**DATA WAREHOUSING & DATA MINING**

**SEMESTER: Spring 2021-2022**

FINAL-TERM ASSIGNMENT

DATA MINING WITH WEKA

**SUBMITTED BY:**

|  |  |
| --- | --- |
| **Student Name** | **Student ID** |
| Das, Proma | 19-39625-1 |
| Shanto, Majharul Islam | 19-39585-1 |
| Ahmed, Sk. Shihab | 19-39575-1 |
| Didar Hossen | 18-39084-3 |

**GROUP NO:** 05

**SECTION:** C

**DEPARTMENT:** CSE

**SUBMITTED TO ||**

**COURSE TEACHER:** TOHEDUL ISLAM

**INTRODUCTION:**

The practice of extracting patterns and other useful information from massive data sets is known as data mining. It's also known as knowledge discovery in databases (KDD). Data mining is utilized in a wide range of research and business applications, including healthcare, education, sales and marketing, product development, and so on. It is an interdisciplinary issue in computer science and statistics with the goal of extracting information from a data collection and changing it into a usable structure for later use. Some of the classification methods used in data mining include KNN, Naive Bayes, K-means clustering, Agglomerative Hierarchical Clustering and Decision Tree. I have chosen “Key Indicator of Heart Disease” to classify the condition of heart by using Three different classifier. Two is from supervised learning which is naive bayes and Decision Tree algorithm. And the other one is Unsupervised learning which is K means clustering algorithm. By using that we can find the best suited classifier for the data set.

**Information about the data set:**

In this report, the used “Key Indicator of Heart Disease”, a CSV data set file, collected from kaggle.com which was used to predict the outcome of the heart.

**The targeted feature is:**

Heart Disease

**The other feature sets are:**

BMI

Smoking

Alcohol Drinking

Stroke

Physical Health

Mental Health

Difficulty Walking

Sex

Age Category

Race

Diabetic

Physical Activity

General Health

Sleep Time

Asthma

Kidney Disease

Skin Cancer

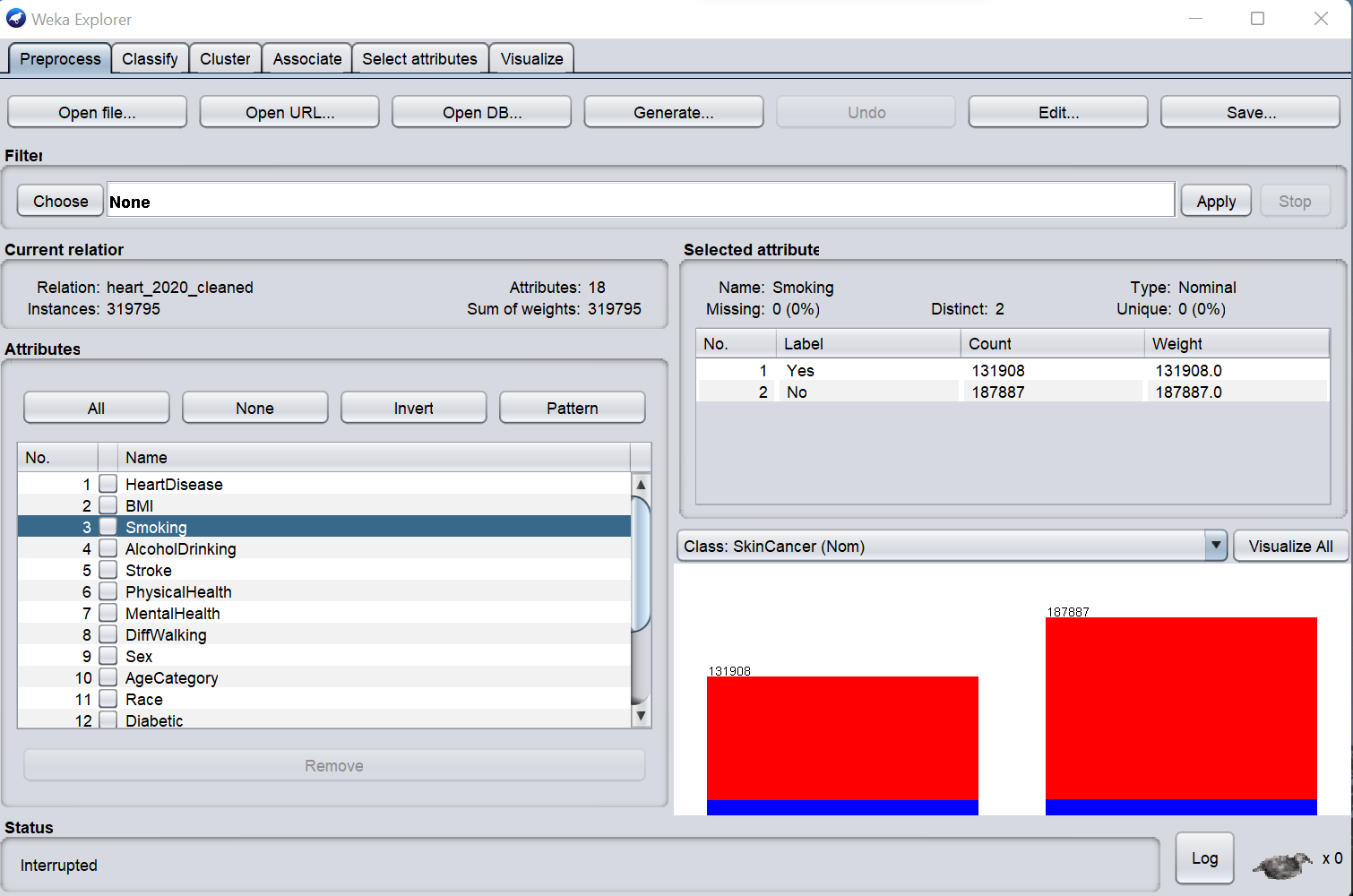
**About the attribute:**

The data set contains 17 attribute and 1 class attribute which is the targeted feature to predict. The class attribute refers that for which reasons this heart disease occurs.

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Representation in Data set** | **Data Type** |
| BMI | Numeric Value | Continuous |
| Smoking | Yes, No | Categorical |
| Alcohol Drinking | Yes, No | Categorical |
| Stroke | Yes, No | Categorical |
| Physical Health | Numeric Value | Continuous |
| Mental Health | Numeric Value | Continuous |
| Diff Walking | Yes, No | Categorical |
| Sex | Male, Female | Categorical |
| **Age Category** | **In Years** | **Continuous** |
| Race | White, Black, American, Asian | Categorical |
| Diabetic | Yes, No | Categorical |
| Physical Activity | Yes, No | Categorical |
| Gen Health | Very Good,Fair,Good,Poor,Excellent | Categorical |
| Sleep Time | In Hour | Continuous |
| Asthma | Yes, No | Categorical |
| Kidney Disease | Yes. No | Categorical |
| Skin Cancer | Yes, No | Categorical |

There is total 319795 instances of these 18 attributes and all these instances were used for classification.

**Here are the graphical details of the attributes:**

****

**Figure 1: Selected Data set**

****

**Figure 2: Details of all the attribute**

**Classifier:**

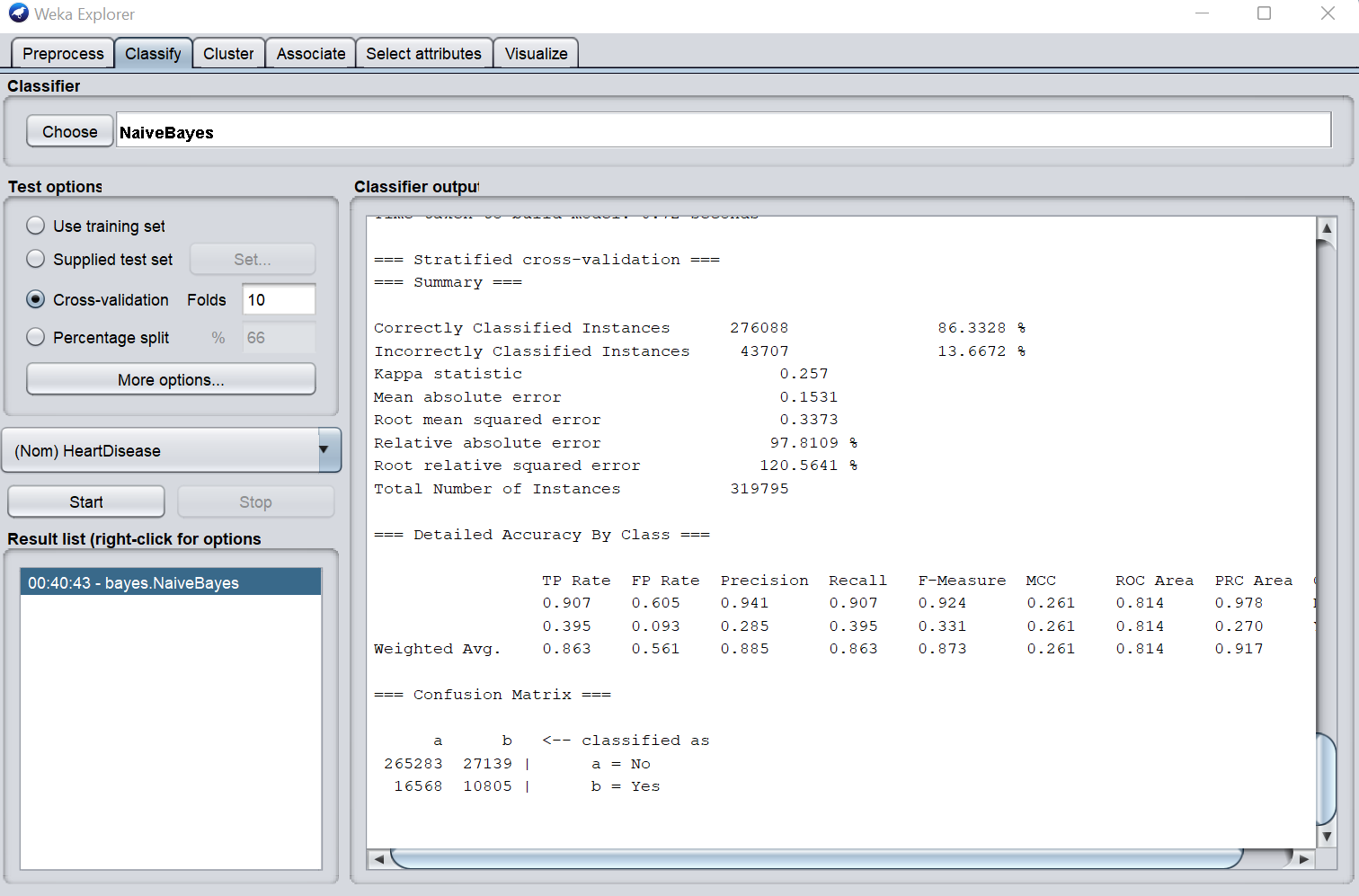
A classifier is a machine learning model that distinguishes between objects based on particular characteristics. To compare the results, three types of classification were applied using the same data. Naive Bayes, Decision Tree classifiers and clustering were employed in this procedure.

**RESULT**

Weka 3.9.5 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.

**For Supervised Learning**

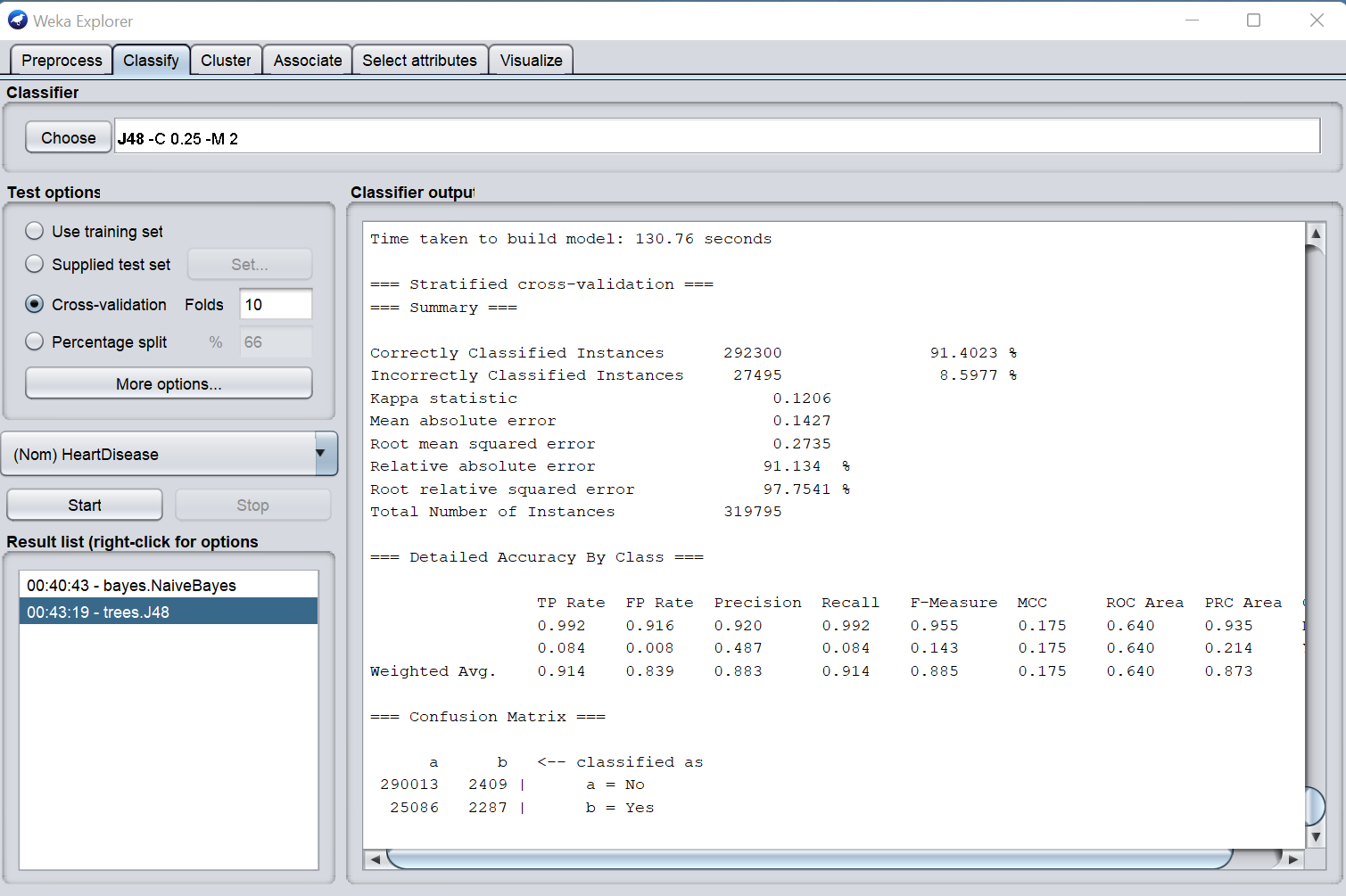
Applying Naive Bayes Classifier: The Naive Bayes method is a supervised learning algorithm for addressing classification issues that is based on the Bayes theorem. It is mostly utilized in text classification tasks that require a large training data set. It's a probabilisticclassifier, which means it makes predictions based on an object's probability. The Naive Bayes format was chosen from the bayes folder for classifying the selected data set.

****

**Figure 3: Naive Bayes Classification**

**Applying Decision tree classifier:**

Decision Tree is a supervised learning technique that may be used to solve both classification and regression problems, however it is most commonly employed to solve classification issues. Internal nodes represent data set attributes, branches represent decision rules, and each leaf node provides the conclusion in this tree-structured classifier. The J48 format was chosen for Decision tree classification while classifying the selected data set.

****

**Figure 4 : Decision Tree Classification**

After applying two types of classification, the highest percentage of correctly classified instances is for Decision Tree classifier is 91.4%. After that comes naive bayes classifier with 86.3% which is lowest. The Decision tree classifier is considered the best classifier for the data set.

**Reason to choose Decision Tree classifier:**

The percentage of instances accurately classified is 91.4%. It would be a better classifier for the data set because it has the highest accurate value. One of the benefits of decision trees is that their outputs are simple to read and understand without the need for statistical expertise. Decision trees demand less data preparation than other decision techniques. Decision trees are more adaptable and simpler to use. Furthermore, because KNN's real-time execution is expensive, Decision tree is faster than KNN. A decision tree also has the advantage of forcing the evaluation of all possible outcomes of a decision and tracing each path to a conclusion. It generates a detailed analysis of the effects along each branch and flags decision nodes that require additional investigation. This improves the accuracy of predictions.

**Here is the summary of the Decision tree classifiers result:**

• Correctly Classified Instances: 292300 91.4 %

• Incorrectly Classified Instances: 27495 8.5977 %

• Kappa statistic: 0.1206

• Mean absolute error: 0.1427

• Root mean squared error: 0.2735

• Relative absolute error: 91.134%

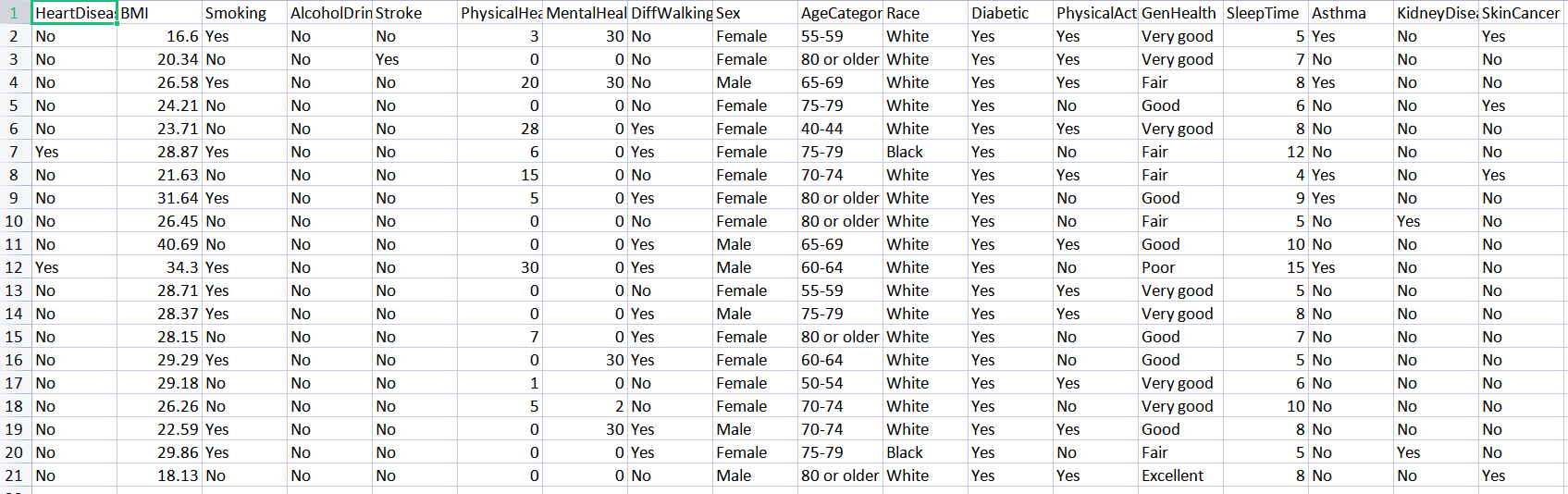
• Root relative squared error: 97.7541 %

• Total Number of Instances: 319795

**PREPARING TEST-DATA SET:**

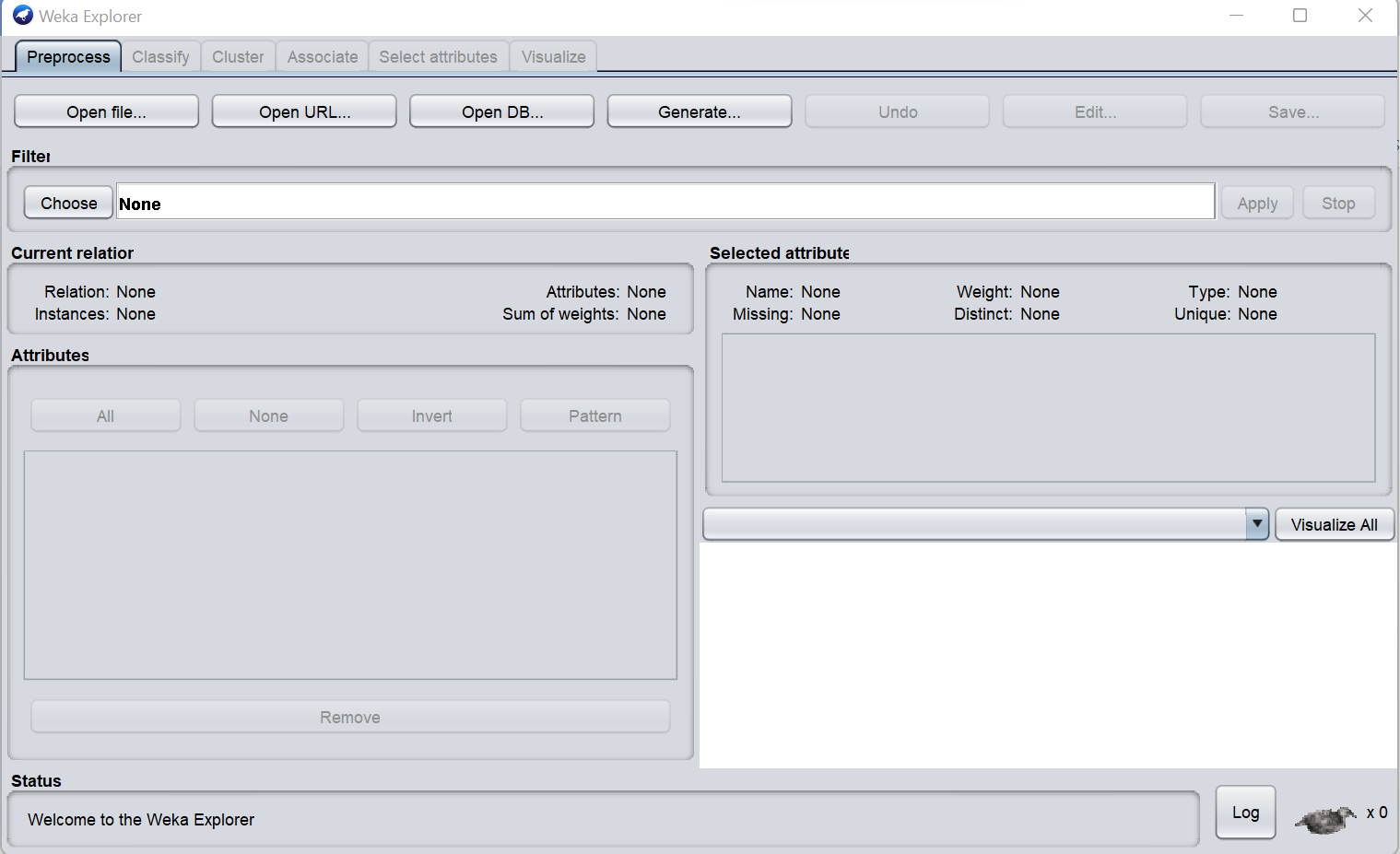
One of the most fundamental methods in machine learning is to train your algorithm on a different and unique training set from the test set for which its accuracy will be measured. A training data set was created using a subset of the referred data set in order to detect machine learning behavior. The model was then put to the test with a test data set, which is a subset of the training data set. Things that were ensured when constructing the test data set were that it was large enough to give statistically relevant findings. It was also indicative of the entire data set. To put it another way, test sets that differed from the training set were not chosen. The appropriate classifier is subsequently selected.

If the test set contains N instances of which are correctly classified, C are correctly classified reductive accuracy, P = C/N. There are 20 instances in this prepared test data set.

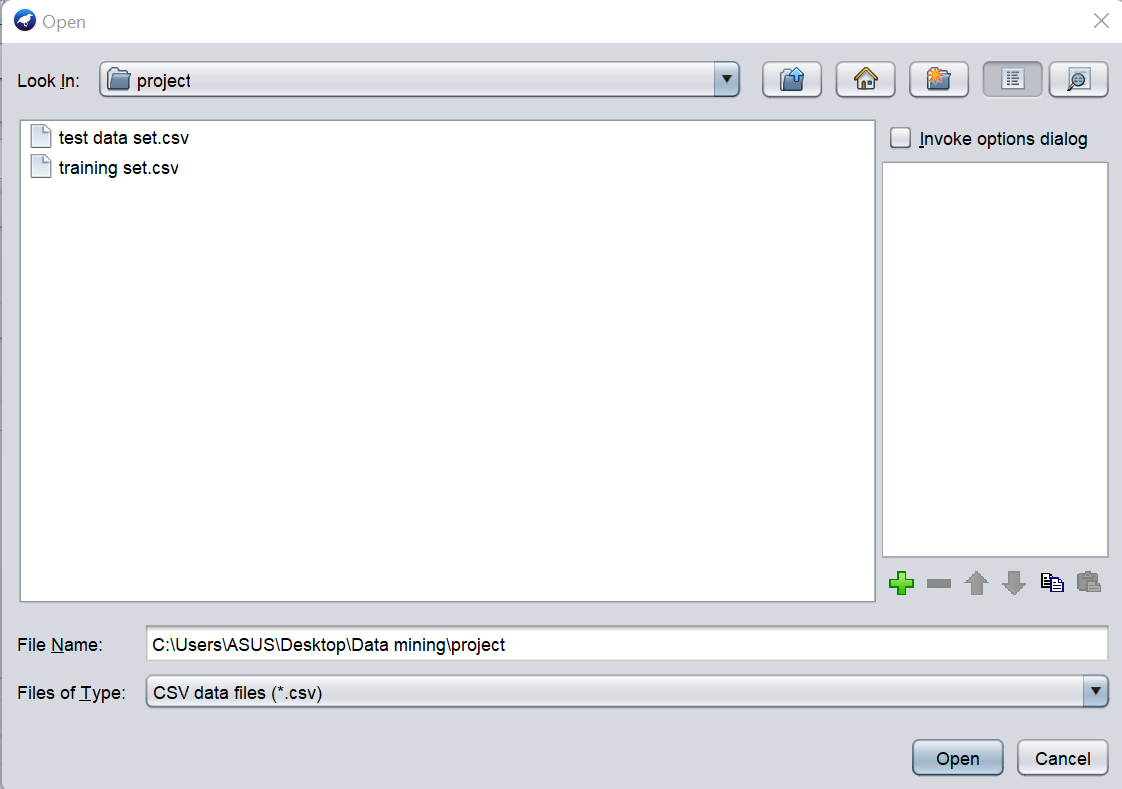
****

**PROCEDURE OF TESTING THE TEST DATASET:**

1. First, Weka 3.9.5 was opened and ‘Explorer’ option was chosen and such window named weka explorer was opened.

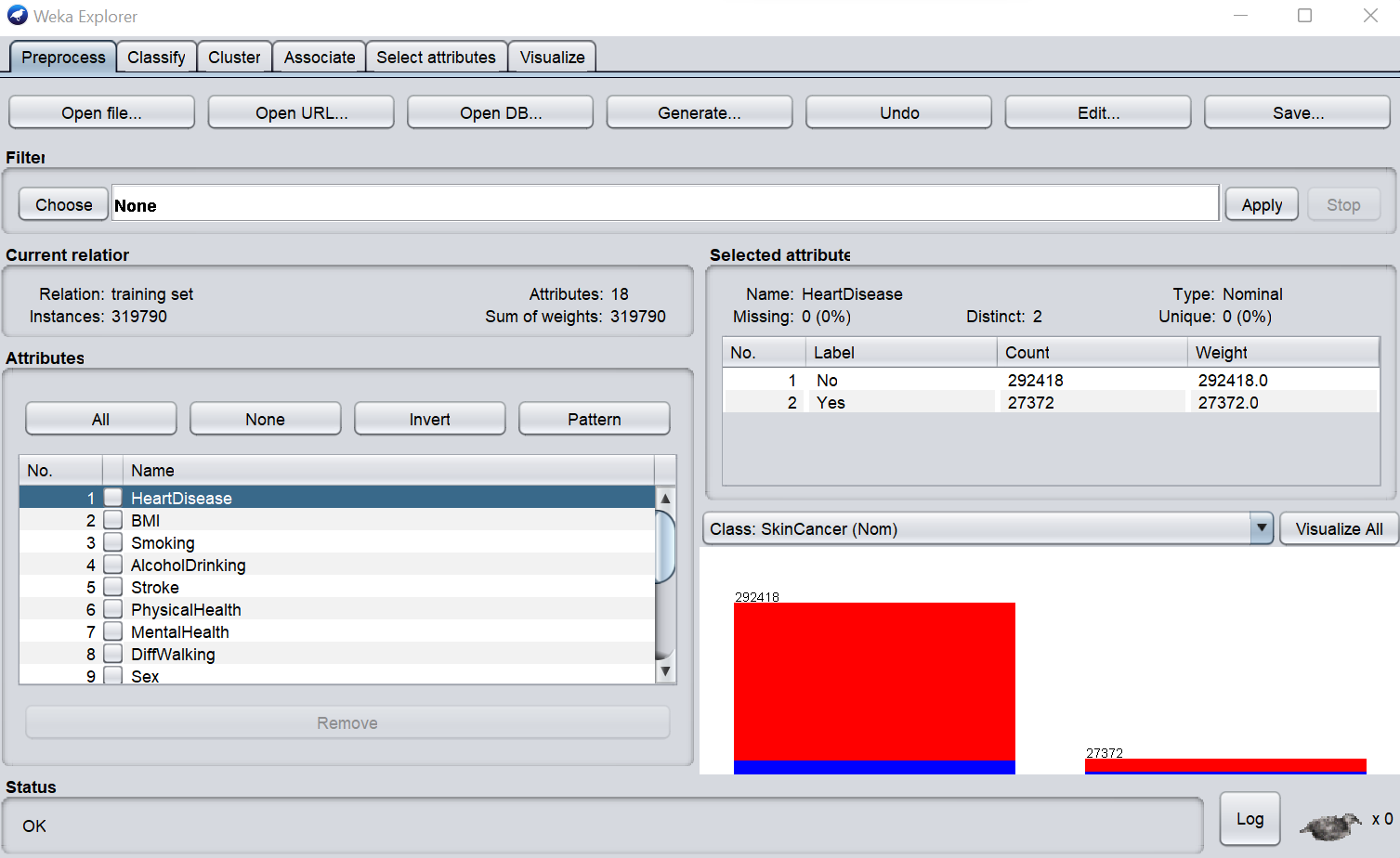
**Figure 5: Weka explorer**

1. Then, the open file option was selected and the extracted CSV file, training data set was selected from the device.

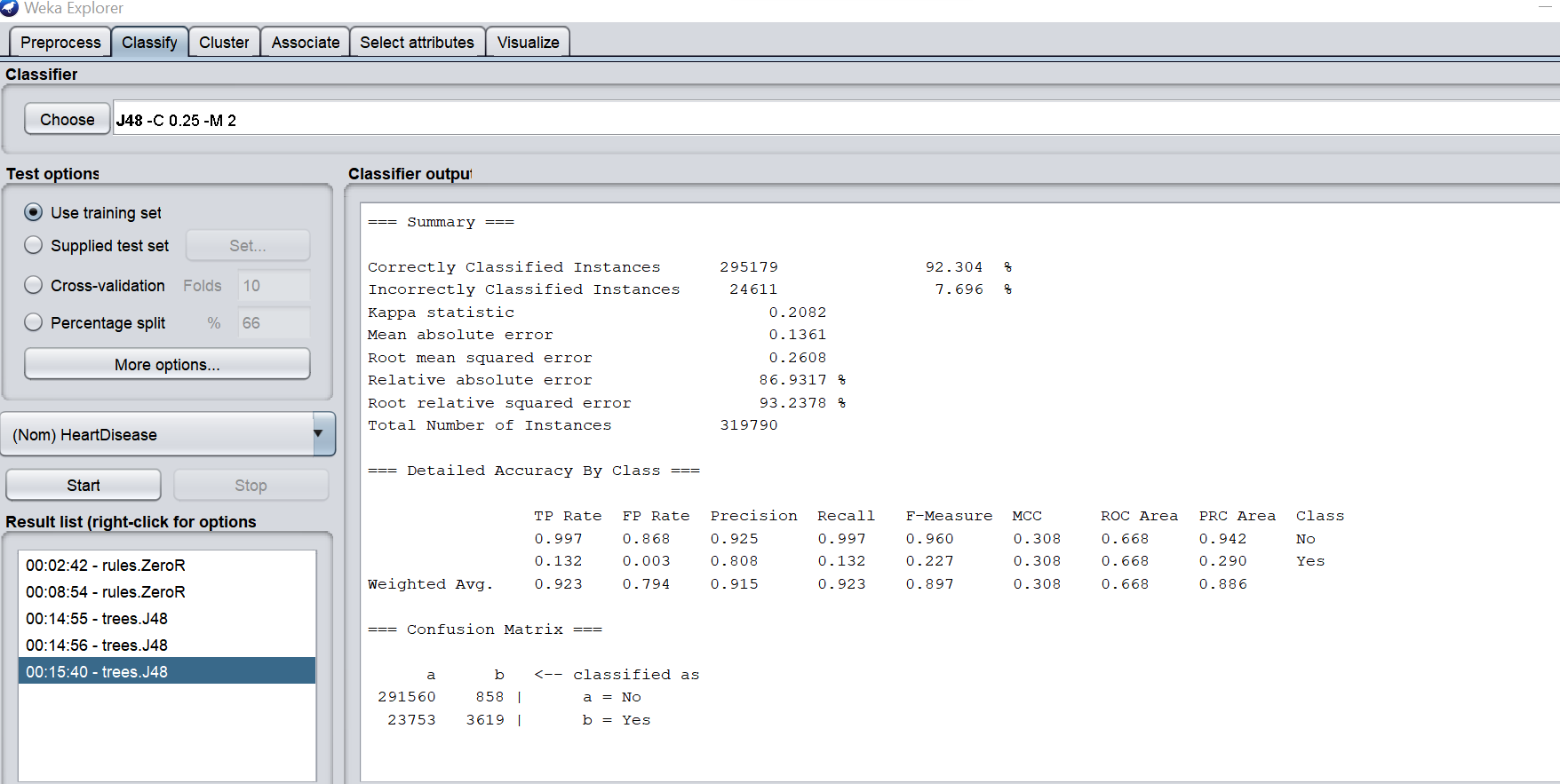
****

**Figure 6: training data select**

1. After the open option was clicked, the details of the data set popped.

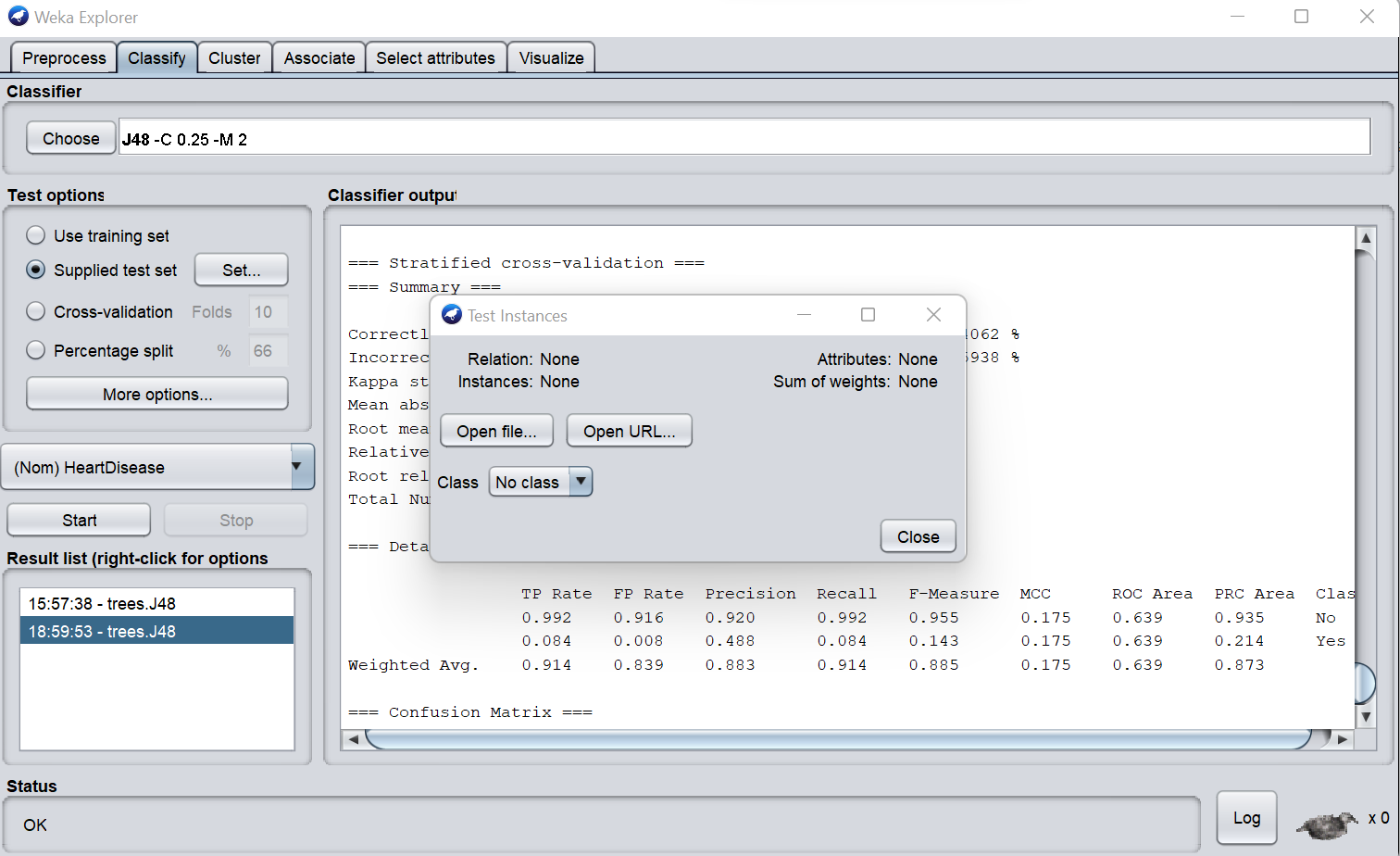
**Figure 7: Training data set**

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

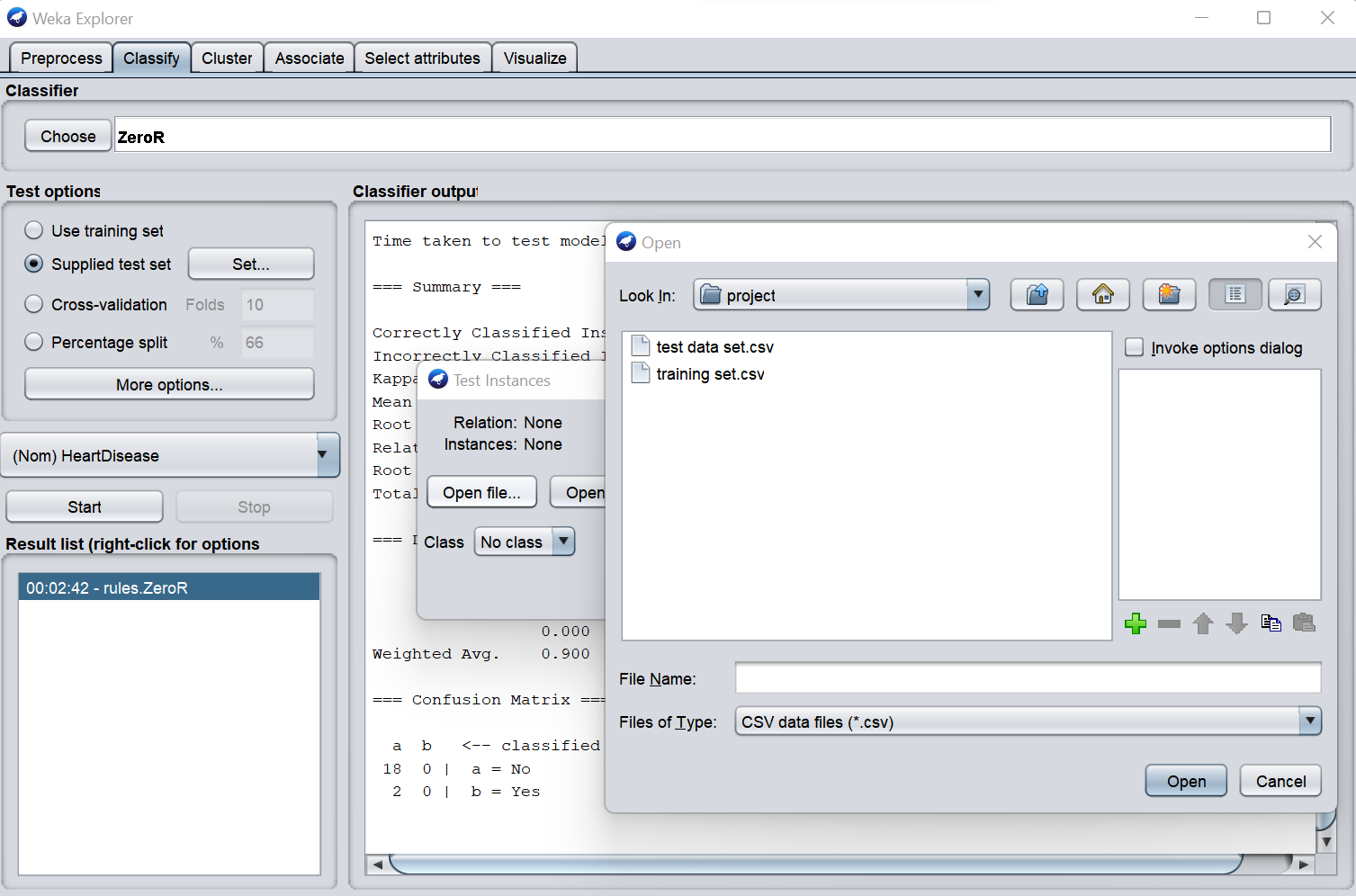
****

**Figure 8: result of training set**

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

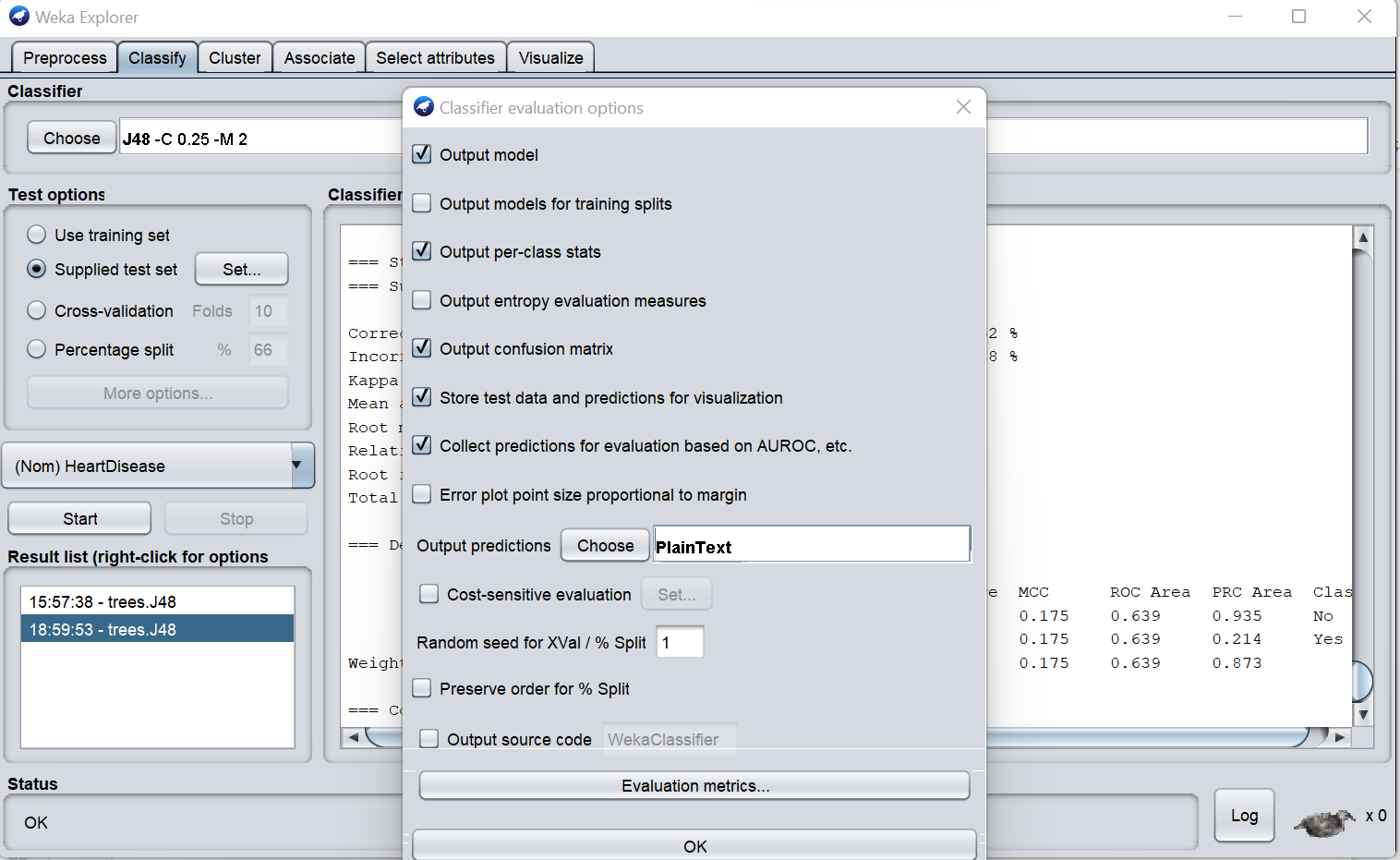
**Figure 9: Supplied test set selected**

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

****

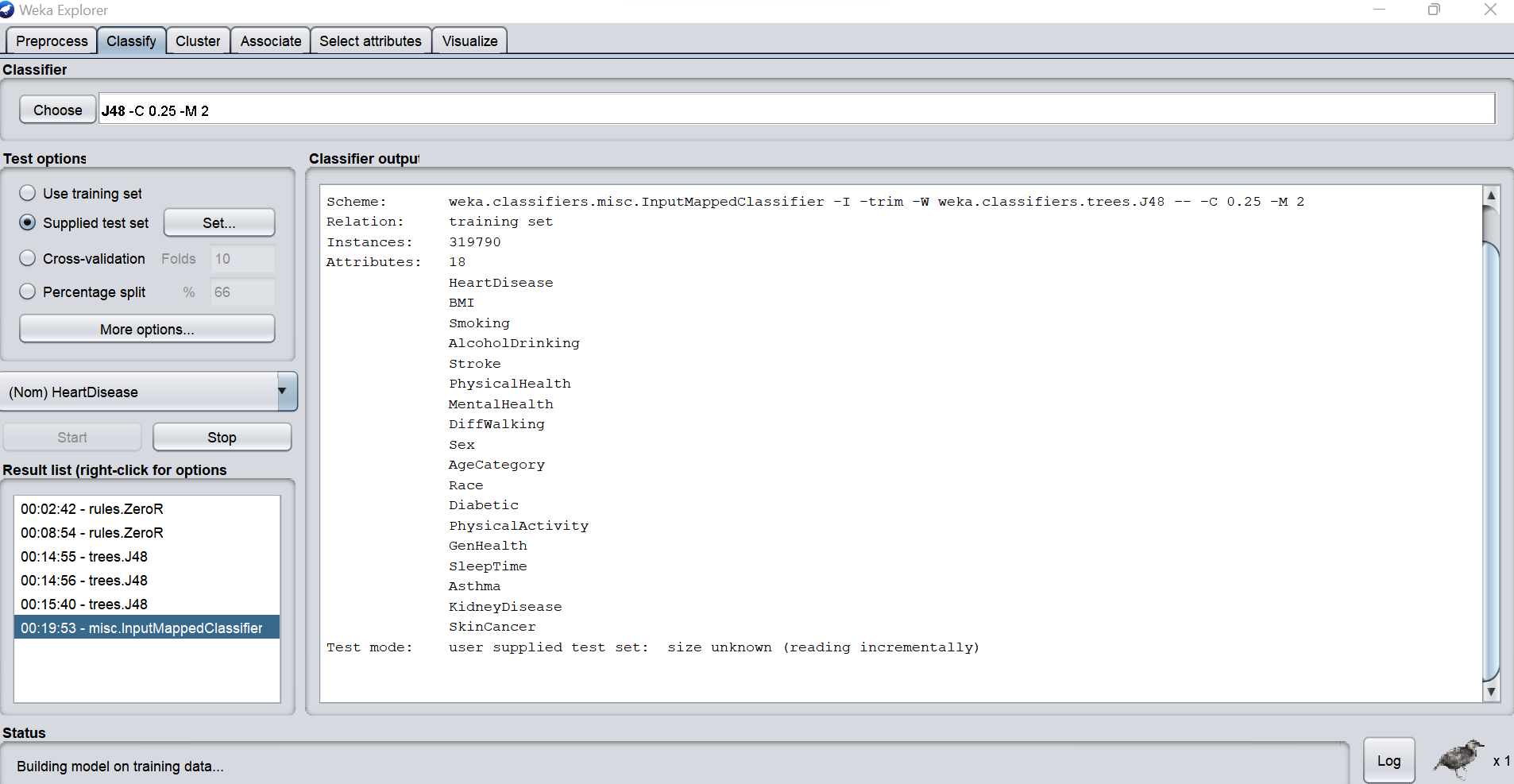
**Figure 10: Test set opened**

7.To make sure the test set works properly; it was made sure that the output predictions were in Plain Text form.

****

**Figure 11: Plain text format chosen for output prediction**

8. Then the start button was clicked.

****

**Figure 12: Test set run**

**RESULT OF TEST-DATA SET MODEL**

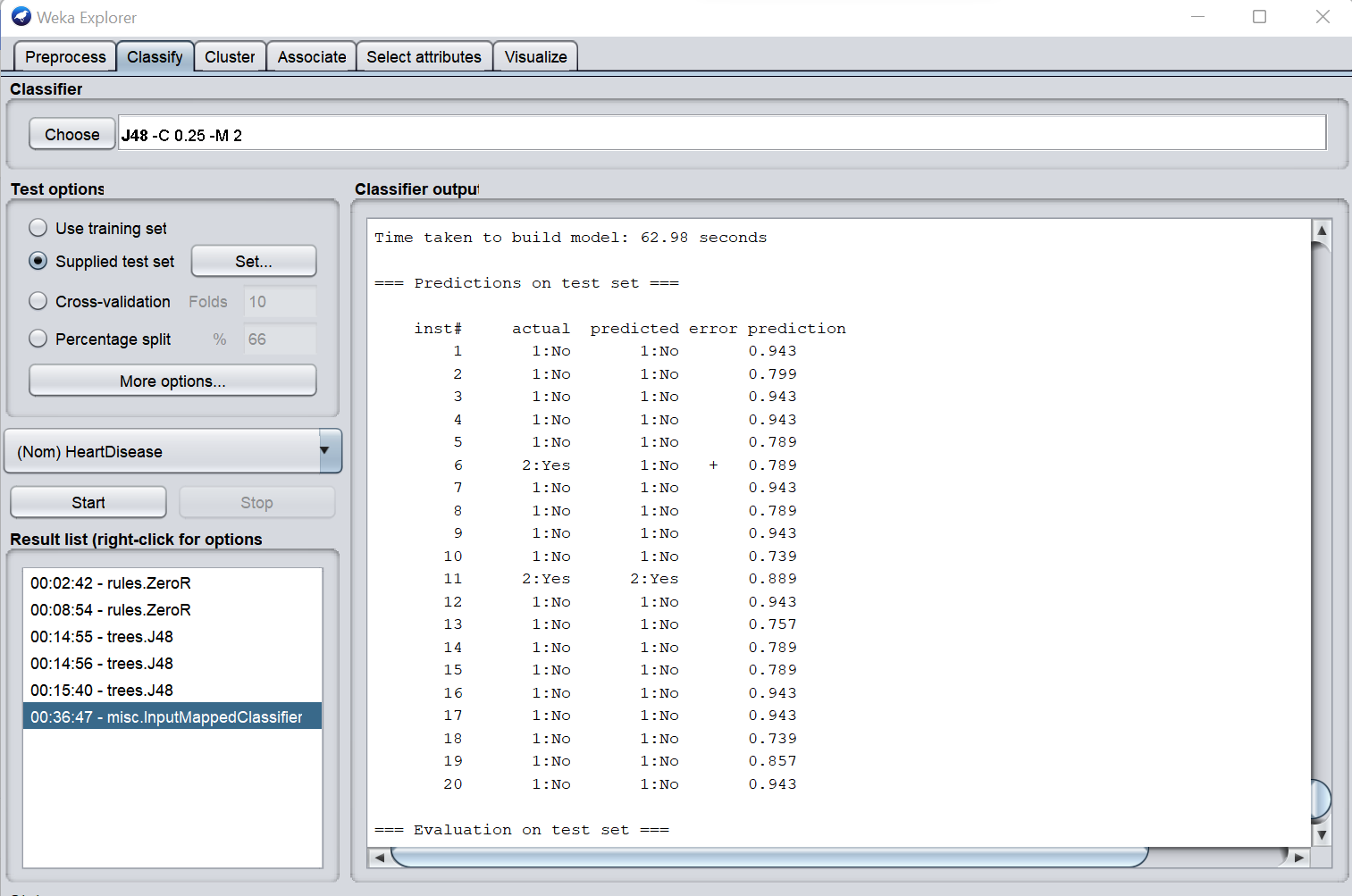
Once the start button was clicked, the output for the test data set came. In the result, the test mode was ‘user supplied test set’ it means the classifier is evaluated on how well it predicts the class of a set of instances loaded from a file which was inputted by the user. The total time taken to build the model was 62.98 seconds and among the 20 instances, there were 1 instance where error was found. Which means in the test model, that 1 instance were not properly classified and the Machine learning model predicted it after finding the error.

In this model, correctly classified instances % Value describes the amount of accuracy correctly classified instances provides by the algorithm. In this case, the percentage is 95% which is quite good.

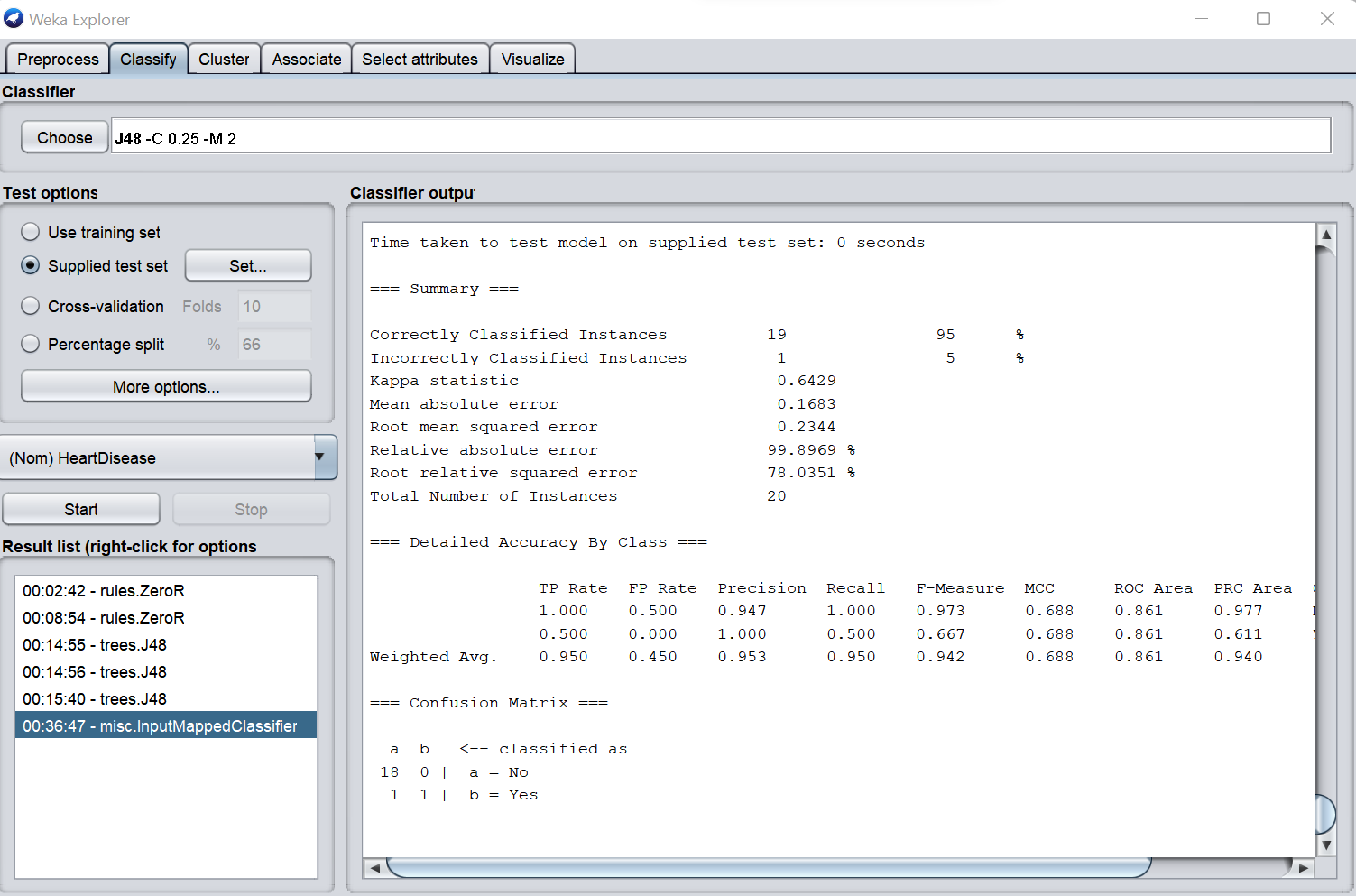
Incorrectly classified instances % Value describes how much incorrect instances are given by the algorithm. In this case, the percentage is 5%.

Mean Absolute Error (MAE): It can define as statistical measure of how far an estimate 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 0.1683.

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 modelled/ estimated. RMSE is used to measure the accuracy. It is ideal if it is small. In this case the RMSE is 0.2 which is ideal.

****

**Figure 13:Prediction result of test set**

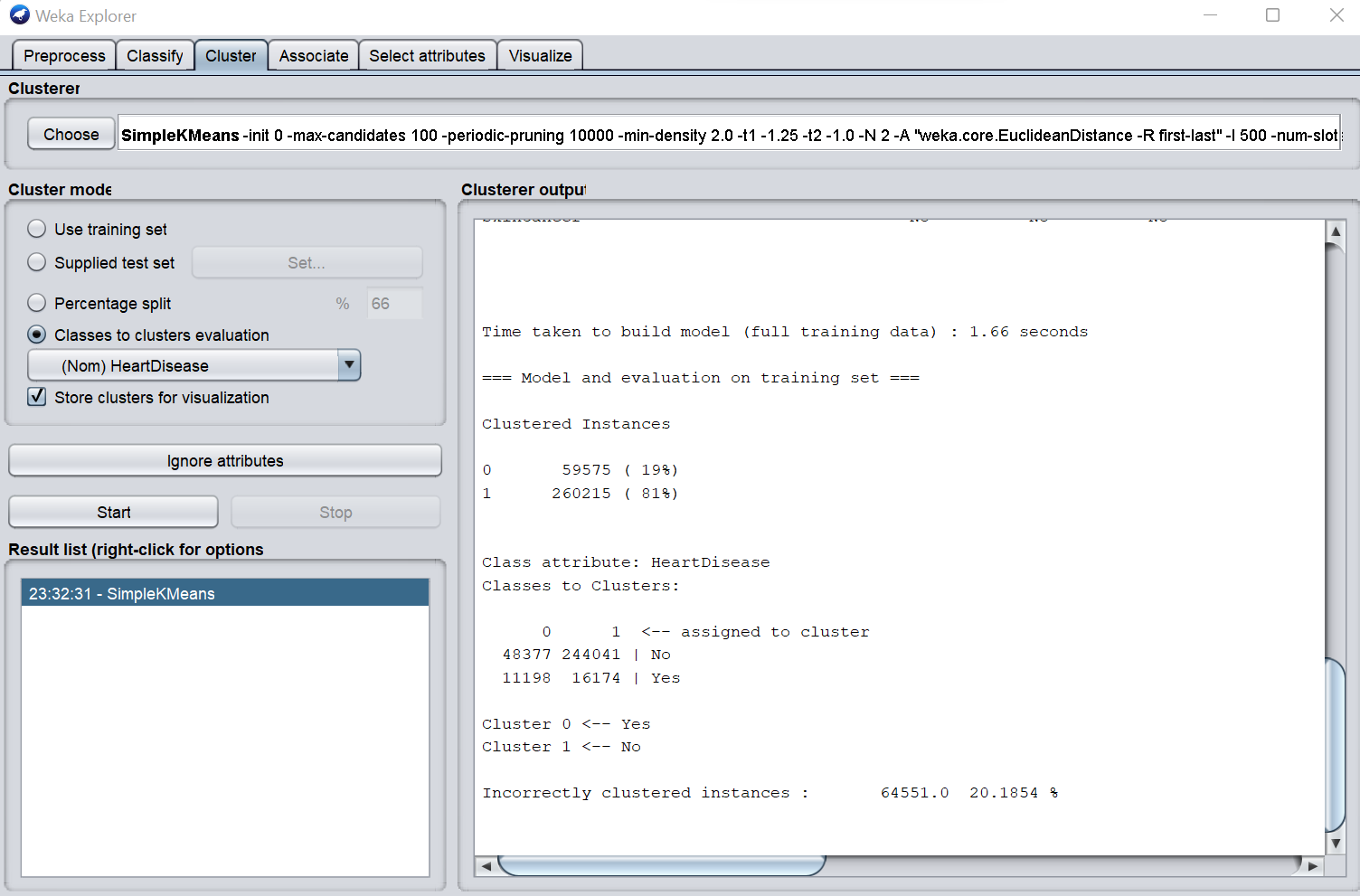
****

**Figure 14: Result of accuracy**

**For Unsupervised Learning:**

Applying K-Means Clustering Algorithm: A clustering algorithm finds groups of similar instances in the entire data set. WEKA supports several clustering algorithms such as EM, Filtered Cluster, Hierarchical Cluster, Simple K Means and so on.

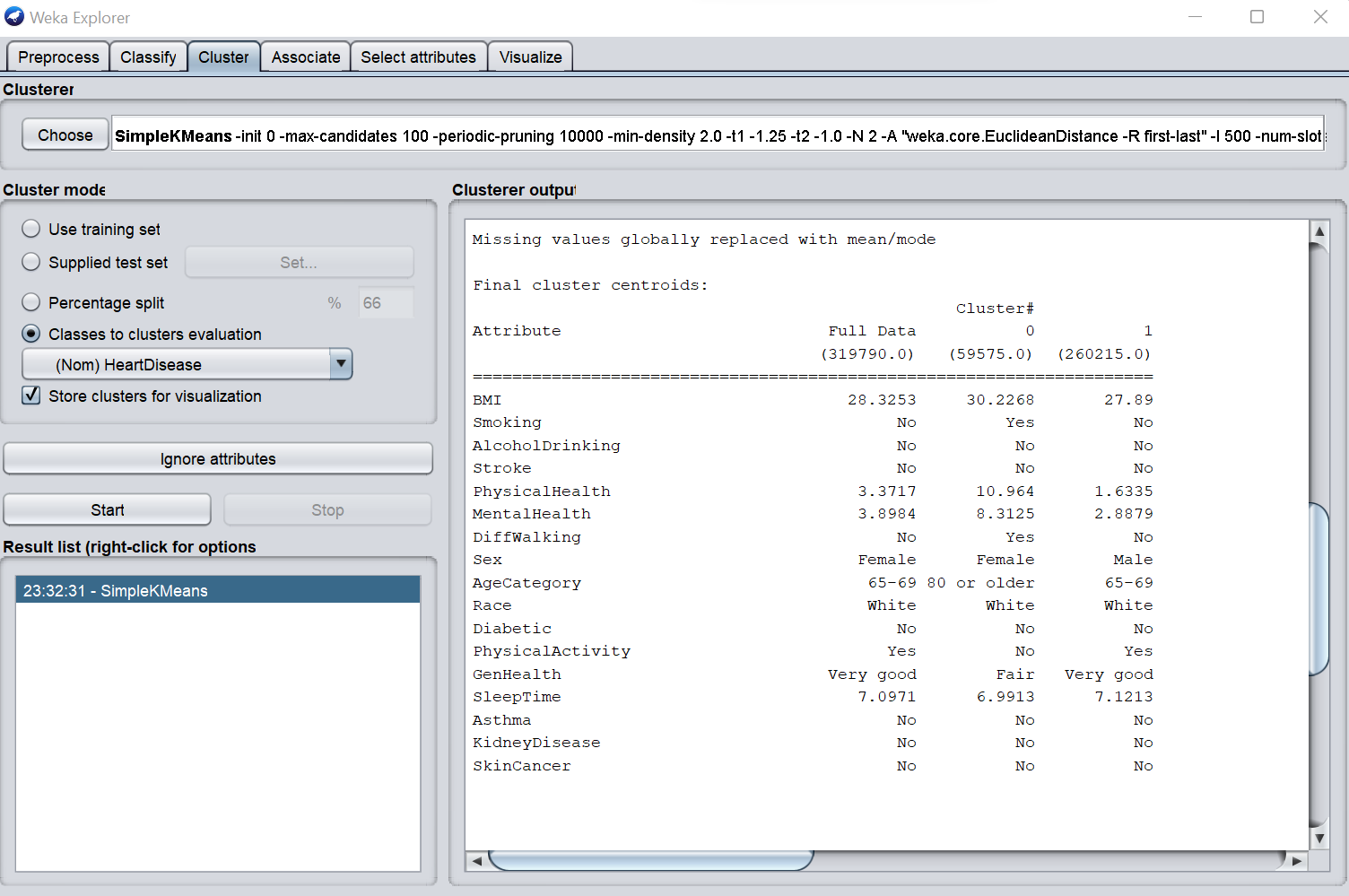
**Examining Output:**

****

**Figure 15: Clustered Instances**

**From the output screen as we see:**

1. There are 2 clusters detected
2. The cluster 0 represent Yes and cluster 1 represent No.

****

**Figure 16: Attribute of clusters**

**DISCUSSION:**

The purpose of this report was to find a suitable classifier for heart disease prediction which will classify the heart disease as accurately as possible and will be able to predict the class from the test data set. After applying three different classifier which are naive bayes and Decision tree algorithm which is from supervised learning. And the other one is k means clustering which is from unsupervised learning. From the supervised learning our best-chosen classifier for the data set is decision tree classifier with 91.4% accuracy. Then a training set extracted from original data set was selected to prepare a machine learning model. A prepared test set with 20 instances was used to test the model and finally the model accuracy is 95% for the prepared test data set. Creating training and test data set is an important concept in data science as it is used to improve generalization and minimizing over fitting. Then for the unsupervised learning we use k means clustering algorithm. As we see in the picture for cluster 0 which is yes the instance are 59575 and the percentage is 19% and then for the cluster 1 which is not the instances are 260215 and the percentage is 81% .Lets discuss the confusion matrix the( true positive) which means they have heart disease which is 48377.Then (false positive) which means they don’t have heart disease but the algorithm says they have which is 11198.Then the (true negative ) which means they did not have heart disease which is 16174.And the (false negative) which means patient has heart disease but the algorithm says they didn’t which is 244041.This also helps to give an unbiased evaluation about the accuracy of the model itself.

**REFERENCES:**

1. Key Indicator of Heart Disease data set: <https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease>
2. Training set and test set: <https://github.com/promadas/Key-indicator-of-heart-disease-dataset.git>