Module 1 Assignment: K-Nearest Neighbors Algorithm

Dhairyav Jatin Shah

College of Professional Studies, Northeastern University

Professor: Justin Grosz

January 15, 2023

**Introduction**

A dataset with census information on the characteristics of US citizens is used in this assignment. We will create a model to distinguish between citizens with low and high incomes by understanding the properties of the dataset and understanding its attributes. This research focuses on analyzing the link between citizens' salaries and other census attributes relevant to it, such as age, number of working hours, number of years of education, etc., to achieve this goal. The data supplier can then classify the citizens into those with high salaries and those with low salaries using the findings.

**Data cleaning**

The dataset contains 48842 records and 15 variables. Out of the 15 variables the data set contains 6 numerical variables and 9 categorical variables. Upon analyzing, we observed abnormalities in values for some columns. We replace these values by null values. Further we treat these null values by replacing them with the mode of the individual column. We have used mode in this step as the columns having abnormal values are categorical i.e., **“Workclass”, “Occupation”,** and **“Native country”**. In the next step upon analyzing the numerical columns we discover that 2 columns **“Capital gains”** and **“Capital Loss”** have more than **44,000 records** of value **“0”** which is almost **90%** of the data. Hence, these columns as well as the column “fnlwgt” will be dropped from the data set from further analysis.

**EDA**

In the process of exploratory data analysis, we understand our target variable i.e., **“Salary”** and its correlation with the other independent variables. First, we plot a count plot to understand the total number of citizens having salary higher than 50k and less than 50k. From fig 1 (in appendix 1), we can see that there are **37,155** records having salary less than 50k and only **11,687** records having salary greater than 50k. From the observed values we can see that the data has more than double records of salary less than 50k compared to salary greater than 50k. There is a possibility that the classification model would be able to classify salaries less than 50k better than salaries greater than 50k in case of overlapping clusters. To understand the numerical variables, we create a descriptive statistic table (table 1 in appendix 2). In the obtained table we can see that the average age of the data set is 38 years with a minimum of 17 years and a maximum of 90 years. The average number of years of education of the data set is 10 which corresponds to some college attended in the education column. But we can see that a minimum of 1 which corresponds to preschool and a maximum of 16 which corresponds to a PhD or a Doctorate degree. The average number of hours people working is 40 hours with a minimum of 1 hour and a maximum of 99 hours in a week. Further we plot a pair plot with the salary as hue to understand the distribution of the numerical variables. From fig 2 (in appendix 1), the obtained pair plot shows an overlapping of values with the 2 salary categories. In the scatter plot of Hours-per-week vs age, we can see a little better cluster compared to other numerical variables. To understand better we now plot a heatmap. Before plotting the heatmap we perform encoding on our target variable to convert it into a numerical variable to include it into the heatmap which would be helpful while fitting the model as well. From fig 3 (in appendix 1) , as we discussed earlier correlation between hours per week and age is the highest also salary is equally correlated with age and hours per week and has a higher correlation with education number. We can use the 3 variables to perform classification of the target variable.

**Analysis**

**Modeling**

After completing the initial analysis, we will perform K-nearest neighbors’ algorithm to classify the salaries. To derive a model, we will now split the data set into 2 parts X and Y and further these data sets into training and testing data sets. To split the data into X and Y variable we select the number of years of education and age in the X variable and Salary in the Y variable. We are going to use 80% of the data to train the model and 20% to test the data. Before we prepare the train and test data set, we will perform scaling on the X variables. Since KNN uses the Euclidean distance between two data points to determine nearest neighbors, data scaling is necessary. Magnitudes affect Euclidean distance. High magnitude features will be given more weight than low magnitude features. Upon splitting the data into train and test sets we obtain 39073 records in the train data set and 9769 records in the test data set. While performing KNN it is important to determine an accurate value of K to obtain the accurate output upon testing. To obtain an optimal value of K we plot a graph of K-values against the error rate. Upon performing this method, we can see in fig 4 (in appendix 1) the error rate is the lowest for K = 4. Now we will fit the KNN model using python library Sklearn. We will fit 3 models to compare the performance of the model using different values of K. The values of K we are going to use are 2, 3 and 4. From the table 4 (in appendix 2) we can see that the accuracy score for the model where k = 2 is 76%, the accuracy score for the model where k = 3 is 75%, and the accuracy score for the model where k = 4 is 77%. Hence, we will use the model where K=4 on the test data set to classify as it has the highest accuracy. Also, from the table we can see that the model is not able to accurately classify the salaries greater than 50k as it is able to classify salaries less than 50k. We can see that it is able to classify salaries below 50k with an accuracy of 86% whereas the accuracy for the classification of the salaries above 50k is only 36%. From table 5 (in appendix 2) we can see the confusion matrix obtained when K=4. In this matrix we can see that we have 6935 true positives and 623 false negatives, 476 false positives and 1732 true negatives. In our case the type 1 error is high as the data is skewed towards the salary less than 50k. Hence the model can predict more true positives compared to false negatives. We have further plot graphs to show the decision boundaries of the classification. From fig 5 (in appendix 1) we can see that when compared to K=2 and K=3, K=4 is able to form a better decision boundary.

**Conclusion**

Based on our analysis and modeling, we can say that classifying the salaries into high income and low income is not easy as the data is skewed towards to the low-income citizens and is able to classify them accurately with an accuracy of 86% but is not able to classify citizens with high - income as the accuracy is 36% only. The overall accuracy of the data is 77% which is again not so good to conclude that the classification of salaries was done correctly. To improve the classification of the target variable we can take a small random sample from the entire data set with same number of records of citizens having high and low income. By doing this we could get a better picture of the classification.

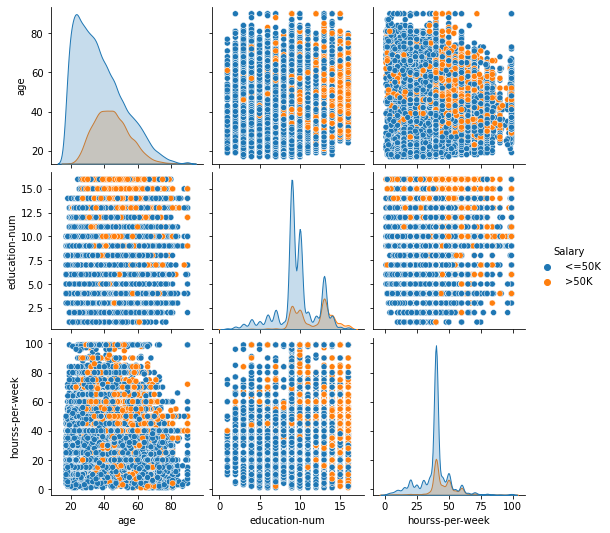
**Reference**

1. Band.A (May 2020) *How to find the optimal value of K in KNN?* <https://towardsdatascience.com/how-to-find-the-optimal-value-of-k-in-knn-35d936e554eb>
2. Datacamp (Aug 2018) *KNN Classification Tutorial using Scikit-learn* <https://www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn>
3. The Click Reader (n.d.) *K-Nearest Neighbours (KNN) Classifier* <https://www.theclickreader.com/k-nearest-neighbours-knn-classifier/>

**Appendix 1**

Chart, bar chart

Description automatically generatedFig 1 : The count plot of variable “Salary”

Fig 2 : Pair Plot

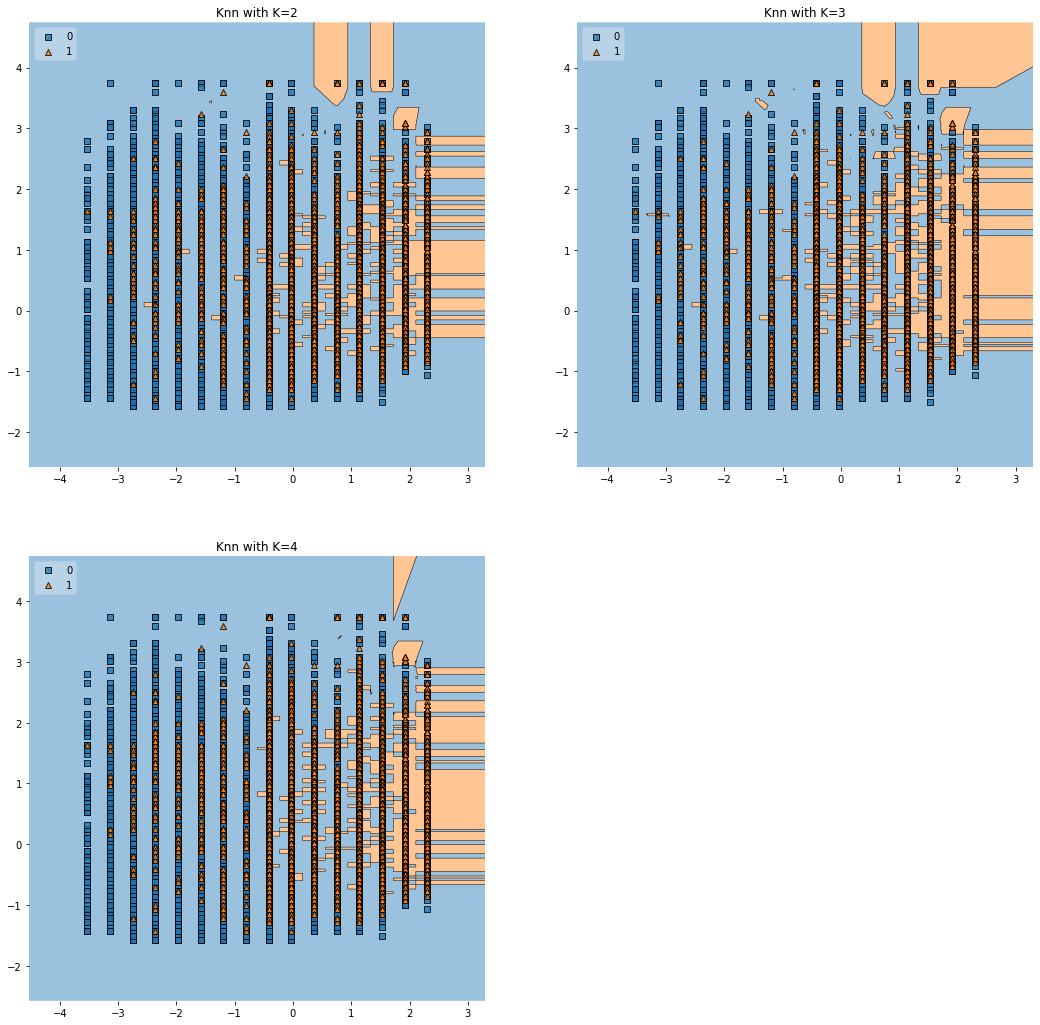
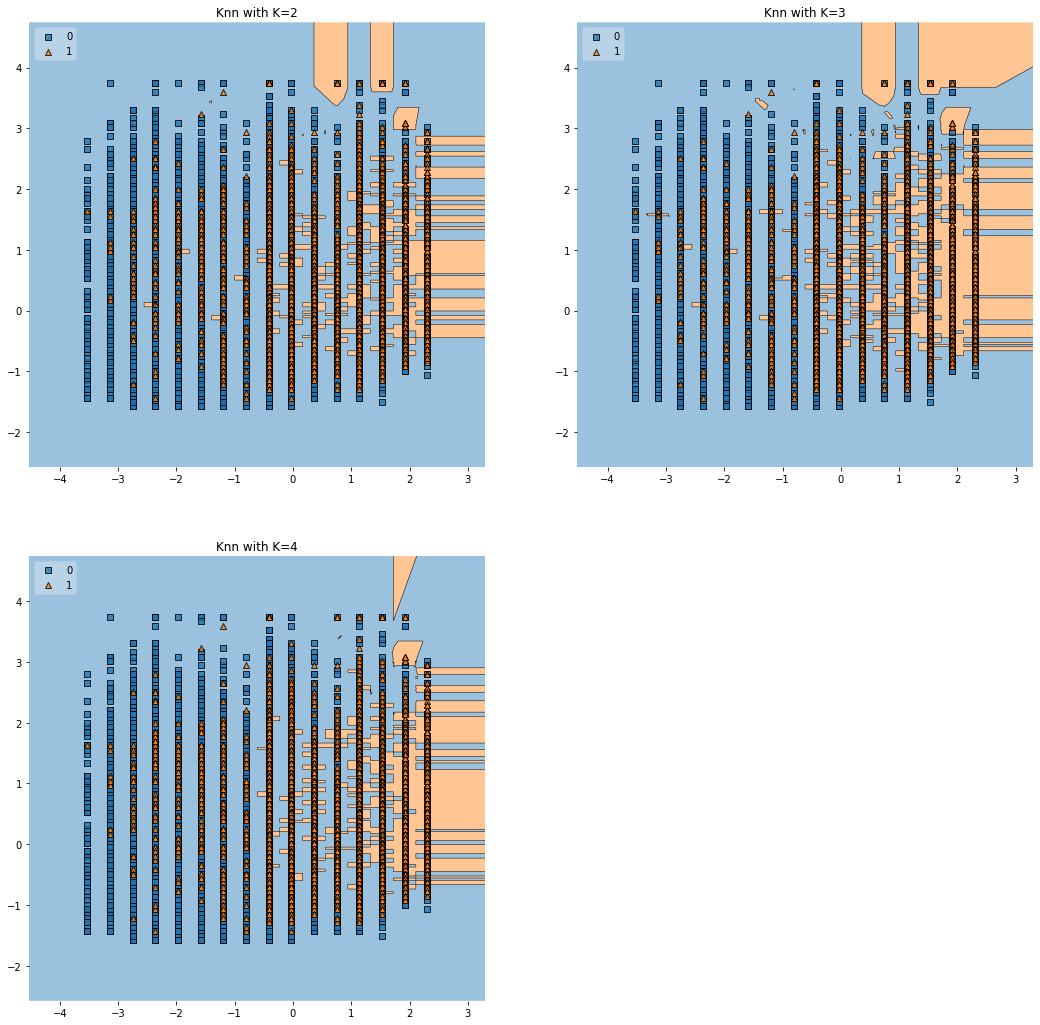
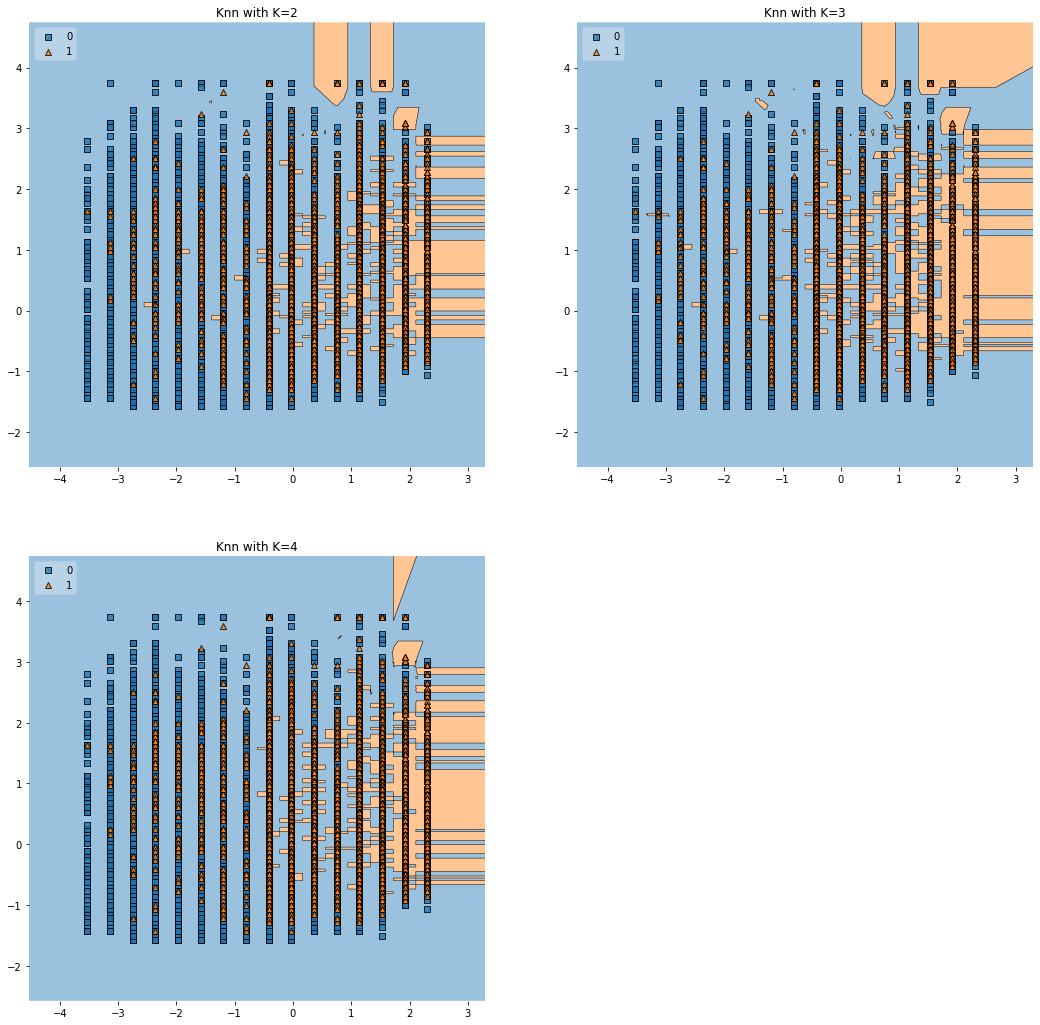
A picture containing chart

Description automatically generatedFig 3: Heatmap

Chart, line chart

Description automatically generatedFig 4: Error Rate vs K-values

Fig 5 : Decision Boundaries



**Appendix 2**

Table 1 : The descriptive statistics of numerical variables.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **count** | **mean** | **min** | **max** |
| **age** | 48842.0 | 38.643585 | 13.710510 | 90.0 |
| **education-num** | 48842.0 | 10.078089 | 2.570973 | 16.0 |
| **hourss-per-week** | 48842.0 | 40.422382 | 12.391444 | 99.0 |

Table 2: Classification Report

|  |  |  |  |
| --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** |
| **K=4** |  |  |  |
| 0 | 0.80 | 0.94 | 0.86 |
| 1 | 0.57 | 0.26 | 0.36 |
| **K=3** |  |  |  |
| 0 | 0.83 | 0.85 | 0.84 |
| 1 | 0.49 | 0.44 | 0.46 |
| **K=2** |  |  |  |
| 0 | 0.80 | 0.92 | 0.86 |
| 1 | 0.52 | 0.26 | 0.35 |

Table 3: Accuracy

|  |  |
| --- | --- |
| **K – values** | **Accuracy** |
| 4 | 77% |
| 3 | 75% |
| 2 | 76% |

Table 5: Confusion Matrix

|  |  |  |
| --- | --- | --- |
|  | **Yes** | **No** |
| **Yes** | 6935 | 479 |
| **No** | 1732 | 623 |