****

**AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH**

**Data Warehouse and Datamining**

**Section: C**

Group Name: Alpha4

**Group Member :**

|  |  |  |
| --- | --- | --- |
| **SL** | **Member Name** | **ID** |
| 1 | Rashedul Haque | 19-39431-1 |
| 2 | Maliha Tabassum | 19-394291 |
| 3 | Tahsin Adiba Nijum | 19-40428-1 |
| 4 | Tahsin Adiba Nijum | 19-40428-1 |

Contents

[Introduction: 2](#_Toc101437736)

[Result: 13](#_Toc101437737)

[Discussion: 15](#_Toc101437738)

# Introduction:

Heart disease data is a dataset that we have used in our project to apply classification. The dataset covers the information about the factors that have the greatest impact on heart disease. The dataset is from [www.kaggle.com](http://www.kaggle.com) It has 299 instances with a total of 18 attributes.

Dataset:

Table, Excel

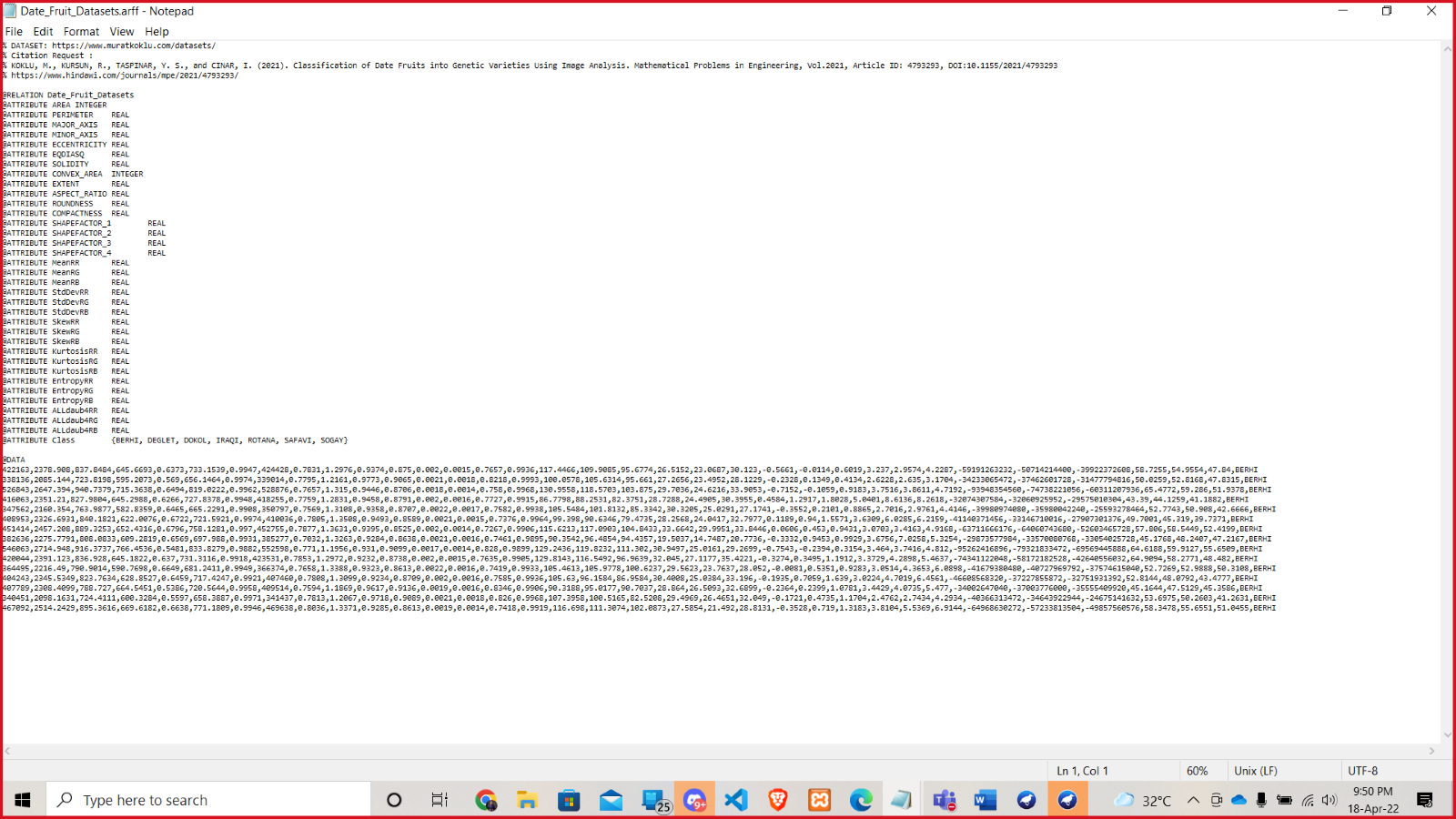
Description automatically generated

The dataset’s first 29 instances are (out of 300) are shown as a sample in the figure. For full dataset please go to <https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease>

Attributes:

1. HeartDisease: Yes/No
2. BMI: 16.6~58.54
3. Smoking: Yes/No
4. AlcoholDrinking: Yes/No
5. Stroke: Yes/No
6. PhysicalHealth: 0~30
7. MentalHealth: 0~30
8. DiffWalking: Yes/No
9. Sex: Male/Female
10. AgeCategory:
11. Race: White, Black, Asian, American Indian/Alaskan Native
12. Diabetic: Yes/No/No, borderline diabetes/Yes (during pregnancy)
13. PhysicalActivity: Yes/No
14. GenHealth: Very good, Good, Fair, Excellent, Poor
15. SleepTime: 2~15
16. Asthma: Yes/No
17. KidneyDisease: Yes/No
18. SkinCancer: Yes/No

We also collect another unsupervised dataset. We collected this dataset from <https://www.kaggle.com/datasets/muratkokludataset/date-fruit-datasets>



**Procedure:**

The following step will be to select categorize from Weka. Weka is a software package that includes visualization tools and algorithms for data analysis and predictive modeling. We will utilize Weka to analyze data for this task.

Graphical user interface, application

Description automatically generated

I have chosen 2 classifications from the WEKA classifiers list. Those are NaiveBayes and J48.

**1.Naive Bayes**

It is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. We must follow the following steps given below to apply naïve bayes classification method.

Step-1: Select the file to apply

Graphical user interface

Description automatically generated

Step-2: Select the Naïve Bayes Classifier

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface

Description automatically generated

Graphical user interface, table

Description automatically generated with medium confidence

Graphical user interface

Description automatically generated

Graphical user interface, text, application

Description automatically generated

**2.Decision Tree-(J48)**

Decision Tree is the classification technique that consists of three components root node, branch (edge or link), and leaf node. Root represents the test condition for different attributes, the branch represents all possible outcomes that can be there in the test, and leaf nodes contain the label of the class to which it belongs. The root node is at the starting of the tree which is also called the top of the tree. J48 is an algorithm to generate a decision tree that is generated by C4.5 (an extension of ID3). It is also known as a statistical classifier. Now we have to do following steps to apply j48 decision tree.

Step-1: Select the dataset and apply

Graphical user interface

Description automatically generated

Step-2: Select J48 Classifier

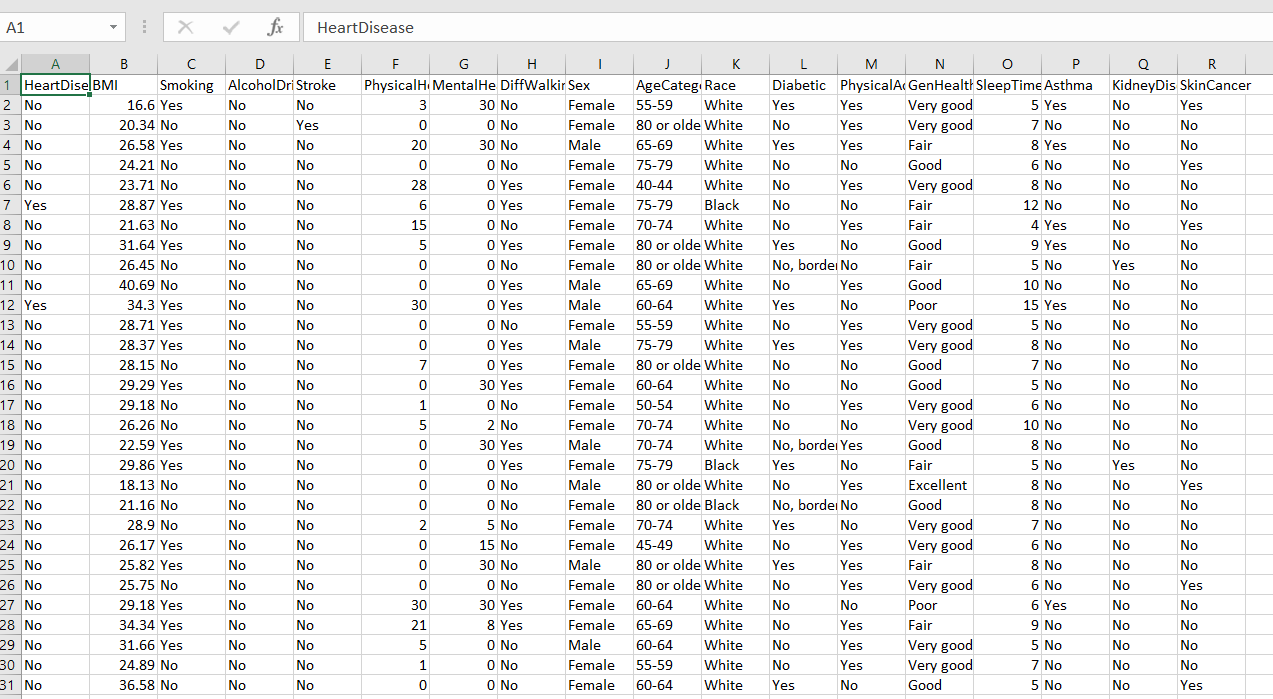
Graphical user interface, text, application, email

Description automatically generated

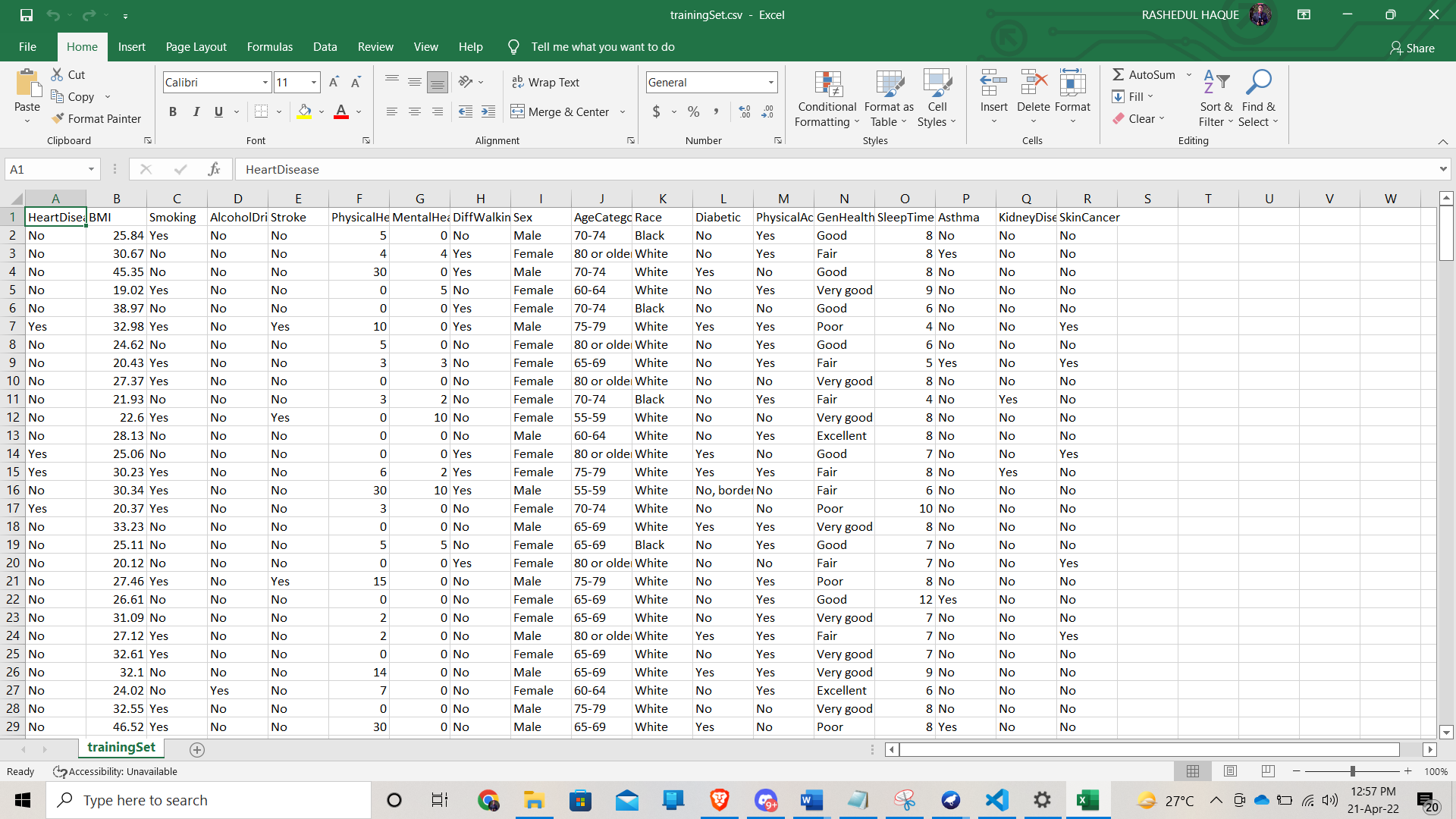
Text

Description automatically generated

Then we will make a test data set from the above dataset



Then we will supplied the test data set in main dataset in weka

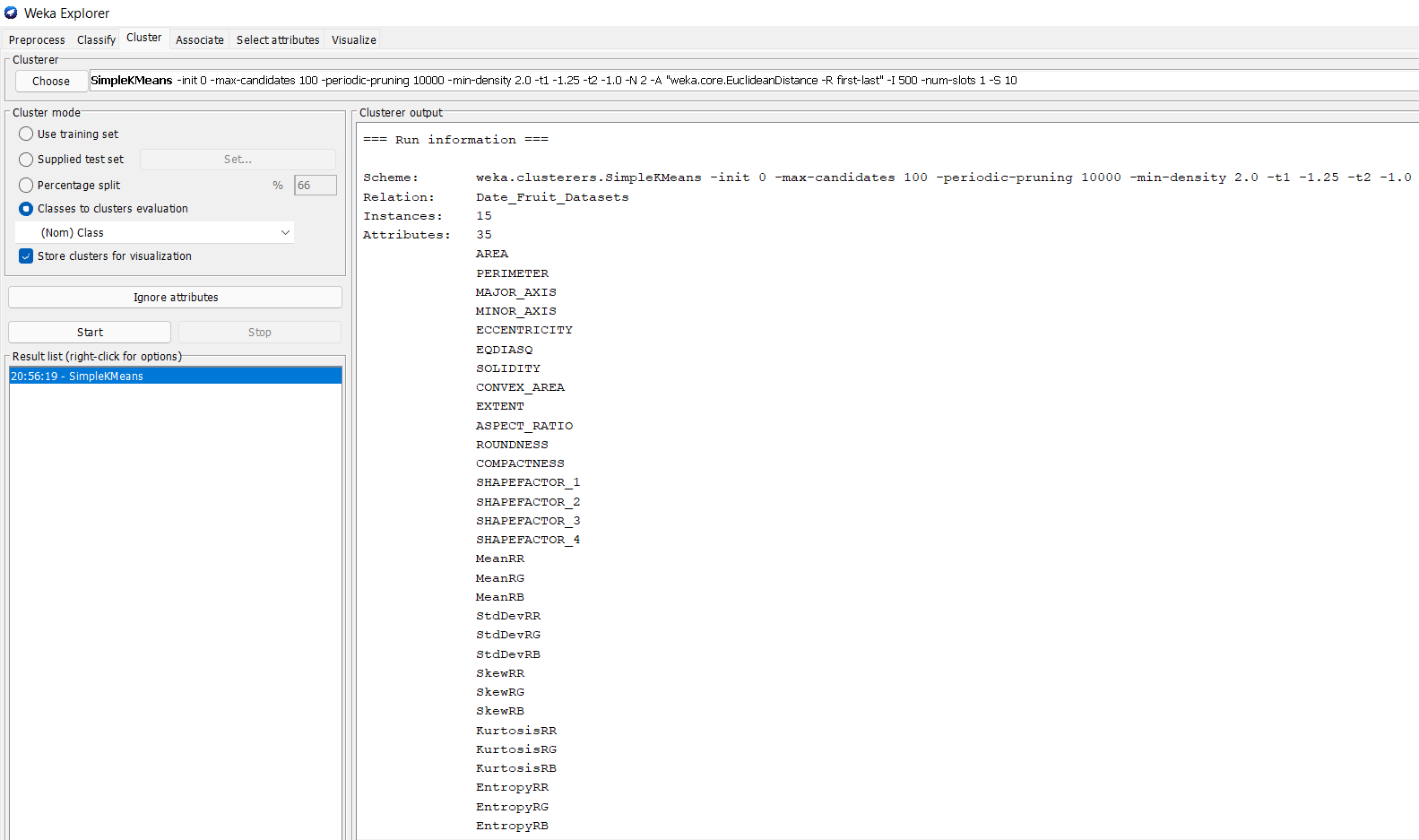


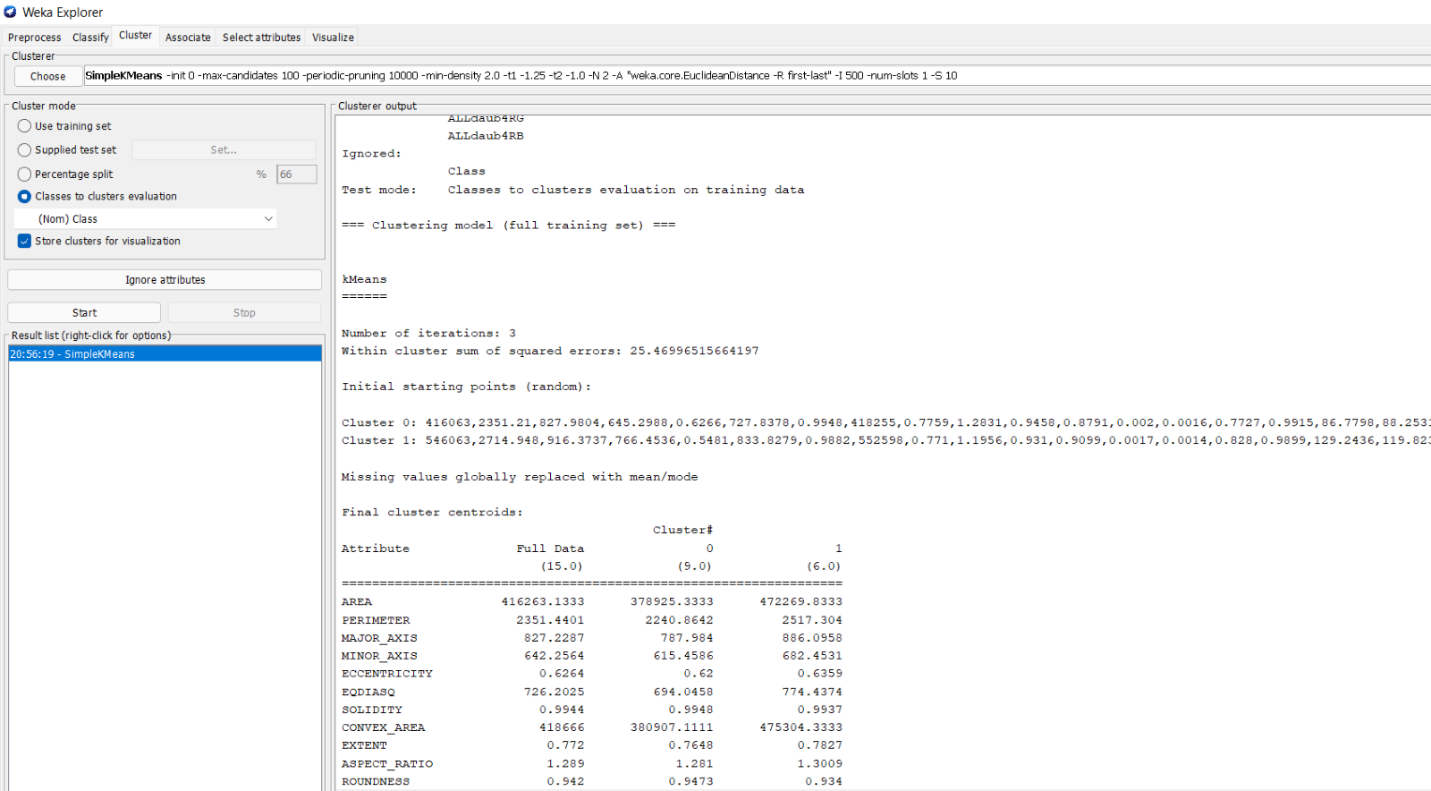
And it is our training set.

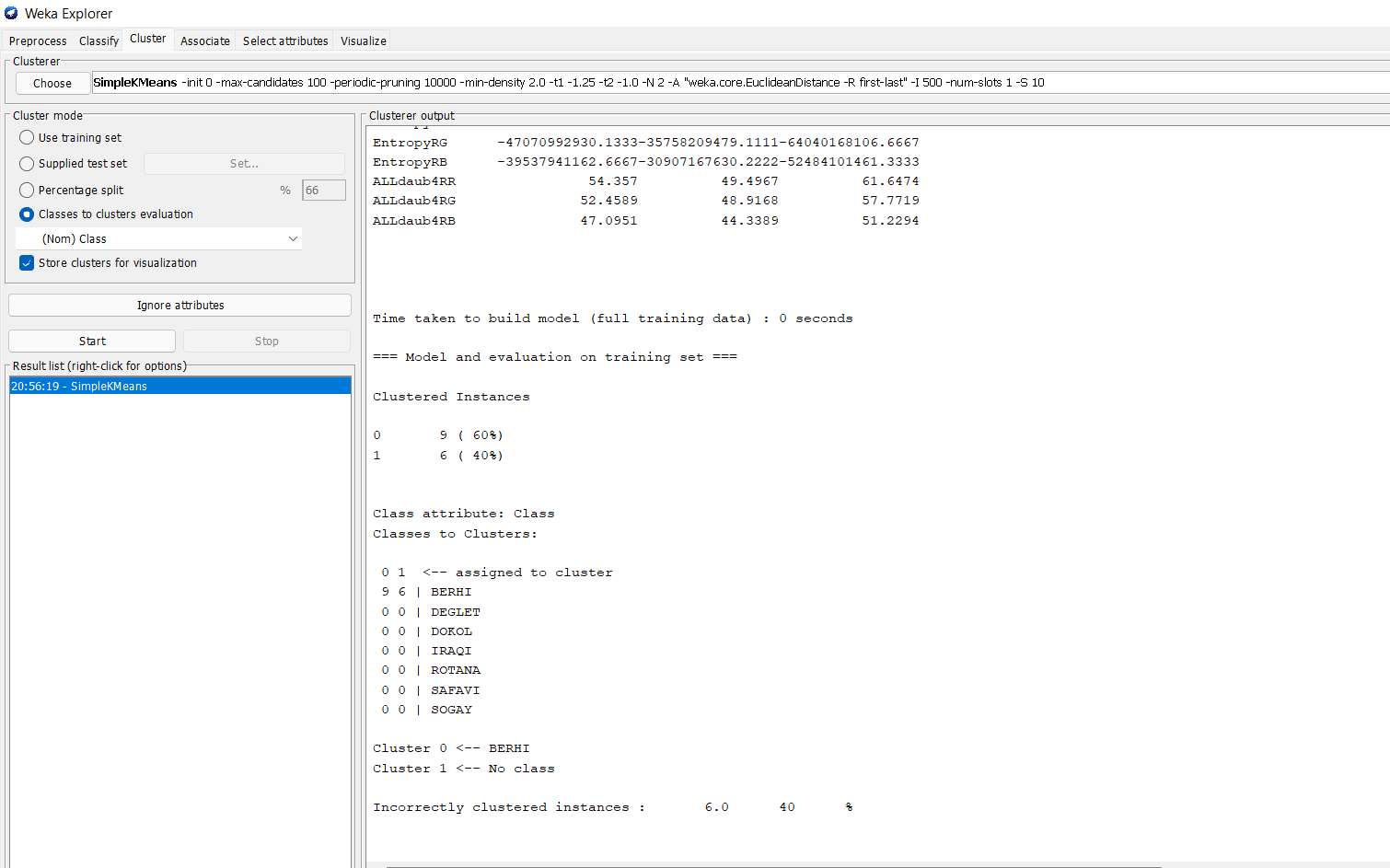
Through this training set we will supplied our test set

For task 3---

Here we applied k means clustering algorithm in weka.



  
We will used k-means algorithm which is available in this cluster then simpleKMeans. There are 15 instances and we had 35 attributes and it has ignored the attributes class. Test mode selected Classes to clusters evaluation on training data. It has taken 3 iterations in order to make this clusters.  K-means  algorithm set to 2 cluster based on specific set on combinational data. Then this part is the final cluster centroids. For the First clustered which is named a 0, for this combination data it generates first centroid and second clustered which is name a 1 this is also  combination data, it generates second centroid.

  
We have total 15 instances. In  the first class there are 9 instances and in the second class we have 6 instances. For that cluster 0 was named at BERHI and clustered 1 was named at no class. We are getting incorrect instances. There are about  6 instances or rows which is wrongly clustered . Incorrectly clustered 40% instances. 

# Result:

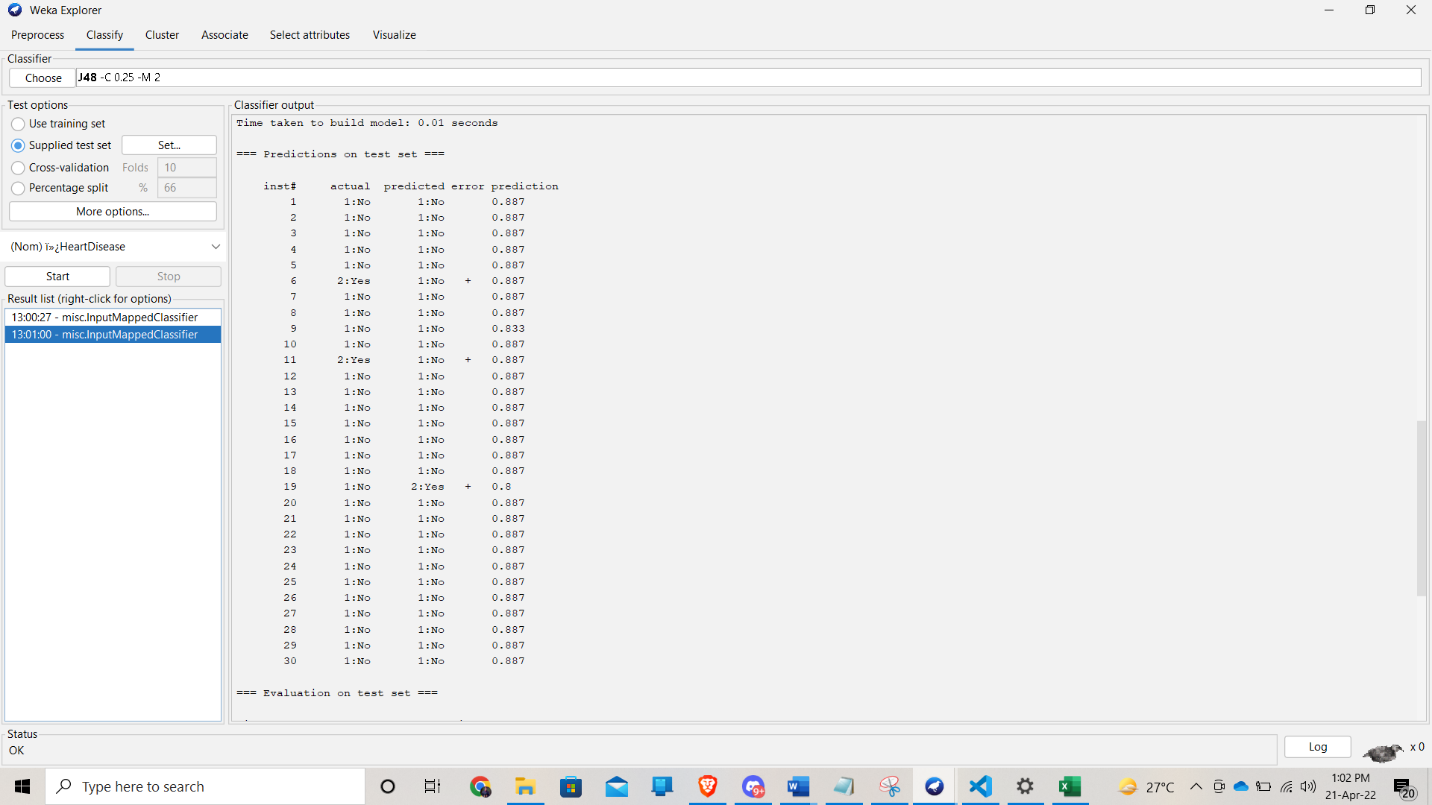
From task 1---

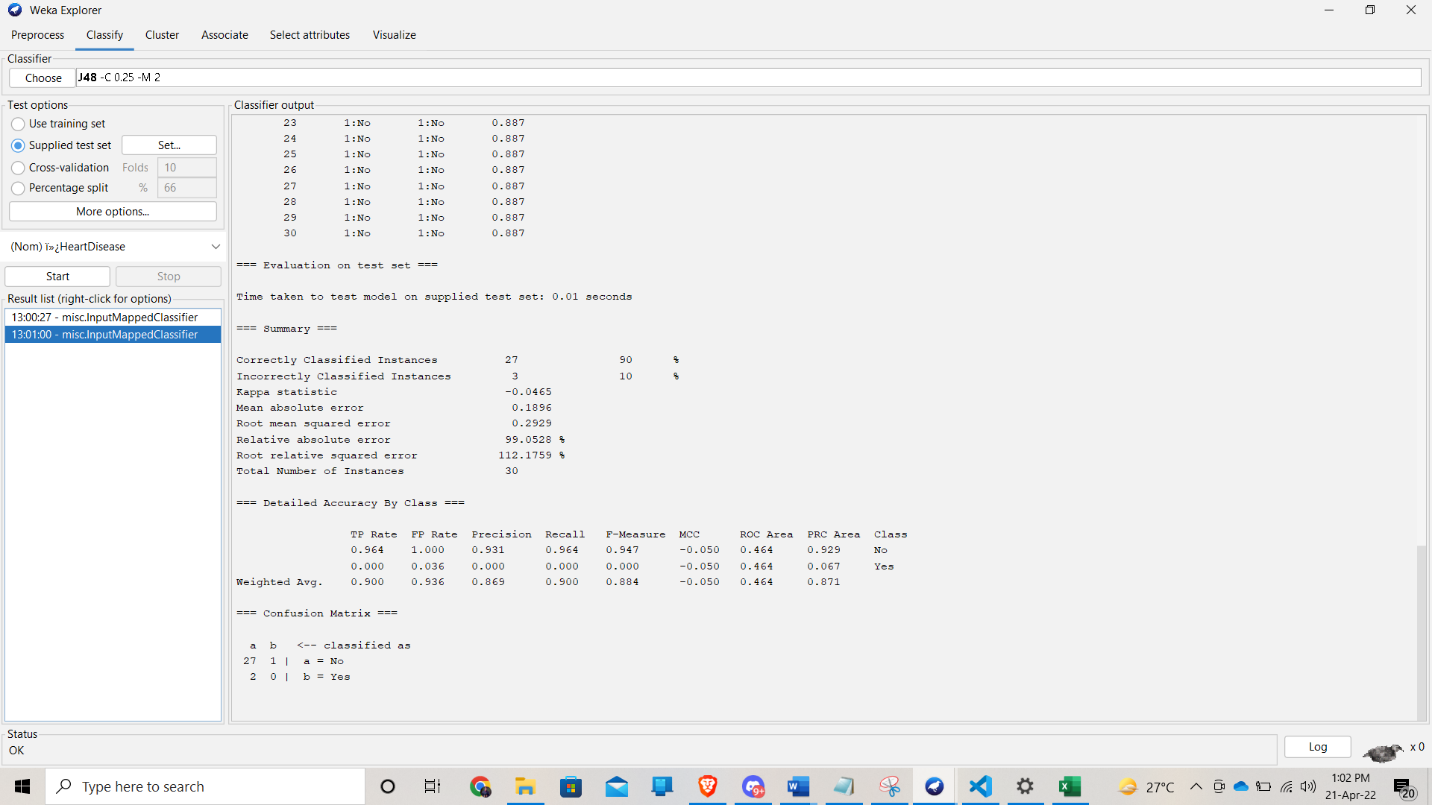
After using Naive Bayes classifiers, the result reveals that 229 instances were successfully identified, with a rate of 76.5886 %. And with 70 instances erroneously categorized, the percentage is 23.4114 %. In J48 trees, 233 cases were successfully identified, with a 77.9264 % success rate. Also, 66 is wrong. The rate of classified incidents is 22.0736 %.

As a result, the accuracy of the J48 classifiers is higher. The number of right cases is significantly higher than the other two. So J48 is the best classifier.

**Best classifier is J48 with 1.3378% higher correctness from Naïve Bayes.**

From Task 2----





From Task 3----

Here we can see 2 cluster. In cluster 0 there 9 instances (60%) and in cluster 1 there 6 instances (40%) those are the incorrectly instances.

# Discussion:

The purpose of this report was to find a suitable classifier for the Heart disease dataset which  
will classify the drug as accurately as possible and will be able to predict the class from test dataset.  
After applying three different classifier which are decision tree and naïve Bayes classifier for the dataset is decision tree classifier with 78% accuracy.A prepared test  
dataset was used to test the model and finally the Models accuracy was 93.33% for the prepared test dataset.

We also choose a data fruits unsupervised dataset. We used k means clustering in this dataset and can cluster 60% instance correctly and 40% incorrectly.