# AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH (AIUB)

## Faculty of Science & Technology

# PROJECT REPORT

## Project Title: Smoking Status Prediction of Brain Stroke

| | | | |
|---|---|---|---|
| **Due Date:** | August 7, 2022 | **Date of Submission:** | August 7, 2022 |
| **Course Title:** | DATA WAREHOUSE & DATA MINING | | |
| **Course Code:** | CSC4285 | **Section:** | C |
| **Semester:** | Summer 2021-22 | **Degree Program:** | BSc in CSE |
| **Course Teacher:** | **AKINUL ISLAM JONY** | | |

**Declaration and Statement of Authorship:**

1. I/we hold a copy of this Assignment/Case-Study/Project, which can be produced if the original is lost/damaged.
2. This Assignment/Case-Study is my/our original work and no part of it has been copied from any other student's work or from any other source except where due acknowledgment is made.
3. No part of this Assignment/Case-Study has been written for me/us by any other person except where such collaboration has been authorized by the concerned teacher and is clearly acknowledged in the assignment.
4. I/we have not previously submitted or currently submitting this work for any other course/unit.
5. This work may be reproduced, communicated, compared, and archived for the purpose of detecting plagiarism.
6. I/we give permission for a copy of my/our marked work to be retained by the Faculty Member for review by any internal/external examiners.
7. I/we understand that Plagiarism is the presentation of the work, idea, or creation of another person as though it is your own. It is a form of cheating and is a very serious academic offense that may lead to expulsion from the University. Plagiarized material can be drawn from, and presented in, written, graphic and visual forms, including electronic data, and oral presentations. Plagiarism occurs when the origin of the source is not appropriately cited.
8. I/we also understand that enabling plagiarism is the act of assisting or allowing another person to plagiarize or copy my/our work.

** Student(s) must complete all details except the faculty use part.*
** Please submit all assignments to your course teacher or the office of the concerned teacher.

## Group Members:

| Sl No | Name | ID | PROGRAM | SIGNATURE |
|---|---|---|---|---|
| 1 | AYESHA AKTER VELEN | 19-40892-2 | BSc in CSE | |
| 2 | SHOCHI AKTER | 19-40235-1 | BSc in CSE | |
| 3 | RAPHA TAHSIN | 19-40944-2 | BSc in CSE | |
| 4 | DIBYA JYOTI HORE | 19-41310-3 | BSc in CSE | |

| *Faculty use only* | | |
|---|---|---|
| FACULTY COMMENTS | **Marks Obtained** | |
| | **Total Marks** | |

## Table Of Contents Page No.

# Project Title

Smoking Status Prediction of Brain Stroke.

# Project Overview

## Abstract

This project is about Smoking Status Prediction of Brain Stroke.. The main purpose of this project is to apply data mining algorithm so that we can visualize the future event depending on the current datasets as well as its accuracy rate.The guideline of this report is to evaluate the feature of parameters used for a variety of datasets.To develop this project, the software named **“”Weka”** is used.

## Objective

Our datset for this project is about Brain Stroke Prediction.From this data set we will predict the smoking status of brain stroke. By doing this project we will have clear knowledge about different tools and classifiers in data mining.There are some objectives for this project.The objectives of this project are-

1. To familiarized with “Weka” tool.
2. To apply different classifications algorithms using “Weka”.
3. To determine the accuracy of the selected data set.
4. To show the comparison between different classifiers.

## Description

In this project, we will go over selecting dataset with minimum 1000 instences and applying different classifier algorithm techniques to predict our target attribute. It also describes numerous additional manufactured strategies used to improve the operation of the program. There are numerous techniques available like including K-NN, Nave Bayes, Decision Tree (ID3, CART), Association Rule and clustering . In this order, we will employ Nave bias algorithm, K- Nearest Neighbour(K-NN) algorithm and Decision Tree to develop our project model.

## Outcome

For this type of analysis we get to know the accuracy rate of different classifiers and can predict which algorithm is better working for our dataset.

# Dataset Overview

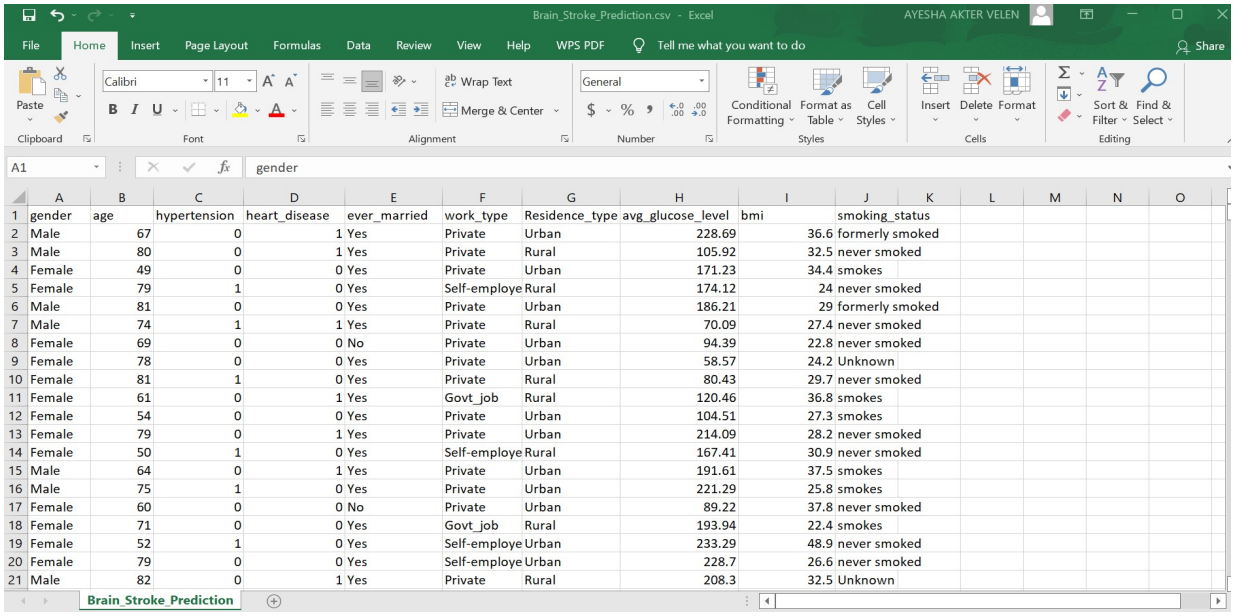
## Source URL

The dataset link for our project is given below:
**https://www.kaggle.com/datasets/zzettrkalpakbal/full-filled-brain-stroke-dataset**

## Dataset Description

We take the Smoking Status of Brain Stroke dataset from Kaggle web site.Our datasets have 10 attributes and 4981 instances. Based on this our prediction of brain stroke smoking status becomes more accurate.The Target variable for this dataset is Smoking status.The attributes of this project are-

- gender: "Male", "Female" or "Other"
- age: age of the patient
- hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
- heartdisease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease 5) evermarried: "No" or "Yes"
- worktype: "children", "Govtjov", "Neverworked", "Private" or "Self-employed" 7)
- Residencetype: "Rural" or "Urban"
- avgglucoselevel: average glucose level in blood
- bmi: body mass index
- smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*



| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | gender | age | hypertension | heart_disease | ever_married | work_type | Residence_type | avg_glucose_level | bmi | smoking_status |
| 2 | Male | 67 | 0 | 1 | Yes | Private | Urban | 228.69 | 36.6 | formerly smoked |
| 3 | Male | 80 | 0 | 1 | Yes | Private | Rural | 105.92 | 32.5 | never smoked |
| 4 | Female | 49 | 0 | 0 | Yes | Private | Urban | 171.23 | 34.4 | smokes |
| 5 | Female | 79 | 1 | 0 | Yes | Self-employe | Rural | 174.12 | 24 | never smoked |
| 6 | Male | 81 | 0 | 0 | Yes | Private | Urban | 186.21 | 29 | formerly smoked |
| 7 | Male | 74 | 1 | 1 | Yes | Private | Rural | 70.09 | 27.4 | never smoked |
| 8 | Female | 69 | 0 | 0 | No | Private | Urban | 94.39 | 22.8 | never smoked |
| 9 | Female | 78 | 0 | 0 | Yes | Private | Urban | 58.57 | 24.2 | Unknown |
| 10 | Female | 81 | 1 | 0 | Yes | Private | Rural | 80.43 | 29.7 | never smoked |
| 11 | Female | 61 | 0 | 1 | Yes | Govt_job | Rural | 120.46 | 36.8 | smokes |
| 12 | Female | 54 | 0 | 0 | Yes | Private | Urban | 104.51 | 27.3 | smokes |
| 13 | Female | 79 | 0 | 1 | Yes | Private | Urban | 214.09 | 28.2 | never smoked |
| 14 | Female | 50 | 1 | 0 | Yes | Self-employe | Rural | 167.41 | 30.9 | never smoked |
| 15 | Male | 64 | 0 | 1 | Yes | Private | Urban | 191.61 | 37.5 | smokes |
| 16 | Male | 75 | 1 | 0 | Yes | Private | Urban | 221.29 | 25.8 | smokes |
| 17 | Female | 60 | 0 | 0 | No | Private | Urban | 89.22 | 37.8 | never smoked |
| 18 | Female | 71 | 0 | 0 | Yes | Govt_job | Rural | 193.94 | 22.4 | smokes |
| 19 | Female | 52 | 1 | 0 | Yes | Self-employe | Urban | 233.29 | 48.9 | never smoked |
| 20 | Female | 79 | 0 | 0 | Yes | Self-employe | Urban | 228.7 | 26.6 | never smoked |
| 21 | Male | 82 | 0 | 1 | Yes | Private | Rural | 208.3 | 32.5 | Unknown |

**Figure: Dataset of the Project**

By depending on the attributes of this dataset we have to predict that persons smoking status prediction for Brain stroke .The brain stroke prediction for smoking status is based on gender.


**Figure: Visualization on Gender attribute**

In our dataset, the number of male are 2074 and the number of female are 2907. The blue color represents Male and the red color represents Female.

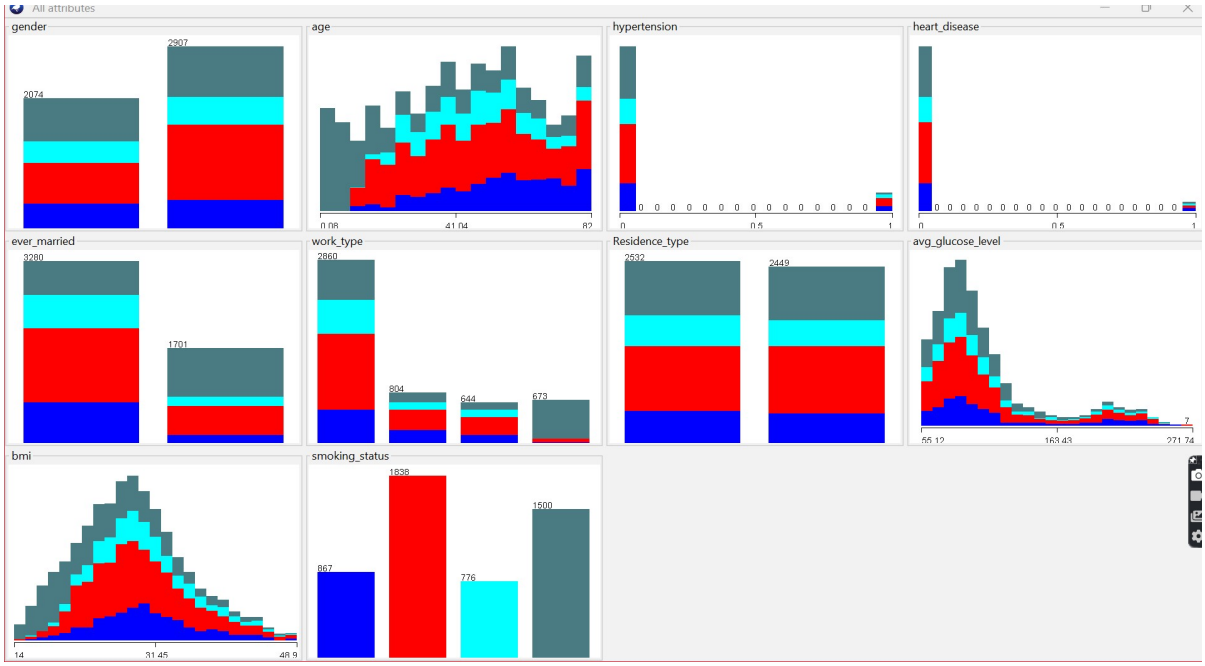



**Figure: Visualization on All attribute**

This is the visualization of all attributes in our dataset.

# Model Development

## Classification

Classification in data mining is a common technique that separates data points into different classes. It allows you to organize data sets of all sorts, including complex and large datasets as well as small and simple ones.

It primarily involves using algorithms that we can easily modify to improve the data quality. This is a big reason why supervised learning is particularly common with classification in techniques in data mining. The primary goal of classification is to connect a variable of interest with the required variables. The variable of interest should be of qualitative type.

The algorithm establishes the link between the variables for prediction. The algorithm we use for classification in data mining is called the classifier, and observations we make through the same are called the instances. There are multiple types of classification algorithms like-

- ❖ Decision trees,
- ❖ K-NN,
- ❖ Naive Bayes,
- ❖ Random Forest,
- ❖ Linear regression

Each algorithm has its unique functionality and application. All of those algorithms are used to extract data from a dataset. Which application we use for a particular task depends on the goal of the task and the kind of data you we to extract.

## Applied Algorithms with Results

To building our project we applied three classification algorithms. They are-

1. **Naive Bayes Algorithm:** Naive Bayes does not employ rules, a decision tree, or any other explicit representation of the classifier. It relies on all attributes being categorical.It is a fast and uncomplicated classification algorithm.However, the Naive Bayes classifier is notorious for being poor at estimation because it assumes all features are of equal importance, which is not true in most real-world scenarios.

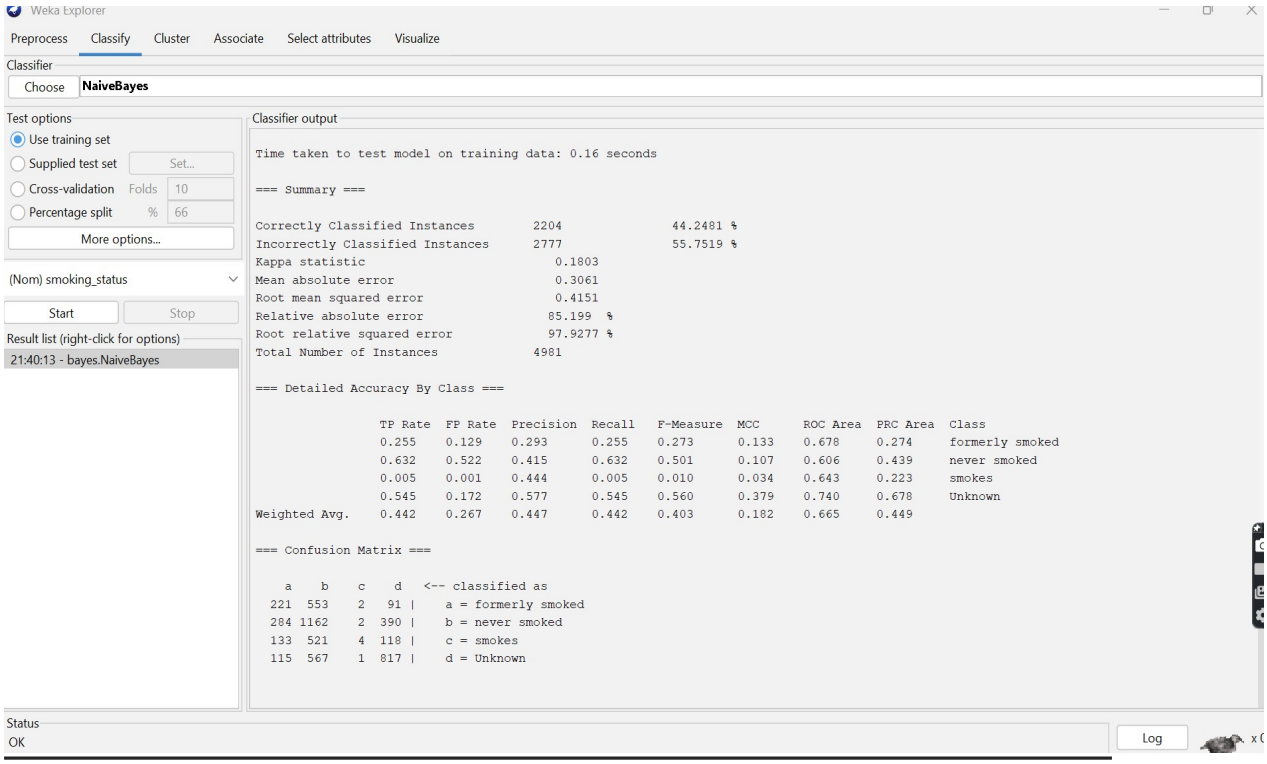



**Figure: Naive Bayes using All Attributes**

There are 4981 instances and 10 attributes in our dataset.Here we use all given attributes to predict our target attribute. In Naive Bayes, with all attributes we get only 2204 instances correctly classified and 2777 instances incorrectly classified. The percentage of accuracy with this dataset is less then 50% which is 44.2481%.

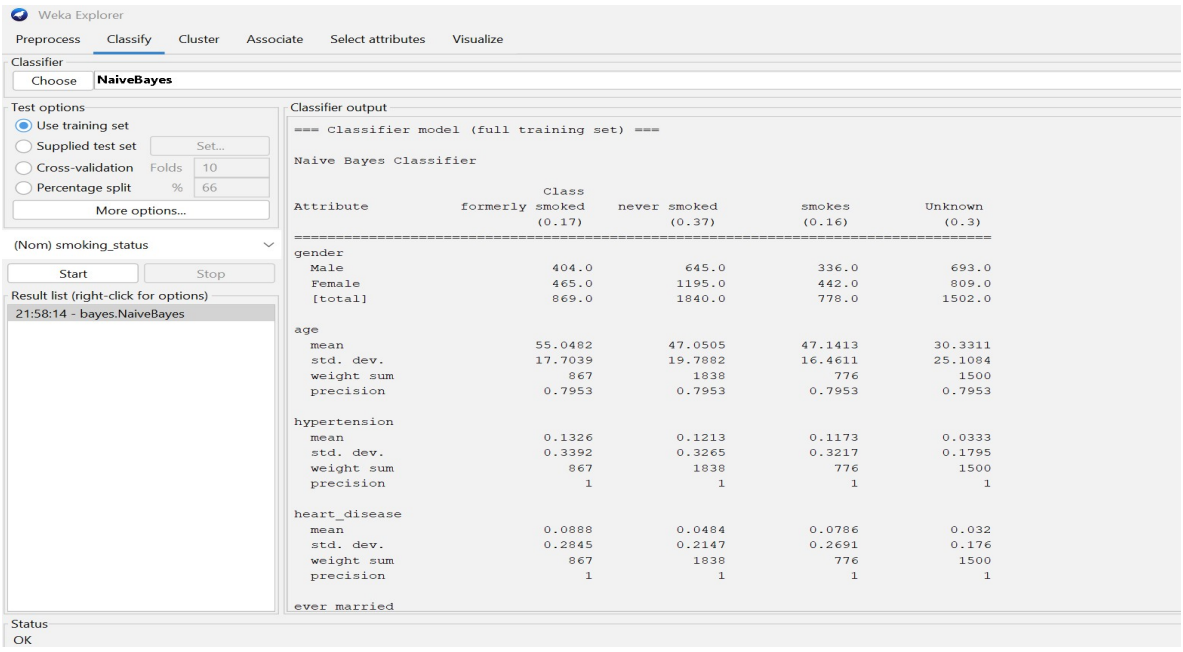



**Figure: Naive Bayes Classifier Model with All Attributes**

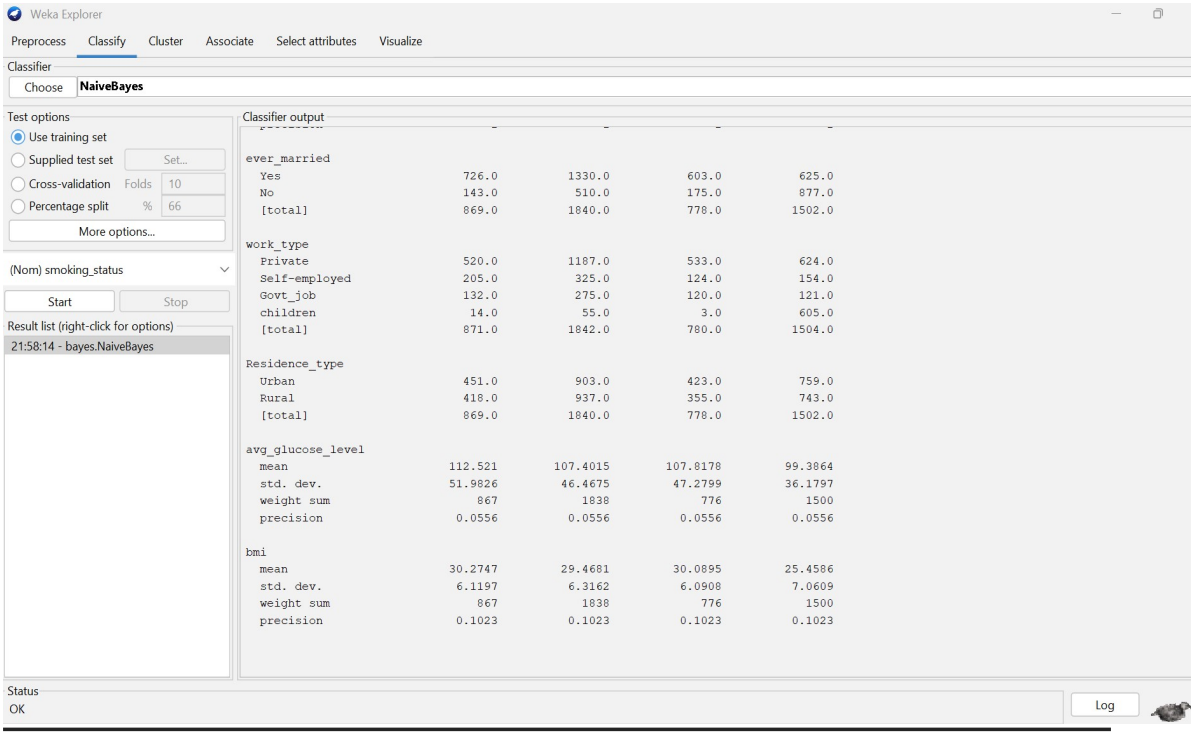



**Figure: Naive Bayes Classifier Model with All Attributes**

The classifier model for Naive Bayes was shown in the two above figures.

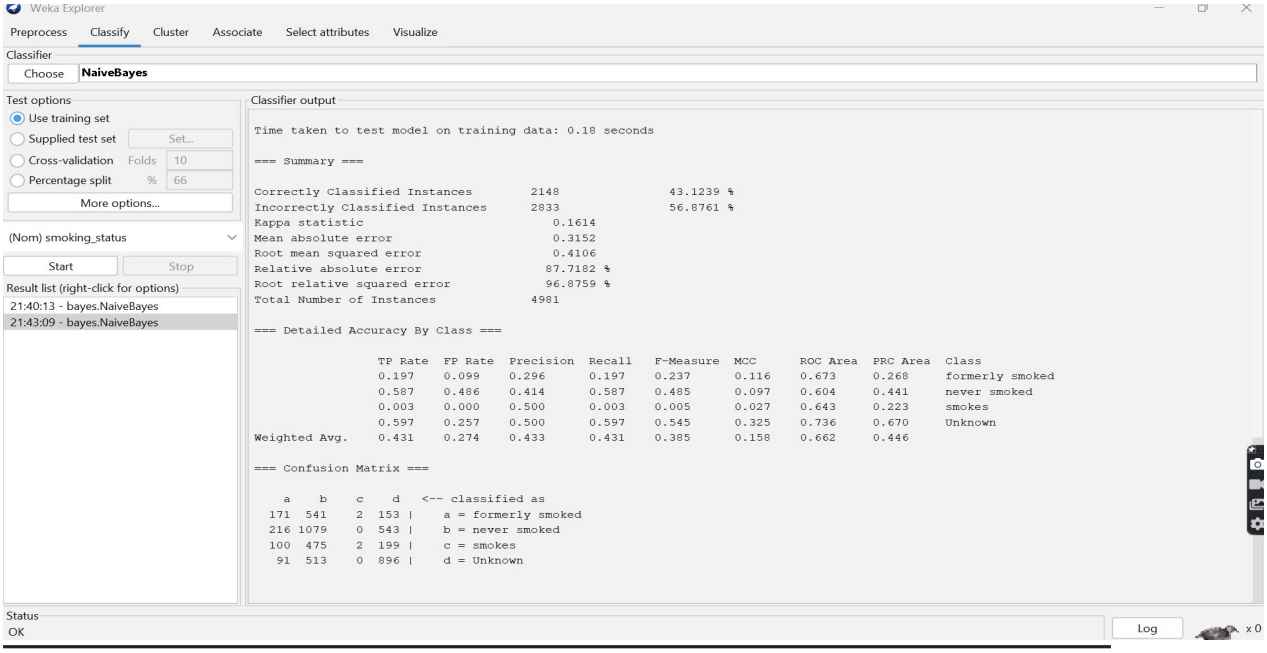



**Figure: Naive Bayes using attributes except(ever_married,work_type,residence_type)**

Here we excluded three attributes which are ever_married,work_type and residence_type for this figure.Using the remaining attributes we get only 2148 instances correctly classified and 2833 instances incorrectly classified.

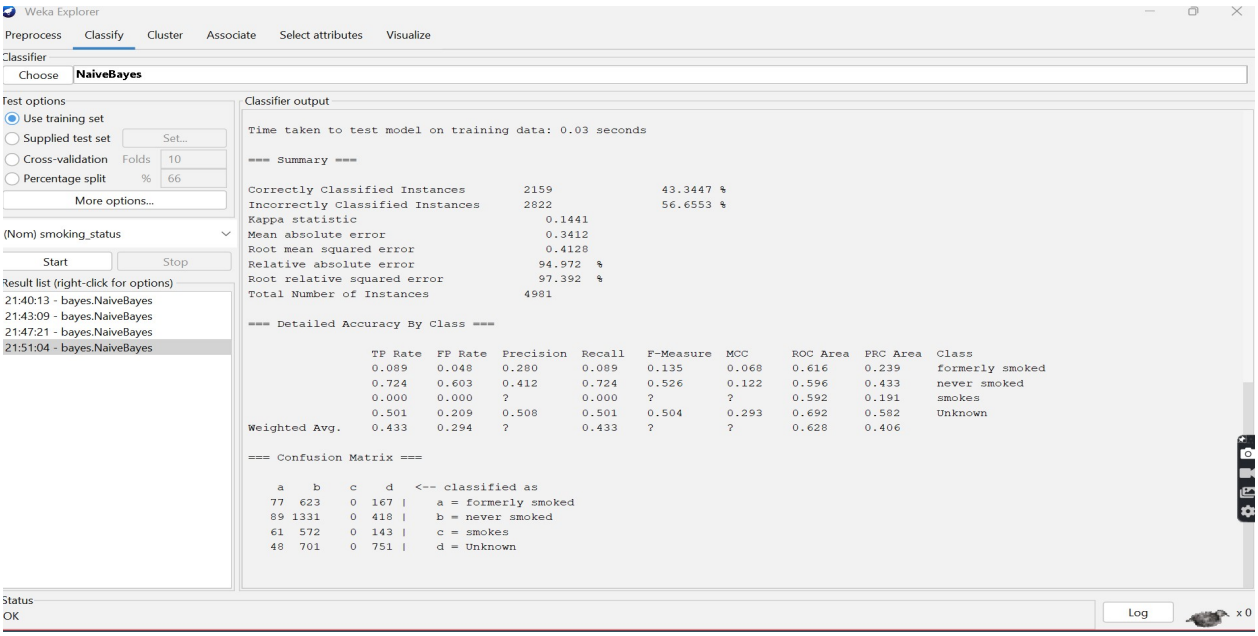



**Figure:Naive Bayes using attributes (gender,heart_disease,bmi,smoking_status)**

Here we included four attributes which are gender,heart_disease,bmi and smoking_status.We get only 2159 instances correctly classified and 2822 instances incorrectly classified.Accuracy of this mode is only 43.3447%.

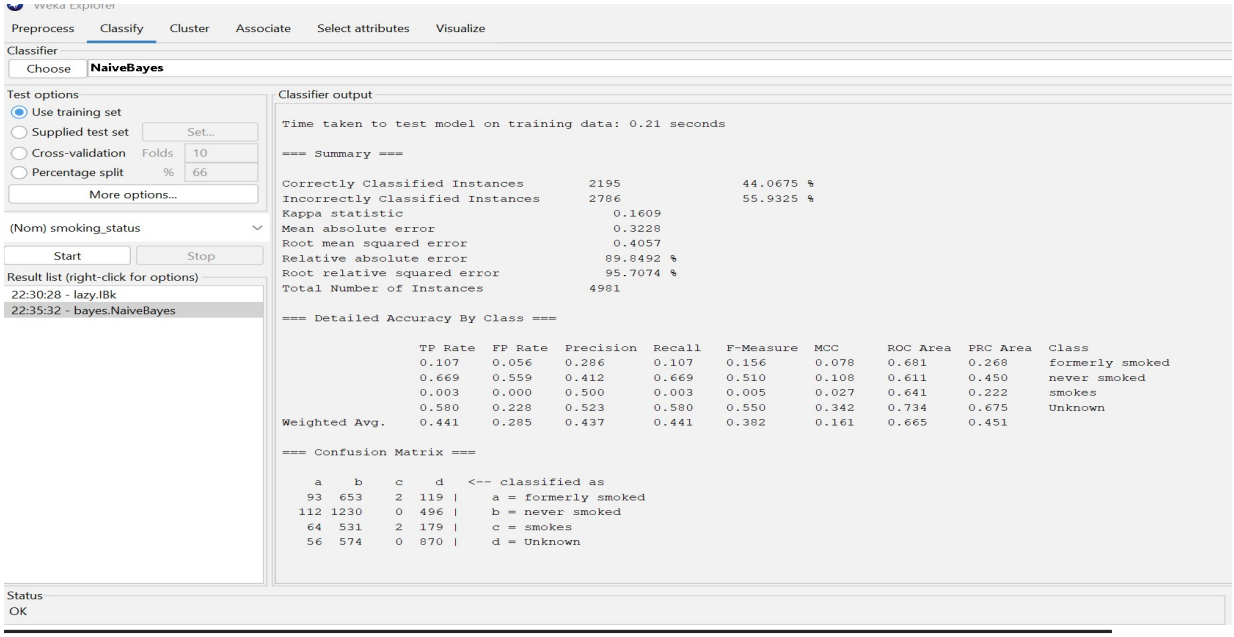



**Figure:Naive Bayes using attributes (gender,age,hypertension,heart_disease,smoking_status)**

Here we included other four attributes which are gender, age, hypertension, heart_disease and smoking_status.We get only 2195 instances correctly classified and 2786 instances incorrectly classified.Accuracy of this mode is only 44.0675%.

2. **<u>K- Nearest Neighbour Algorithm(K-NN):</u>**Nearest Neighbour classification is mainly used when all attribute values are continuous, although it can be modified to deal with categorical attributes.It uses 'feature similarity' to predict the values of new data points which further means that the new data point will be assigned a value based on how closely it matches the points in the training set.It's quite an expensive algorithm as finding the value of k takes a lot of resources. Moreover, it also has to calculate the distance of every instance to every training sample, which further enhances its computing cost.

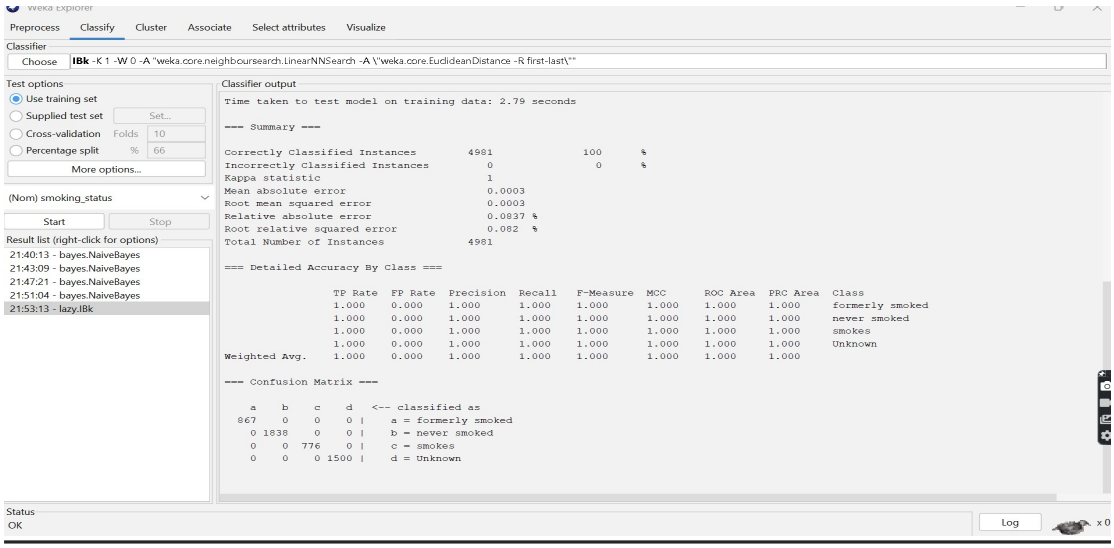



**Figure:K-NN using All Attributes**

There are 4981 instances and 10 attributes in our dataset.Here we use all given attributes to predict our target attribute. By using K-NN algorithm with all attributes we get 100% accuracy of correctly classified instances where the value of K=1.It means that all of the instances are correctly classified.

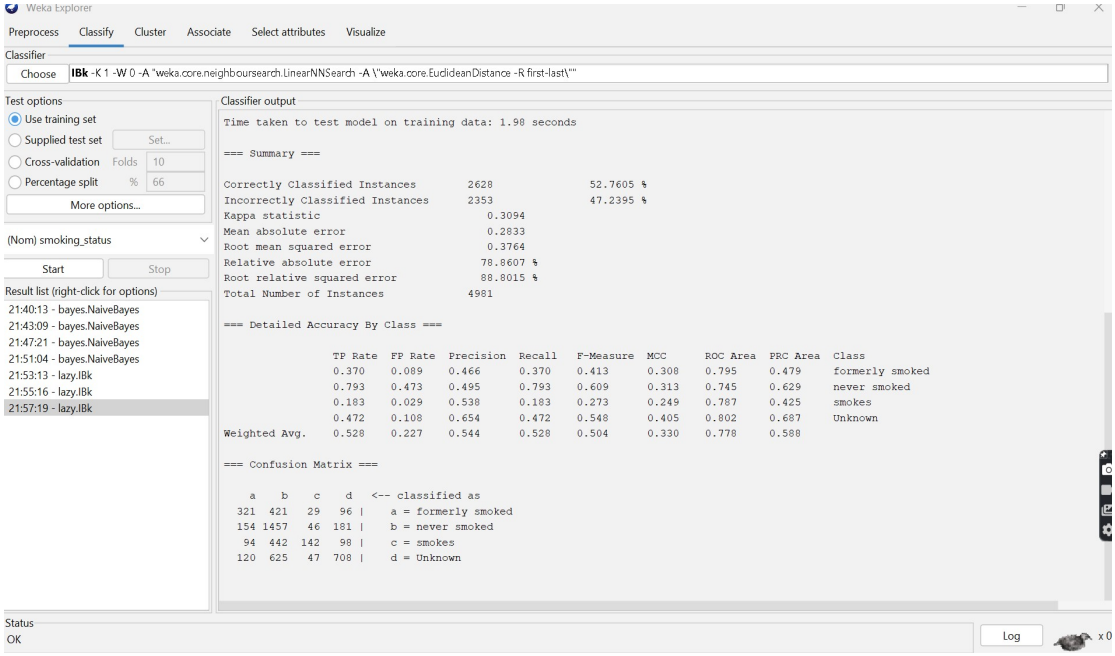



**Figure: K-NN using attributes**
**(gender,heart_diseases,bmi,smoking_status)**

In this figure we included four attributes which are gender,heart_diseases,bmi, and smoking_status.The percentage of correctly classified instances is more then 50% which is 52.7650%.

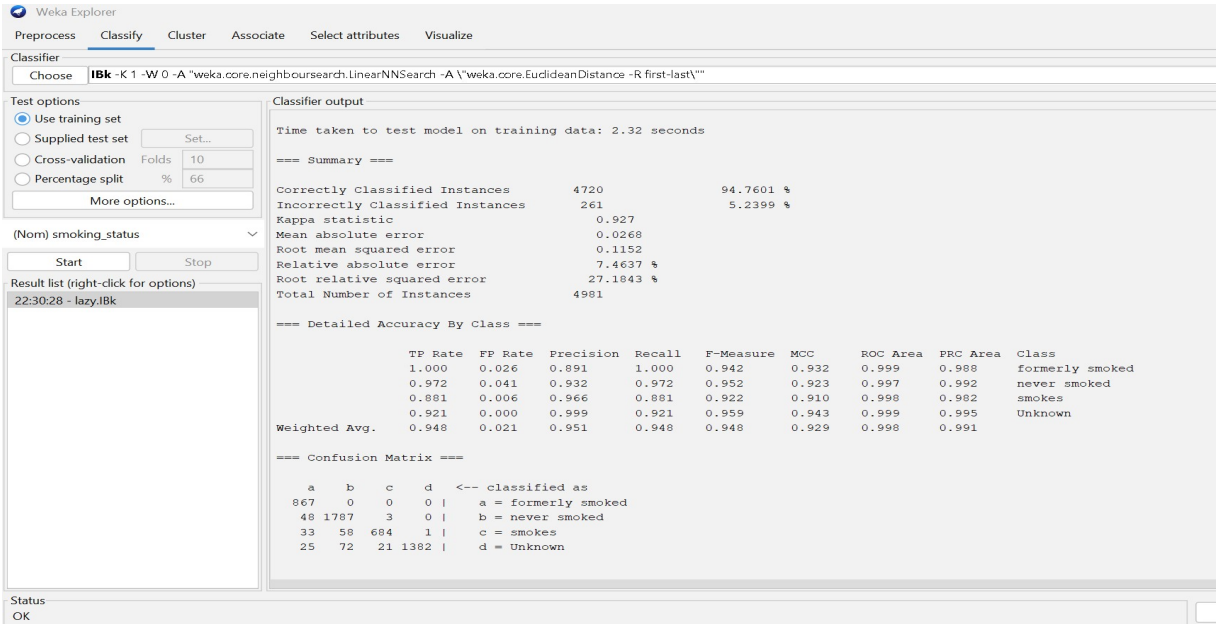



**Figure: K-NN using attributes except(gender,age,bmi)**

Here we excluded four attributes which are geder,age and bmi. The accuracy rate of this mode is 94.7601%.

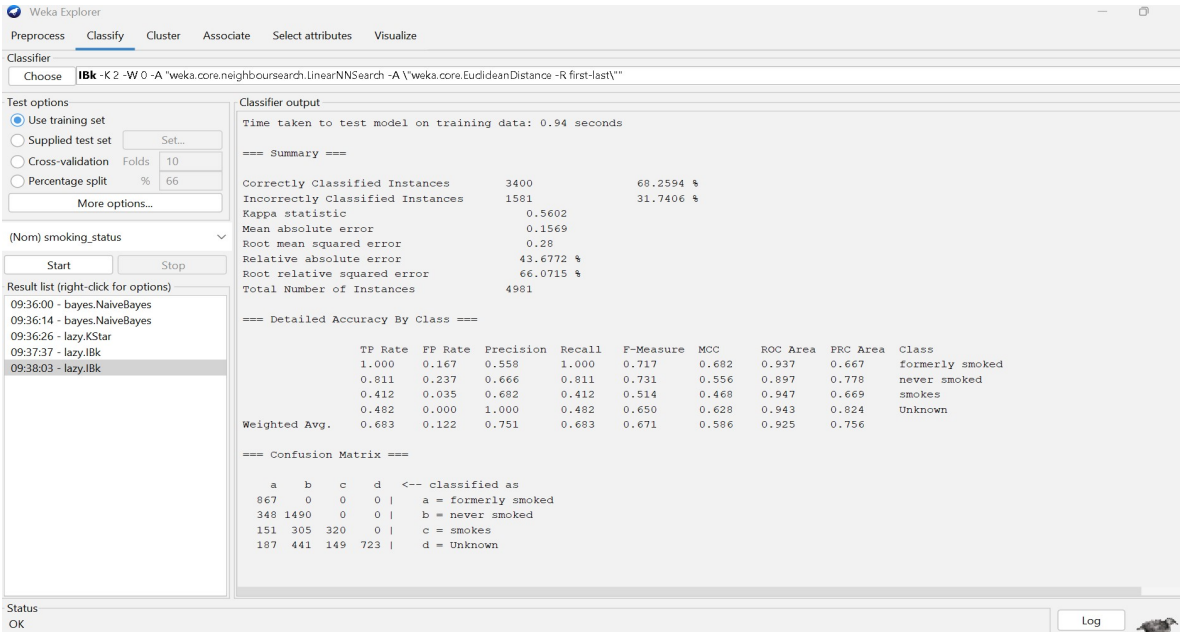



**Figure: K-NN using attributes except(age,ever_married,ave_glucose_level)**

Here the attributes age,ever_married and ave_glucose_level were excluded. The value of k for this figure is 2.We get 3400 instances correctly classified and 1581 instances incorrectly classified.Accuracy of this mode is only 68.2594%.

3. **Decision Trees:**A decision tree is a tree-like graph with nodes representing the place where we pick an attribute and ask a question, edges represent the answers to the question and the leaves represent the actual output or class label. Decision trees classify the examples by sorting them down the tree from the root to some leaf node, with the leaf node providing the classification to the example. Each node in the tree acts as a test case for some attribute, and each edge descending from that node corresponds to one of the possible answers to the test case. This process is recursive in nature and is repeated for every sub-tree rooted at the new nodes.It would predict which classes a new data point would belong to according to the created decision tree. Its prediction boundaries are vertical and horizontal lines.

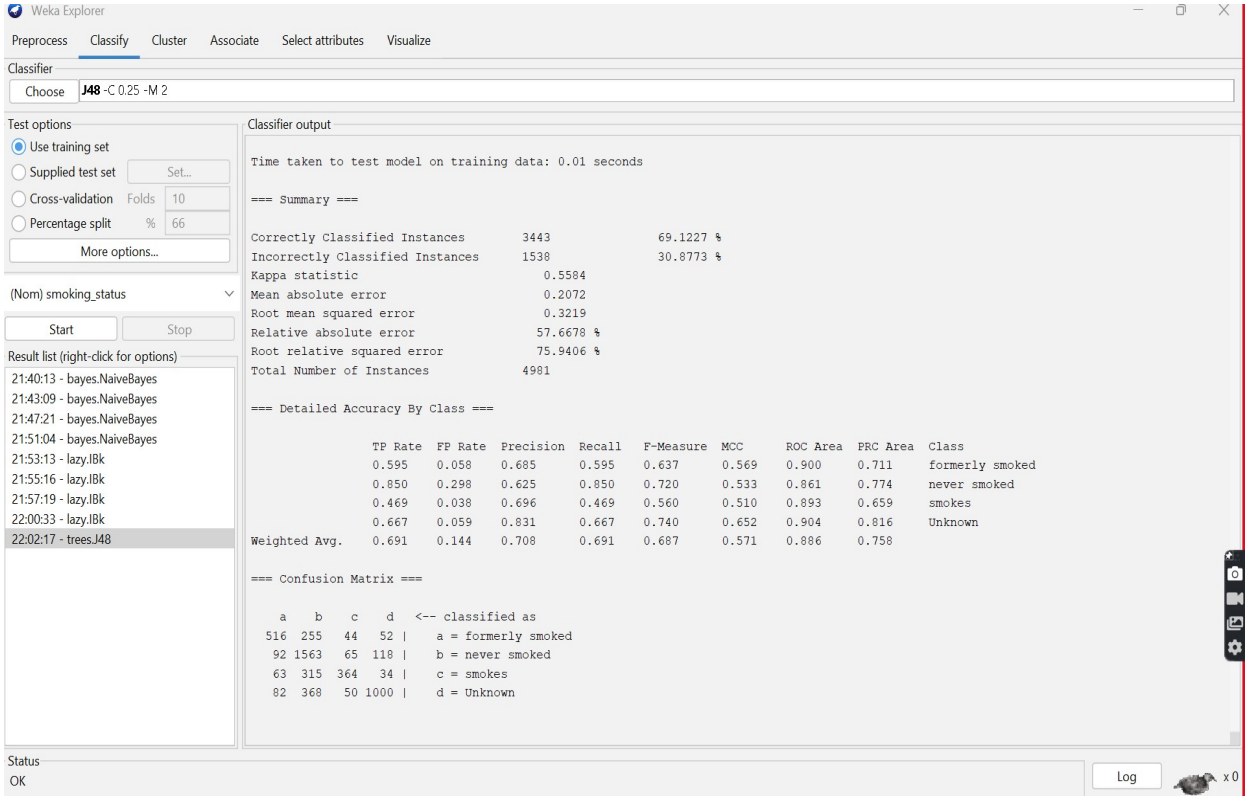



**Figure:Decision Tree using All Attributes**

Here we use all given attributes to predict our target attribute. By using Decision tree algorithm with all attributes we get 69.1227% accuracy of correctly classified instances.

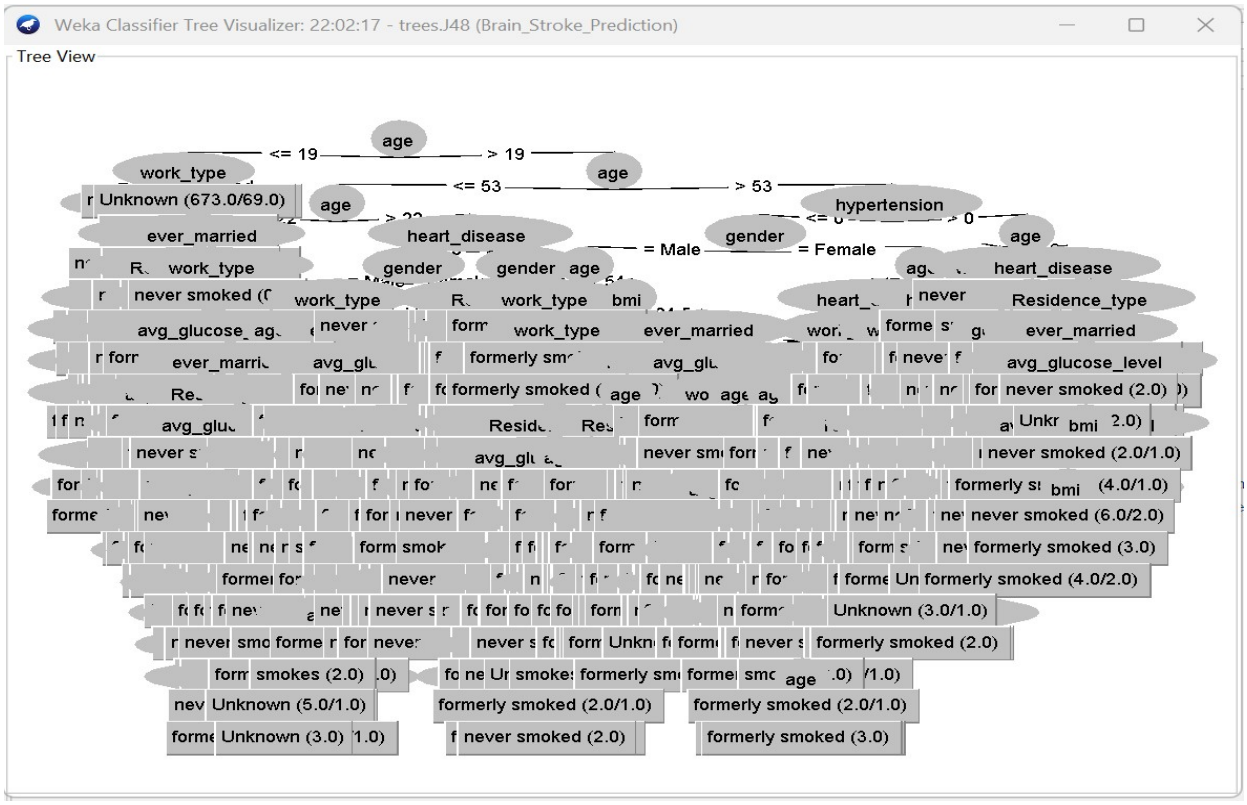



**Figure: Tree view with All Attributes**

This figure shows us the decision tree view of our dataset using all attributes.

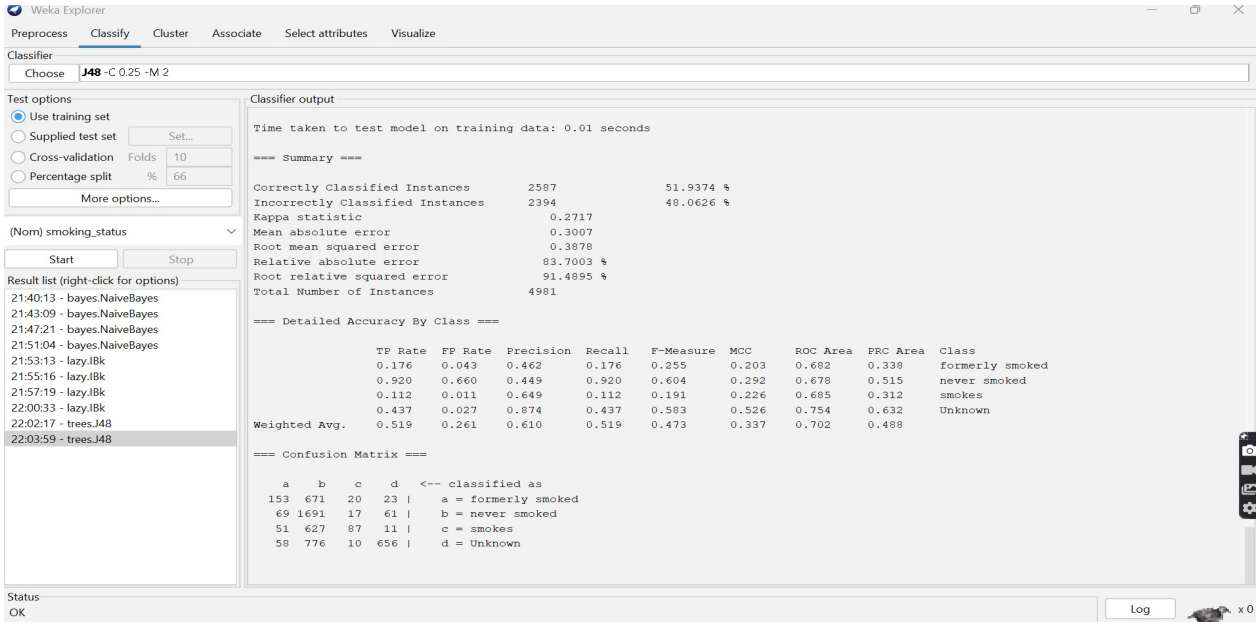



**Figure: Decision Tree using attributes except(age,hypertension,ever_married,)**

Here we uses seven attributes out of ten attributes. The excluded attributes are age, hypertension and ever_married. The accuracy mode is a little more then 50% for this figure which is 51.9374.

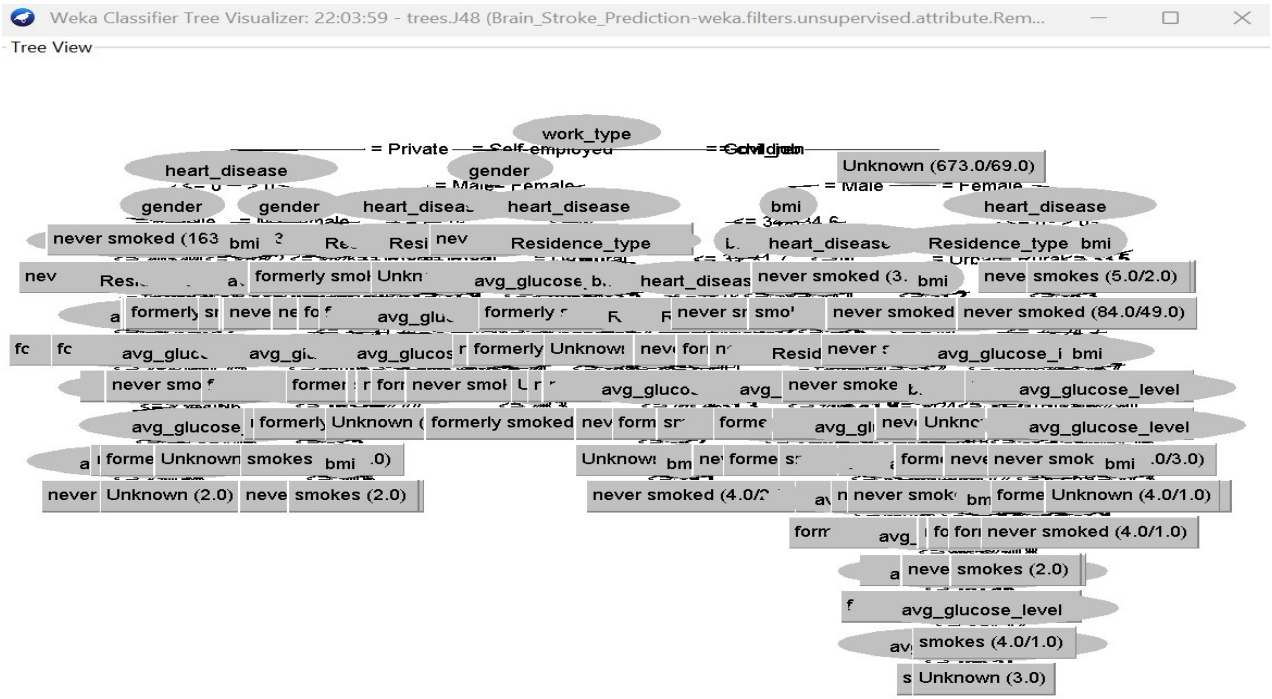



**Figure: Tree view using attributes except(age,hypertension,ever_married,)**

This figure shows us the decision tree view of our dataset with excluding age, hypertension and ever_married attributes.

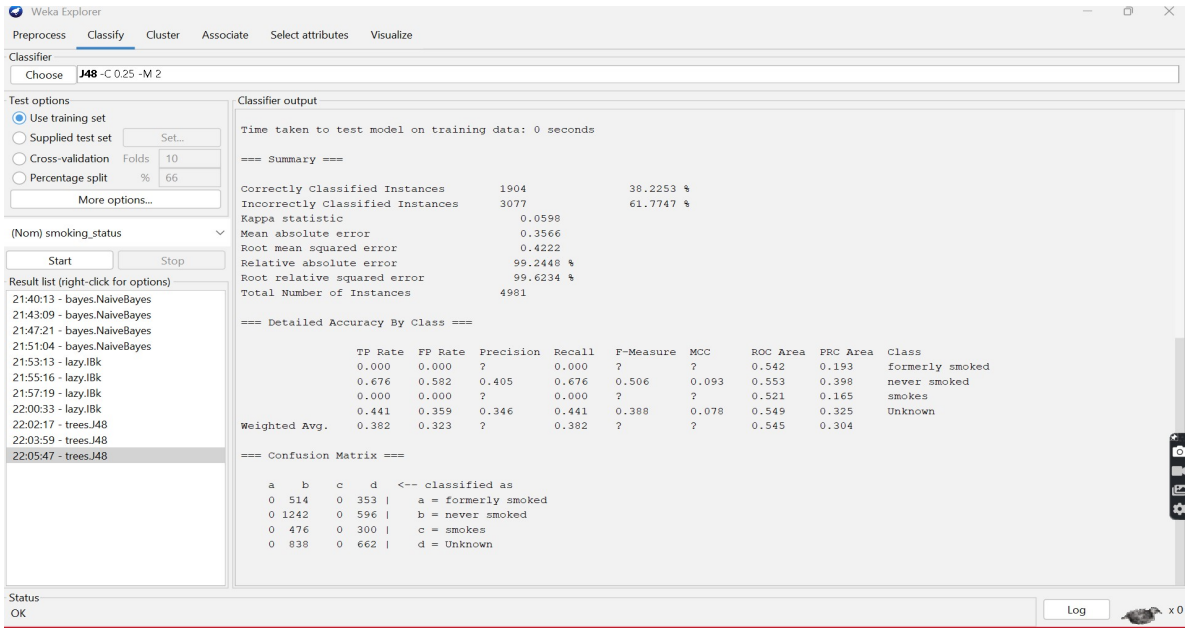



**Figure: Decision Tree using attributes (gender,heart_diseases,smoking_status)**

We only use three attributes for this figure. The amount of correctly classified instances are 1904 and incorrectly classified instances are 3077.

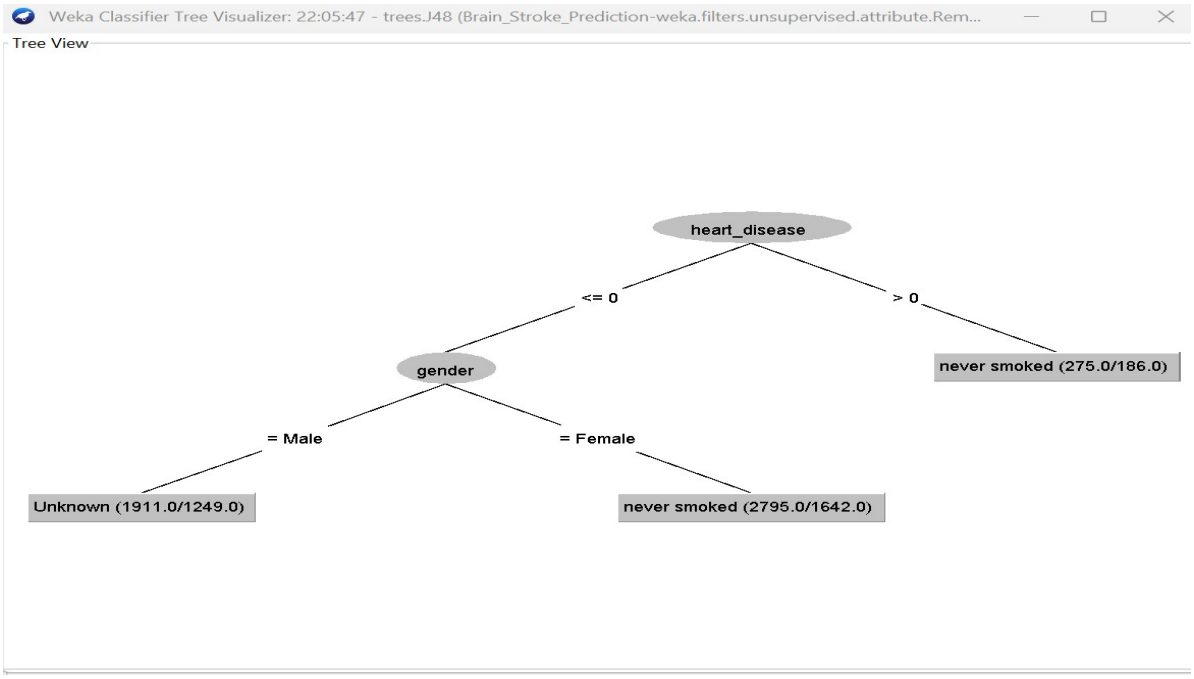



**Figure: Tree view using attributes (gender,heart_diseases,smoking_status)**

This decision tree was visualized using only three attributes to get our target attribute which are gender, heart_diseases and smoking_status.

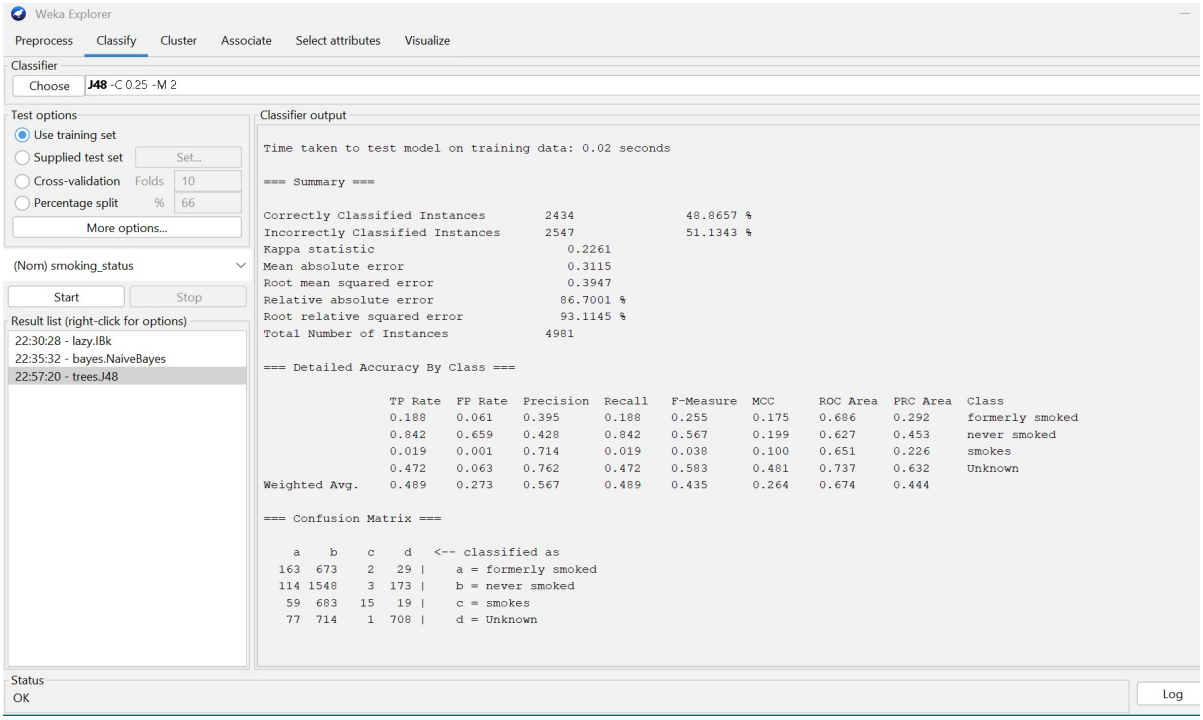



**Figure:Decision Tree using attributes (gender,age,hypertension,heart_diseases,smoking_status)**

We get accuracy of 48.8657% using six attributes which are gender,age,hypertension,heart_diseases,smoking_status. It means that the amount incorrectly classified instances are higher then the correctly classified instances.

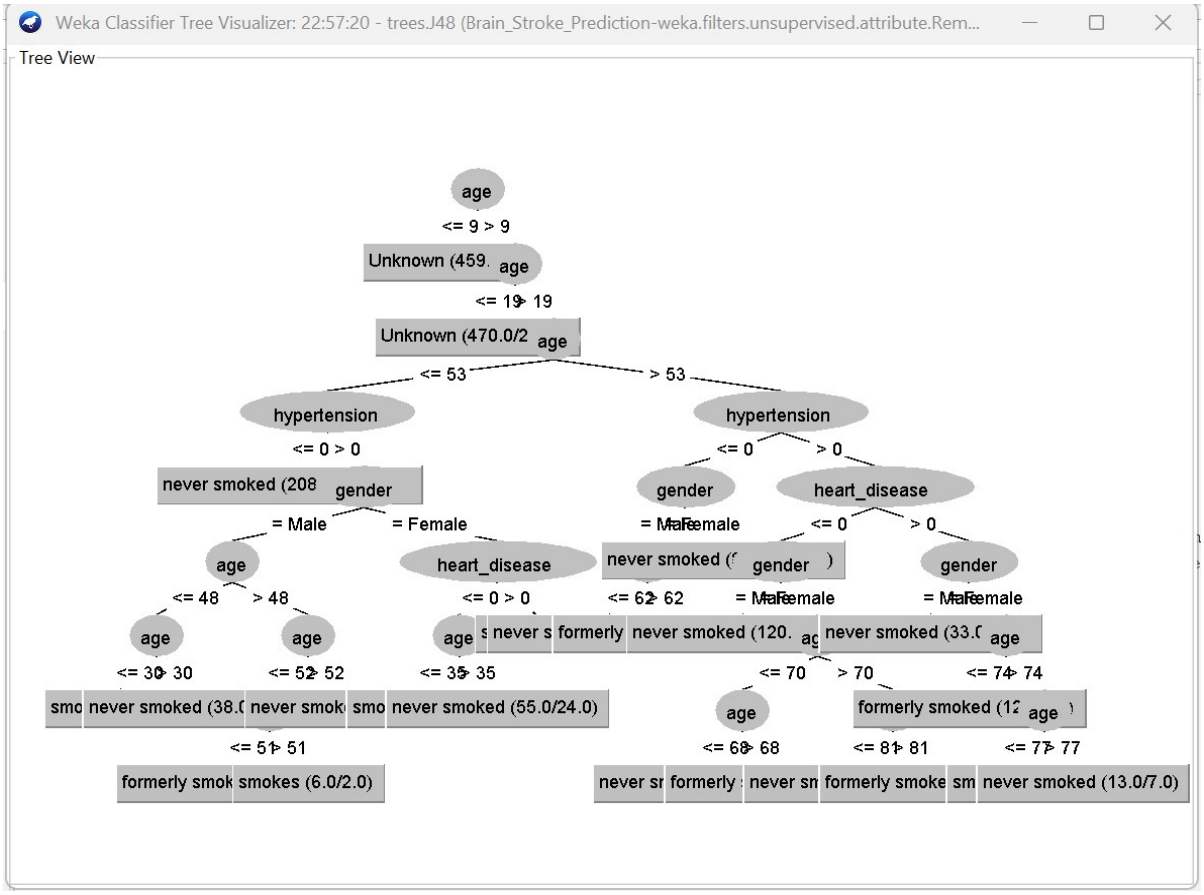



**Figure: Tree view using attributes (gender,age,hypertension,heart_diseases,smoking_status)**

This decision tree was visualized excluding only four attributes to get our target attribute which are residence_type, work_type, ave_glucose_level and bmi.

# Plotting of Applied Algorithms

## Naive Bayes Plotting:

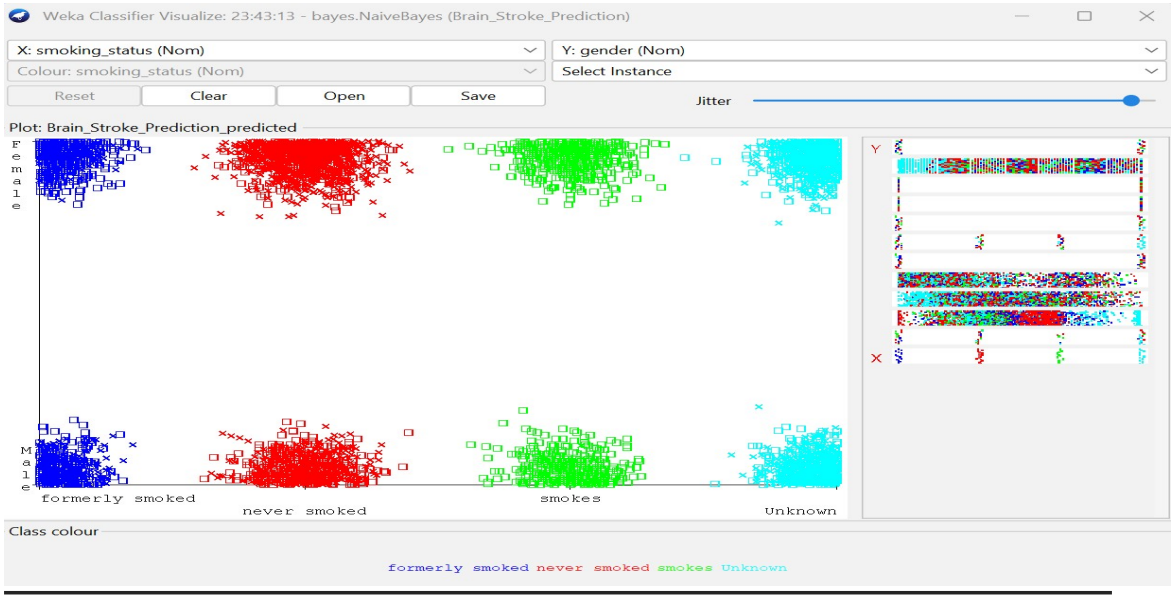



**Figure: Naive Bayes Plotting(Smoking status VS Gender)**

For this plotting the X-axis represents smoking_status and Y-axis represents gender. This plotting shows that the higher rate of female are in class of never smoked.


**Figure: Naive Bayes Plotting (Smoking status VS Age)**

For this plotting the X-axis represents smoking_status and Y-axis represents age. This plotting shows that the higher rate of formerly smoked is in between the age of 41 to 82.


**Figure: Naive Bayes Plotting (Smoking status VS ave_glucose_level)**

For this plotting the X-axis represents smoking_status and Y-axis represents ave_glucose level. The ave_glucose_level between 55.22 to 163.43 is higher for all classes.


**Figure: Naive Bayes Plotting (Smoking status VS residence_type)**

For this plotting the X-axis represents smoking_status and Y-axis residence_type.

## K- Nearest Neighbour Algorithm(K-NN) Plotting:

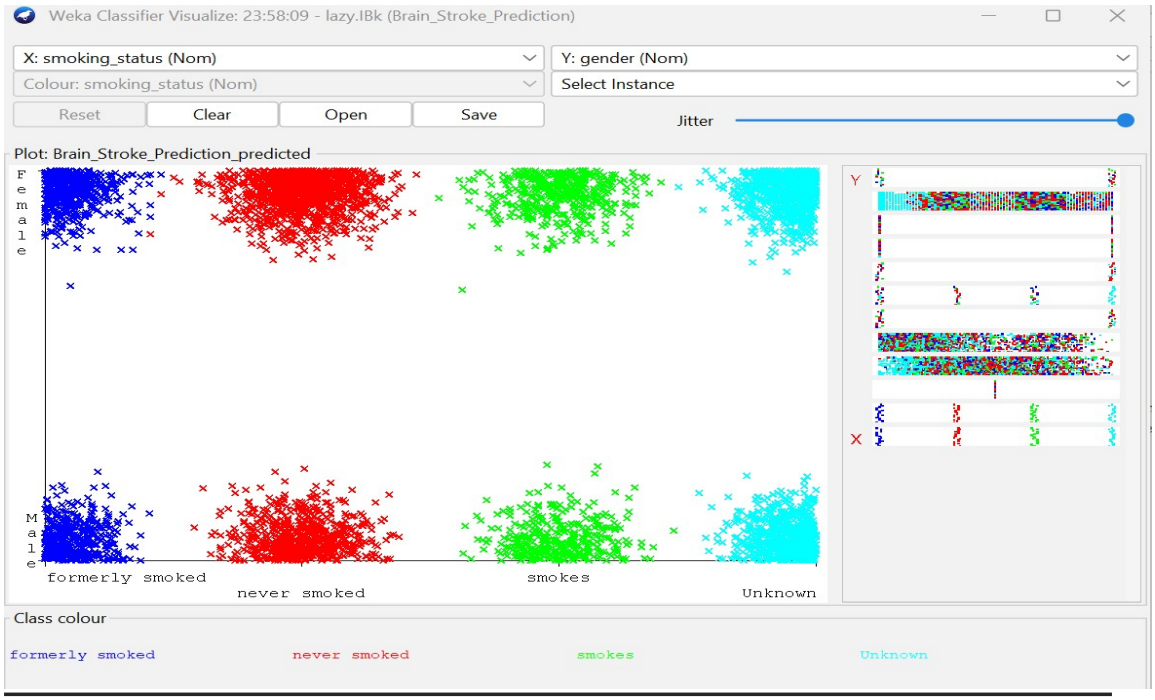



**Figure: K-NN (smoking_status VS gender)**

For this plotting the X-axis represents smoking_status and Y-axis gender. The class formerly smoked of male is less then female.


**Figure: K-NN Plotting (smoking_status VS residence_type)**

For this plotting the X-axis represents smoking_status and Y-axis residence_type. The smoking status of urban people is greater then the rural people.


**Figure: K-NN plotting (smoking_status VS bmi)**

For this plotting the X-axis represents smoking_status and Y-axis bmi. The smoking status for this plot is higher for all classes where bmi is in between 14 to 31.45.


**Figure: K-NN Plotting (smoking_statusVS predicted smoking_status)**

For this plotting the X-axis represents smoking_status and Y-axis represents predicted smoking_status.The higher amount of predicted smoking status is belong to the class of never smoked.

## Decision Tree Plotting

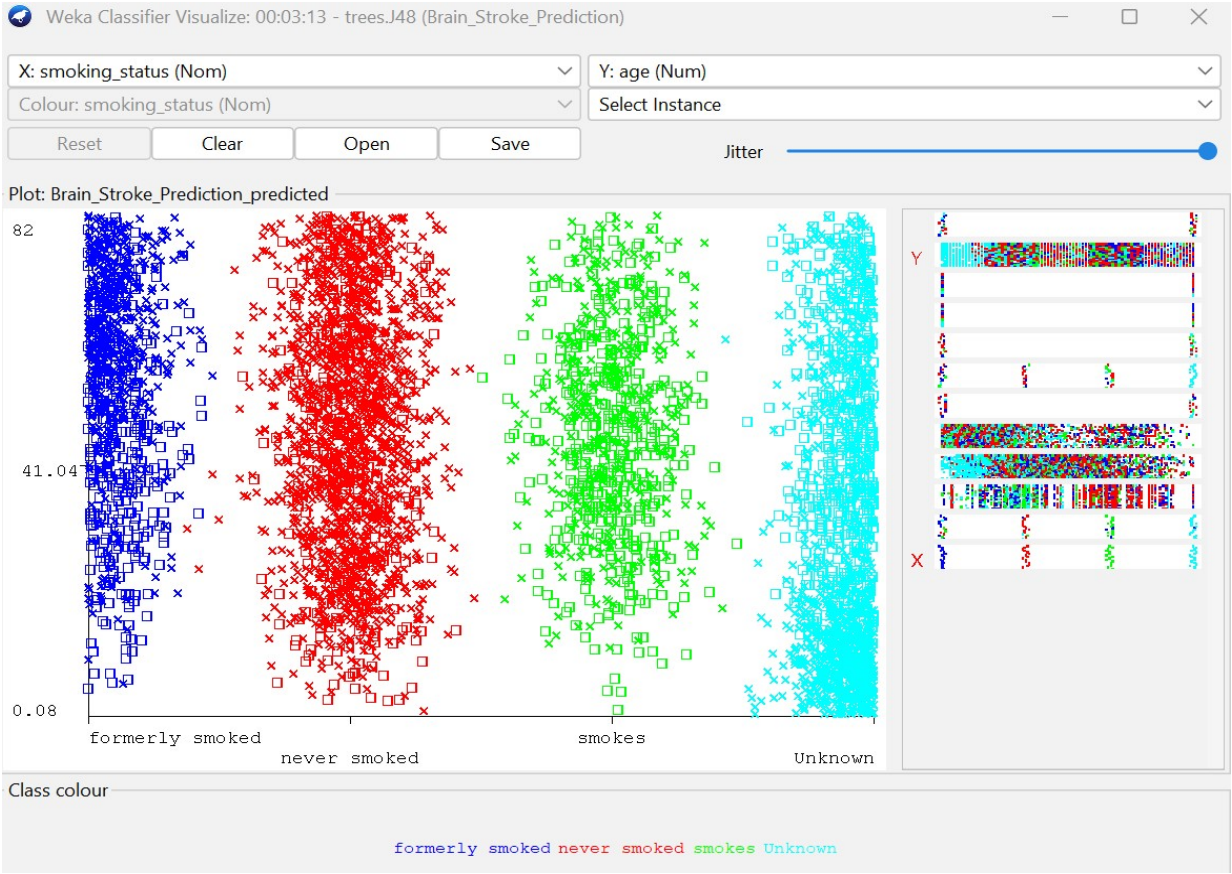



**Figure: Decision tree Potting(smoking_status VS age)**

For this plotting the X-axis represents smoking_status and Y-axis represents age.This plotting shows that the lower rate of all classes is in between the age of 0.08 to 41.04.


**Figure: Decision tree Potting(smoking_status VS heart_diseases)**

For this plotting the X-axis represents smoking_status and Y-axis represents heart_diseases.For all classes a few number of people have heart diseases.


**Figure: Decision tree Potting(smoking_status VS hypertension)**

For this plotting the X-axis represents smoking_status and Y-axis hypertension. This plotting shows us that maximum number of people does not have hypertension.


**Figure: Decision tree Potting(smoking_status VS predicted_smoking_status)**

For this plotting the X-axis represents smoking_status and Y-axis represents predicted smoking_status. This shows us that the higher amount for predicted smoking status is in class of never smoked.

## Discussion

We selected a dataset for our project with 4982 instances and 10 attributes. After applying three algorithms with all attributes we get,

**Table-1: Data table accuracy rate using All Attributes**

| Applied Algorithm | Correctly Classified Instances | Percentage Of Accuracy | Incorrectly Classified Instances | Percentage Of Inaccuracy |
|---|---|---|---|---|
| Naive Bayes | 2204 | 44.2481% | 2777 | 55.7519% |
| K-Nearest Neighbour (K-NN) | 4981 | 100% | 0 | 0% |
| Decision Tree | 3443 | 69.1227% | 1538 | 30.8773% |

From the data table we can see that the K-Nearest Neighbour(K-NN) is giving 100% accuracy,the decision tree gives about 69.1227% of accuracy and the naive bayes gives us 44.2481% accuracy. Overall, the K-NN is more accurate then any other algorithoms where we get 100% accuracy which means all instances are correctly classified.K-NN is used when all the attributes are continuous. It can also be altered to deal with the categorical attribute as well. The K-NN algorithm uses 'feature similarity' to predict the values of any new data points.We conducted this project excluding different attributes to get a better understanding which algorithm is better.From above we see that the Naive Bayes algorithm gives us percentage of accuracy between 43.1239% to 44.2481%.The K-NN algorithm gives us percentage of accuracy between 52.7605% to 100%.On the other hand, Decision Tree gives us percentage of accuracy between 38.2253% to 69.1227%.We can say that the accuracy rate of K-NN algorithm is higher then other two algorithms. Finally we can come to the end point that our prediction accuracy in K-NN algorithm is best for our project dataset.

# Conclusion

After evaluating all three algorithms, we determined that K-NN achieves the maximum accuracy in measuring the investigated dataset. We achieved our goal by applying the algorithms to predict smoking status for brain stroke. We have to test the accuracy of our prediction by varying the values in our dataset with Nave Bias, decision tree, and K-NN. The accuracy of our forecast with K-NN matched the majority of the cases and was the most accurate of the three algorithms. We achieved our goal and had an excellent solution of predicting the smoking status of brain stroke by using the Weka tool, datasets, and testing alternative values with the algorithms.