Northeastern University

ALY6020: Predictive Analytics

Assignment 1: Understanding Income Inequality through KNN Implementation

Abhigna Ramamurthy, 002982276

College of Professional Studies, Northeastern University

Professor: Justin Grosz

Jan 15, 2022

Introduction

In this assignment census dataset is considered with records of high and low income with attributes such as personal (age, marital status, relationship), educational, economic (occupation, capital gain/loss, salary), and geographical (native country). These values is aimed to help understand the income inequality and differences between >50K income earners as opposed to <50K income earners. The main business questions revolve around using machine learning models to better understand what contributes most to inequality in salary, thus shedding light on how policies can be modified to ensure equal pay across the US. In this report, a comprehensive exploratory analysis on the hidden patterns and relationships between the features are discussed along with applying the results to social aspects of salary gap.

Data cleaning

The dataset contains 48,842 records and 15 features with 6 integer values and rest are object values. The first step in data profiling and cleaning was to investigate the missing/invalid values. The first and foremost thing to clean up along with loading the excel file to a data frame in python was to add appropriate column names as they were missing. Based on the initial summary statistics, there were no missing data in the dataset, however there were invalid data of "?" in workclass and occupation columns. As these columns correlated with each other which means for every "?" value in workclass column had a "?" value in occupation column too. This can be attributed to unknown work or unemployment. As these values represented less than 5% of the whole data, I removed columns with this invalid value. Columns that were removed as part of data cleanup based on relevance were fnlwgt, education (as education_num provided a continuous variable to work with), occupation, and relationship.

Next step in data cleaning was to take care of categorical variables into analyzable data formats. As majority of the records showed “Private” value for workclass column, the feature was converted into a Boolean value from a categorical value with a value of “1” if Private or “0” is other. Similarly, marital status was changed to “1” if the person was never married, and “0” for other cases, and native country was converted to Boolean of “1” if birth country was the United States and “0” for other values. Salary was also converted to Boolean value of “1” for >50K and “0” for <50K. Sex and Race columns were one hot encoded for further analysis. Overall, we finally had 46,043 records and 15 features to work with.

Exploratory Data Analysis

In this section, exploratory approach to analyzing, summarizing, and identifying patterns, relationships, and anomalies are discussed for the census dataset. With approximately 90% of peoples’ country of birth being the US, the census has about 70% of male and 30% of female population. 85% of whom are from white race, and rest divided among the other races. The main feature of this analysis is the salary category, which has been converted to Boolean of 1 for >50K and 0 for <50K. From the violin chart below, which is a hybrid of box plot and density plot shows we have 75% of population under low-income category.

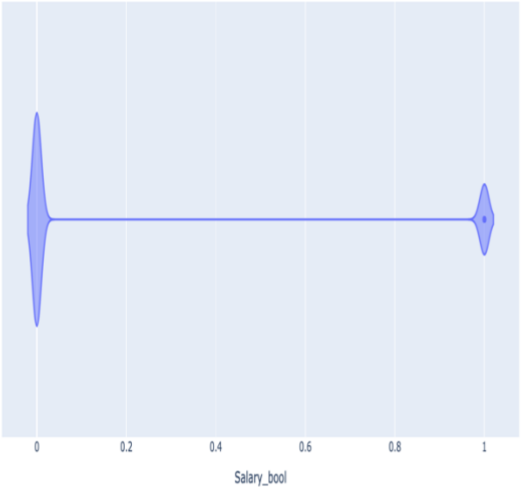


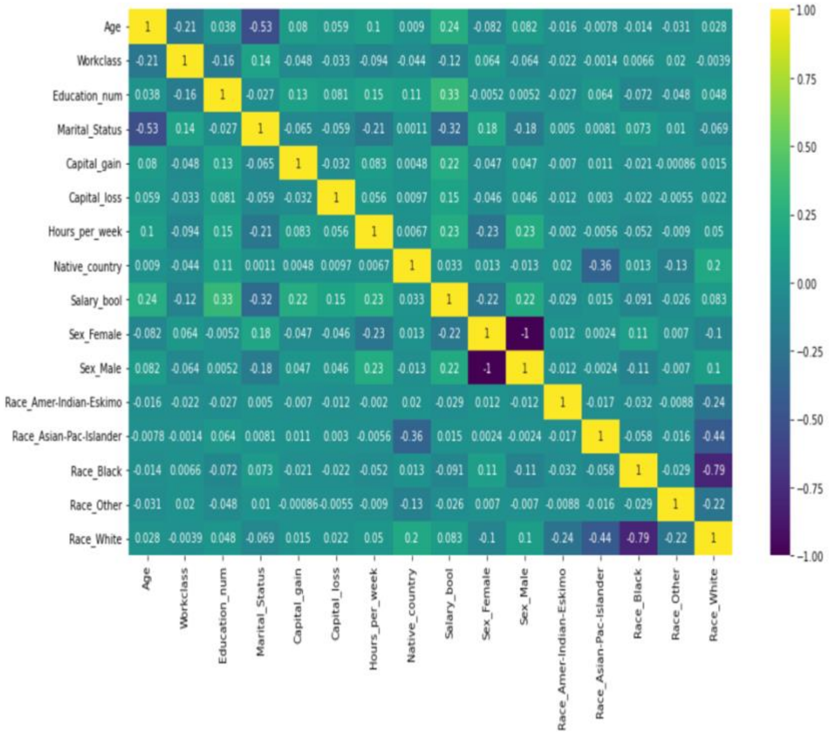
Figure 1: Violin plot of salary distribution

Histograms help understand the overall distribution of data in the given dataset. From Figure 1 histogram in Appendix 1 we can deduce that majority of the census records falls under 24-47 years of age with a right skewed tail with overall data containing ages from 17-90 years. Most people aged between 24-47 fall under high income category (>50K) than low income (<50K).

Workclass also showed that majority of the US citizens work in private sector, followed by various government employments, then self-employed and a small group of people are without pay or never employed as shown by Figure 2 in Appendix 1. Furthermore, the box plot of Figure 3 in Appendix 1 shows that median age of people in private sectors in much lower than that of government employees, or self-employed, indicating majority of youth are working in private sector jobs. We can also see that of high paying occupation are executive managerial or profession specific jobs in Figure 4, Appendix 1.

Majority of the US citizens have high school or college as their highest degree as shown in stacked bar chart of Figure 4 in Appendix 1. Interestingly we can see close to equal proportions of low- and high-income earners in bachelor's or master's degree holders. Whereas professional school and doctorates have more high-income earners than low income. This shows that higher the education level, higher the chances of falling under high-income category.

Correlation heatmap shows in Figure 2 below shows the correlation between the various numerical values of the dataset. It is evident from the heatmap that salary category has the highest correlation with the education which is in line with previous analysis of higher the education degree, higher the chances of falling in higher-income category. We can also see positive correlations with age. Older people tend to have higher capital gains which is in line with higher salary category. Another socially unfortunately correlation is that male gender identified people tend to fall more into high-salary earners, as compared to female population. Salary also has high negative correlation with marital status which indicates that married in any manner including separated/divorced/widowed have higher income than never married.



	Age	Workclass	Education_num	Marital_Status	Capital_gain	Capital_loss	Hours_per_week	Native_country	Salary_bool	Sex_Female	Sex_Male	Race_Amer-Indian-Eskimo	Race_Asian-Pac-Islander	Race_Black	Race_Other	Race_White
Age	1	-0.21	0.038	-0.53	0.08	0.059	0.1	0.009	0.24	-0.082	0.082	-0.016	-0.0078	-0.014	-0.031	0.028
Workclass	-0.21	1	-0.16	0.14	-0.048	-0.033	-0.094	-0.044	-0.12	0.064	-0.064	-0.022	-0.0014	0.0066	0.02	-0.0039
Education_num	0.038	-0.16	1	-0.027	0.13	0.081	0.15	0.11	0.33	-0.0052	0.0052	-0.027	0.064	-0.072	-0.048	0.048
Marital_Status	-0.53	0.14	-0.027	1	-0.065	-0.059	-0.21	0.0011	-0.32	0.18	-0.18	0.005	0.0081	0.073	0.01	-0.069
Capital_gain	0.08	-0.048	0.13	-0.065	1	-0.032	0.083	0.0048	0.22	-0.047	0.047	-0.007	0.011	-0.021	-0.00086	0.015
Capital_loss	0.059	-0.033	0.081	-0.059	-0.032	1	0.056	0.0097	0.15	-0.046	0.046	-0.012	0.003	-0.022	-0.0055	0.022
Hours_per_week	0.1	-0.094	0.15	-0.21	0.083	0.056	1	0.0067	0.23	-0.23	0.23	-0.002	-0.0056	-0.052	-0.009	0.05
Native_country	0.009	-0.044	0.11	0.0011	0.0048	0.0097	0.0067	1	0.033	0.013	-0.013	0.02	-0.36	0.013	-0.13	0.2
Salary_bool	0.24	-0.12	0.33	-0.32	0.22	0.15	0.23	0.033	1	-0.22	0.22	-0.029	0.015	-0.091	-0.026	0.083
Sex_Female	-0.082	0.064	-0.0052	0.18	-0.047	-0.046	-0.23	0.013	-0.22	1	-1	0.012	0.0024	0.11	0.007	-0.1
Sex_Male	0.082	-0.064	0.0052	-0.18	0.047	0.046	0.23	-0.013	0.22	-1	1	-0.012	-0.0024	-0.11	-0.007	0.1
Race_Amer-Indian-Eskimo	-0.016	-0.022	-0.027	0.005	-0.007	-0.012	-0.002	0.02	-0.029	0.012	-0.012	1	-0.017	-0.032	-0.0088	-0.24
Race_Asian-Pac-Islander	-0.0078	-0.0014	0.064	0.0081	0.011	0.003	-0.0056	-0.36	0.015	0.0024	-0.0024	-0.017	1	-0.058	-0.016	-0.44
Race_Black	-0.014	0.0066	-0.072	0.073	-0.021	-0.022	-0.052	0.013	-0.091	0.11	-0.11	-0.032	-0.058	1	-0.029	-0.79
Race_Other	-0.031	0.02	-0.048	0.01	-0.00086	-0.0055	-0.009	-0.13	-0.026	0.007	-0.007	-0.0088	-0.016	-0.029	1	-0.22
Race_White	0.028	-0.0039	0.048	-0.069	0.015	0.022	0.05	0.2	0.083	-0.1	0.1	-0.24	-0.44	-0.79	-0.22	1

Figure 2: Correlation matrix

Analysis

Data Pre-processing

Upon data cleaning and exploratory data analysis, I did several steps to get the data ready for modelling which includes multicollinearity check with variance inflation factor method, train-test split (80-20) of dataset and standard scaling to handle difference in scales of the features. Multicollinearity step using VIF helped eliminate highly correlated values that could have derailed the model analysis. Female sex, and Native country were the two features that were eliminated at this step. Male sex feature was highly negatively correlated with female sex variable, hence removal of that reduced the collinearity. Native country was also removed as majority of the data indicated the US, hence was not providing any new insight in the analysis.

Data Modeling

In this section, I want to discuss the K-Nearest Neighbor algorithm implemented for the census dataset. The dependent variable for this analysis is the salary value of high-salary (>50K) and low salary (<50K). I first created a base model with value of neighbors (K) as 5 and later used GridSearchCV and Error rate methods to determine optimal K value for the analysis. GridSearchCV is an iterative method to find optimal value of hyperparameters for an algorithm by running all the possible scenarios. In this case, I ran values of K from 1 to 25 to get the optimal value with best results to be K of 24. Third method I used was to see error rate reduction for various K values as shown below in Figure 3 which shows error rate versus K values. As we can see that error rate reduces consistently after K value of 15. This is a good indication that value of 15 can provide optimal results for our data. A similar graph for accuracy rate was also done as shown in Figure 1 of Appendix 2. The same K value of 15 is showcasing the start of steady increase in accuracy rate value.

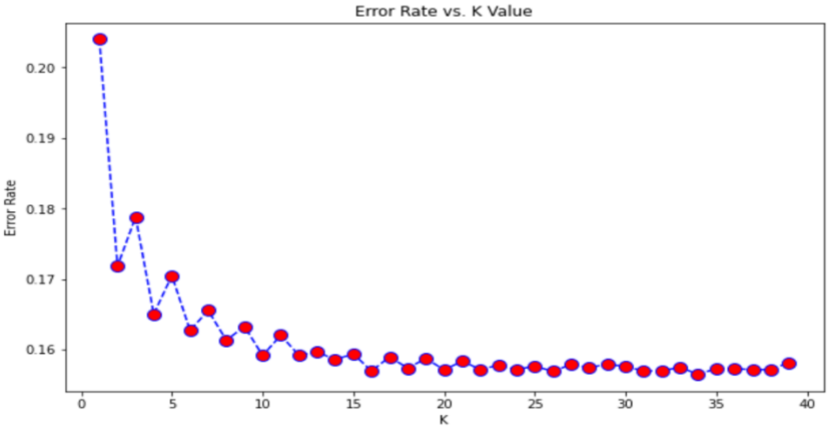


Figure 3: Error Rate versus K value.

KNN K-value	*F1 score*	*Precision*	*Recall*	*AUC*	*Accuracy*
K=5 (Base model)	0.59	0.65	0.54	0.72	86%
K=24 (GridSearchCV)	0.59	0.73	0.50	0.72	83%
K=15 (Error rate)	0.60	0.70	0.53	0.73	83%

Table 1: Benchmarking metrics used in the three methods

Model analysis and output

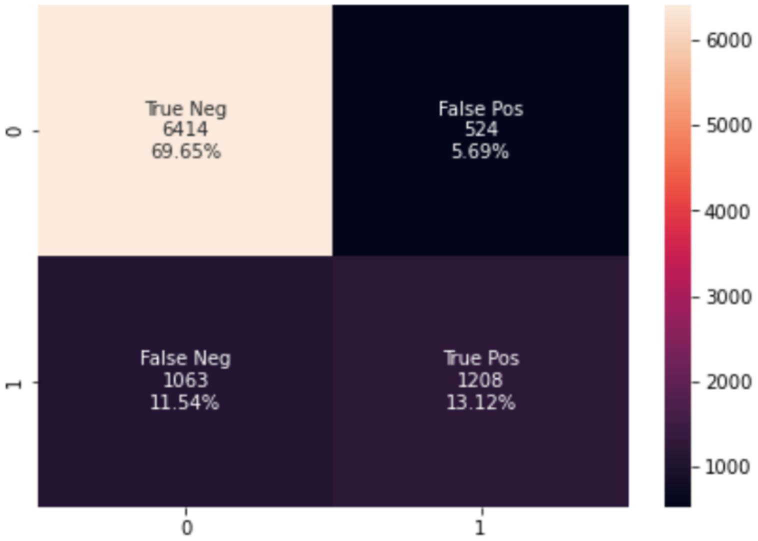


Figure 4: Confusion matrix for K=15

From the above table which shows the benchmarking metrics for various K values, K of 15 works best giving optimal results comparatively. Hence error rate method is chosen. The confusion matrix for this method is shown below and it is a summarized version of performance of the KNN classification model with K as 15. The model predicted 70% people to fall into low-salary category and 13% into high-income category correctly. The false positives, that is actual value is low-salary, but predicted is high-salary is as low as just 6%, and false negative of actual high-income and predicted low-income is 12%. Overall, the performance of the model in predicting the correct category of salary is good. F1 score of the model is quite good with 0.7 value. The model provides a precision of 0.7 which means the model makes correct predictions 70% of times, and recall of 0.53, which means if higher the salary, the model can predict it 53% of the times. AUC of 0.72 means that the classifier performance is better than random, and it can distinguish positive and negative examples to some extent but not perfectly.

Conclusion

Based on the above analysis, we can say with certainty that there is a difference in the features of low-income and high-income groups. Although some of the features overlap, higher education, older age, higher capital gain, and higher hours per week contribute to high salary in United States. Promoting higher education or even professional education along with healthy lifestyle can be a major focus for equal pay campaign. After examining the results of KNN model, we can say for certain that the model can bucket the low and high salary populations correctly for majority of the cases, thus providing an optimized outlook into salary income classification.

Reference

Accuracy, Precision, Recall & F1-Score - Python Examples - Data Analytics. (2022). Data Analytics. https://vitalflux.com/accuracy-precision-recall-f1-score-python-example/

Allibhai, E. (2018, September 26). Building a k-Nearest-Neighbors (k-NN) Model with Scikit-learn. Medium. https://towardsdatascience.com/building-a-k-nearest-neighbors-k-nn-model-with-scikit-learn-51209555453a

An Introduction to Interpretable Machine Learning with LIME and SHAP. (2021). Datasset to Mindset. https://www.data4v.com/an-introduction-to-interpretable-machine-learning-with-lime-and-shap/

Krish Naik. (2019, June 18). K Nearest Neighbour Easily Explained with Implementation [Video]. YouTube. https://www.youtube.com/watch?v=wTF6vzS9fy4

Appendix 1 Exploratory Data Analysis

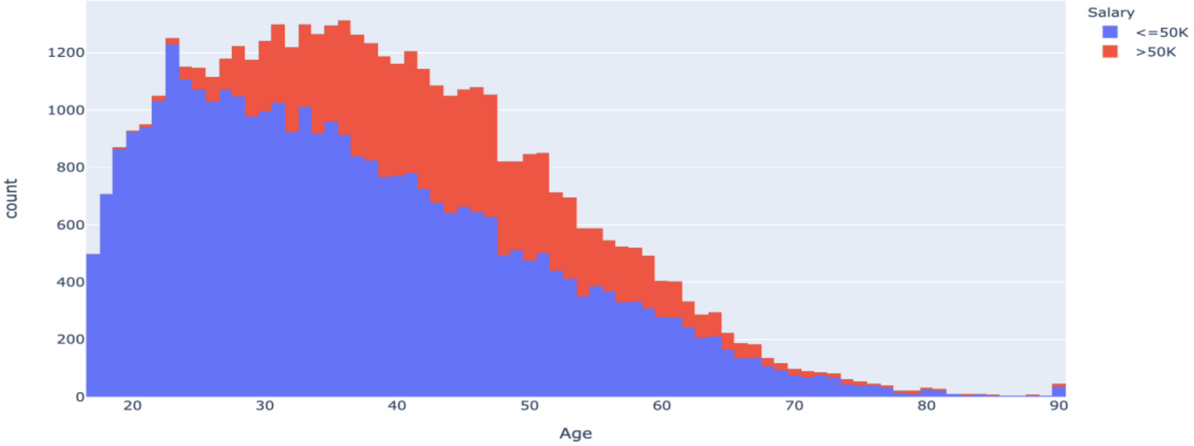


Figure 1: Histogram of Age categorized into high (>50K) and low (<50K) income earners.

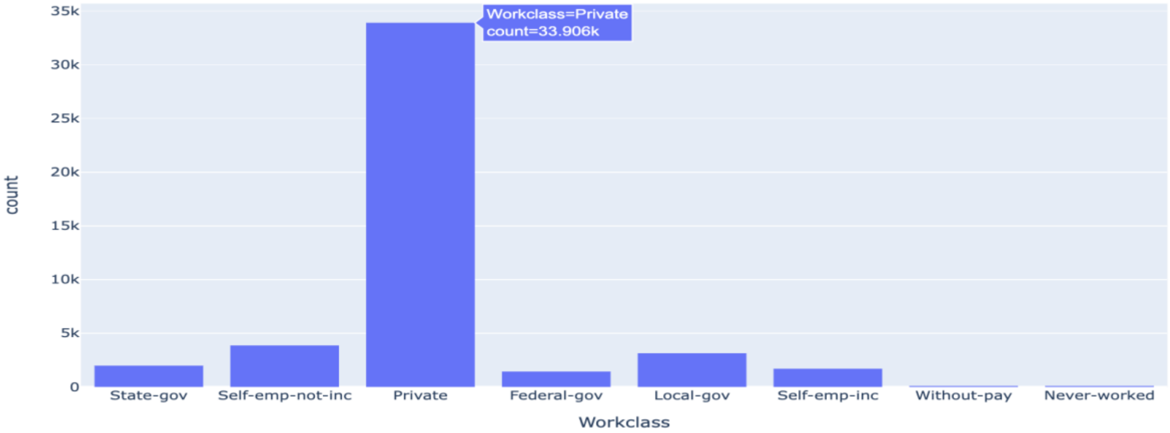


Figure 2: Histogram of workclass distribution

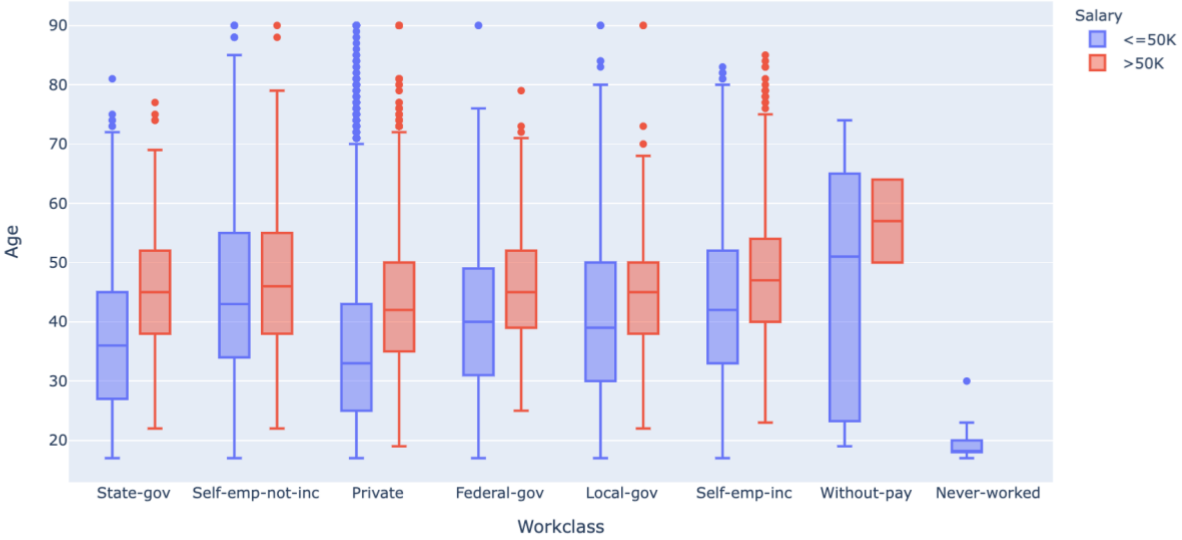


Figure 3: Box plot of workclass by age for the two salary categorizations.

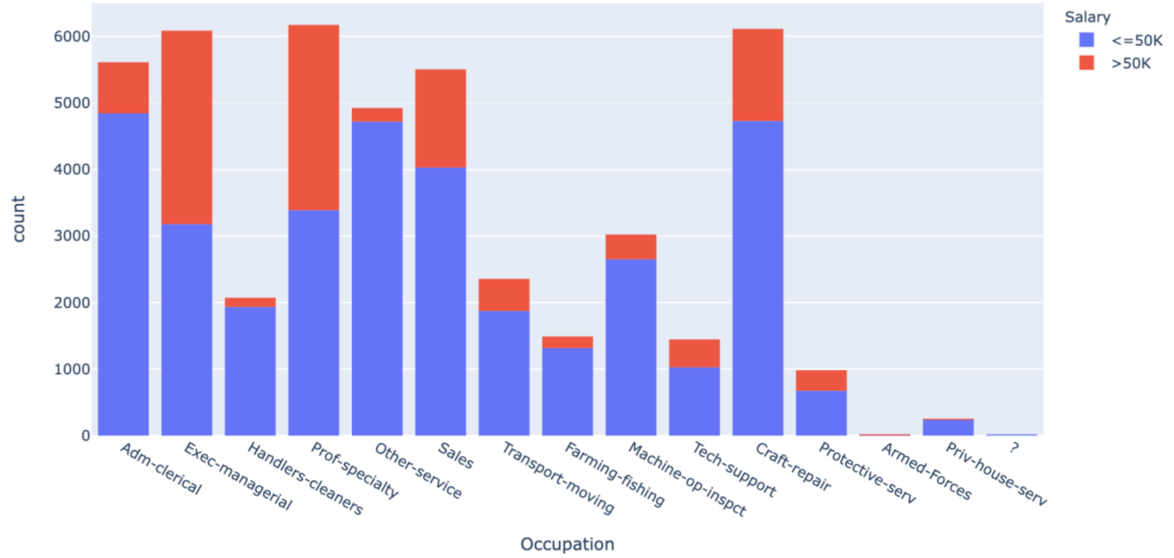


Figure 4: Occupation histogram with salary categorization.

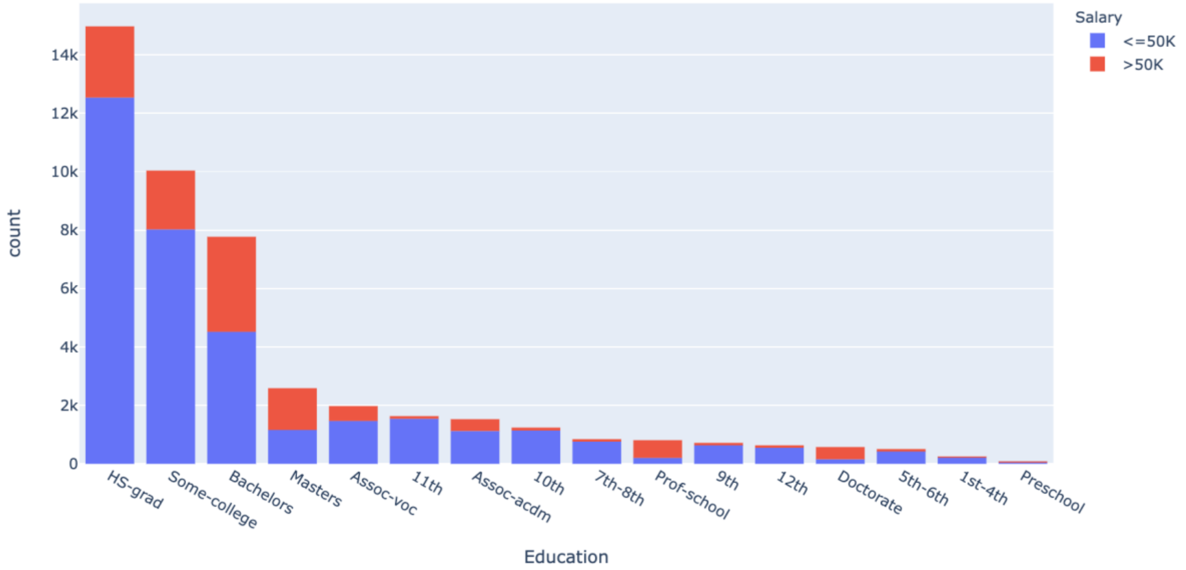


Figure 5: Stacked bar graph of education categorized by salary

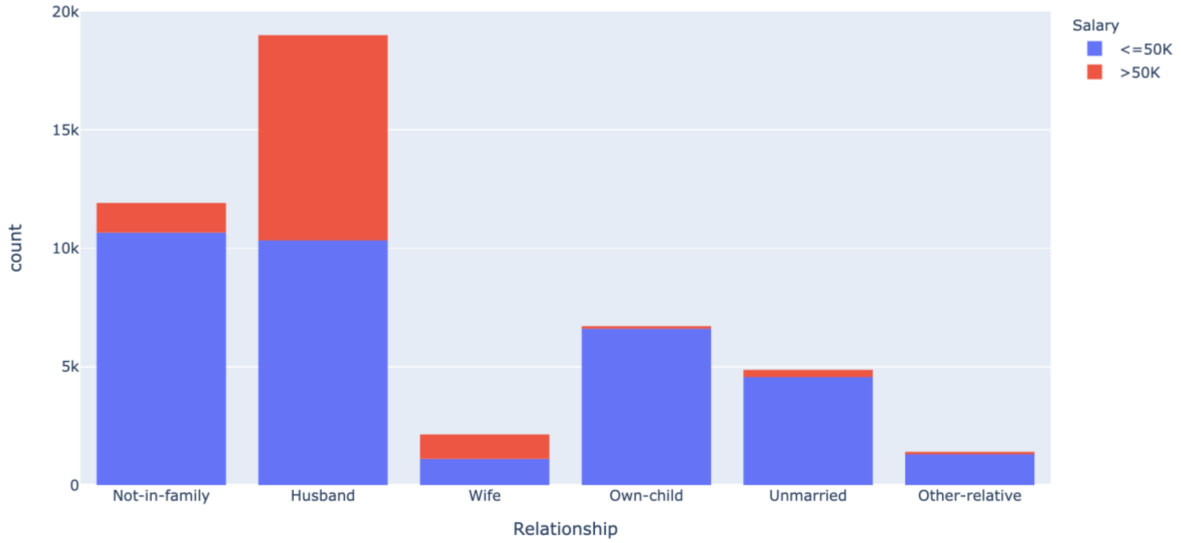


Figure 6: Bar graph of relationships in the dataset

Appendix 2 Model Analysis

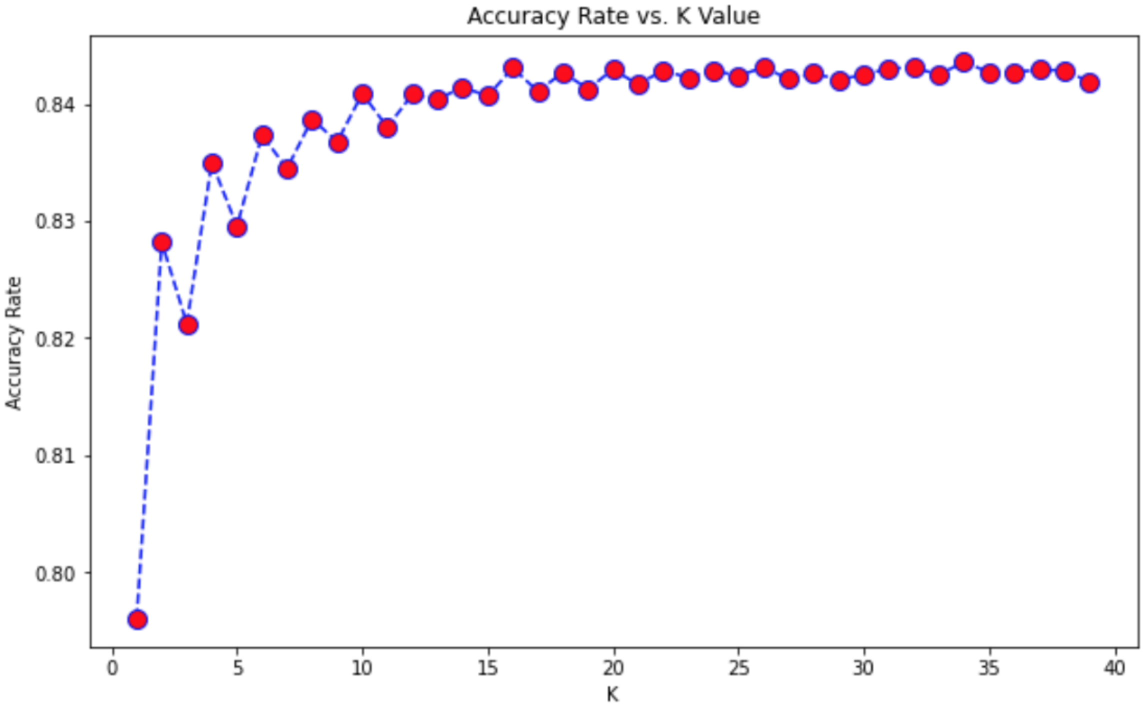


Figure 1: Accuracy rate versus K value