# DATA WAREHOSING & DATA MINING

**SEMESTER:** SUMMER 2021-2022

**FINAL PROJECT**

**SECTION: B GROUP:**

**SUBMITTED BY:**


| STUDENT NAME | STUDENT ID |
|---|---|
| MAYESHA ABDULLAH | 19-39911-1 |
| NILOY KUNDU ADHAR | 19-39883-1 |
| NILA MAHAZABEN TURNA | 19-39964-1 |
| MD. ISHMAM MIR | 19-39925-1 |


**SUBMITTED TO:**

**TOHEDUL ISLAM**

**INTRODUCTION:**

Data mining is the process of discovering patterns and other valuable information and knowledges from large amount of data. It is also known as knowledge discovery in data (KDD). Data mining is used in many areas of research and business, including healthcare, education, sales and marketing, product development, etc. It is a computer science and statistics multidisciplinary topic with the general purpose of extracting information from data collection and transforming the information into an accessible structure for subsequent use. KNN, Naive Bayes, and Decision Tree are some of the classification algorithms used in data mining. We have chosen the "Heart failure prediction dataset" to classify the type by using two different classifiers and find the best-suited classifier for the dataset. Among the classifiers we have chosen Naive Bayes and Decision Tree(J48) to classify the dataset.

**Information about the dataset:** In this report, we used "Heart failure prediction dataset", it is a CSV dataset file which was collected from **Kaggle.com,** was used to predict the outcome of the heart failure that might be accurate for the patient according to their health condition.

The targeted feature is:

- HeartDisease

The other feature sets are:

- Age
- Sex
- ChestPainType
- RestingBP
- Cholesterol
- FastingBS
- RestingECG
- MaxHR
- ExerciseAngina
- Oldpeak
- ST_Slope

There are total 918 instances among of these 12 attributes and all these instances were used for classification. Here are the graphical details of the attributes:

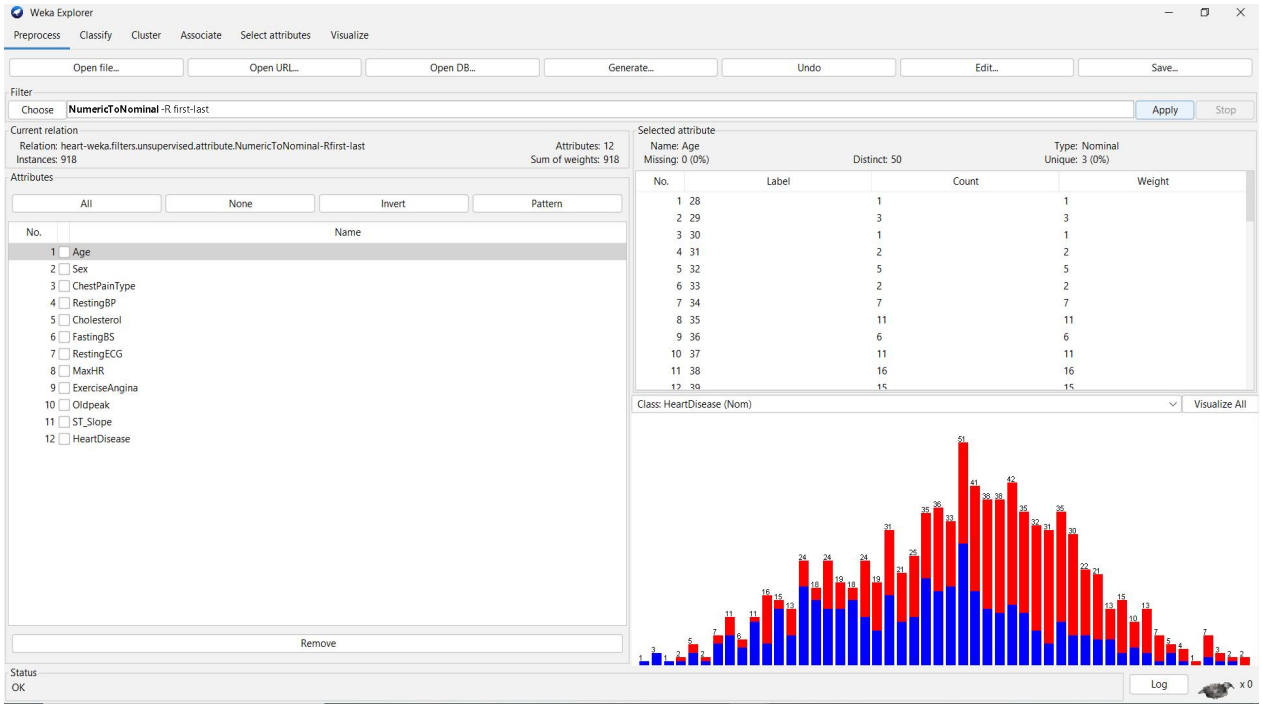



*Figure 1: Selected dataset*

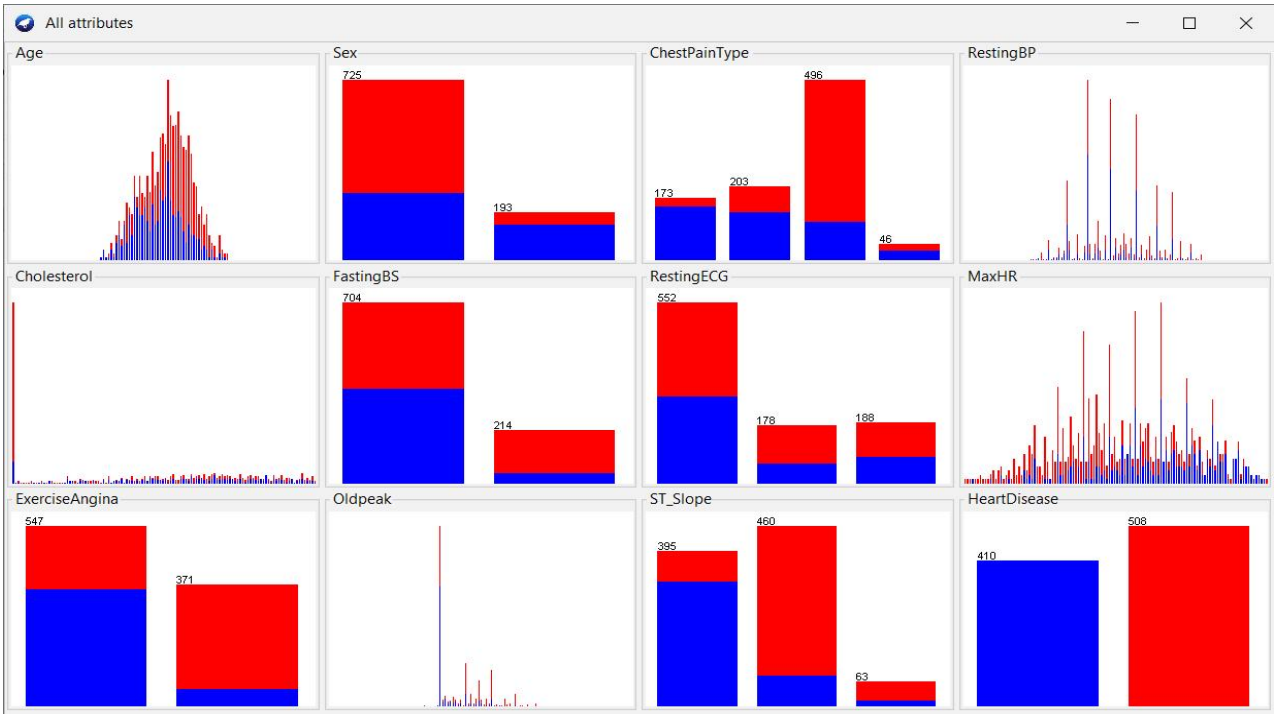



*Figure 2: Graphical representation of all attributes*

**Selected classifiers:** A classifier is a machine learning model that is used to discriminate different objects based on certain features. Two kinds of classifiers have been used with the same data to compare the result. In this process, Naïve Bayes, and Decision Tree(J48) Classifiers were used.

**RESULT OF THE CLASSIFIERS:**

Weka 3.8.6 version software was used to construct the classifier. Weka is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

**Applying Naïve Bayes classifier**: Naïve Bayes algorithm is a supervised learning algorithm, which is based on the Bayes theorem and is used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. While classifying the selected dataset, the NaiveBayes format was selected from the Bayes folder.

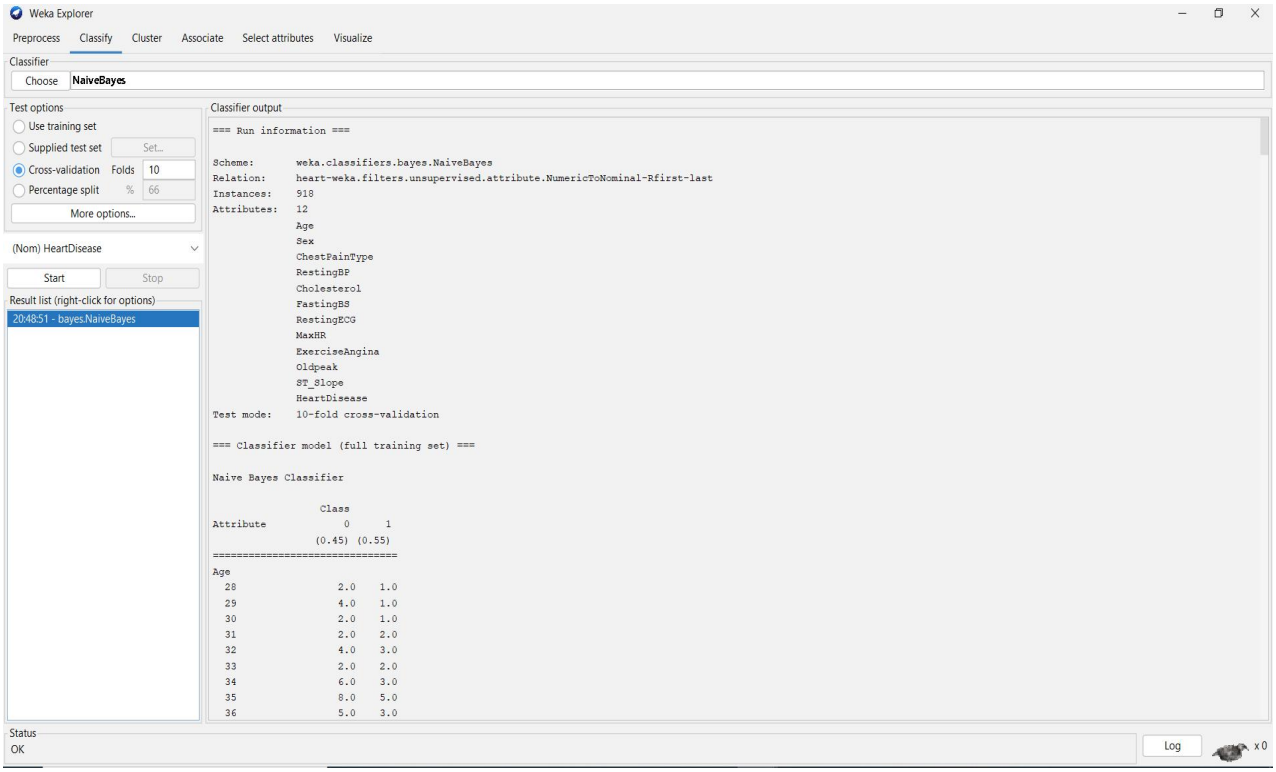



*Figure 3: Naive Bayes Classification*

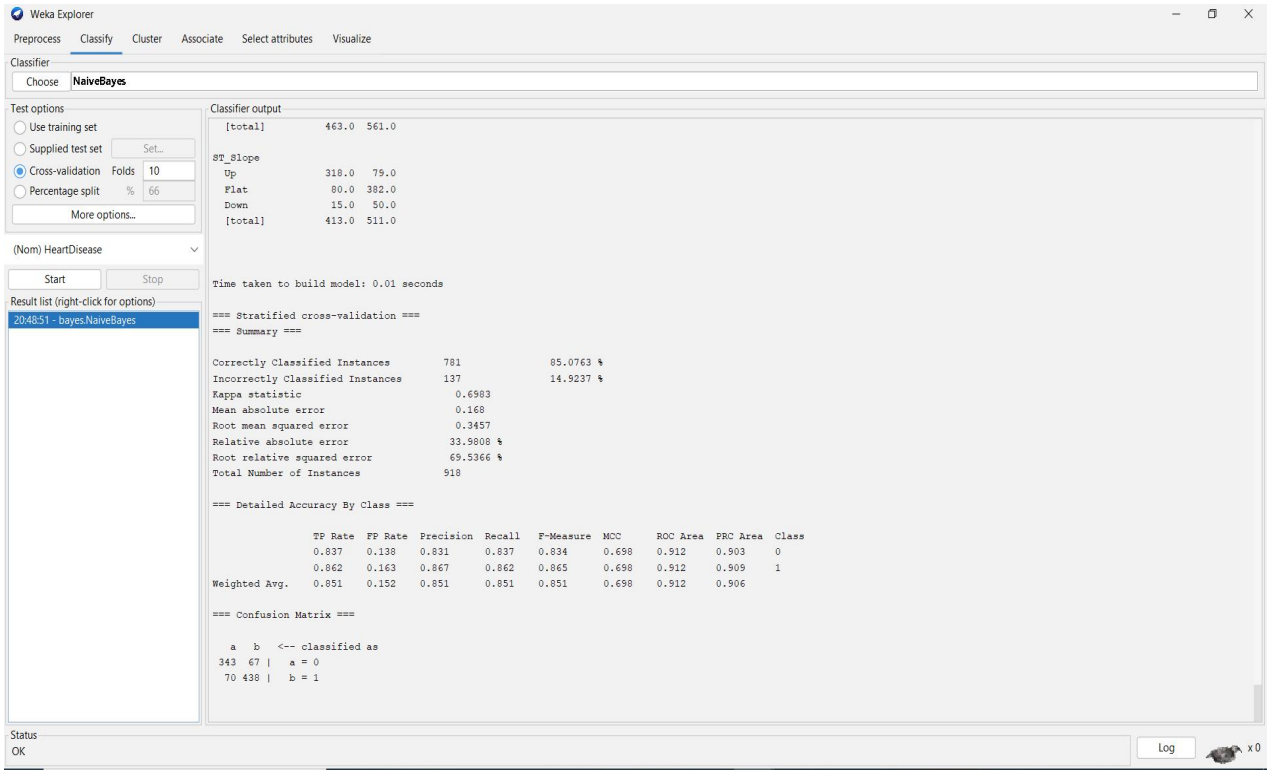



*Figure 4: Naive Bayes Classification*

## Applying decision tree classifier:

A decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. While classifying the selected dataset, the J48 format was selected for Decision tree classification.

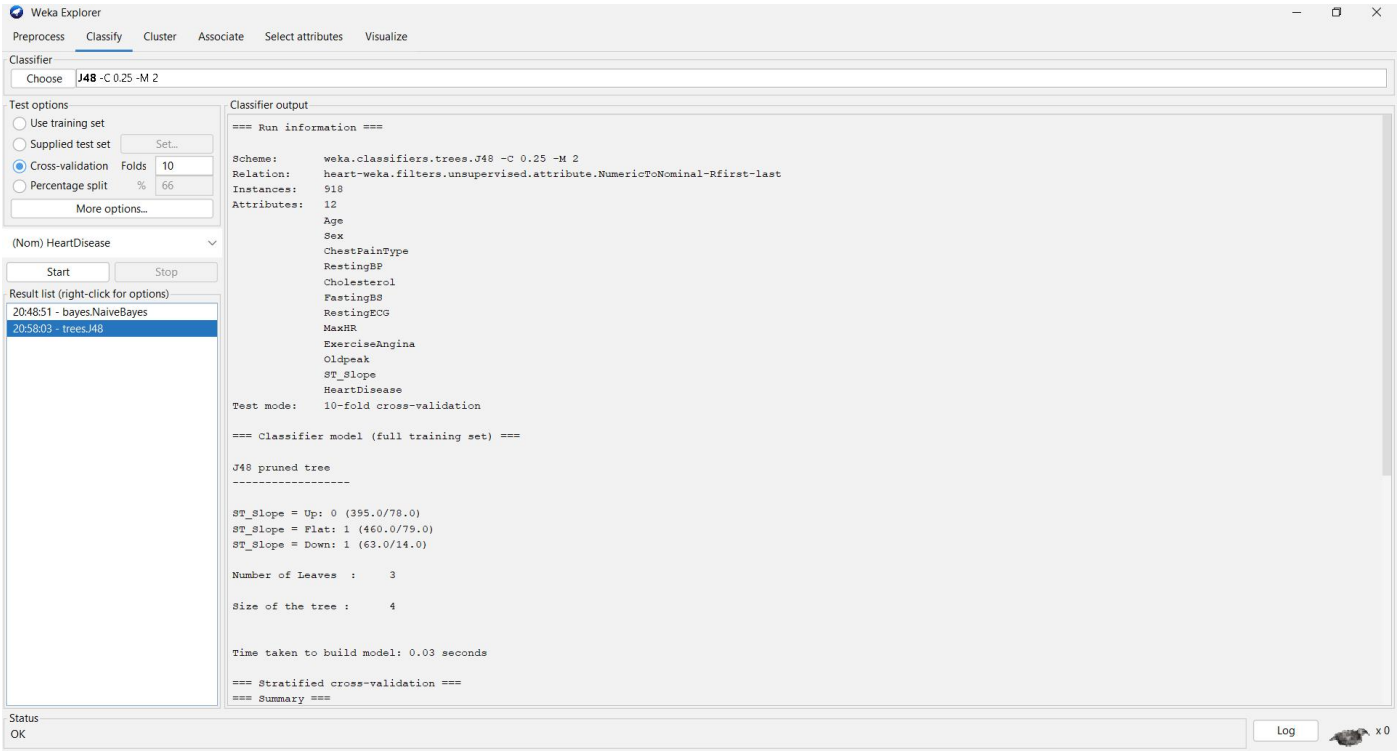



***Figure 5: Decision tree(J48) classification***

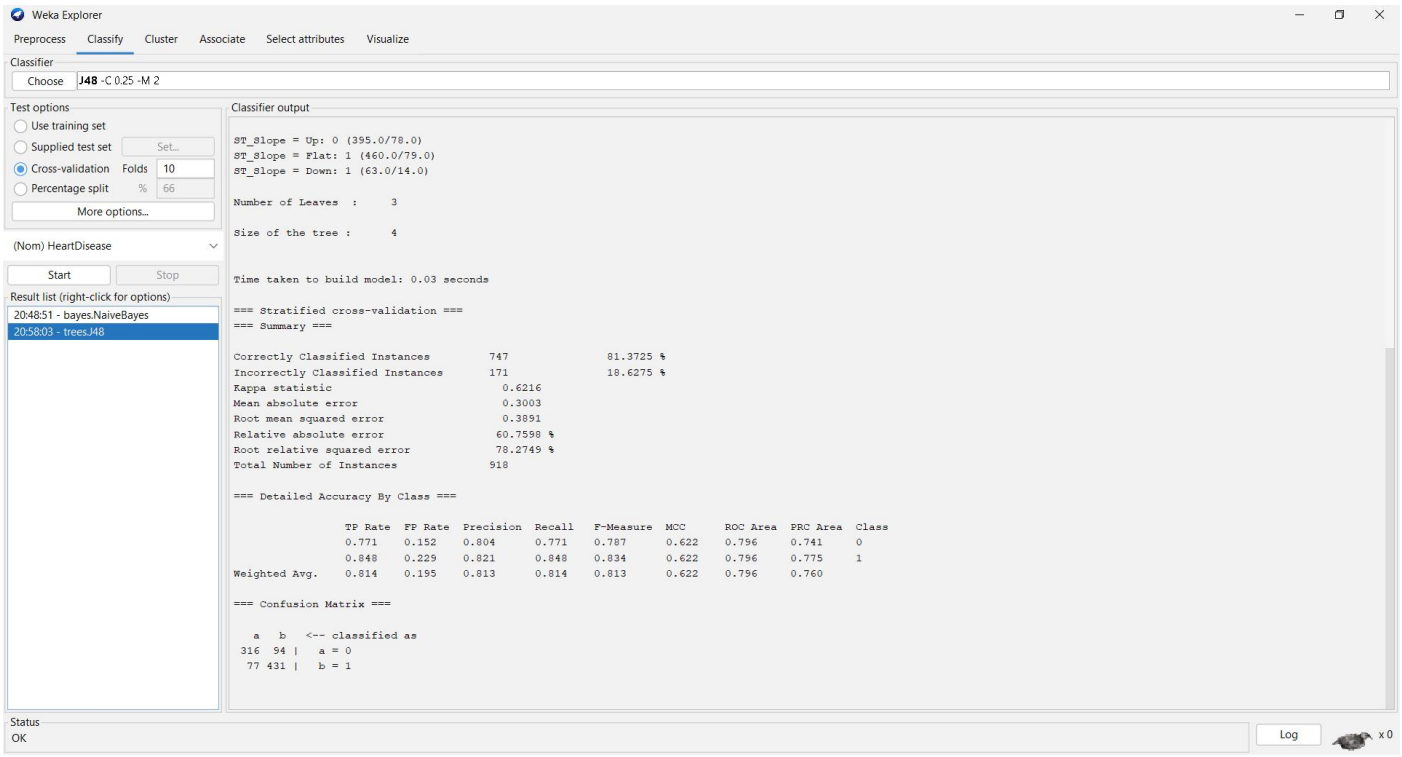



***Figure 6: Decision tree(J48) classification***

After applying two types of classifiers, the highest percentage of correctly classified instances is for the naïve Bayes classifier with 85.0763%. After that comes the Decision tree(J48) classifier with 81.3725%. The Naïve Bayes classifier is considered the best classifier for the dataset.

Reason to choose Naïve Bayes classifier is: From the obtained result of the two classifiers, it is clearly seen that Naïve Bayes has the highest percentage of correctly classified instances which is 85.0763%. As it has the most accurate value so, it would be more suitable classifier for the dataset.

**The advantages of Naïve Bayes is→**

- It is simple and easy to implement.
- It doesn't require as much training data.
- It handles both continuous and discrete data.
- It is highly scalable with a number of predictors and data points.
- It is fast and can be used to make real-time predictions.

**Here is the summary of the Naïve Bayes classifiers result:**

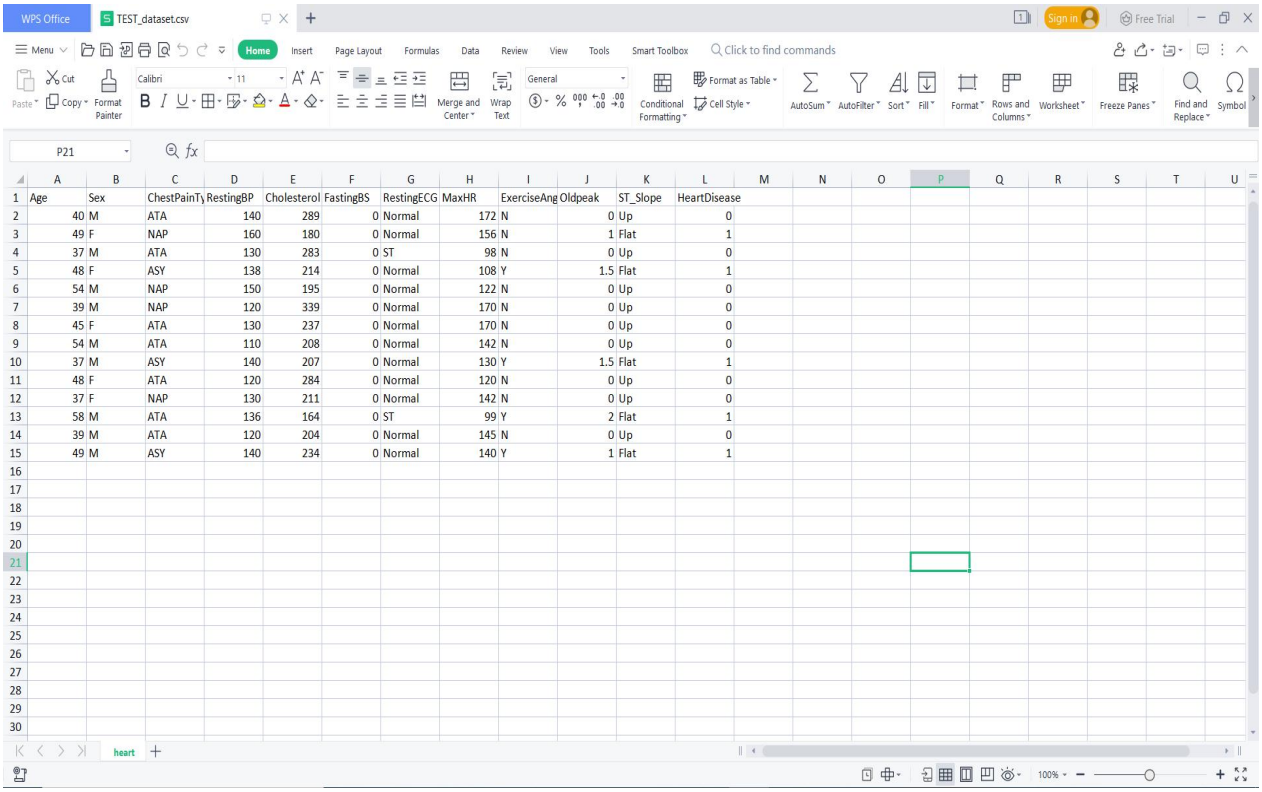
- Correctly Classified Instances : 781 85.0763 %
- Incorrectly Classified Instances : 137 14.9237 %
- Kappa statistic : 0.6983
- Mean absolute error : 0.168
- Root mean squared error : 0.3457
- Relative absolute error : 33.9808 %
- Root relative squared error : 69.5366 %
- Total Number of Instances : 918

## DISCUSSION:

In this task 1 section at first, we have chosen a proper dataset that has proper context and content and is also applicable for two of our selected classifiers. After that we selected two classifiers from Weka software that is applicable to our dataset with proper classification and accuracy. After applying these classifiers to our dataset, we got two different correctly classified instances, which are 85.0763%, 81.3725%. Also, with the value of some parameters like IP rate fp rate recall f measure, etc., and with the help of the confusion matrix we have selected our best classifier. Confusion matrix helps us to identify how good my algorithm is doing. That's how we have completed task 1 and found the best classifier which is the **Naïve Bayes classifier**.

## INTRODUCTION:

One of the most important mechanisms in machine learning is to train your algorithm on a training set that is separate and distinct from the test set for which will be gauged its accuracy. To detect a machine learning behavior, a training dataset has been set which was extracted subset from the referred dataset. Then this model had been tested with a test dataset which is a subset to test the trained model. While preparing the test dataset, things that were made sure are that the dataset was large enough to yield statistically meaningful results. Also, it was representative of the data set as a whole. In other words, the test sets with unusual characteristics than the training set were not chosen. The suitable classifier is then used to predict the classification for the instances in the test set. If the test set contains N instances of which C is correctly classified, C is correctly classified Predictive accuracy, P = C/N. There are 14 instances in this prepared test dataset.



| Age | Sex | ChestPainTy | RestingBP | Cholesterol | FastingBS | RestingECG | MaxHR | ExerciseAng | Oldpeak | ST_Slope | HeartDisease |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 40 | M | ATA | 140 | 289 | 0 | Normal | 172 | N | 0 | Up | 0 |
| 49 | F | NAP | 160 | 180 | 0 | Normal | 156 | N | 1 | Flat | 1 |
| 37 | M | ATA | 130 | 283 | 0 | ST | 98 | N | 0 | Up | 0 |
| 48 | F | ASY | 138 | 214 | 0 | Normal | 108 | Y | 1.5 | Flat | 1 |
| 54 | M | NAP | 150 | 195 | 0 | Normal | 122 | N | 0 | Up | 0 |
| 39 | M | NAP | 120 | 339 | 0 | Normal | 170 | N | 0 | Up | 0 |
| 45 | F | ATA | 130 | 237 | 0 | Normal | 170 | N | 0 | Up | 0 |
| 54 | M | ATA | 110 | 208 | 0 | Normal | 142 | N | 0 | Up | 0 |
| 37 | M | ASY | 140 | 207 | 0 | Normal | 130 | Y | 1.5 | Flat | 1 |
| 48 | F | ATA | 120 | 284 | 0 | Normal | 120 | N | 0 | Up | 0 |
| 37 | F | NAP | 130 | 211 | 0 | Normal | 142 | N | 0 | Up | 0 |
| 58 | M | ATA | 136 | 164 | 0 | ST | 99 | Y | 2 | Flat | 1 |
| 39 | M | ATA | 120 | 204 | 0 | Normal | 145 | N | 0 | Up | 0 |
| 49 | M | ASY | 140 | 234 | 0 | Normal | 140 | Y | 1 | Flat | 1 |

*Figure 7: Test dataset*

## PROCEDURE OF TESTING THE TEST DATASET:

1. First, Weka 3.8.6 was opened and the ‘Explorer’ option was chosen and such window named Weka explorer was opened.

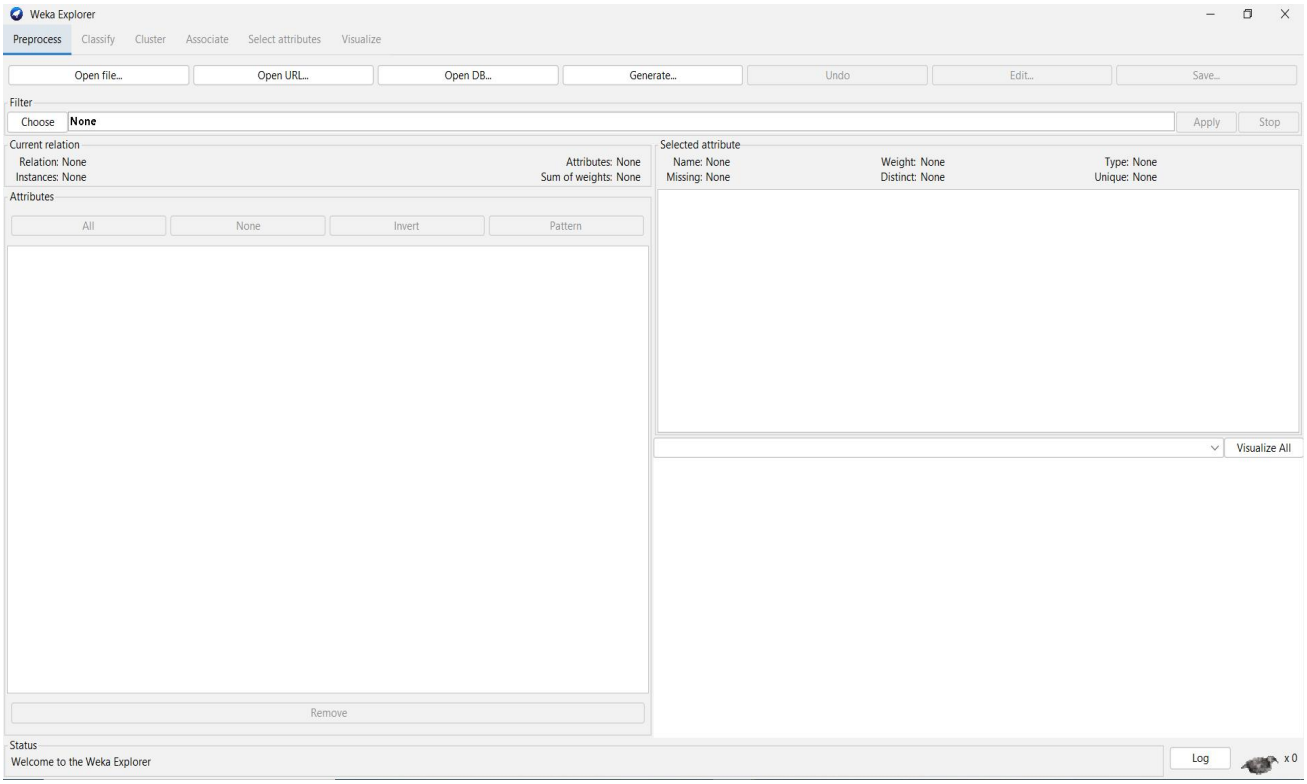



*Figure 8: Weka explorer*

2. Then, the open file option was selected and the extracted .csv file, training dataset was selected from the device.

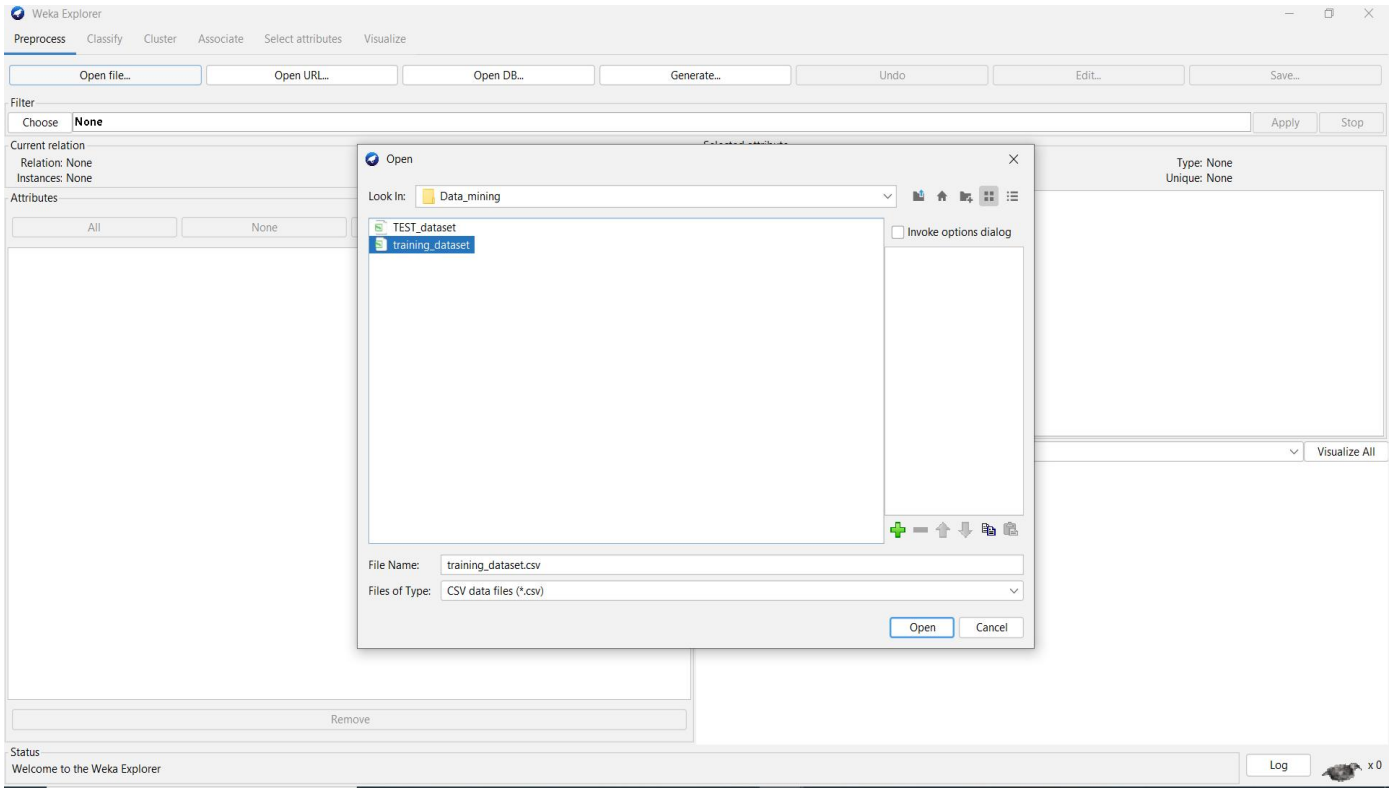



*Figure 9:Training dataset*

3. After the open option was clicked, the details of the dataset popped.

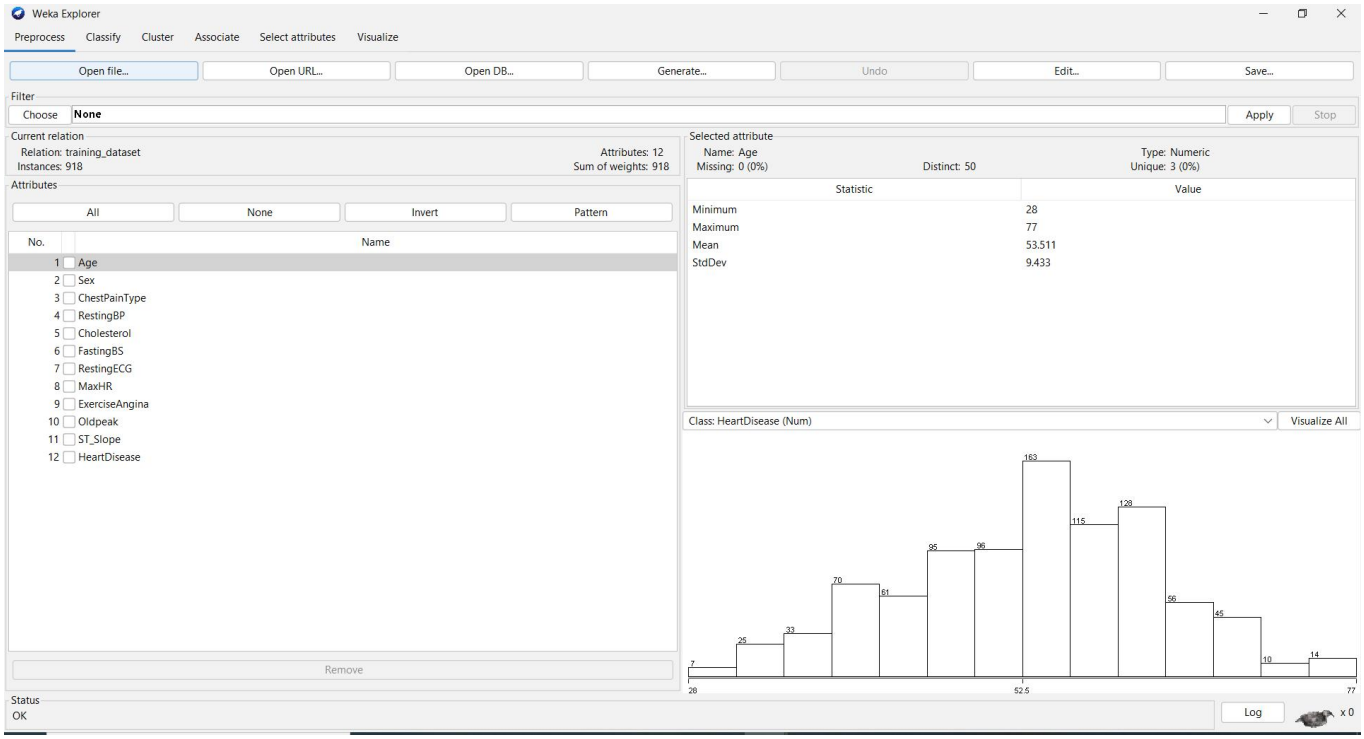



*Figure 10: Training dataset visible*

4. Then the preferred classifier (Decision tree classifier) for the training dataset was selected. Then from the test options, the use training set was chosen and the start option was selected.

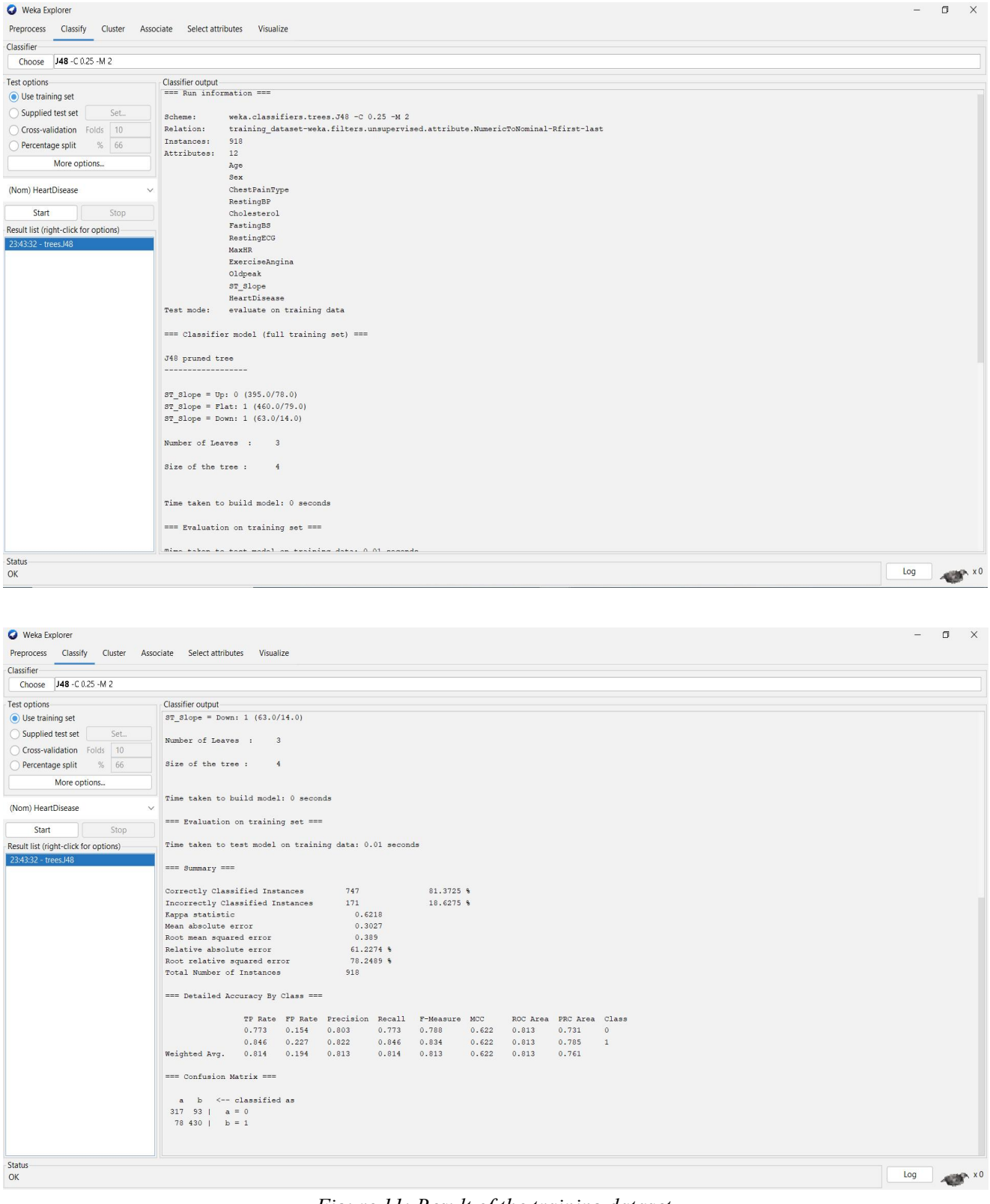



*Figure 11: Result of the training dataset*

5. To input the test set, from the test options, supplied test data was selected and then a window named test instances came.

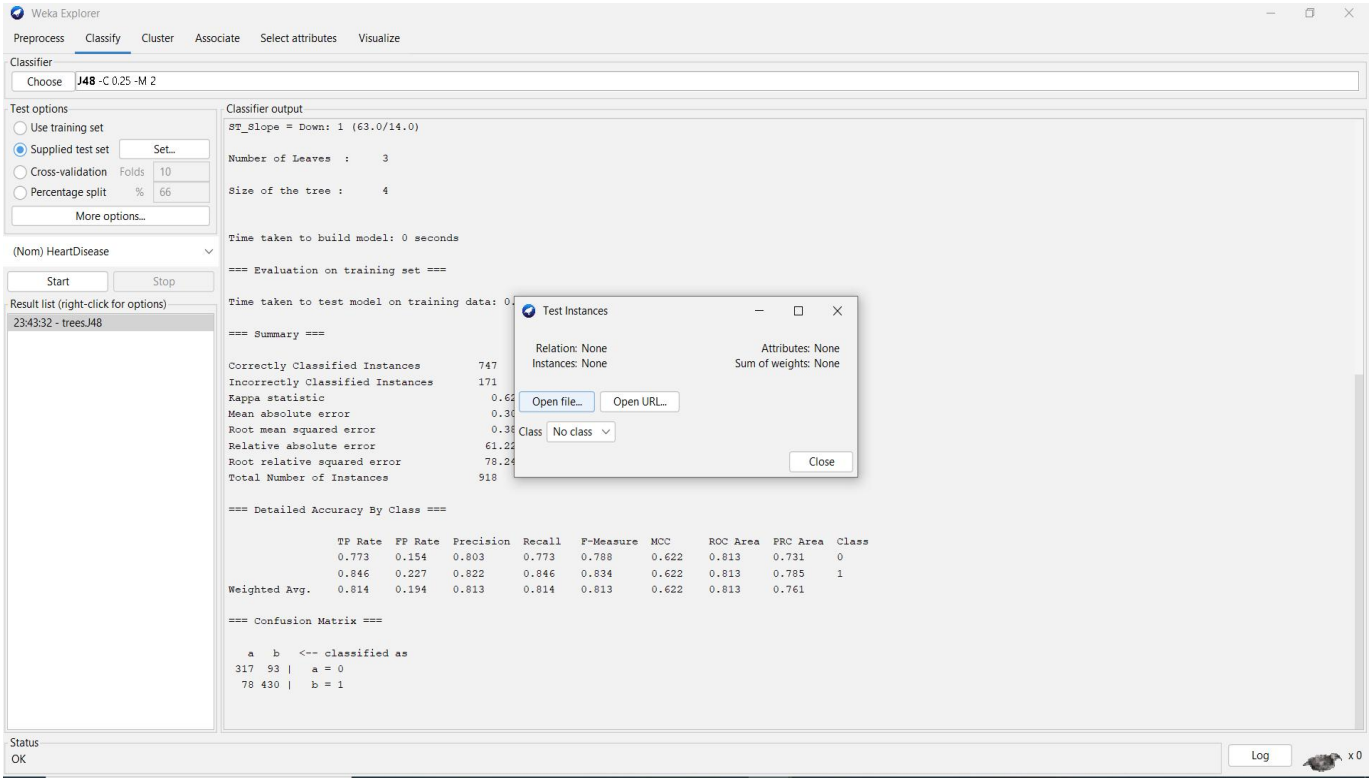
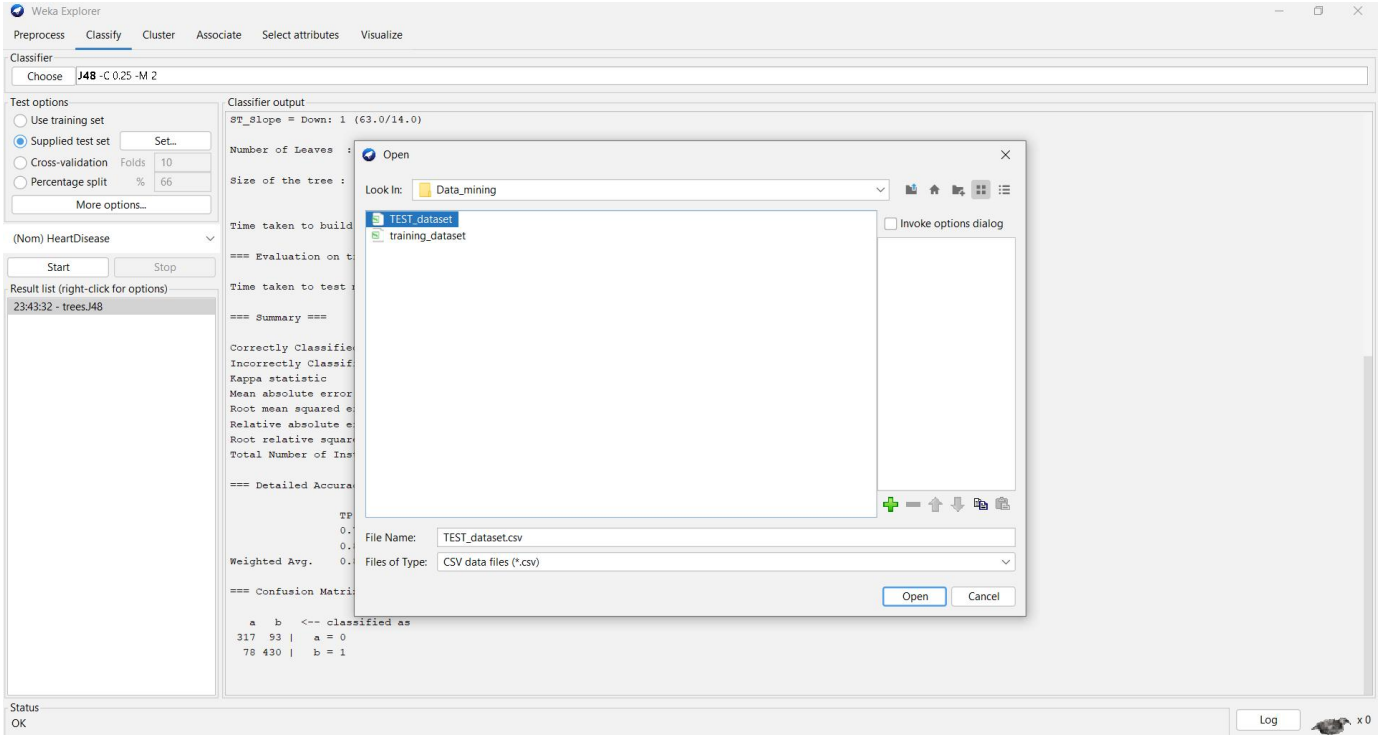



*Figure 12: Supplied test dataset selected*

6. Then the open file option was chosen to insert the test data and then the dataset was opened.



7. To make sure the test set works properly, it was made sure that the output predictions was in PlainText form.

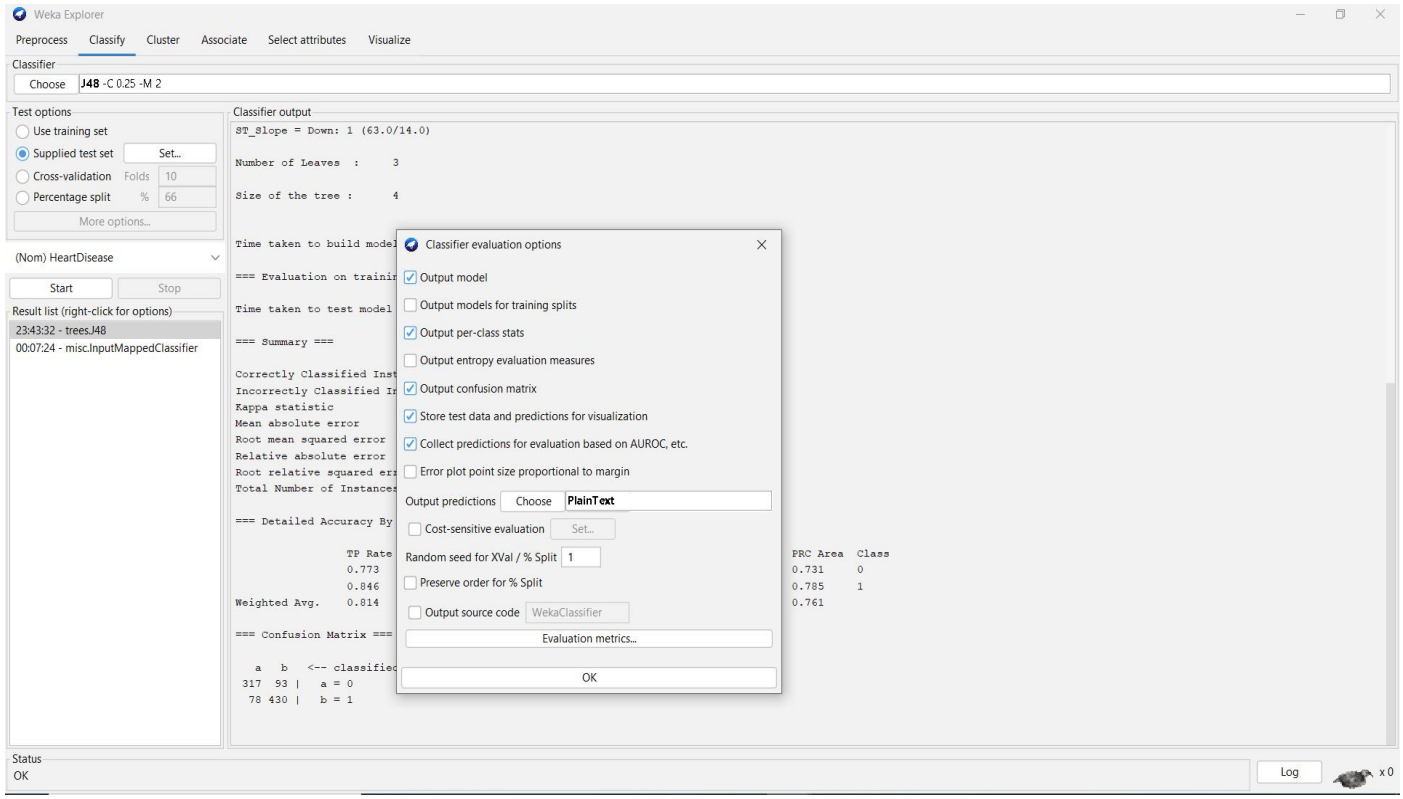
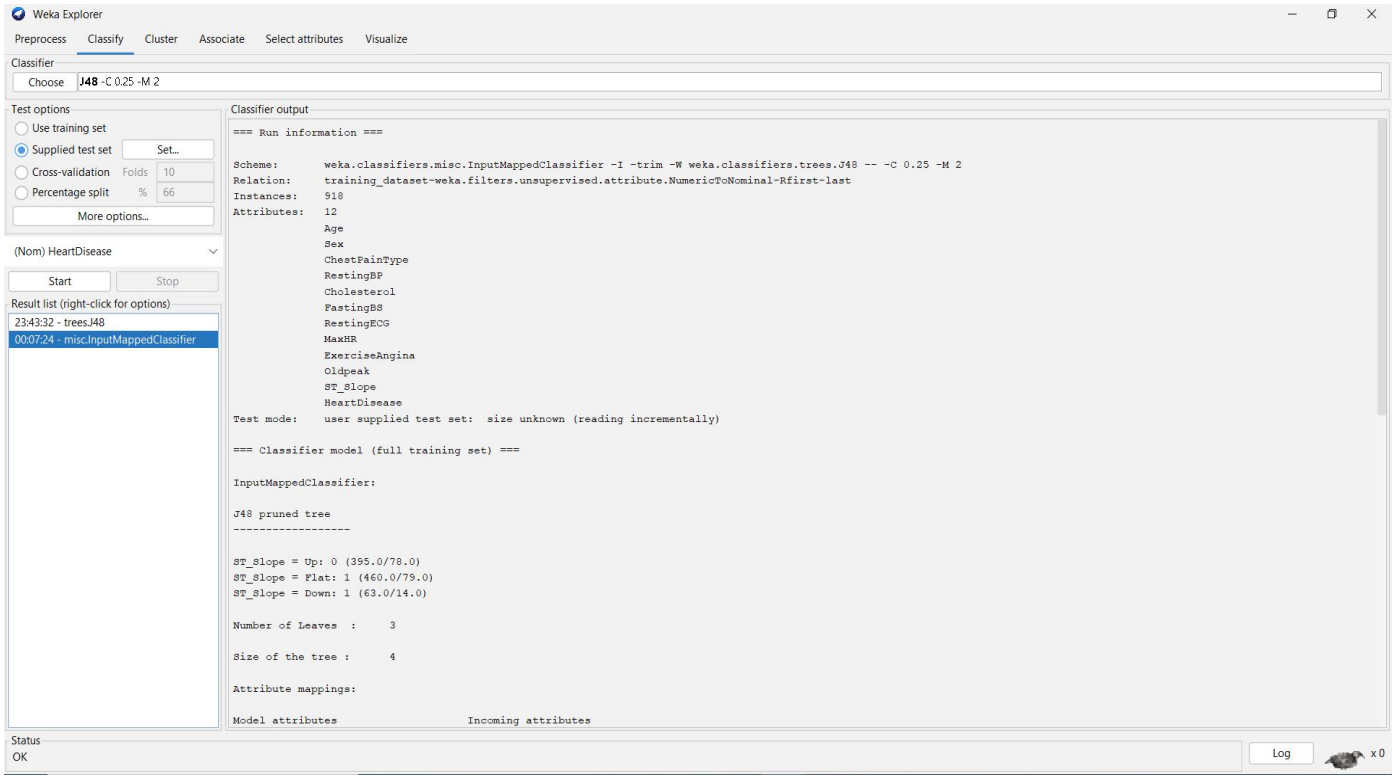



*Figure 14: PlainText format chosen for output prediction*

8. Then the start button was clicked.



## RESULT OF TEST-DATASET MODEL:

Once the start button was clicked, the output for the test dataset came. In the result, the test mode was ‘user-supplied test set’ which means the classifier is evaluated on how well it predicts the class of a set of instances loaded from a file that was inputted by the user. The total time taken to build the model was 0.0 seconds and among the 14 instances, there were 0 instances where the error was found. In this model, the Correctly classified instances % Value describes the amount of accuracy of correctly classified instances provided by the algorithm. In this case, the percentage is 100%. Incorrectly classified instances % Value describes how much incorrect instances are given by the algorithm. In this case, the percentage is 0%.

**Mean Absolute Error (MAE):** It can define as a statistical measure of how far an estimate is from actual values i.e. the average of the absolute magnitude of the individual errors. It is usually similar in magnitude but slightly smaller than the root mean squared error. In this model the MAE is 0.168

**Root Mean-Squared Error (RMSE**): The Root Mean Square Error (RMSE) calculates the differences between values predicted by a model / an estimator and the values observed from the thing being modeled/ estimated. RMSE is used to measure the accuracy. It is ideal if it is small. In this case, the RMSE is 0.3457 which is ideal.

## DISCUSSION:

The idea behind training and test sets is to test the generalization error. Separating data into training and testing sets is an important part of evaluating data mining models. Since the test set data already contains known values for the property I want to predict, it is easy to determine if the model's assumptions are correct. So here I carefully separate the data into test/training and apply them in weka after that it gives the error table and accuracy rate that’s how I get my proper result and accuracy model.

## INTRODUCTION:

Unsupervised Learning is a machine learning technique in which the users do not need to supervise the model. Instead, it allows the model to work on its own to discover patterns and information that was previously undetected. It mainly deals with the unlabelled data. Unsupervised learning is helpful for finding useful insights from the data. Unsupervised learning is much similar to human learning to think by their own experiences, which makes it closer to the real AI. Unsupervised learning works on unlabeled and uncategorized data which makes unsupervised learning more important.

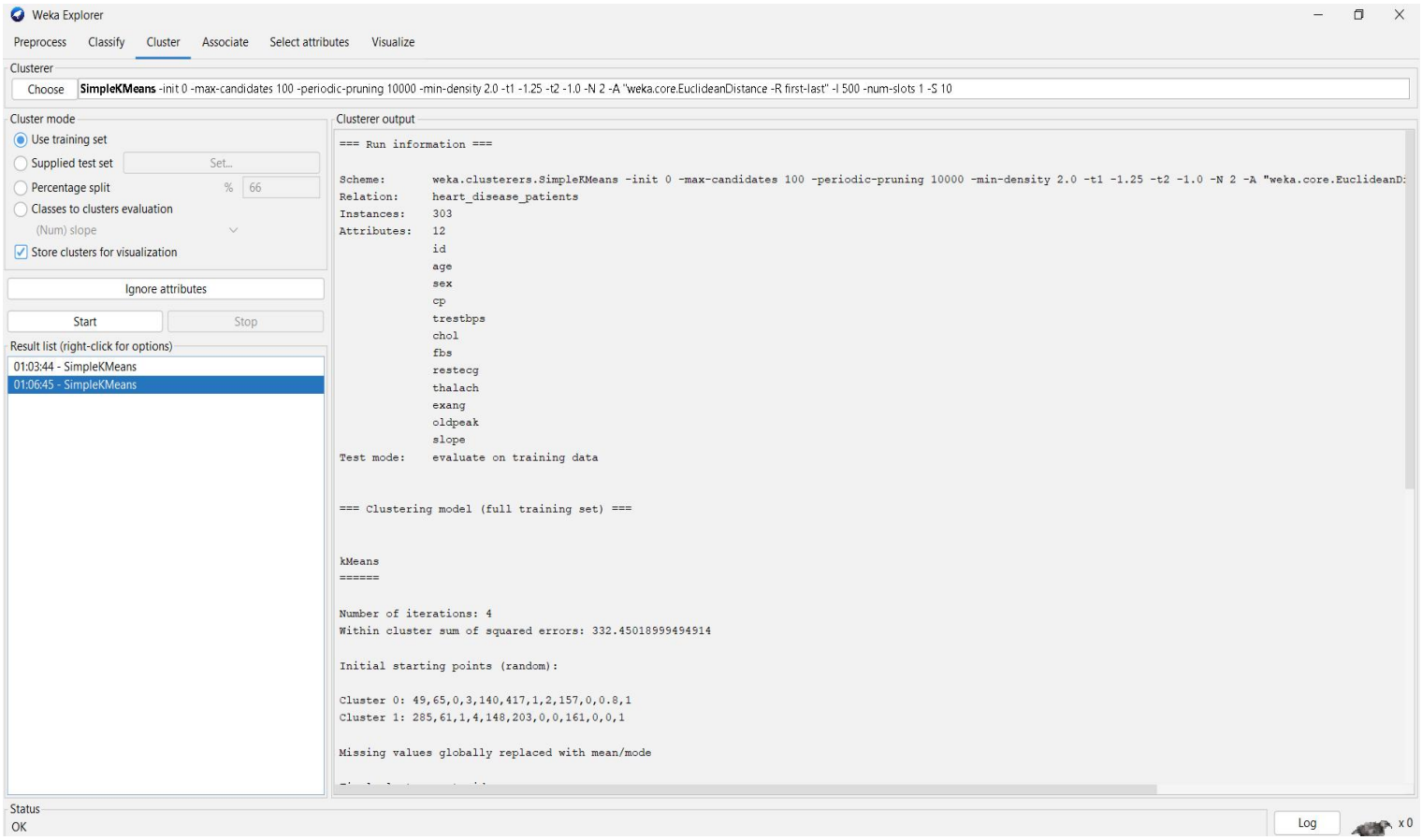
We have chosen "Heart Disease Patients" dataset. We will also use K means clustering Algorithm.

About the dataset in this report, we used "Heart Disease Patients", a CSV dataset file collected from **Kaggle.com**.

## RESULT:

Applying K means clustering Algorithm.

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on. It allows us to cluster the data into different groups and is a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Num) slope
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
01:03:44 - SimpleKMeans
01:06:45 - SimpleKMeans
Clusterer output
=== Run information ===
Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanD
Relation: heart_disease_patients
Instances: 303
Attributes: 12
id
age
sex
cp
trestbps
chol
fbs
restecg
thalach
exang
oldpeak
slope
Test mode: evaluate on training data
=== Clustering model (full training set) ===
kMeans
======
Number of iterations: 4
Within cluster sum of squared errors: 332.45018999494914
Initial starting points (random):
Cluster 0: 49,65,0,3,140,417,1,2,157,0,0.8,1
Cluster 1: 285,61,1,4,148,203,0,0,161,0,0,1
Missing values globally replaced with mean/mode
Status
OK
Log
x 0

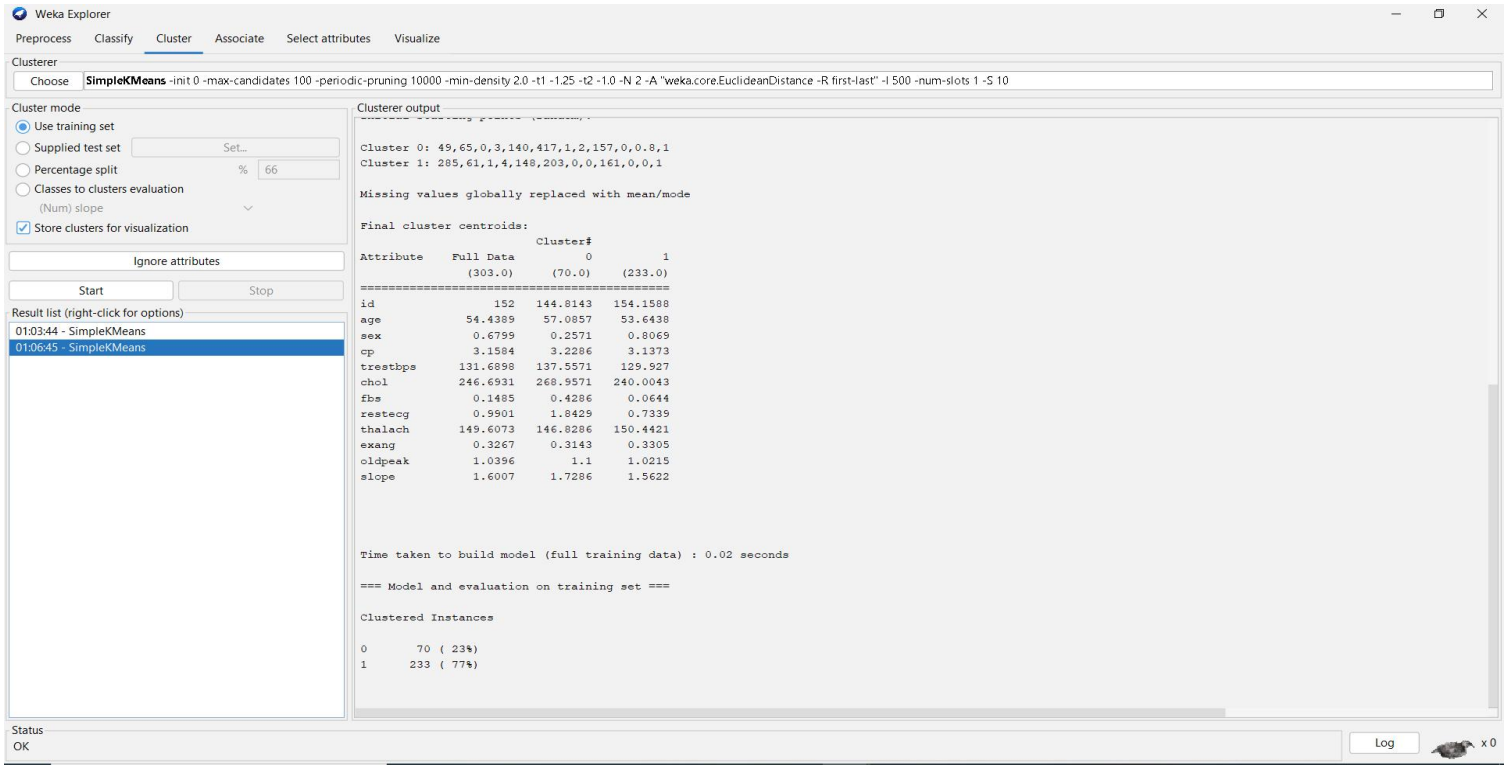



*Figure 16: K means clustering algorithm*

Here is the summary of the K means clustering Algorithm result:

Number of iterations: 4

Initial starting points (random):

Cluster 0: 49,65,0,3,140,417,1,2,157,0,0.8,1

Cluster 1: 285,61,1,4,148,203,0,0,161,0,0,1

Time taken to build model (full training data) : 0.02 seconds

Clustered Instances

0 70 (23%)

1 233 (77%)

## DISCUSSION:

We have chosen a proper unsupervised dataset. We used a csv file and then apply K means Clustering Algorithm. By default, the value of K is 2 so there are two results.

For 0 instances it is 23% whereas for 1 is 77%. By that we have completed K means clustering for unsupervised data.

## REFERENCE:


1. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction
2. https://www.kaggle.com/datasets/kingabzpro/heart-disease-patients
3. https://www.investopedia.com/terms/d/datamining.asp
4. https://en.wikipedia.org/wiki/Data_mining