [3]. We also picked education level based on the premise that higher education attainment often leads to better-paying jobs, which can be backed up by the U.S. Census Bureau’s findings that college-educated workers earn over $40,000 more than those without a college degree [4]. Using the KNN model we hoped to explore the relationships between income levels and how these features commonly believed to correlate to income can successfully capture these trends. Specifically, how these factors can predict whether someone makes less than or more than $50,000. To start the model configuration, we split the data using a 75% to 25% ratio and then normalized the data before running the model. For our project, we tested both standardizing and normalizing. Normalizing the data by scaling each feature to a range between 0 and 1 gave us better results. It was crucial to normalize it because each variable had a different scale. For example, age ranged from 17 to 90, education from 1 to 16, and hours per week from 1 to 99. Normalization was needed since KNN relied heavily on distance calculations so evenly scaling each variable was important for enhancing the model's performance. We then tried various k values for the KNN algorithm and found that setting k to 5 gave the best results. The model yielded 761 true positives which represent where our model identified individuals who make more than $50,000, 5145 True Negatives representing where the model predicted individuals who made less than $50,000, 497 false positives which represent where our model incorrectly labels individuals who make more than $50,000, and 1137 false negatives which shows where the model failed to identify actual individuals who make more than $50,000.

The model had an accuracy of about 78.33%. This decently strong result confirmed that our model is fundamentally solid and useful for practical income classification. This supports the hypothesis that age, hours worked per week, and education levels play a role in classifying whether individuals will make

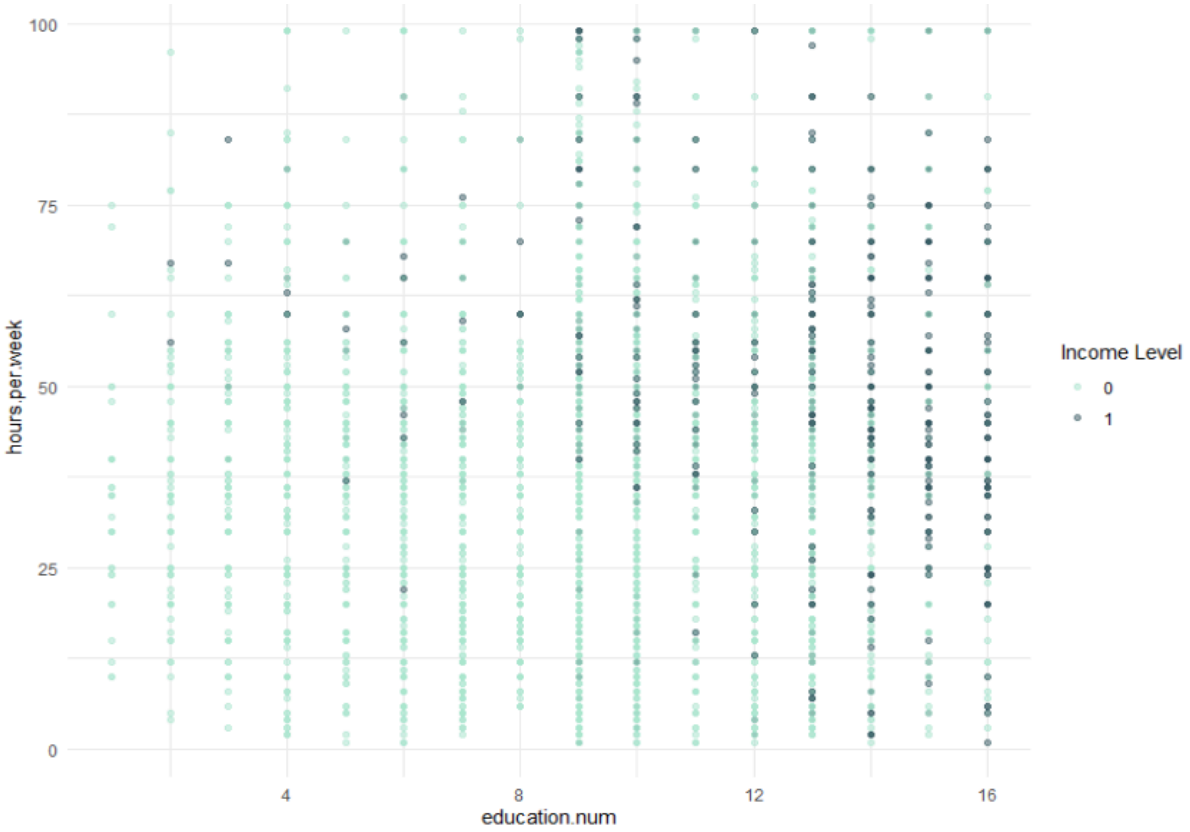
more or less than $50,000. To further our analysis we created scatter plots to support our KNN Model and gain deeper insights into how well our chosen predictors forecast income.



The Education vs. Age scatter plot is where we can observe the distribution of individuals across different ages and education levels segmented by income level. From this plot, we can see different income brackets, the lighter color being individuals who make less than $50,000 and the darker color being individuals who make more than $50,000. We see an increase in the density of higher income levels as the education level increases, indicating a consistent trend across different age groups that higher education correlates with higher income.



The Hours Per Week vs. Age scatter plot shows the relationship between the number of hours worked and the age of individuals, segmented by income level. From this plot, we can see different income brackets, the lighter color being individuals who make less than $50,000 and the darker color being individuals who make more than $50,000. From this plot, we can see that hours per week vary across different ages but younger individuals around 20-30 years old and older individuals older than 60 show a wide range of working hours. There is a clustering of high-income individuals for middle-aged individuals from 30 to 60 years old and 40-60 hours work weeks, which could be due to the typical 40-hour week indicating someone is working full time.



The Hours Per Week vs. Education Level scatter plot shows the relationship between the number of hours worked per week and the level of education an individual has segmented by income level. From this plot, we can see different income brackets, the lighter color being individuals who make less than $50,000 and the darker color being individuals who make more than $50,000. There is a visible trend that those with high education levels often work in the 40-60-hour range. In contrast, lower education levels show a more scattered distribution of working hours and lower income levels.

From the KNN model and deeper relationship analysis, it was clear that education level was the primary predictor of higher income. Age and hours worked also contributed, though to a lesser extent. Our KNN model successfully captured these trends, showing it's a dependable method for predicting income based on these factors. The KNN model and its findings can help individuals make better career choices by understanding how age, education, and work hours could impact their income levels.

Now that we've explored the limitations of logistic regression in capturing outcomes, we continued with the KNN algorithm to classify data points based on their proximity to the k nearest neighbors. While we struggled with our logistic regression, our KNN model successfully yielded better results. To further analyze our dataset, we chose to utilize the decision tree classification technique.

4.3. Decision Tree Models

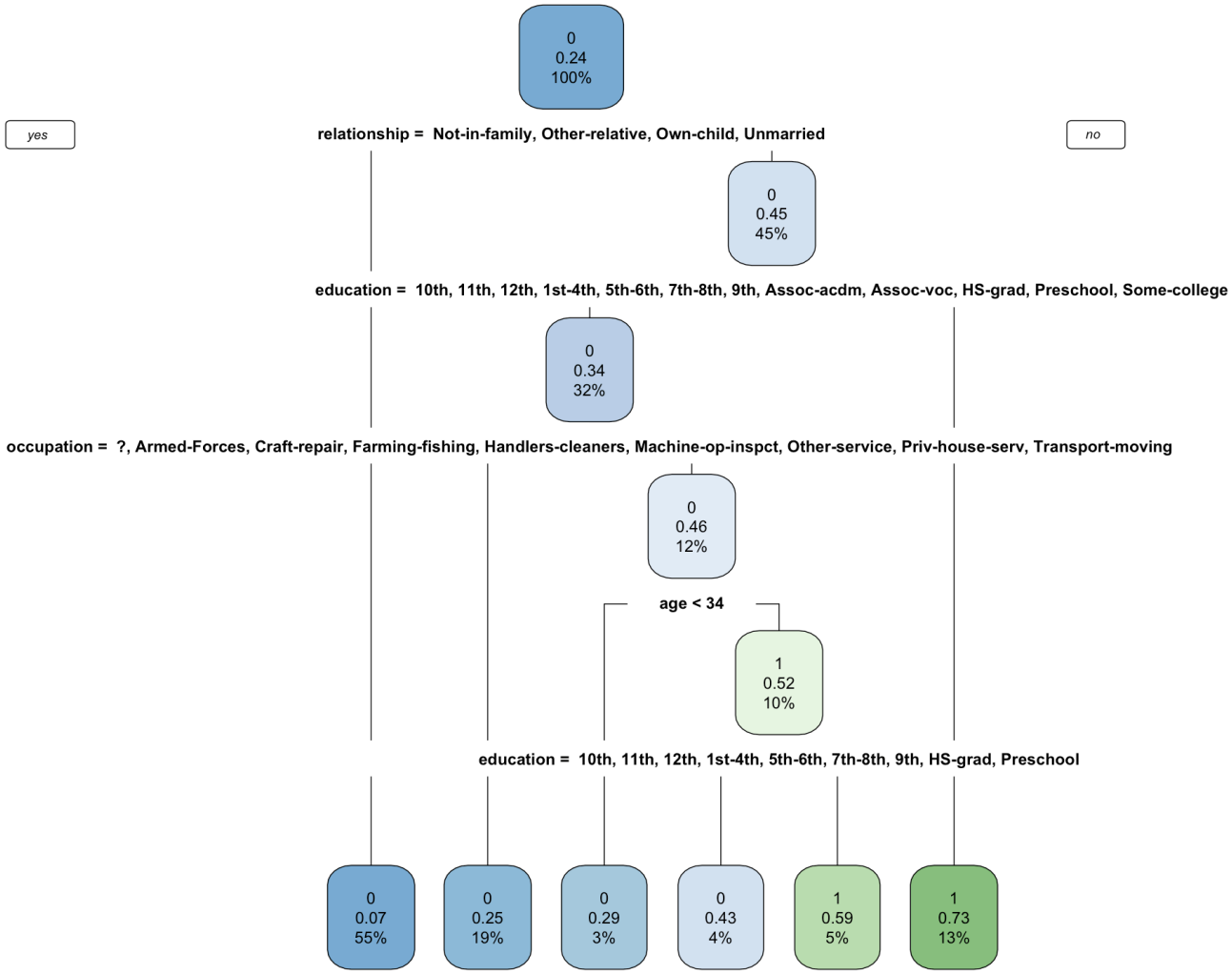
Decision tree models provide a valuable framework for analyzing the complex interplay of demographic and socioeconomic factors that influence an individual's income. We used the 'rpart' package, a widely trusted tool for fitting decision tree models. The 'rpart' package is particularly suited for handling large datasets and complex relationships between variables, offering straightforward implementations. The decision tree model was fitted using demographic and socioeconomic factors such as education level, occupation, age, sex, marital status, and work hours as predictors of income levels.

```
Call:
rpart(formula = Income.num ~ age + workclass + education + marital.status +
    occupation + relationship + hours.per.week + native.country +
    race + sex, data = train, method = "class")
  n= 24420

          CP nsplit rel error    xerror       xstd
1 0.12546721      0 1.0000000 1.0000000 0.01135538
2 0.01342168      2 0.7490656 0.7490656 0.01021201
3 0.01000000      5 0.7088005 0.7327557 0.01012442

Variable importance
  relationship marital.status      education     occupation            sex            age hours.per.week
            29             28             12             11              9              8              3
```

The output shows the complexity parameter (CP), relative error (rel error), cross-validation error (xerror), and cross-validation standard deviation (xstd) of the decision tree. These metrics helped determine the optimal size and complexity of the tree. The variable importance section ranks the predictor variables based on their importance in the model. According to the variable importance rankings derived from the decision tree model, several key factors emerged as significant determinants of income levels. This includes relationship status, marital status, education level, occupation, gender, age, and hours worked per

week. Each factor played a crucial role in predicting an individual's income level, shedding light on the complex interplay of demographic and socioeconomic variables.

Relationship and marital status stood out as the most influential variables, indicating their significant impact on income levels. Whether an individual was married, single, divorced, or in another relationship strongly influenced their income. This suggests that household dynamics and familial responsibilities play a pivotal role in shaping earning potential. Education level is closely followed in importance, underscoring the correlation between education and income. Individuals with higher levels of education tended to earn more, highlighting the value of advanced degrees or specialized training in accessing higher-paying job opportunities. Education served as a gateway to economic advancement, with higher education generally leading to higher income levels. Occupation ranked next in importance, emphasizing the pivotal role of job choice in determining income outcomes. Certain occupations offer higher salaries or income potential compared to others, reflecting the diverse range of career paths available and their respective financial rewards. Occupation choice significantly influenced earning potential, with individuals in certain fields enjoying greater income stability and growth prospects. Gender, or sex, also emerged as an important predictor of income, signaling the presence of gender-based disparities in earnings. Differences between male and female earners contributed to income inequality, highlighting the need for gender equity initiatives to address wage gaps and promote equal opportunities in the workforce. Age ranked lower in importance compared to other variables but still contributes significantly to predicting income levels. Age-related factors such as career progression, experience, and seniority influenced earning potential, with older individuals often commanding higher incomes due to accumulated expertise and tenure in their respective fields. Hours worked per week, while the least important variable, still exerted some influence on income levels. Individuals who worked longer hours may generally have higher incomes, reflecting the correlation between dedication, commitment, and financial remuneration. However, the impact of hours worked per week on income outcomes varies depending on other factors such as occupation type and employment conditions.



The decision tree model utilized in this analysis seeks to forecast income levels by incorporating various demographic and socioeconomic factors as features. At the outset, the root node focused on relationship status, where categories such as Not-in-family,b Other-relative, Own-child, and Unmarried were identified as the most predictive for lower incomes, demonstrating a clear pattern with a 100% split indicating uniform income levels within these groups.

As the analysis progressed down the decision tree, the significance of education level in shaping income outcomes became increasingly evident. Individuals with educational backgrounds spanning from completion of the 10th grade to attainment of Associate's degrees or vocational training typically demonstrated lower incomes; whereas those with some college experience or higher educational

achievements were more inclined to higher incomes. Within the cohort characterized by lower educational attainment, age emerged as a discernible factor influencing income levels. Particularly noteworthy is the consistent trend towards lower income brackets observed among individuals under 34 years old within this subgroup. This finding suggests that age dynamics intersect with education to impact earning potential, possibly reflecting challenges faced by younger individuals in accessing higher-paying job opportunities or advancing in their careers. Furthermore, occupation emerged as a significant influencer of income outcomes. Certain professions such as armed forces, farming, cleaning services, and transportation-related work are associated with lower pay scales, underscoring the importance of occupational choice in determining income levels. These findings underscore the multifaceted nature of income determination, where demographic and socioeconomic factors interact to shape earning potential.

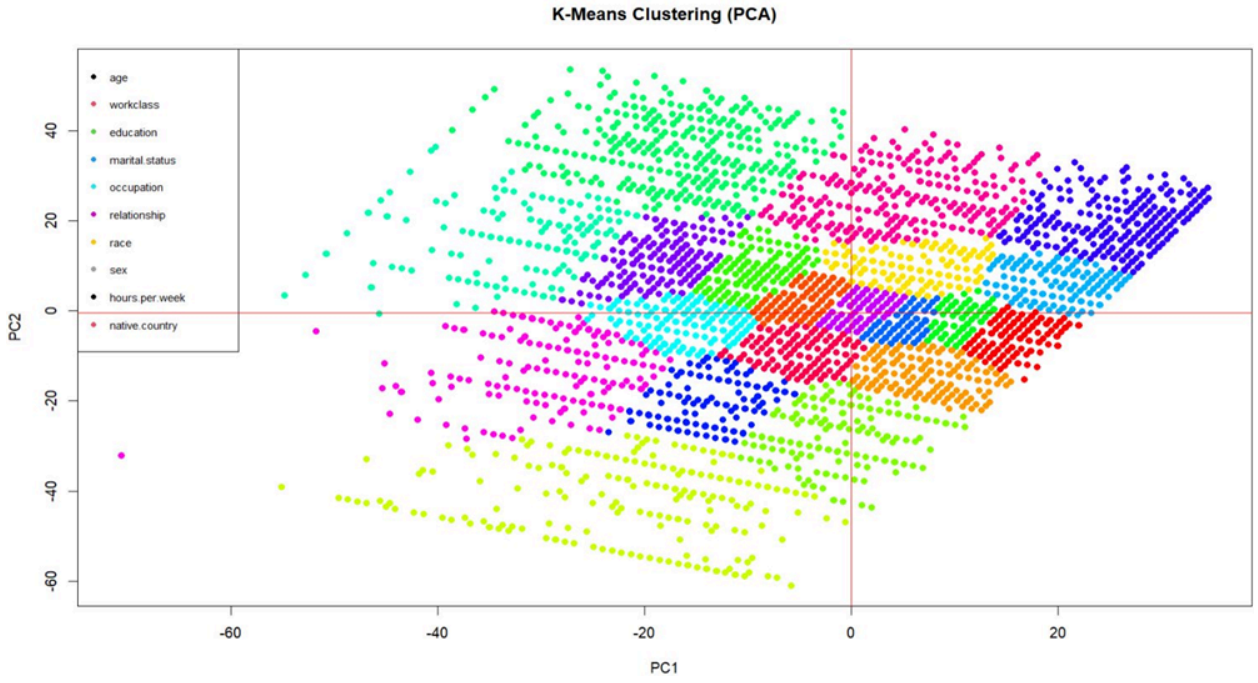
The decision tree model provided valuable insights into the predictors of income levels. To enhance the precision of our project, we also complemented the insights gained with clustering methods. By reducing the dimensionality of our data and uncovering latent patterns, clustering enriches our understanding of income predictors, contributing to a more accurate and comprehensive analytical framework.

4.4. Clustering

We applied a K-Means Clustering Model alongside PCA to our dataset. This aided with understanding the data by grouping data into clusters of similarity. However, there were some challenges in this process: the data consisted of mostly categorical values, preventing us from effectively utilizing K-Means, and the size of the dataset made it a challenge to gain insights from the clusters.

To address these issues, we converted the categorical values into qualitative data. By doing so, a model matrix was created by converting the values to a set of dummy variables with numerical values - also known as one-hot encoding. With the data now having numeric values, it allowed K-Means to perform more accurately and effectively.

The next step was to conduct PCA for its dimensionality reduction capabilities. In doing so, it reduces the dataset to a manageable size without having an impact on important information to identify patterns and variations in the data while enhancing visualizations. After performing K-Means, we identified the relationships between points that were not detectable originally. Through the distribution of the clusters, we based our insights on the similarities and differences among the data points.



The scatter plot represents the model produced going through PCA and K-Means clustering. The axis ticks represent standard deviation, indicating how much the points deviate from the mean in terms of their impact on income. The red lines represent the median for PC1 and PC2. The reason the median was used was because the data originally was categorical. The true mean can be influenced by extreme values after conversion caused by the lack of a numeric value, leading to inaccurate insights. The median prevented potential issues caused by these extremity values, allowing it to be more reliable and representative of the results produced compared to using the mean.

Principal Components 1 and 2 (PC1/PC2) represent the variance captured in the dataset. PC1 captured the largest variability in the data – the major data points responsible for changes in the dataset, and PC2

accounted for the second largest variance that was not captured in PC1. An example to explain these components is that experience and education were assumed to be the most important variables in determining income, which would be included in PC1; and PC2 captures other variables that are not included in PC1 but play a role in determining income. To explain the points, assume there are two individuals with bachelor's degrees, but due to the difference in their PC2 value, they are paid differently because the higher earner could be working in a sector that pays more. In general, the influence of the points in affecting income is dependent on its relative position to the median. PC1 and PC2 accounted for 98.9% of the variance captured in the data, indicating their variables were most responsible for determining income.

> PC1 includes age, hours.per.week, relationship.num, marital.status.num: marital.status Never-married, marital.status Married-civ-spouse, relationship: Own-child, Income.num, workclass: Private, sex.num.
>
> PC2 includes hours.per.week, age, occupation.num, relationship.num, workclass.num, marital.status.num, sex.num: sex Male, Income.num, marital.status: Married-civ-spouse

We then identified the top variables of each component. It was discovered that both components shared common variables, meaning these specific variables had the most impact on wages: age, hours per week, marital status, relationship, sex, and occupation, with workclass and education having a lesser effect. Age is an important variable because it plays a role in career advancement and income levels. It is likely for younger individuals to work in entry-level positions while older individuals are found in more senior roles due to their experience and developed skills. Hours per week affect income as longer hours worked are associated with higher earnings. It could also indicate the type of employment, full-time versus part-time positions. Marital status and relationships provide insights into individuals as those married or are parental figures have additional financial responsibilities to fulfill, so it's expected that they look for higher-earning jobs to provide support. However, those who are unmarried could have the ability to focus

more on developing their skills to advance their careers since they have fewer obligations. Sex shows that gender inequality is still present in the workforce as men earn more than women for similar work. Workclass and occupation represent the employment sector and the roles of the individuals. For instance, executive-managerial positions are associated with higher earnings, due to their responsibilities and the role they play in the business, compared to other fields - consisting of common occupations that offer lower wages such as the restaurant industry. Private sectors are more likely to provide higher pay, benefits, and career advancement opportunities compared to those who are self-employed or work in the public sector. Education provides individuals access to specialized professions, allowing them to obtain knowledge and develop the skills essential for career advancement.

Together, these variables provide a better understanding of the demographic and socioeconomic factors responsible for income disparities. We continue to utilize PCA methods to further analyze the dataset.

4.5. PCA Methods

The final method we used is Principal Component Analysis (PCA). The reason we chose this method was because we wanted to find out what factors can predict income in the future. At the same time, our group aimed to analyze how various demographic and socioeconomic factors influence an individual's income. Principal Component Analysis (PCA) has been very useful to identify the most significant variables that discover the variance in our income datasets.

We identified the variables such as education level, occupation, gender, race, and relationships that we believed might influence income. Using the scale() function, the purpose was to standardize our numerical data by subtracting the mean and dividing by the standard deviation. This step is very important for PCA, which is the foundation of the scale level in terms of our datasets. The normalization steps ensure that each variable contributes equally to the analysis.

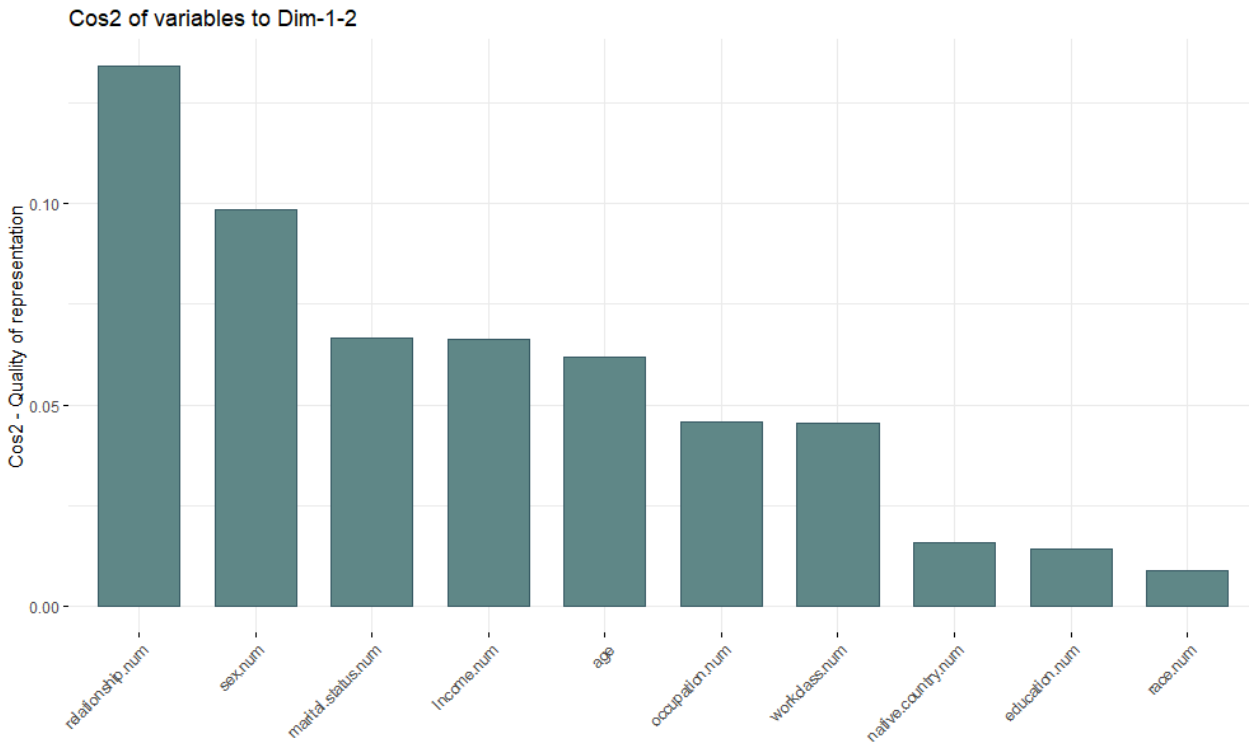
workclass.num	marital.status.num	occupation.num	relationship.num	race.num	sex.num	native.country.num	Income.num
2.9359517	0.9478313	-1.4790301	-0.2612446	0.3850415	0.6927947	0.2649196	-0.5756818
1.8876507	-0.3872683	-0.7345332	-0.8857223	0.3850415	0.6927947	0.2649196	-0.5756818
-0.2089512	-1.7223678	-0.2382018	-0.2612446	0.3850415	0.6927947	0.2649196	-0.5756818
-0.2089512	-0.3872683	-0.2382018	-0.8857223	-2.0110019	0.6927947	0.2649196	-0.5756818
-0.2089512	-0.3872683	0.7544608	2.2366662	-2.0110019	-1.4433813	-5.3039463	-0.5756818
-0.2089512	-0.3872683	-0.7345332	2.2366662	0.3850415	-1.4433813	0.2649196	-0.5756818

Next, we performed the correlation matrix on the normalized data to minimize the dimensionality. We wanted to make the datasets transform from a high-dimensional into a low-dimensional while retaining some meaningful information about the original datasets as well as capturing the most variance in the data.

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Standard deviation	0.6566579	0.3868625	0.3498713	0.3345691	0.30835196	0.30385914	0.20426880	2.967518e-09
Proportion of Variance	0.4128897	0.1433075	0.1172121	0.1071834	0.09104355	0.08840979	0.03995396	8.432229e-18
Cumulative Proportion	0.4128897	0.5561972	0.6734093	0.7805927	0.87163625	0.96004604	1.00000000	1.000000e+00

From the professional table above, we used prcomp() function to perform PCA. We then obtained the summary of our data object, which includes information about the proportion of variance explained by each principal component. From our professional table, we find out that, the first principal component explains almost 40% of the total variance. This implies that about one-third of the data in the set of 8 variables can be represented by just the first principal component. The second one explains 14% of the total variance. The cumulative proportion of Component 1 and Component 2 explains nearly 54% of the total variance.



The PCA results indicate that relationship status and sex were the most influential factors affecting an individual's income. Native country, race, and occupation also played significant roles but were less important compared to the first two factors. Therefore, our dataset classification showed minimal impact compared to other variables. These results can help our group members focus on the real-world critical demographic and socioeconomic factors when it comes to predicting income. Further studies could explore the underlying reasons behind these influences and consider additional variables that might contribute to a more comprehensive understanding of income determinants.

5. Conclusion

Analyzing our dataset with data mining methods allowed us to investigate how different demographic and socioeconomic factors influence an individual's income. Our report detailed our findings for logistic regression, KNN, decision trees, clustering, and unsupervised learning methods. For example, with the logistic regression analysis, we conclude that age, race, and gender significantly impacted the likelihood of an individual earning an income greater than $50,000. We determined that older individuals,

Asian-Pacific-Islander, White, and males had higher odds. These results show the importance of addressing inequality in the workforce and promoting fairness and equal opportunities among all demographics and socioeconomic groups. However, factors like race and gender do not unduly influence the income potential. The practical implications of our analysis have the potential to inform decision-making processes across many areas like career planning, policy development, business strategy, financial planning, and education policies. The insights obtained can enable policymakers and organizations to focus on areas of needed intervention to foster economic development and address inequality within the workforce. These findings can also drive policymakers to emphasize the importance of education to further promote economic growth. Based on these findings businesses can use these insights to improve hiring and compensation strategies. By identifying these critical factors, policymakers and stakeholders can develop targeted interventions to address income disparities and foster economic empowerment for individuals across diverse demographic groups.

References:

[1] https://www.kaggle.com/datasets/nimapourmoradi/adult-incometrain-test-dataset?select=adult_test.csv

[2] https://taxfoundation.org/data/all/federal/average-income-age/

[3] https://www.weforum.org/agenda/2022/09/working-hours-america-income-economy/

[4] https://www.aplu.org/our-work/4-policy-and-advocacy/publicuvalues/employment-earnings/#:~:text=College%2Deducated%20workers%20enjoy%20a,less%20education%20continues%20to%20widen