[redacted]

Table of Contents

Contents

Abstract

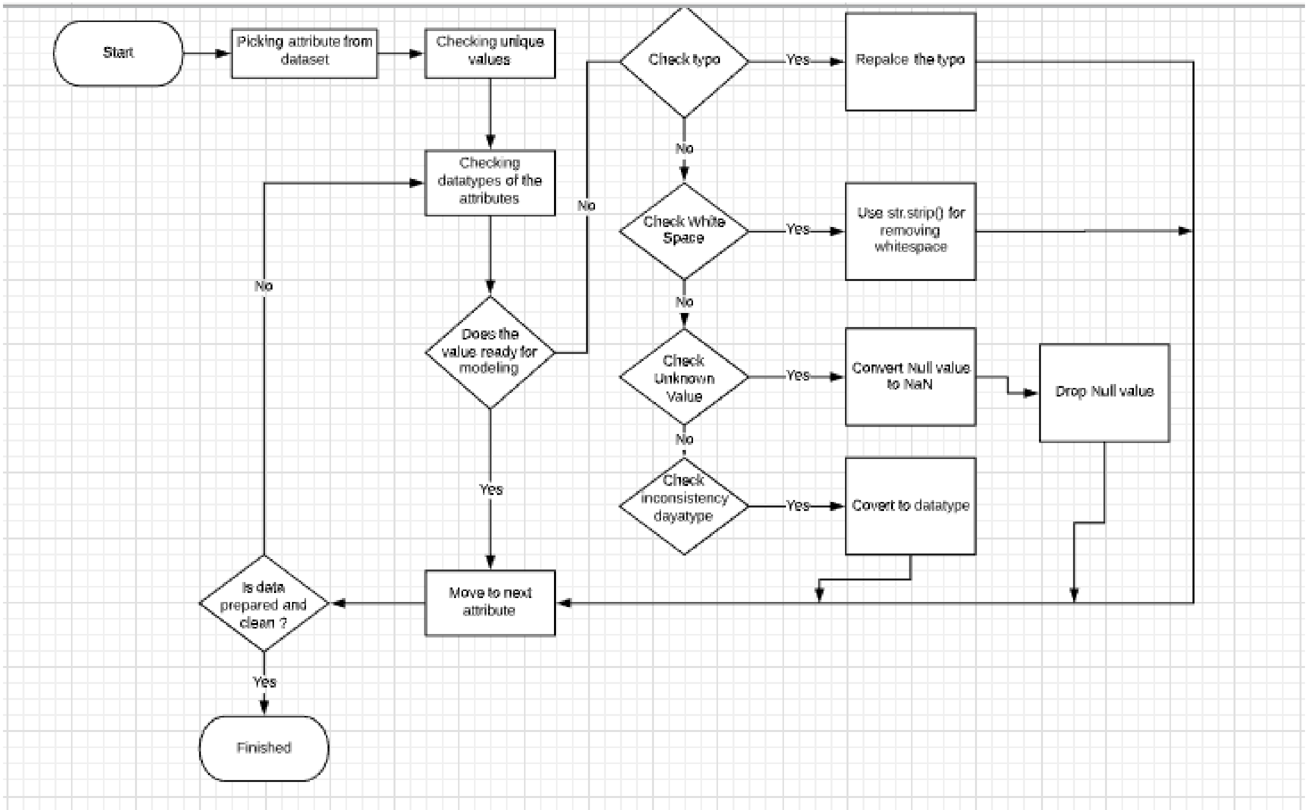
The goal of the assignment was to choose a database from https://archive.ics.uci.edu/ml/datasets.php for Problem Formulation, Data Acquisition and Preparation for data modeling. From UCI Machine Learning Repository, I choose the dataset named adult data set ‘adult.csv” by Ronny Kohavi and Barry Becker. Firstly, the dataset contains a lot of missing value and whitespace error. For data explosion and model purpose, I prepare the data and free from whitespace and typo. After cleaning the data, I exploration the data relation between columns.

Introduction:

The main purpose of the choosing adult dataset to compare the lifestyle of people. The dataset contains the salary, race, sex, marital status, country and so on. The dataset is force on the human lifestyle. Every human has different job, race, nationality and others attribute. This dataset is about the salary of adult’s employee annually based on different attribute. The dataset has attributes of work class, age, education, martial-status, salary, sex, race, capital-gain: capital-loss, hours-per-week, native-country and occupation. The target salary values >50,000 and <=50, 000.per year. Following the target value data in process in 3 way. -Data cleaning, Data Exploration and Data modeling.

First, I load the csv file name adult.csv using pandas library. After cleaning the data, the finding is that some attitude has problem with whitespace -Like ‘Education’, ’Sex’,” Race”, “Salary”, “Relationship”. I used str.strip() function for remove the all-whitespace of categorical data. Following that Work Class attribute had case sensitivity issue. I used str.lower() to get rid of the problem. After that, I had to deal with typos only NativeCountry had type value called ‘South’ which is not including any continent or county. I used all ‘?’ value with Nan using dropna() function.

Workflow of Project


Start
Picking attribute from dataset
Checking unique values
Check typo
Yes
Repalce the typo
No
Checking datatypes of the attributes
No
Check White Space
Yes
Use str.strip() for removing whitespace
No
No
Does the value ready for modeling
Check Unknown Value
Yes
Convert Null value to NaN
Drop Null value
No
Yes
Check inconsistency dayatype
Yes
Covert to datatype
Is data prepared and clean ?
Move to next attribute
Yes
Finished

Data Exploration:

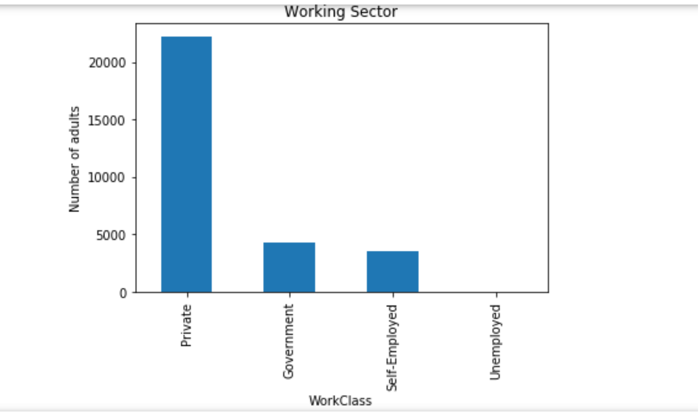
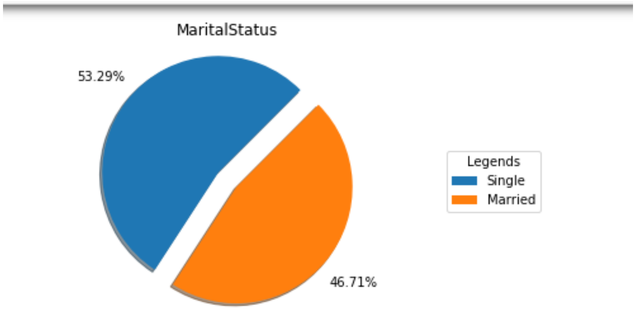


Figure: 1

Figure 1 displays the working sector of adults. It shows that how many adults working in the different sectors like- Private, Government, self-employed and unemployed. According to the bar charts, more than 20000 adults are working in the private companies which is highest of all. The workers in government sector and self-employed in both less than the private sector which is below 5000 adults.



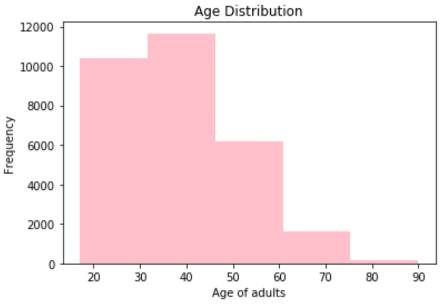
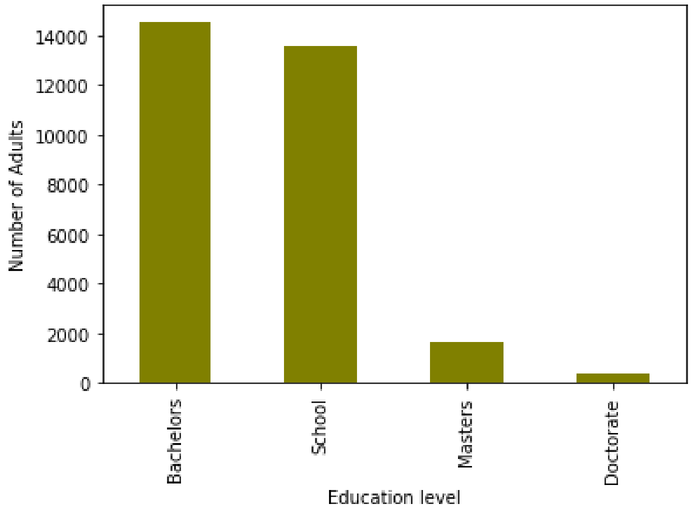


Figure:2

Figure: 3

According to Figure 2, the pie chart shows the martial status of adults in the adult dataset. Following this, we can see that majority (53.29%) of adults are single and 46.71% adults are married.

According to Figure 3, the histogram shows the age of adults. First of all, most of the adults is between age of 35 to 45 which is close to 12000. The 2nd highest is age between 20 to 35 which is around 10000.



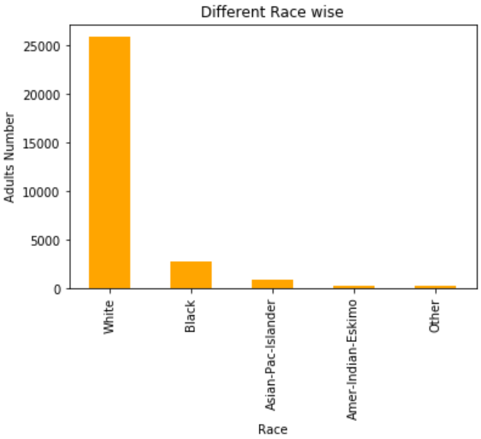
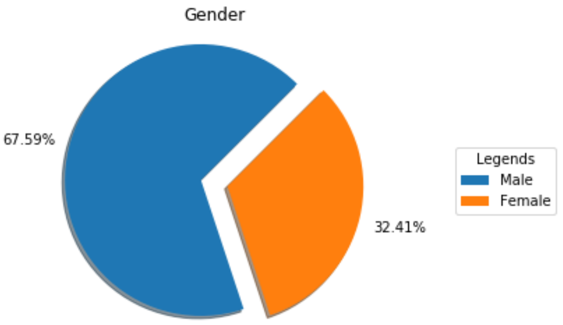


Figure: 4

Figure: 5

Figure 4 bar charts shows that adults education levels. The highest education level of the adult is doctorate. The highest number of adult complete their bachelor's degree which is around 14000. The 2nd highest number adults who complete their school level which is more than 12000. The lowest number of adult's complete master and doctorate degree which is below 2000.

According to Figure 5, the bar chart contains adult in different race. The highest number of adults is white which is more than 25000.The least number of adults is from American Indian which is below 1000. In the dataset, 2000 adults are in black race which is the 2nd highest after white.



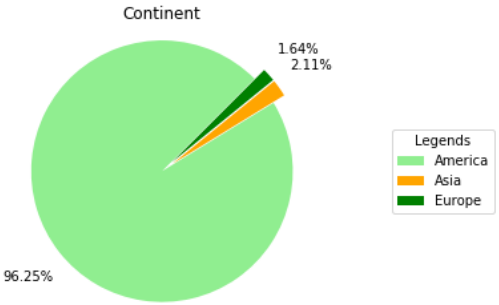


Figure: 6

Figure: 7

In Figure 6, the pie chart divide between Gender- Male & Female. 67.59% are male in the adult dataset and 32.41% is Female. In Figure 7, the pie chart shows continent of adults. The most of number of the people are from America (North and South) continent which is 96.25%. 2.11% people belong to Asia and 1.64% people from Europe.

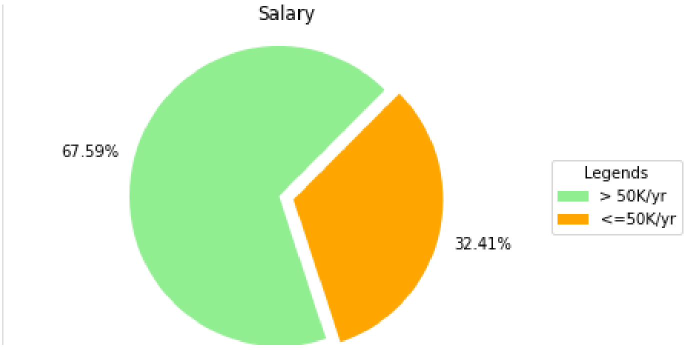


Figure:8

According to Figure 8, the pie chart contains based on salary of adult in adult dataset. 67.59% adults are earned more than 50,000 annually. 32.41% adults earn below 50,000 in that year.

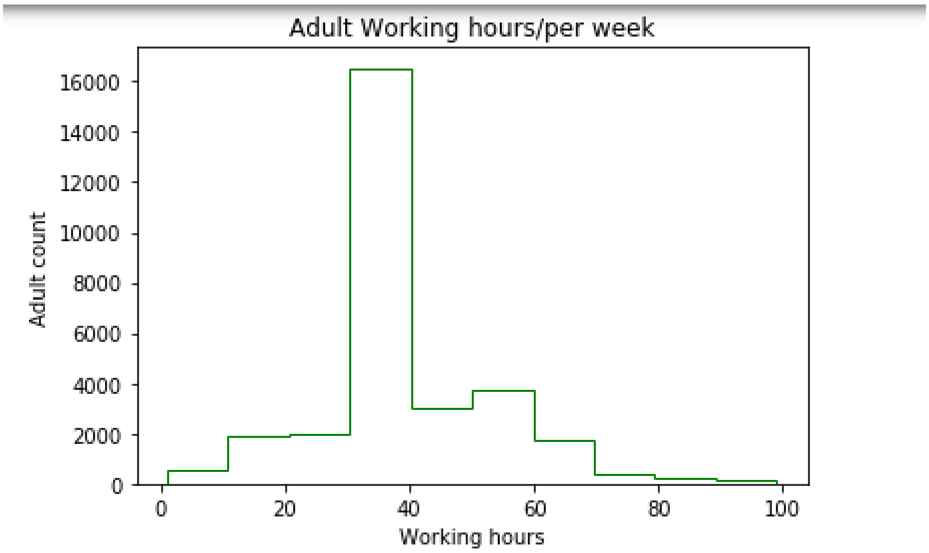
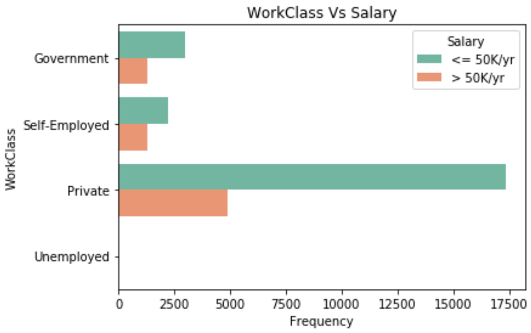


Figure: 9

According to figure 9, the graph shows the adults working hours/per week basis. Most of the adults work 3=25-40 hours per week in their jobs which is around 16000 adults. As full-time job is 40 hours a week, adults who has full time job work 40 hours a week mostly. Some proportion of the adults doing last then 20 hours a week which is below 2000.



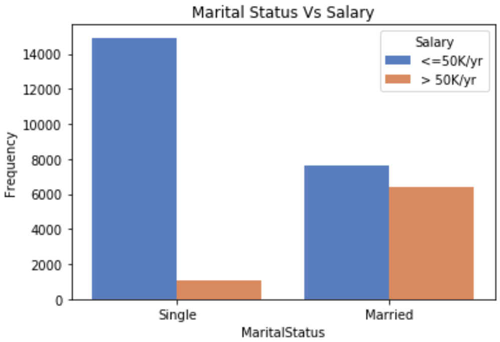


Figure :10

Figure :11

According to Figure 10, the bar is comparison between salary and work class. The bar charts of comparison focusing question that Which type of working class getting more paid. We can see from the chart is private work class getting more paid than any other working class. The amount of salary of private sector is five times than government works and others in amount of below 50000 per year.

In figure 11, the bar charts are a comparison between marital status a salary. The bar charts follow a research question is does material status depends on salary or not. From the comparison, we can see that, most of the single people earning is below 50,000 annually. Married people salary almost same between more and less than 50000 annually.

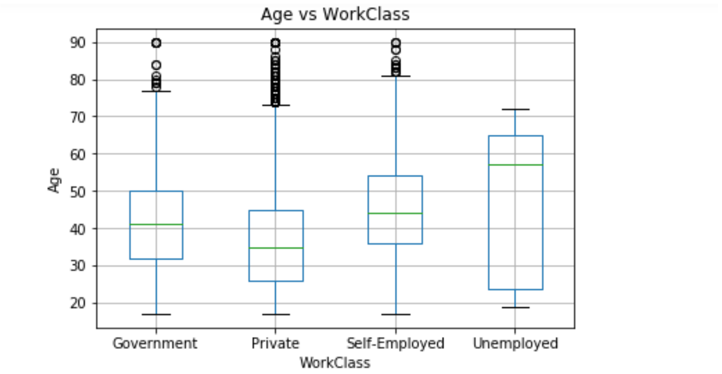
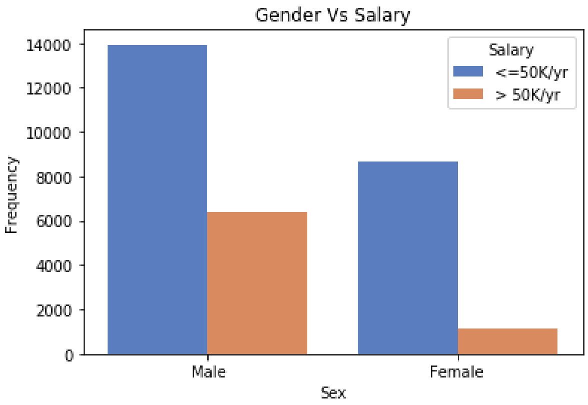


Figure: 12

The boxplot of comparison between age and work class end up a question which is does age matter in work class. From the figure 12 we can see that middle age people are mostly work in private sector. Also, people over 30m mostly work in government sector. The unemployed work class people are mostly over 50.



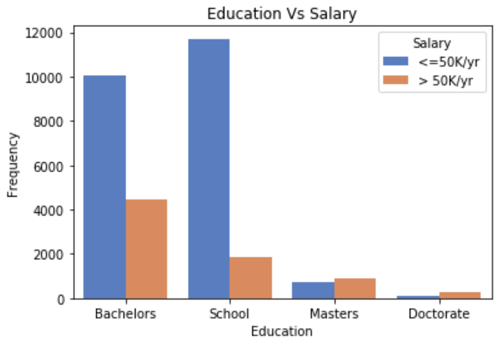
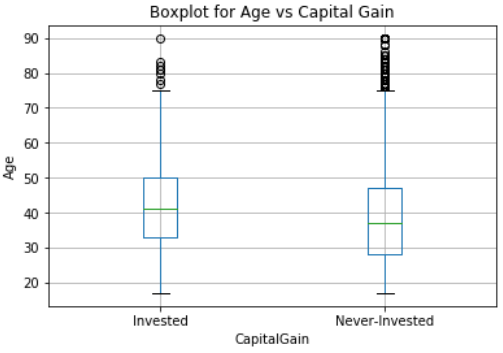


Figure: 13

Figure :14

In Figure 13, the comparison between gender and salary shows in the bar chart. Firstly, the question stands with gender type earn most. It shows that in both salary type male is more than the female.

In Figure 14, the comparison between education and salary. Firstly, this bar chart of comparison stands the question that does more education means more salary. we can see from the figure 14 that most of the adult who complete bachelor’s degree earns below 50000 annually. Adults who complete their masters the salary is not much difference.



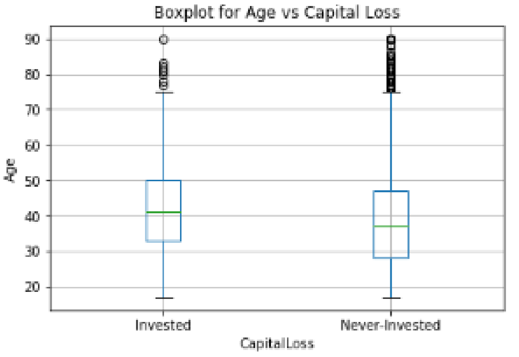


Figure :15

Figure:16

The question stands comparing above two Figure 15 and Figure 16 boxplot that people who choosing to invest can be in capital gain or capital loss. From the boxplot we can justify that middle age adults around 40 who invest with anything for capital gain. On the other hand, the middle-aged adult who invested is aged 41.

Data Modeling

In data modeling, I used K-Nearest Neighbour (KNN) for classification. First of all, label the dataset salary 0 and 1 binary number for salary range. Classification meanly use for to find out data accuracy level. Firstly, for modeling purpose, add categorical data covert numerical value. I use manual mapping for that. Some data already contains numerical value. I divide the dataset data and target for modeling. I split according to requirement Suite1: 50% for training and 50% for testing, Suite2: 60% for training and 40% for testing and Suite 3: 80% for training and 20% for testing. I start with k=1 in KNN means. The result is low accuracy level. Then, I increased the k value for finding accuracy.

The 3 different suite the confusion matrix is:

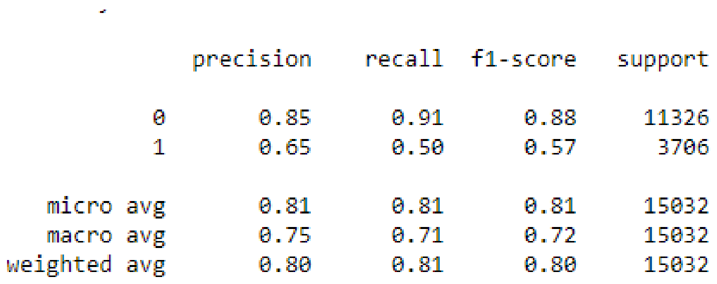
KNN 60%-40%	0	1
0	8416	642
1	1548	1420

KNN 50%-50%	0	1
0	10343	983
1	1857	1849

KNN 80%-20%	0	1
0	4160	407
1	695	751

The classification report:

Suite 1: 50% for training and 50% for testing

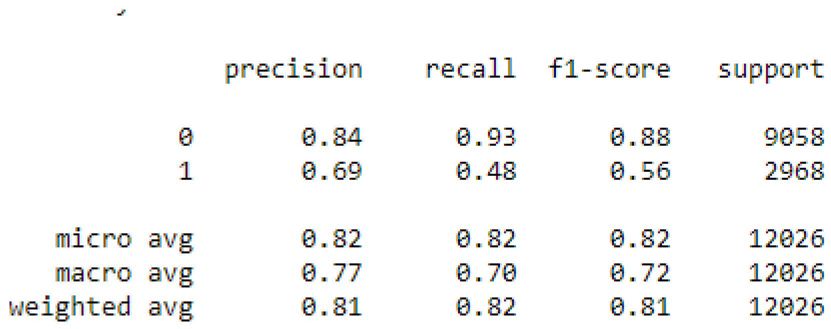


```
              precision    recall  f1-score   support

           0       0.85      0.91      0.88     11326
           1       0.65      0.50      0.57      3706

   micro avg       0.81      0.81      0.81     15032
   macro avg       0.75      0.71      0.72     15032
weighted avg       0.80      0.81      0.80     15032
```

Suite 2: 60% for training and 40% for testing

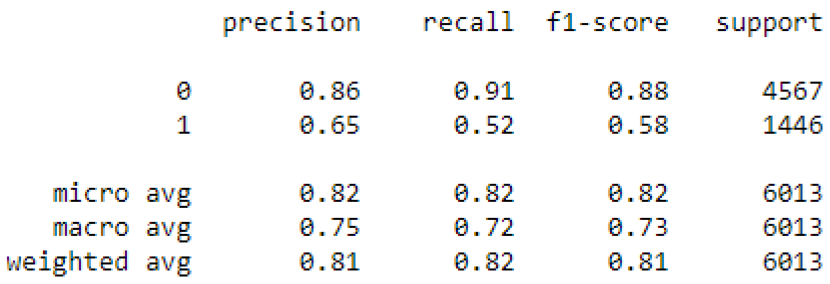


```
              precision    recall  f1-score   support

           0       0.84      0.93      0.88      9058
           1       0.69      0.48      0.56      2968

   micro avg       0.82      0.82      0.82     12026
   macro avg       0.77      0.70      0.72     12026
weighted avg       0.81      0.82      0.81     12026
```

Suite 3: 80% for training and 20% for testing



```
              precision    recall  f1-score   support

           0       0.86      0.91      0.88      4567
           1       0.65      0.52      0.58      1446

   micro avg       0.82      0.82      0.82      6013
   macro avg       0.75      0.72      0.73      6013
weighted avg       0.81      0.82      0.81      6013
```

Discussion:

For the classification report, we can see that the adult dataset is not balanced. F-1 score value is closed than recall and precision. We can see from increasing split ratio, we get the low value at a time. When we increase the spit value o 50%:50% to 60%:40% and 80%:20%, the precision is decreasing.

Conclusion:

In adult data test, comparing the founding of KNN the accuracy: 0.82 and the f-1 score shows the most content results. The dataset has major problem is some attribute the value is not that much clear. However, after doing all modeling, more than 50000 salary earning is more accurate than others.