**Income prediction using Adult data**

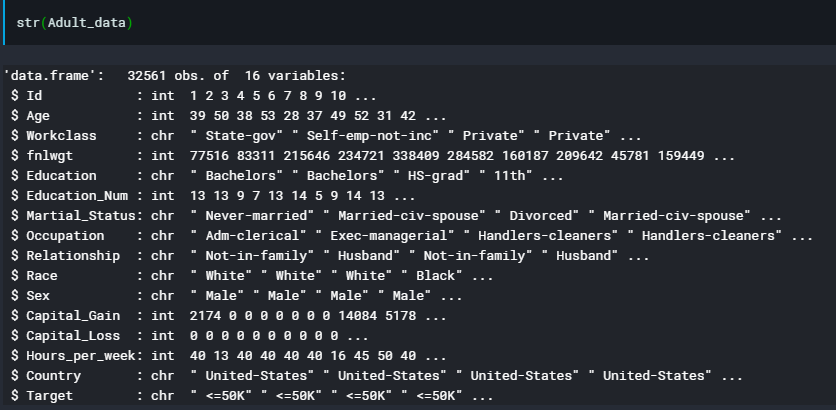
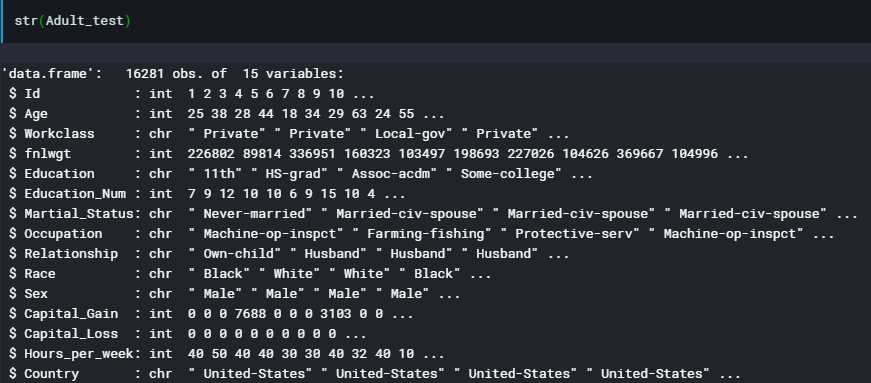
**TABLE OF CONTENTS**

1. **Abstract**
2. **Data Overview**
3. **Data Visualization**
4. **Data Summary**
5. **Preprocessing Steps**
6. **Model Fitting – Logistic Regression**
7. **Model Fitting – Random Forest**

**H. Conclusion**

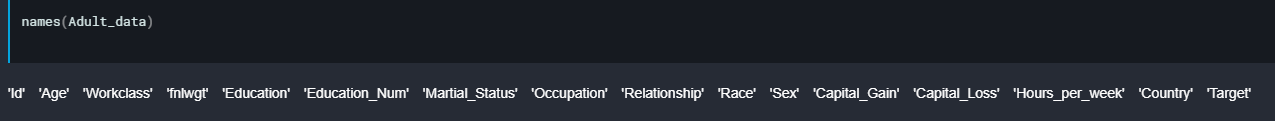
1. **Abstract**

The Adult dataset is from the 1994 Census database. It is also know as “Cencus Income” dataset. Details of this dataset can be found at [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/adult). In this in class competition which is [IT405\_KHAS](https://www.kaggle.com/c/it405-khas) on [Kaggle](https://www.kaggle.com/) platform , we used pre-prepared train and test set. In our initial stages , we preprocessed the data and developed understanding of the data and its useful features that explain the variances by doing various types of exploratory analysis. Then , we implemented some classification algorithms for predicting.After that, we upload our prediction results to kaggle.

**B. Data Overview**

The dataset used in this project has 32561 record and a binomial label indicating a salary <50K or >50K USD. %75 of the record in the dataset have a class label of <=50K and test dataset contains 16,281 records.

There are 14 columns. You can check names of column the Adult dataset screenshot at the below.



**C. Data Visualization**

I tried to understand income relation with age and gender.Firstly,I created histogram of age by target group.After that I created histogram of age by gender group.

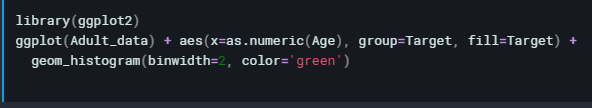


Figure 1-Codes histogram of age by Target group.

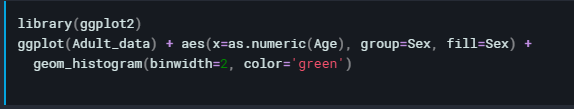


Figure 2-Codes histogram of age by gender group.

Table 1

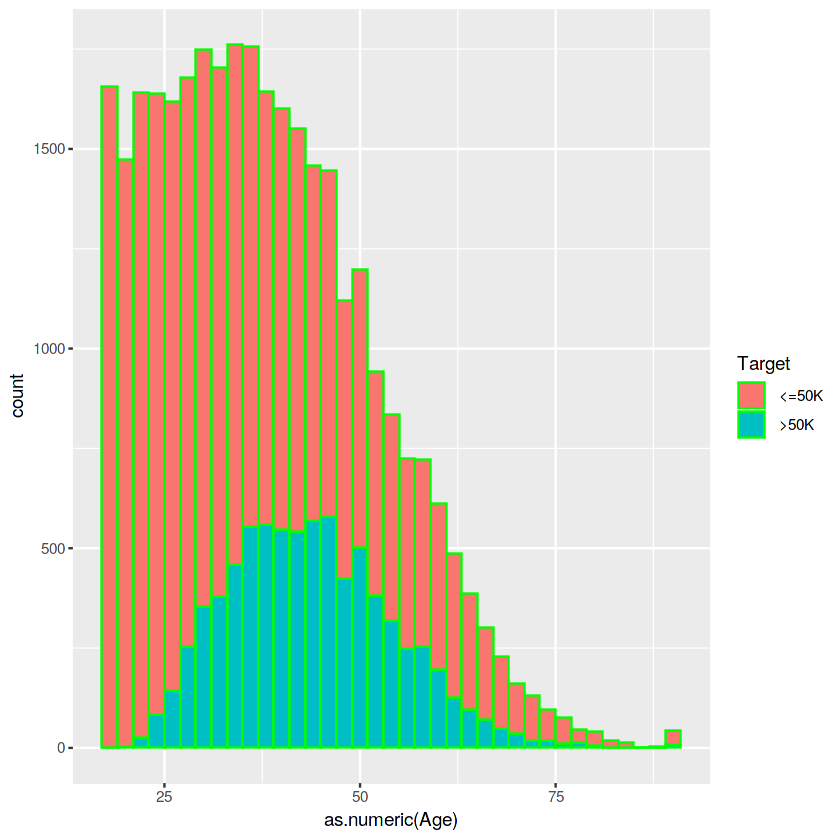
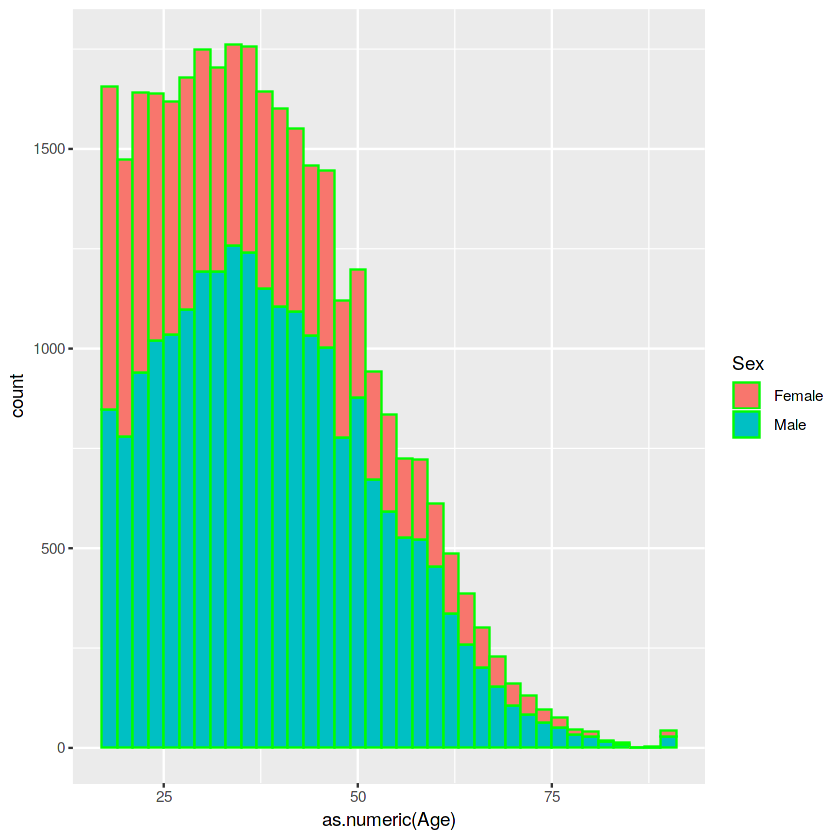
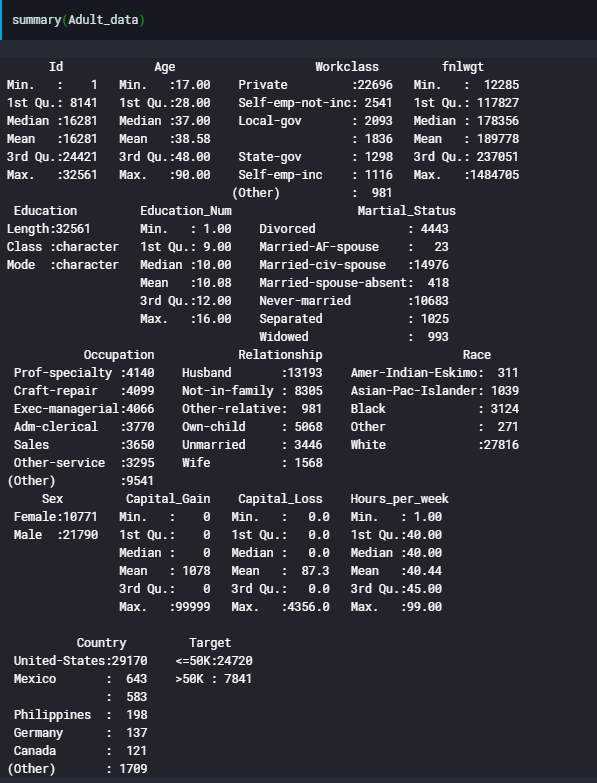


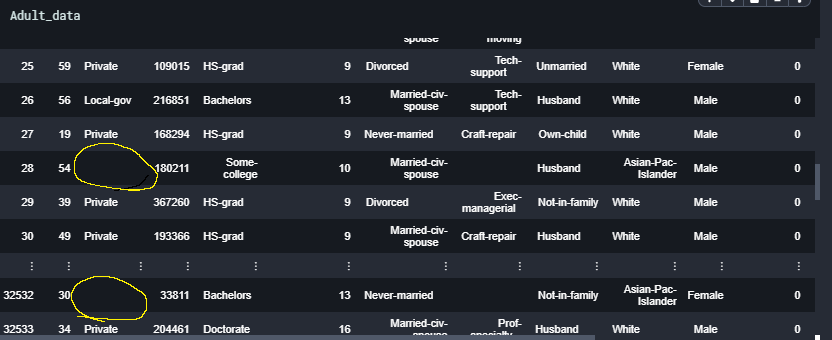
Table 2



**D. Data Summary**

****

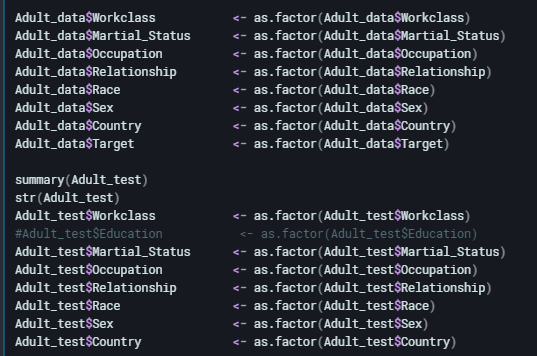
We can see a lot statistics information about our data such as min,max, mean etc.

There some missing values in our dataset. You can see these missing values picture at the below.

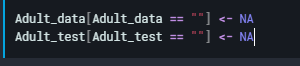
We need to handle these missing values. I will show how i handled these values.

**E. Preprocessing Steps**

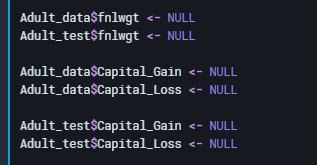
There are some missing values and noisy data in our dataset. Also some columns has no correlation such as capitalgain, capitalloss ,fnlwgt .

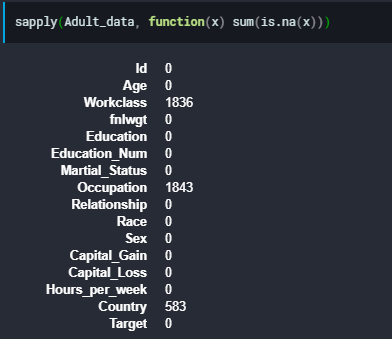
Firstly i converted chr columns to factor.

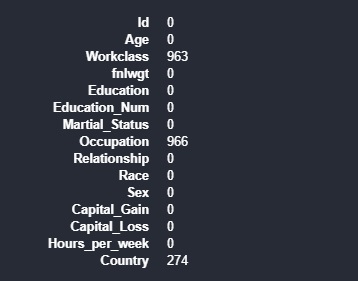
Then i replace missing values to “NA” values.

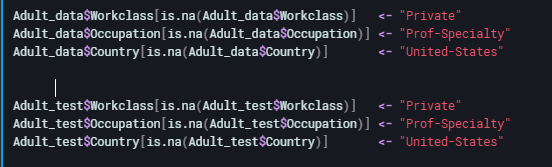


Also i extracted no correlation columns.

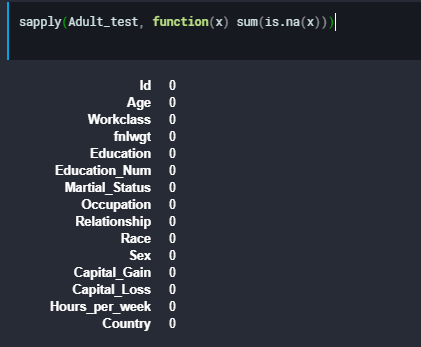


Then i check “NA” values.

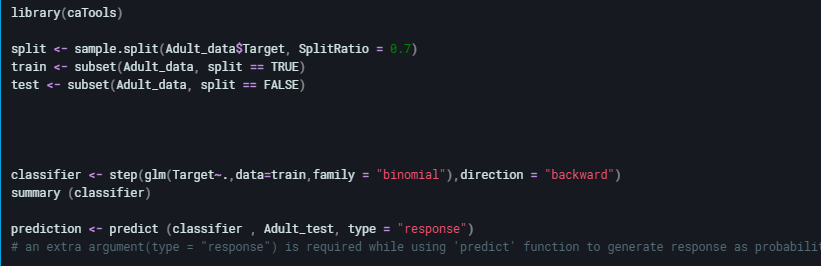


After that i replaced “NA” values to mode values.

Now there is no NA value.

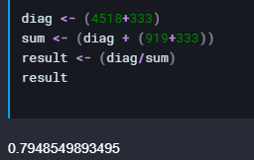
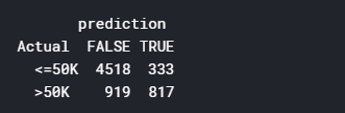


**F. Model Fitting – Logistic Regression**

After preprocessing data we can implement our classification model which is Logistic regression.

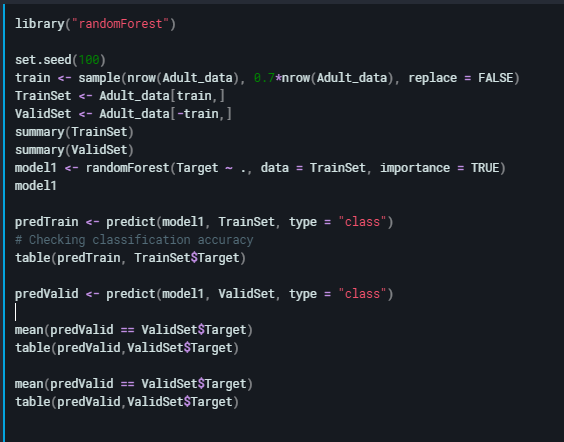
Then check accuracy.



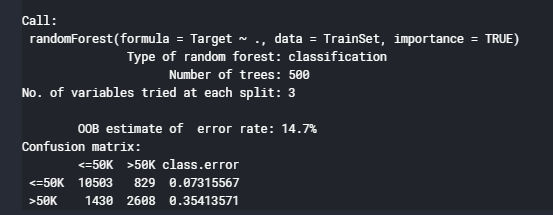


**G. Model Fitting – Random Forest**

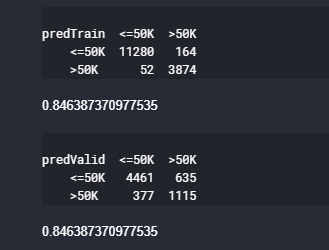
We created our Random Forest classification model and calculate accuracy.



RandomForest 1

Our error rate 14.7. It is not perfect but not bad. I tried reduce error rate while changing seed and split ratio but it is less error rate that i reach.

Random Forest Accuracy.



**H. Conclusion**

To Conclude , I tried to using some classification models to predicting income by using Adult Cencus data. Random Forest is faster and more succesfull than Logistic Regression.