

American International University-Bangladesh (AIUB)  
**Department of Computer Science and Engineering   
Faculty of Science &Technology (FST)  
Summer 21-22**

**DATA WAREHOUSING AND DATA MINING**

**Section: C**

**Project Title: Brain Stroke Classification**

**Submitted by:**

|  |  |
| --- | --- |
| **Name** | **ID** |
| 1. MD.JAWADUL HASAN | 20-41958-1 |
| 1. JOY MATUBBER | 20-41959-1 |
| 1. NIHAN SHAHRIAR PALOCK | 20-43167-1 |
| 4. ARAFAT AL KHAIR FUAD | 20-43607-1 |

**Submitted to:**

**DR. AKINUL ISLAM JONY**

**Assistant Professor**

**Department of Computer Science**

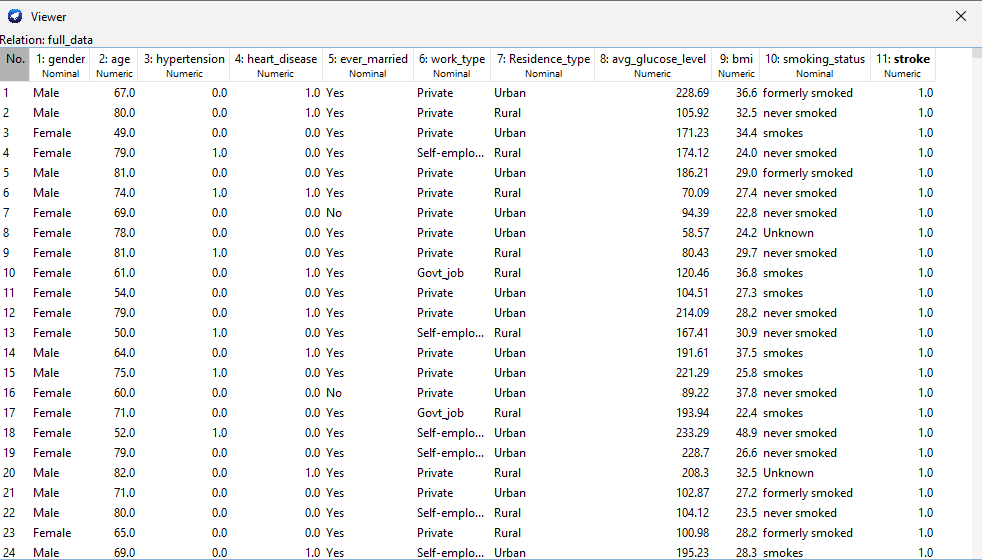
**Project Overview:** A stroke is a condition when there is insufficient blood supply to the brain. Strokes can be ischemic or hemorrhagic, which refers to a lack of blood flow or bleeding. A strong headache might potentially be a symptom of a hemorrhagic stroke. After a stroke, symptoms and signs frequently start to show very quickly. Blood pressure is the most significant risk factor for stroke. High blood pressure can also cause blood clots to form in the arteries leading to the brain, blocking blood flow and potentially causing a stroke. Other risk factors include hypertension, heart disease, and smoking which can affect blood pressure and that causes brain stroke. In this project, we have collected a dataset containing ten different attributes to predict whether a person could have a brain stroke or not. Building a reliable model with higher accuracy may help people predict their stroke possibilities, which will surely help to take some prevention.

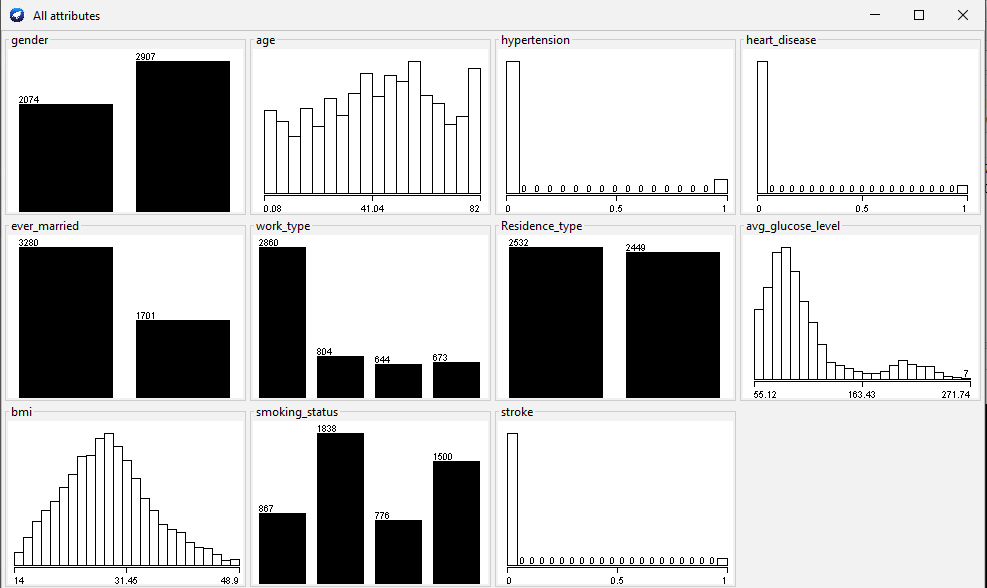
**Dataset Overview:**

The dataset is a brain stroke prediction dataset containing 10 attributes and a class variable named stroke, with values between 0 and 1. Here 1 denotes that the person has predicted brain cancer, and 0 means not.

The first attribute is gender which is a nominal type where 52% are Male, and 48% are Female. The second attribute, age, is a numeric type. The third and fourth attributes are numeric attributes containing a value between 0 and 1.0 if the patient doesn't have hypertension and heart disease and 1 if the patient has hypertension and heart disease. The fifth attribute, married status, is a nominal value where "Yes" denotes married and "No" denotes not married. Among the collected data, 74% are married, and 26% are not married. The sixth, seventh and 10th attributes are nominal attributes. The 10th attribute has some value as "unknown" meaning these values couldn't be collected. The 8th and 9th attribute is numeric. The dataset contains total 4981 rows and there is fortunately no missing values except the 10th attribute "smoking\_status".

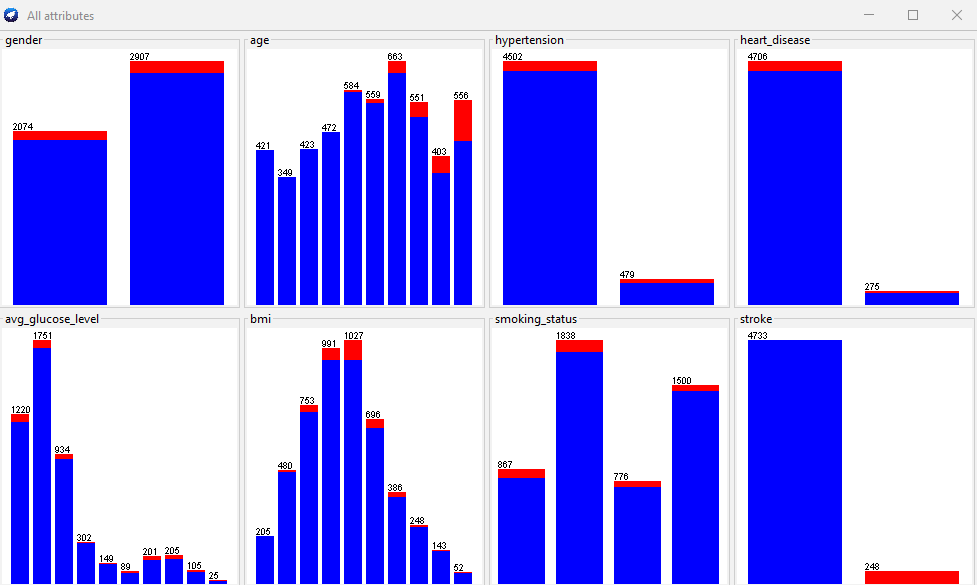
Dataset URL: <https://www.kaggle.com/datasets/zzettrkalpakbal/full-filled-brain-stroke-dataset?select=full_filled_stroke_data+%281%29.csv>





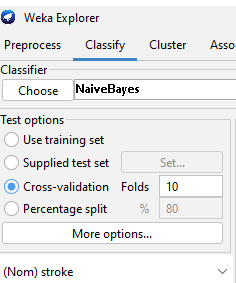
**Model Development:**

Data preprocessing**:** Before applying classification, it is needed to pre-process the data. The dataset contains both continuous values and nominal categorical values. But to use the Naive Bayes algorithm, all the attributes need to be converted to categorical values. Here we have converted all the attributes to nominal values. And ever\_married,work\_type, and Residence\_type are dropped. Because this attribute is less likely related to predicting our target attribute.

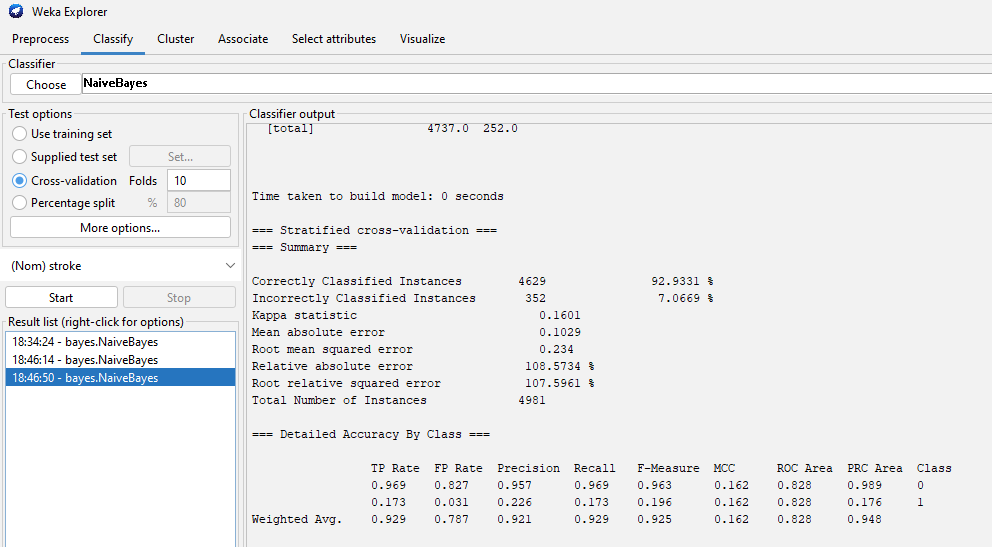


***Applying Naive Bayes Classification:***

**Using K fold cross validation**

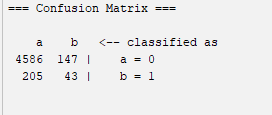


At first, we have chosen K fold cross validation and have settled K=10. We got an accuracy 92.93% after choosing this test option.



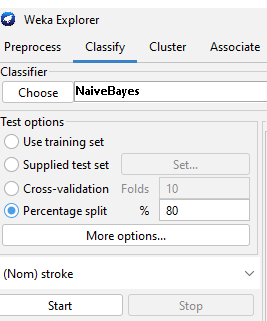
**Predictive accuracy:** 92.93%

**Confusion Matrix:**

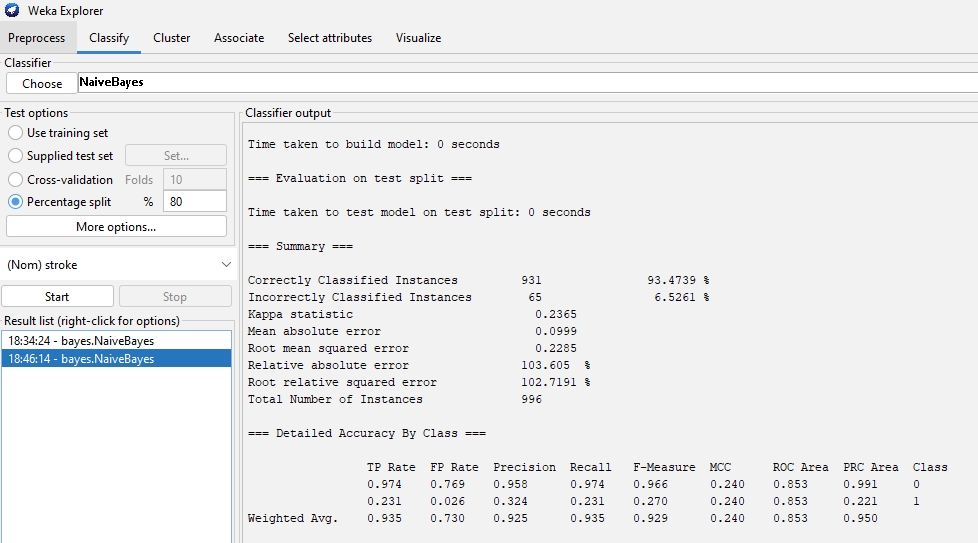


Here output 0 is predicted 4586 times correctly, and 147times it has predicted wrong value. And 1 is predicted correctly 205 times, and 43 times it has predicted the incorrect value

**Using Train test:**

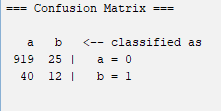


Then have split the dataset into train and test dataset into 80% and 20% ratios. We got an accuracy 93.47% after choosing this test option.



**Predictive accuracy:** 93.47%

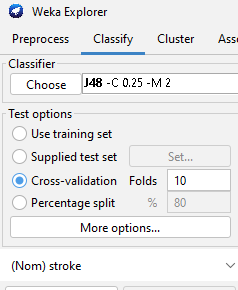
**Confusion Matrix:**



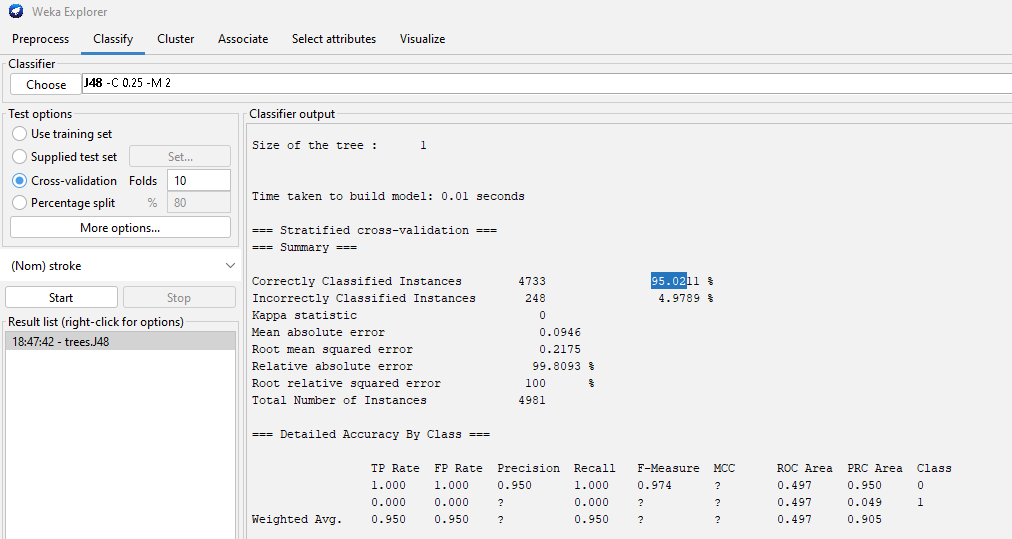
Here output 0 is predicted 919 times correctly, and 25times it has predicted wrong value. And 1 is predicted correctly 40 times, and 12 times it has predicted the incorrect value

***Applying Decision Tree Classification:***

**Using K fold cross validation**

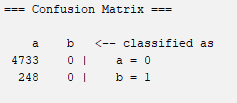


At first, we have chosen K fold cross validation and have settled K=10. We got an accuracy 95.02% after choosing this test option.



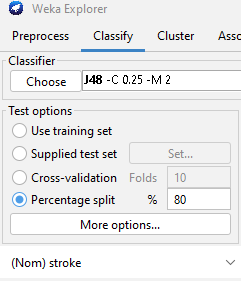
**Predictive accuracy:** 95.02%

**Confusion Matrix:**

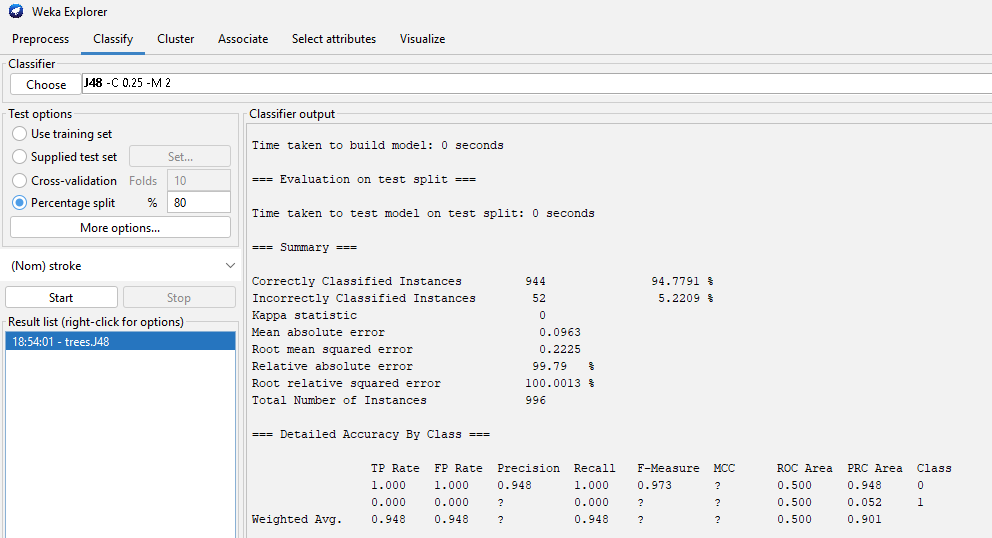


Here output 0 is predicted 4733 times correctly, and 0 times it has predicted wrong value. And 1 is predicted correctly 248 times, and 0 times it has predicted the incorrect value. There could be some instances which remains unclassified due to missing branches of the tree.

**Using Train test:**

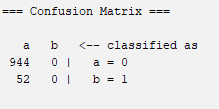


Then have split the dataset into train and test dataset into 80% and 20% ratios. We got an accuracy 94.77% after choosing this test option.



**Predictive accuracy:** 94.77%

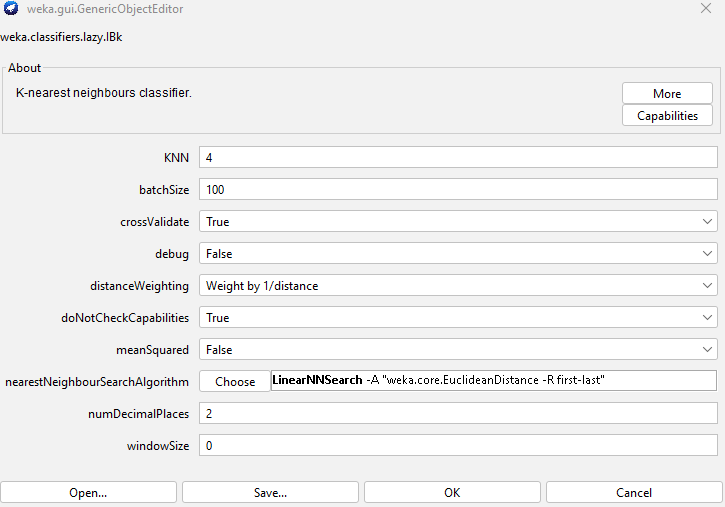
**Confusion Matrix:**



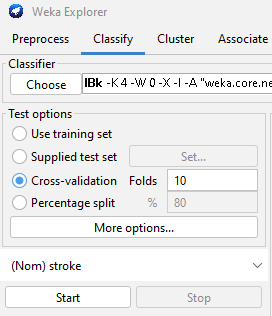
Here output 0 is predicted 944 times correctly, and 0 times it has predicted wrong value. And 1 is predicted correctly 52 times, and 0 times it has predicted the incorrect value. There could be some instances which remains unclassified due to missing branches of the tree.

***Applying Nearest Neighbor Classification Algorithm***

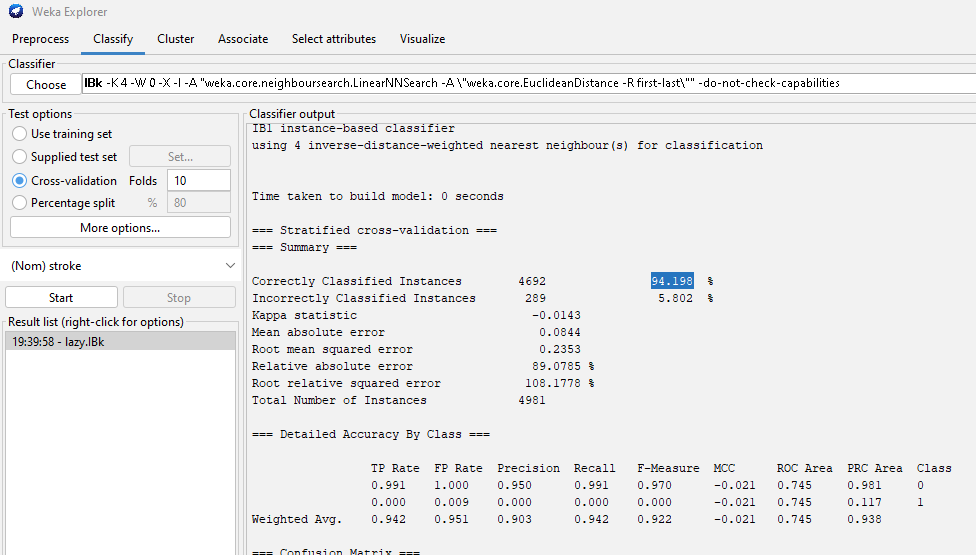
K=4 is assigned.



**Using K fold cross validation**

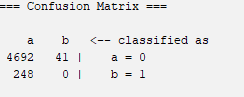


At first, we have chosen K fold cross validation and have settled K=10. We got an accuracy 94.198 % after choosing this test option.



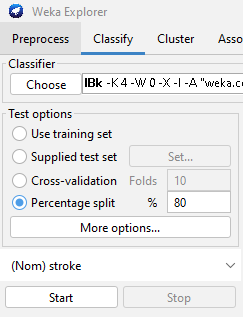
**Predictive accuracy:** 94.198%

**Confusion Matrix:**

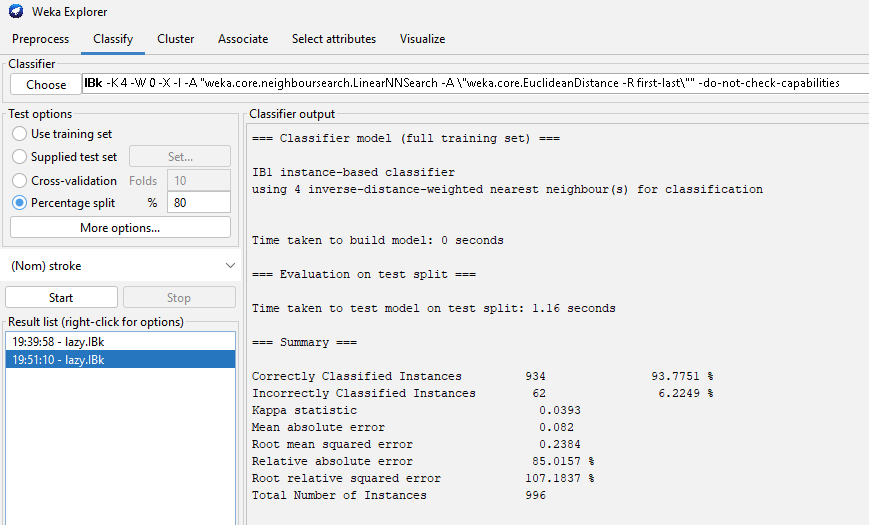


Here output 0 is predicted 4692 times correctly, and 41 times it has predicted wrong value. And 1 is predicted correctly 248 times, and 0 times it has predicted the incorrect value

**Using Train test:**

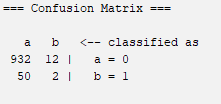


At first, we have chosen K fold cross validation and have settled K=10. We got an accuracy 93.77 % after choosing this test option.



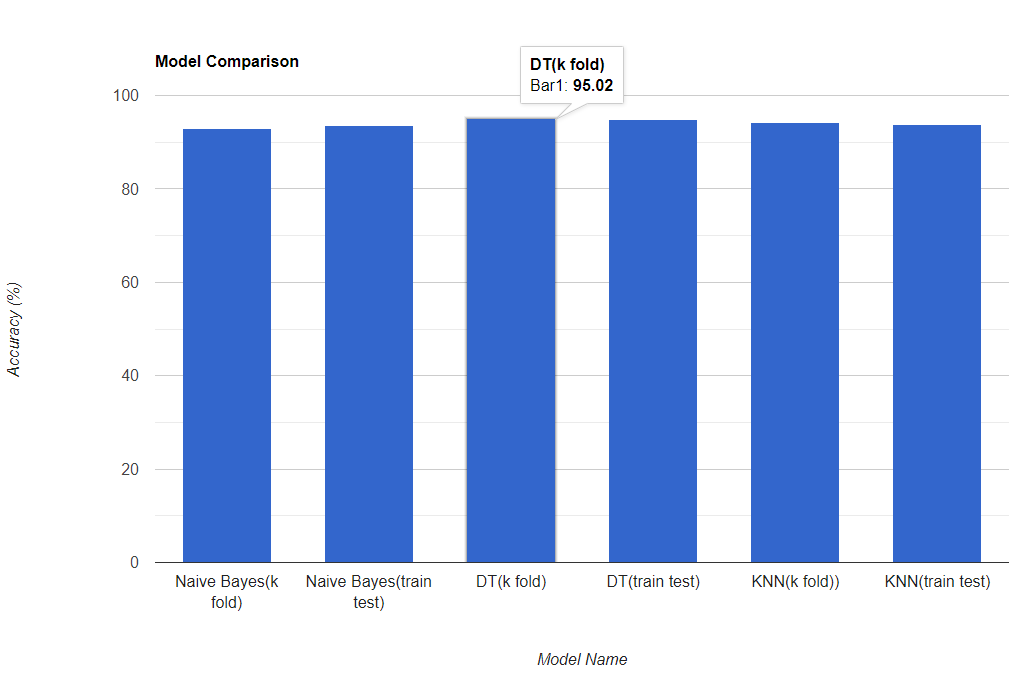
**Predictive accuracy:** 93.77 %

**Confusion Matrix:**

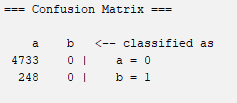


Here output 0 is predicted 932 times correctly, and 12times it has predicted wrong value. And 1 is predicted correctly 50 times, and 2 times it has predicted the incorrect value

**Discussion and Conclusion:**

****

We have used three classification algorithms with k fold and cross-validation and the train and test split method. So far, we got the highest accuracy from the Decision tree algorithm with k fold cross validation with K=10. And the accuracy is 95.02%. So we are considering the decision tree as the best model among them. Now we will observe the confusion matrix of this model.



Here output 0 is predicted 4733 times correctly, and 0 times it has predicted wrong value. And 1 is predicted correctly 248 times, and 0 times it has predicted the incorrect value. The model predicted 248 False Negative which means that we did not predict 248 cases which are going to have stroke.

And the same result is for the other two algorithms so far. False negative value is always high for the KNN and Naive Bayes as well.

Now, if we look at the data, we can see the imbalance of the data. So the model prediction is not going to be acceptable. But it may be possible to get a better result using another algorithm except for these three using these data.

