DATA MINING

REPORT ON WEKA

Assignment-1

Submitted by
Shubham Sharma (1001964524)
Lavanya Srinivasan (1002040671)
Prem Atul Jethwa (1001861810)

Introduction:

WEKA is an open-source tool that aids users in data preprocessing, mining, and visualization. It is known for implementation of a lot of machine learning algorithms and applying them to real-world data mining problems.

The homepage for this tool is shown below:

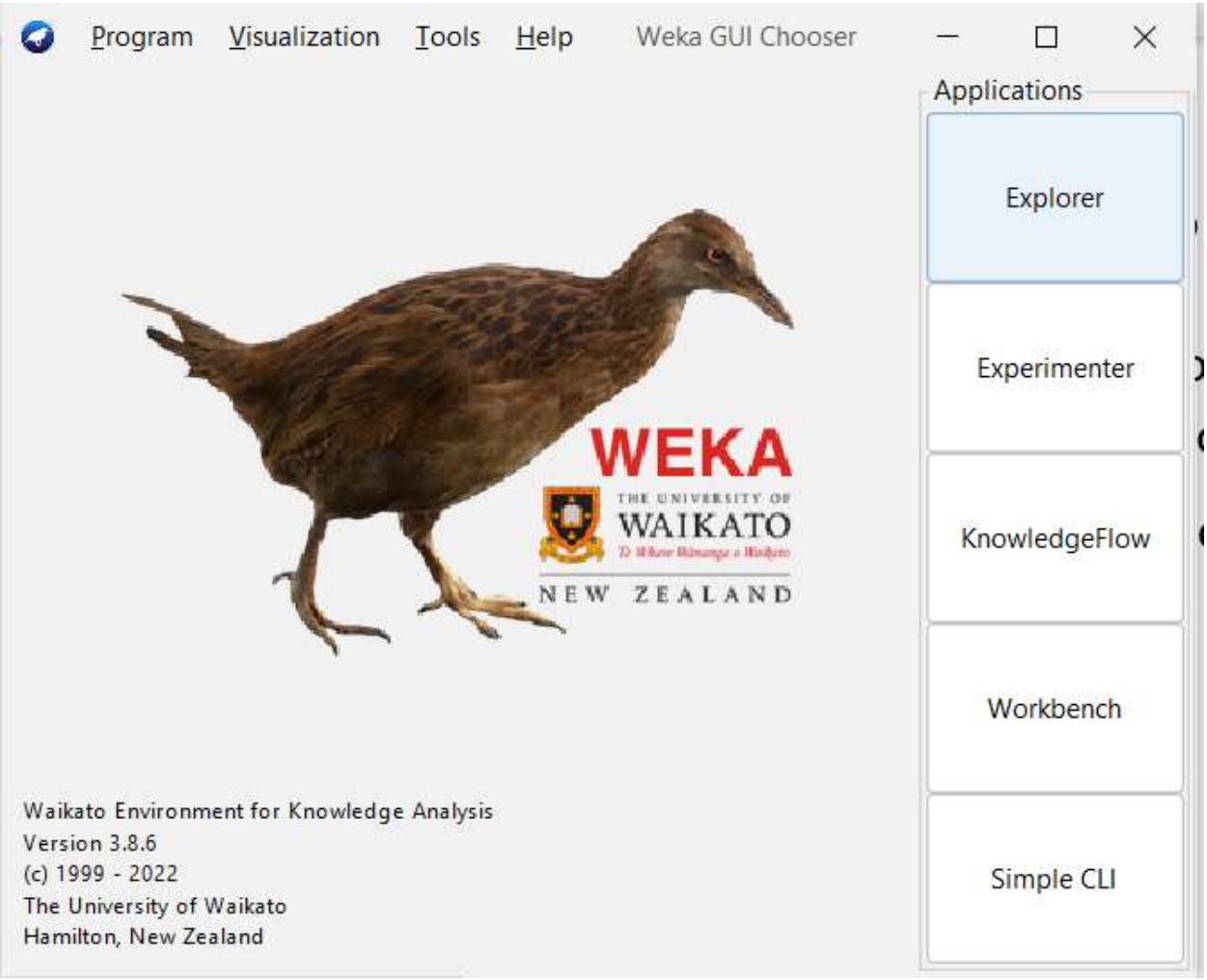

Figure 1: WEKA Homepage

In order to load data in WEKA, we need to click on the Explorer tab on top right and then navigate as Open file... → Go to the directory containing dataset → Change the file type from .arff to .csv and select the WEK Dataset as below:

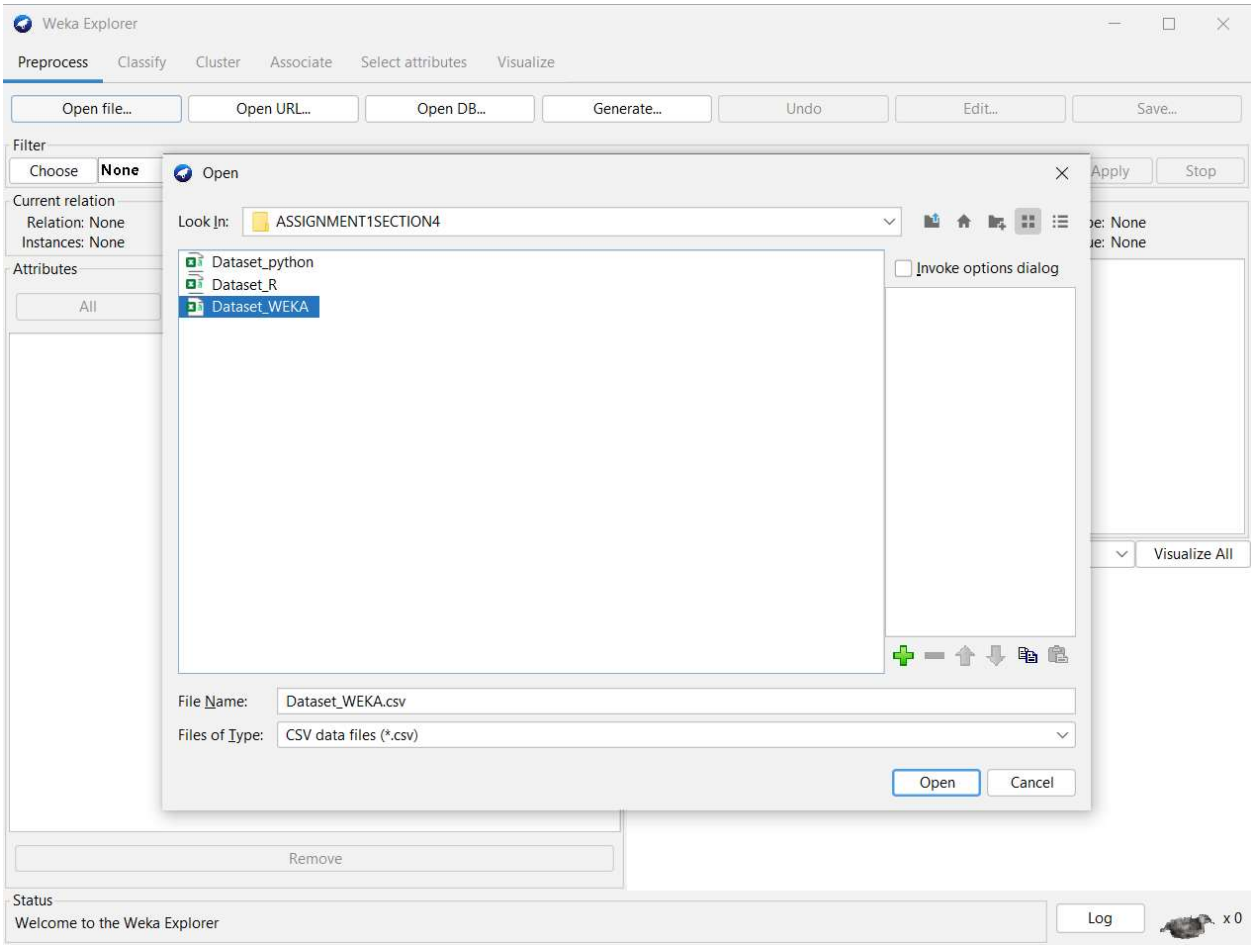

Figure 2: Retrieving Data

Once the file is opened, the WEKA tool does some default analysis and generates results such as the list of attributes and also visualizes them as graphs based on values they have as below:

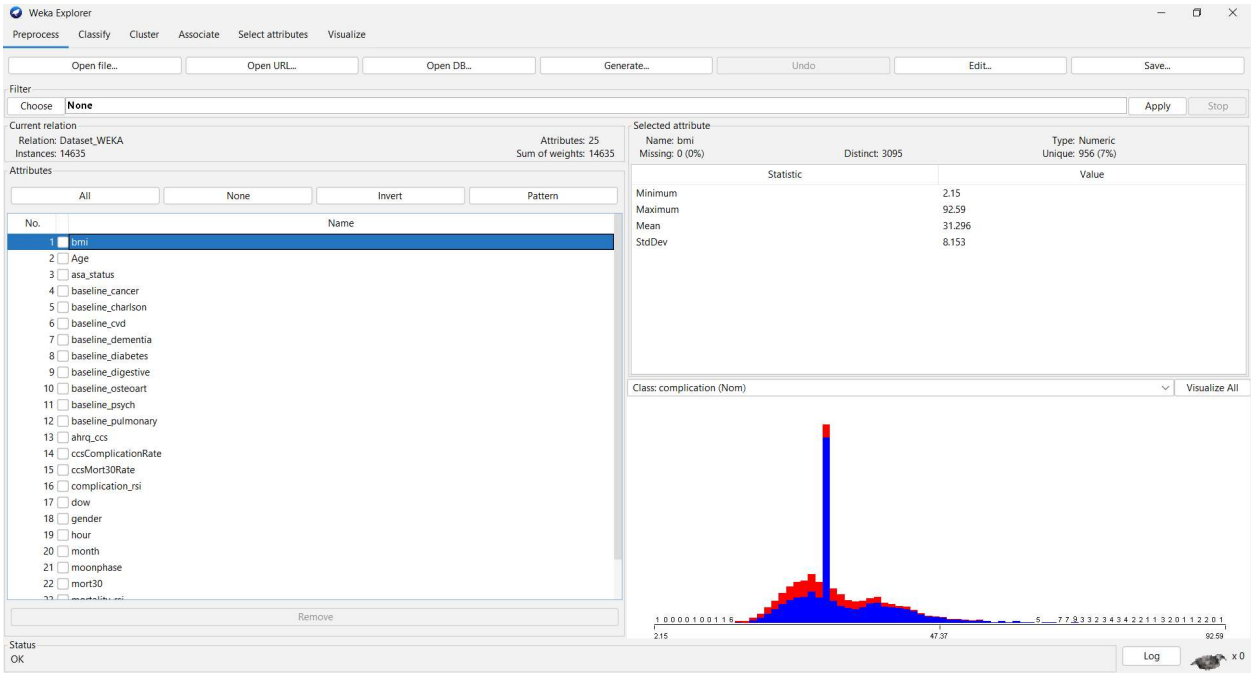

Figure 3: List of attributes

WEKA recognized a total of 25 attributes and 14635 Instances. It also provided basic statistics for each attribute. In the image, as the BMI attribute is selected, the minimum value is shown to be 2.15, Max – 92.59, Mean – 31.296 and Standard Deviation- 8.153.

These analysis and visualizations mark the first step of preprocess in WEKA.

Data Classification:

WEKA can perform classification on a given Dataset based on a variety of available classifiers:

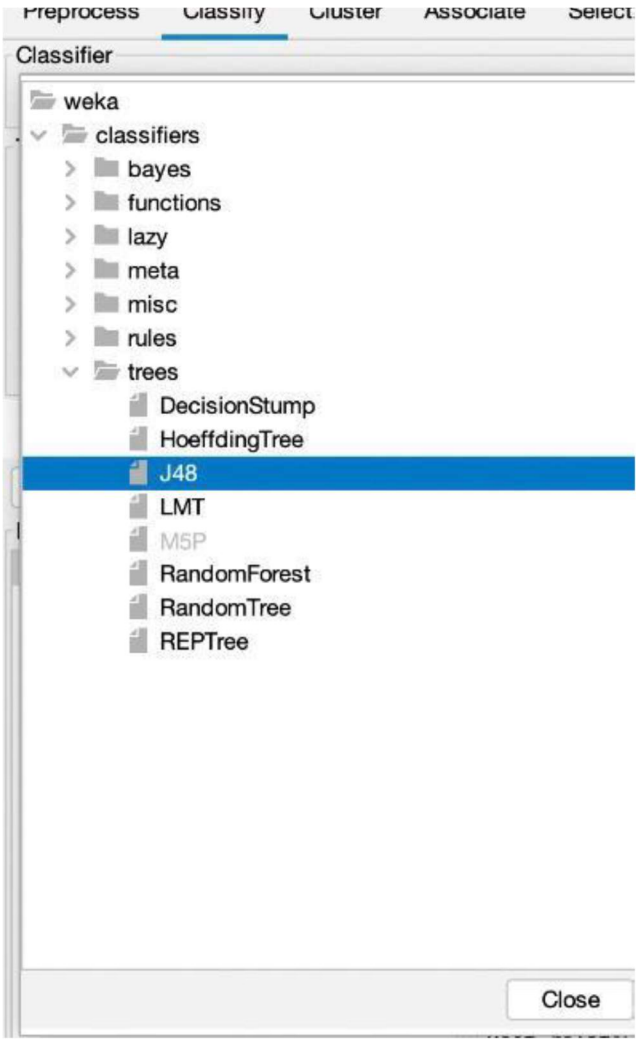

Figure 4: Classifiers in WEKA

Since we don't have a separate training set, we would try using cross validation or Percentage split to divide existing data into training and testing data.

The first classifier to be considered are trees. **Decision trees** are also known as **Classification and Regression Trees (CART)**. They

work by learning answers to a hierarchy of if/else questions leading to a decision. These questions form a tree-like structure, and hence the name.

The J48 is considered to be the best classifier in WEKA and hence, we will try to classify the Dataset using this algorithm in WEKA. We tried classifications based on default parameters:

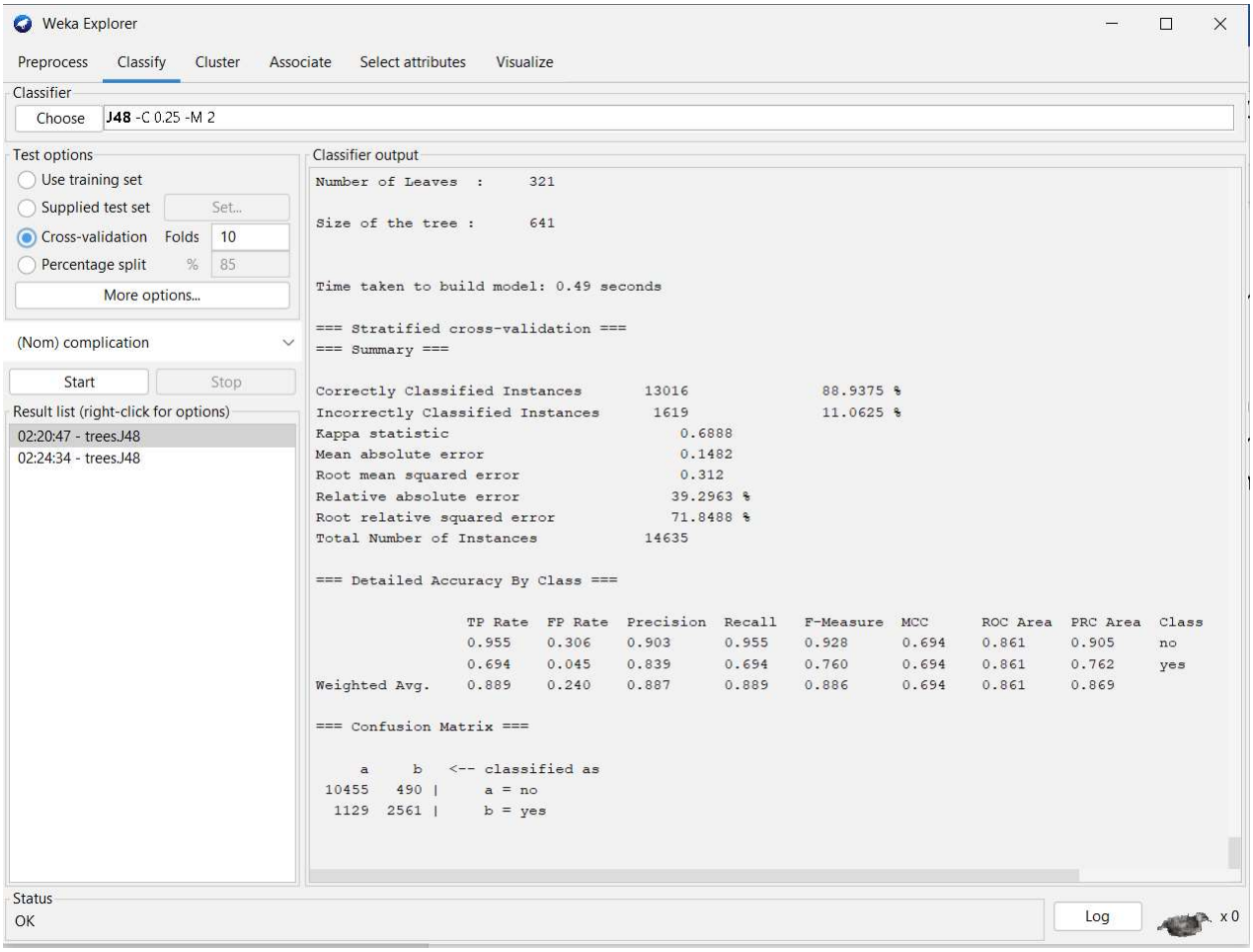

Figure 5: J48 classifications based on default parameters.

We can also divide the Dataset to use percentage of the same as training set and the remaining part for testing set. We tried doing an 85:15 split to generate the following classification:

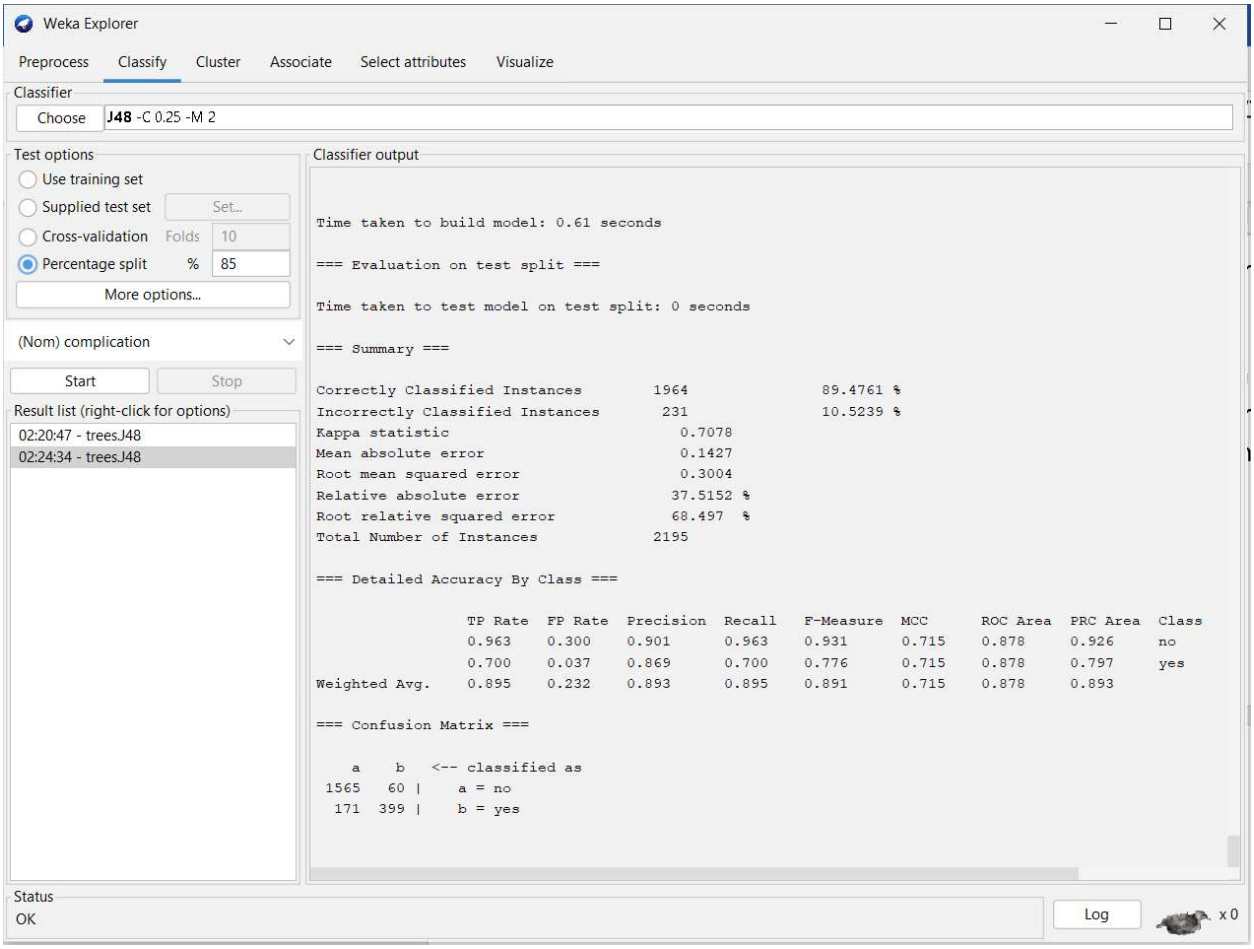

Figure 6: J48 classifications based on Percentage Split.

To visualize the classification into a decision tree, we can right click on the results generated on the result list and select visualize tree as follows:

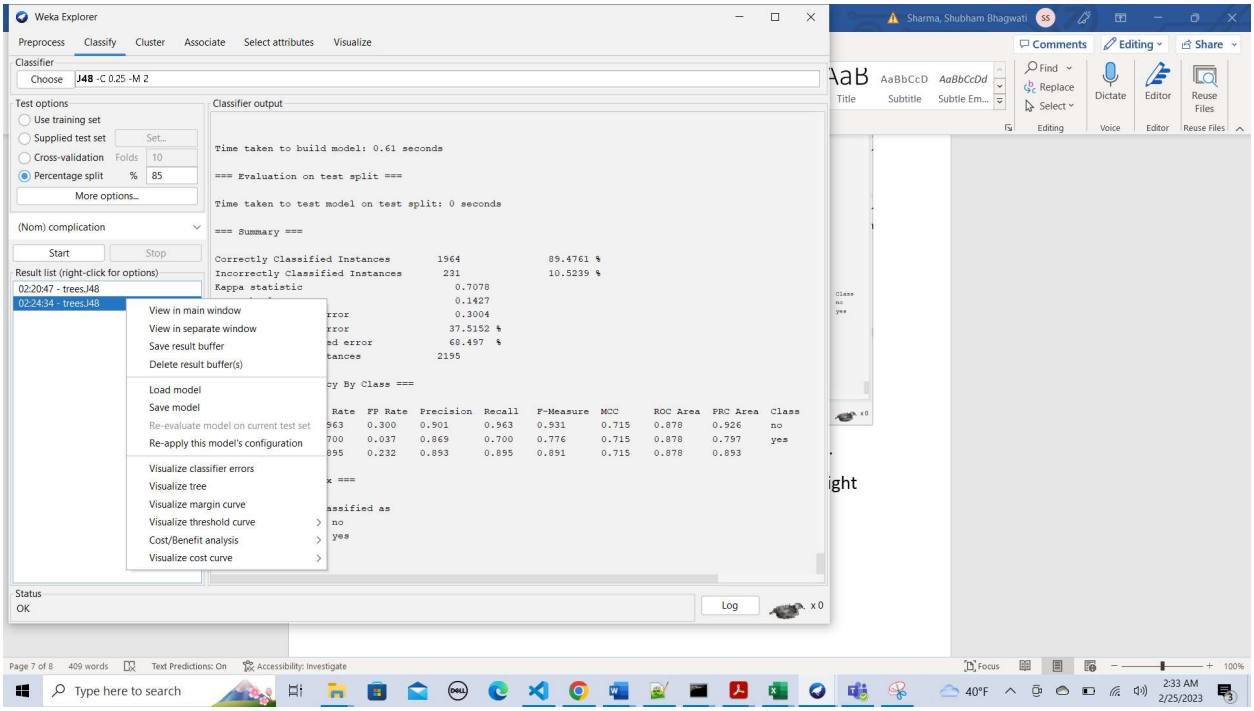

Figure 7: Generate Tree visualization.

This will generate the tree structure as follows:

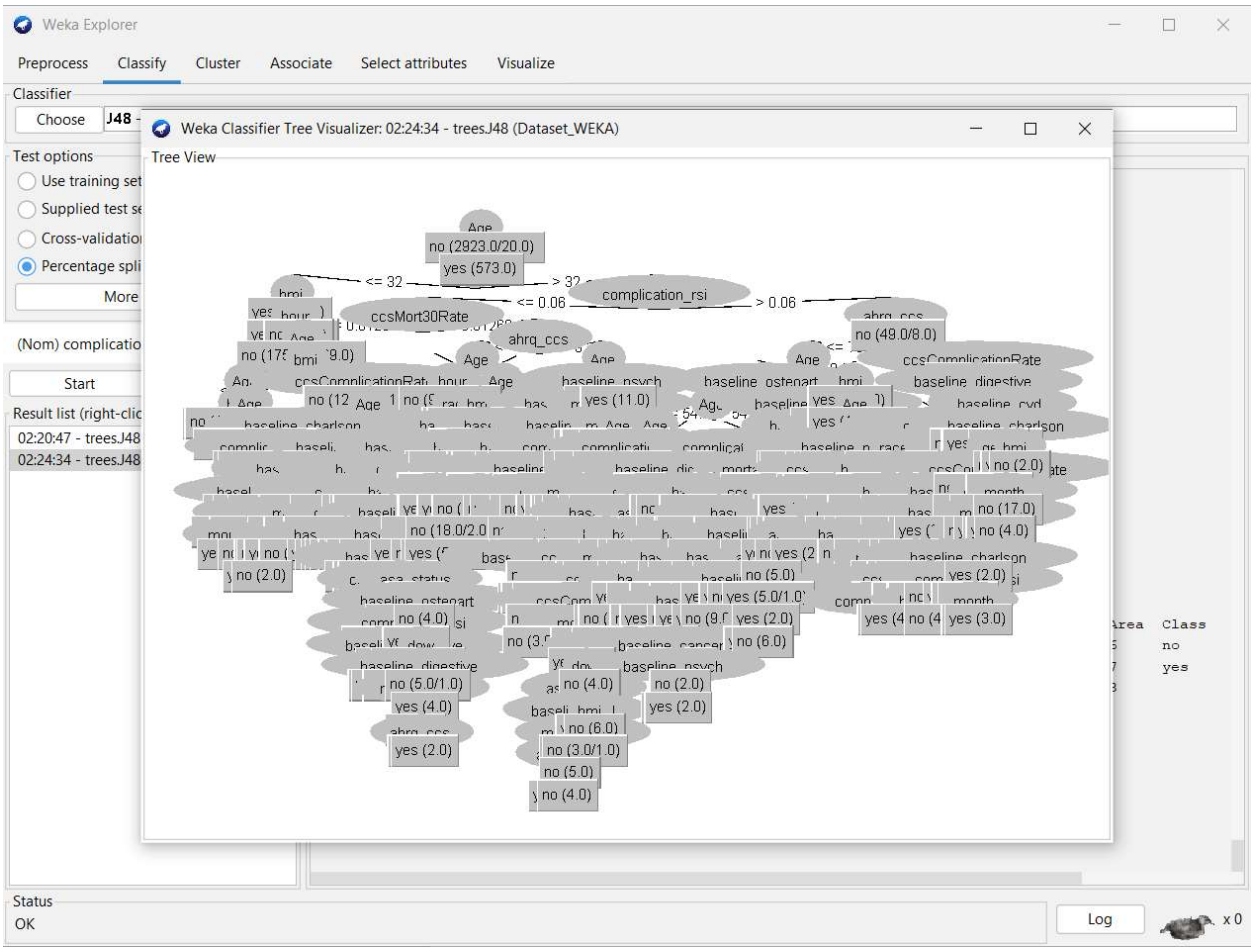

Figure 8: J48 visualized Tree.

We also tried classifying the Dataset using RepTree. It builds a decision tree using gain/variance information and only sorts numeric attributes once:

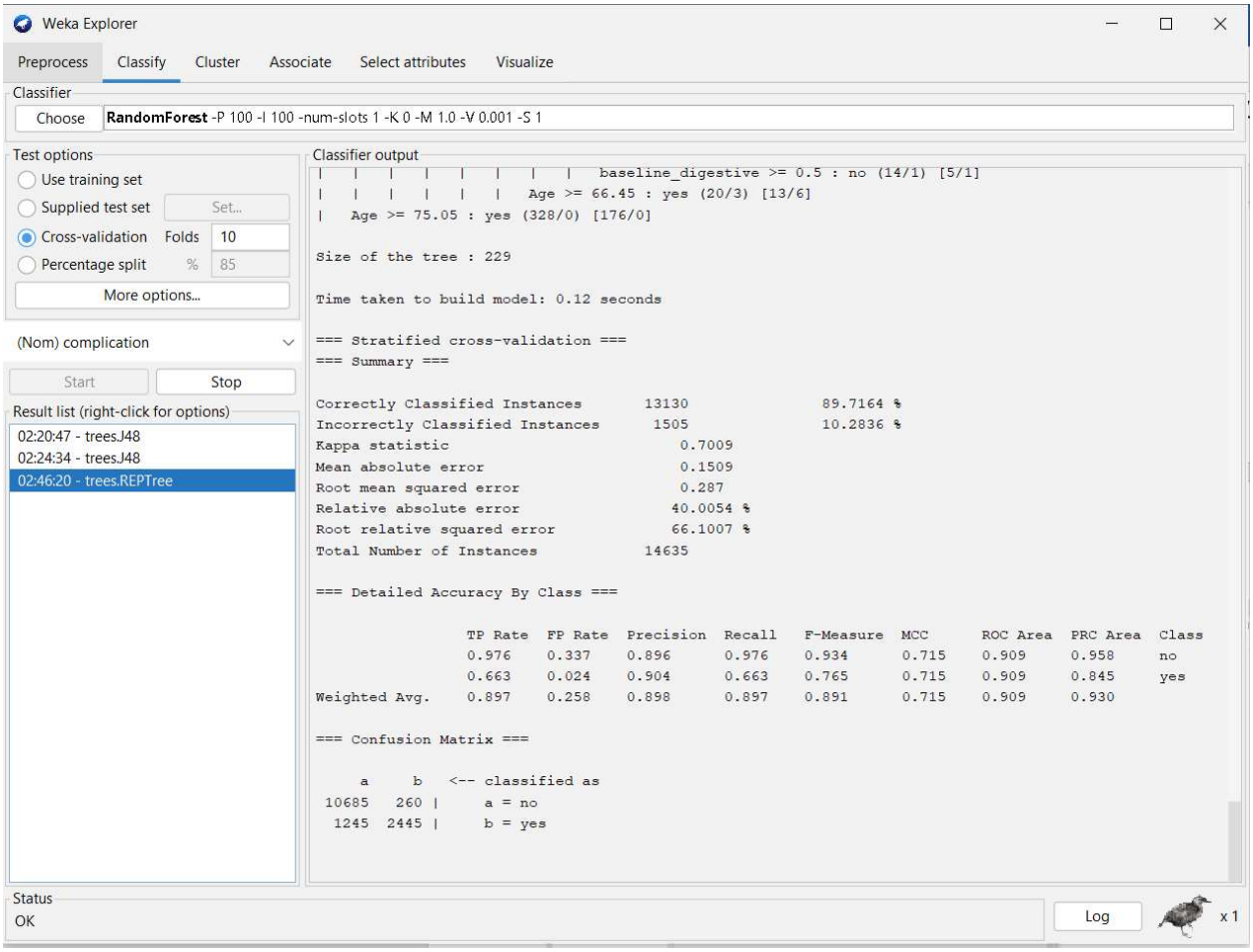

Figure 9: RepTree classifications based on default parameters.

Clustering:

Clustering helps us to get a better understanding of the dataset by dividing it into groups of similar kind of data.

We selected the EM(Expectation Maximization) clustering that assigns a probability distribution to each instance which

indicates the probability of it belonging to each cluster from the group.

EM clustering algorithm was able to detect 5 group in the given dataset. It also provided the mean and std dev statistics for each attribute in a given cluster:

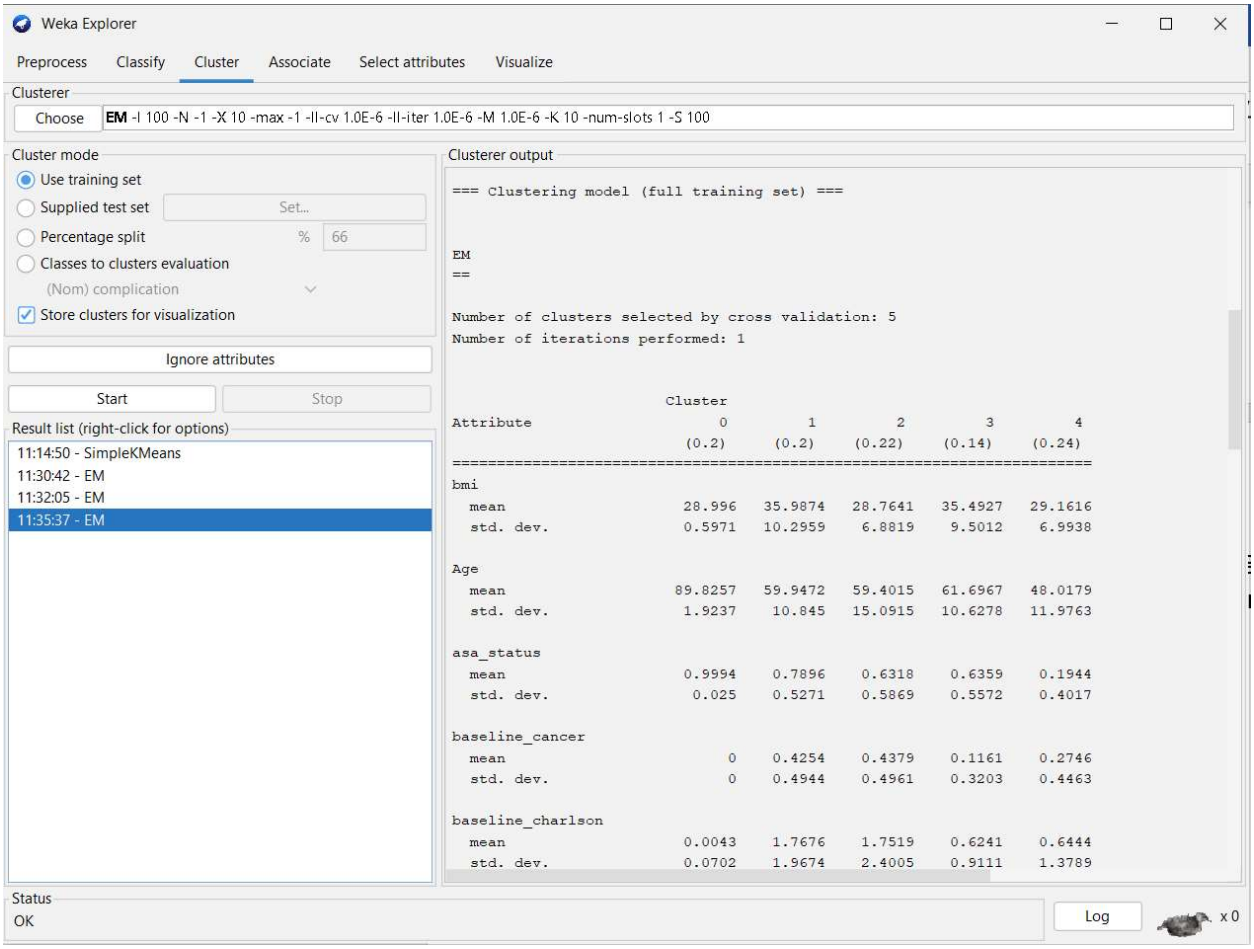

Figure 10: EM clustering on Dataset.

It also gave a summary of percentage data belonging to each dataset:

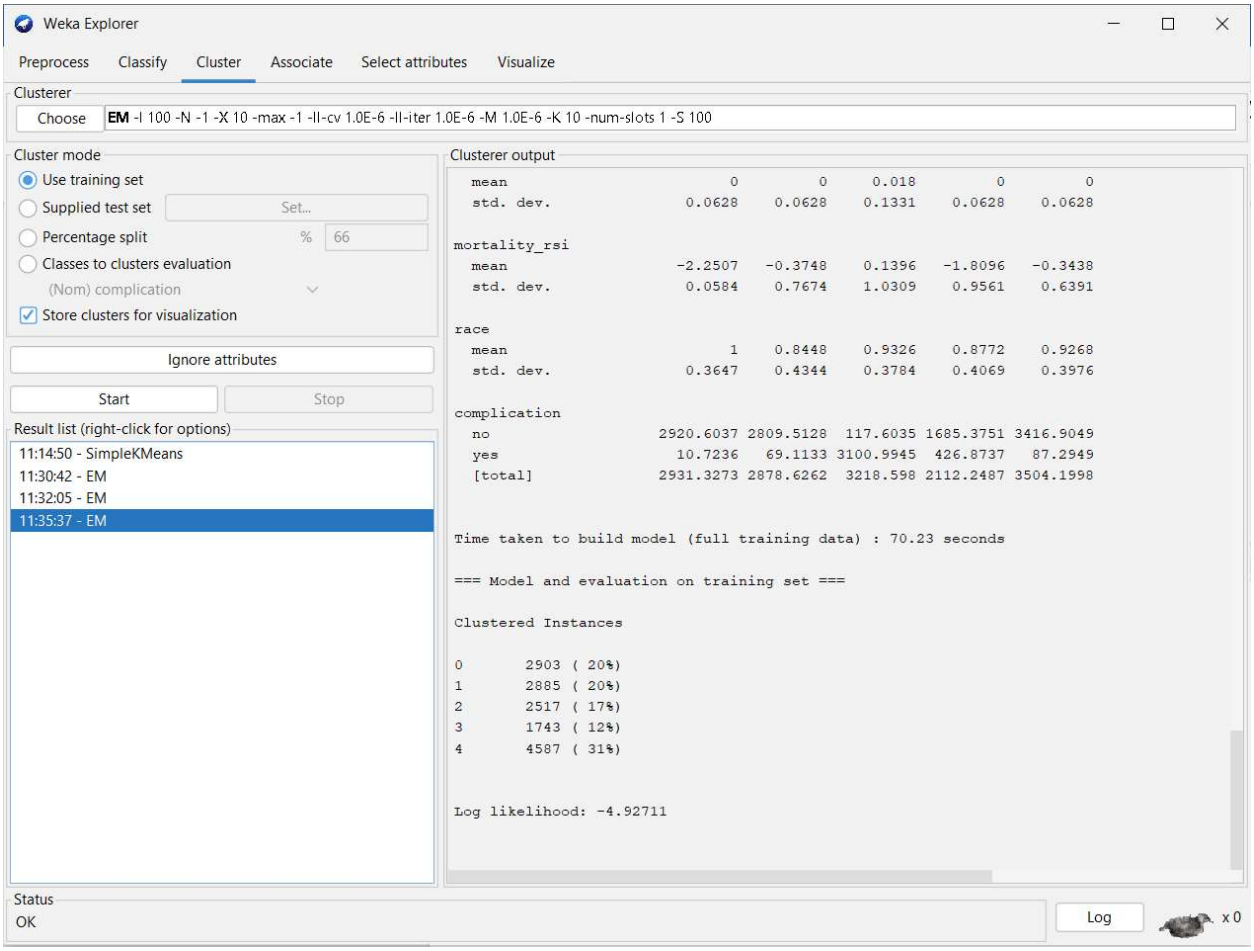

Figure 11: Summary of EM clustering.

Glimpse of Data:

We can also visualize the clustered data into a graph. We selected the Age vs BMI for plotting the graph and found the following result based on different colors used for each cluster groups:

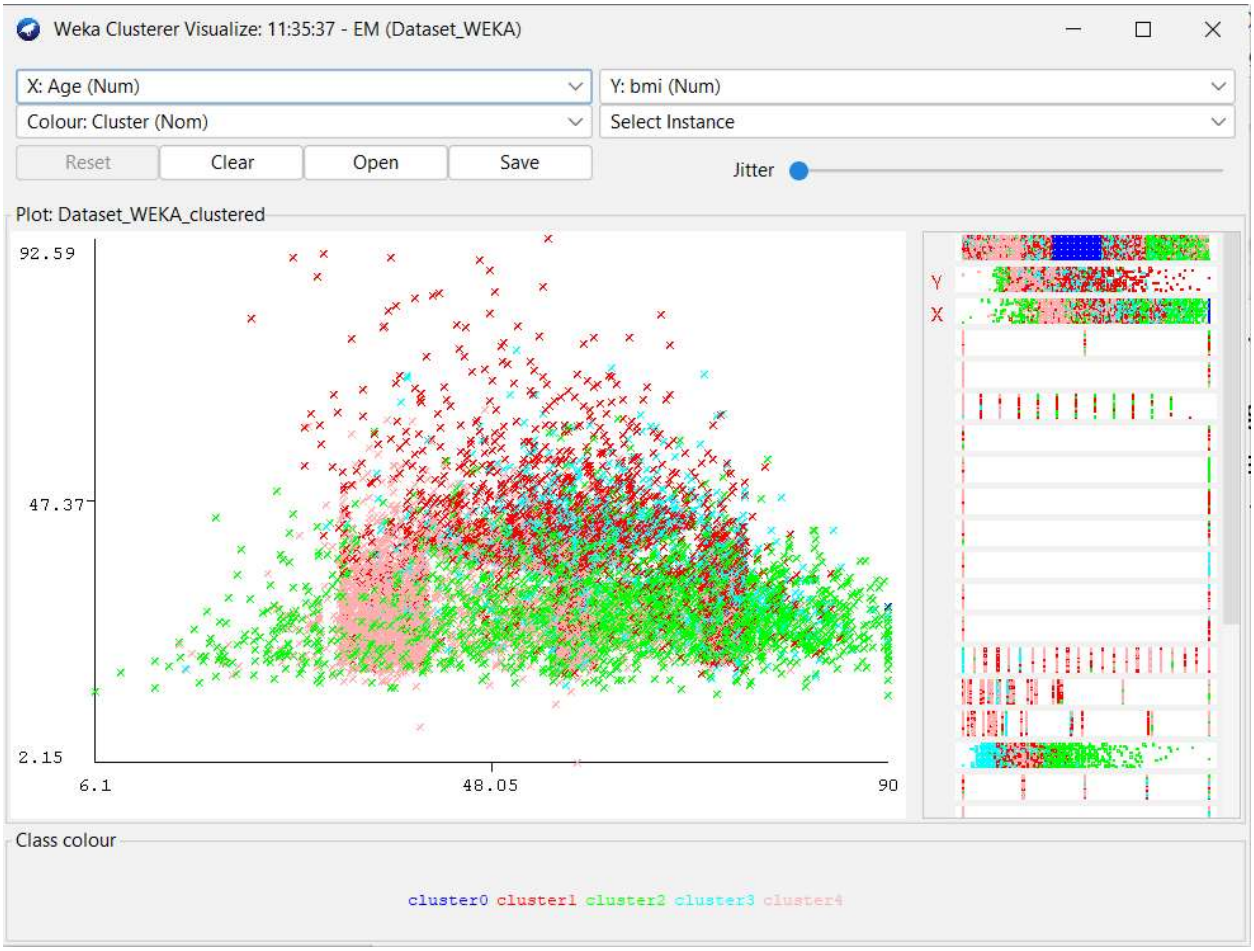

Figure 12: Age vs BMI plot for EM clustering.

From the above graph, we can observe that the BMI for most middle-aged people ranges 20-35.

Visualization:

WEKA provides visualization interface for getting graphical insights on the given Dataset. It can be used to solve complex issues. The jitter feature adds random noise to the data points in order to spread it out and might aid in detecting hidden datapoints.

We can visualize the data based on any attribute combinations for the X and Y axis.

Here we retrieved an Age vs Mortality graph:

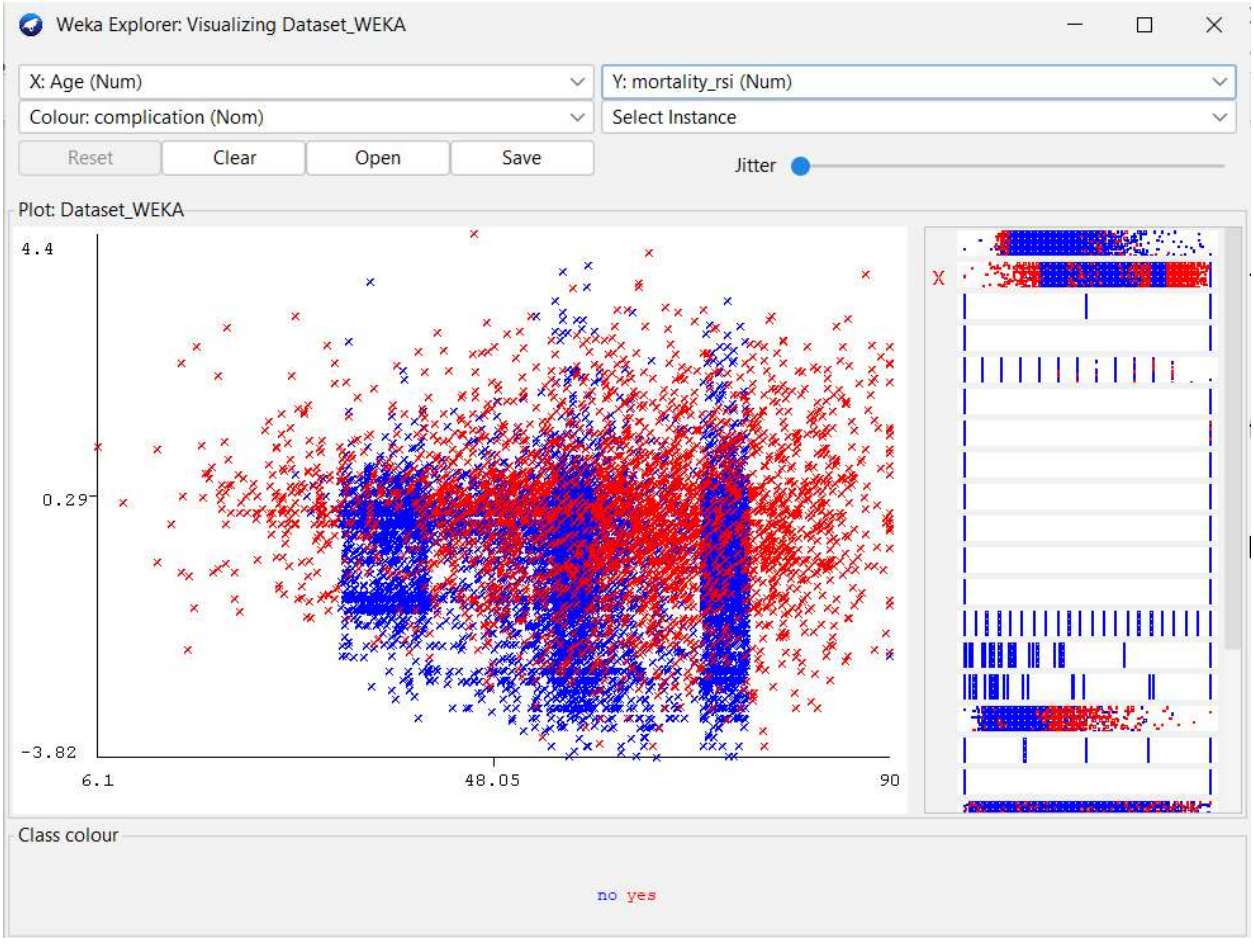

Figure 13: Age vs Mortality visualization graph.

The high-level conclusion drawn from the graph could be that the mortality data is dense between the age groups 60-80.

Conclusions:

- WEKA provides advanced techniques to process a dataset and also visualize the same.
- Easy to get a rough idea about every attribute by preprocessing.
- We can generalize the data and cluster them into groups of similar data types.
- Can follow a visualized graph for clear insights on related attributes by plotting the required attributes in the X-axis and Y -axis.
- High BMI is one of the common factors in people with complications. Even age plays an important factor for determining these complications.

Contribution by team members:

Python – Lavanya Srinivasan

R language – Prem Atul Jethwa

Weka – Shubham Sharma

References:

1) https://www.analyticsvidhya.com/blog/2020/03/decision-tree-weka-no-coding/
2) https://www.tutorialspoint.com/weka/what_is_weka.html
3) https://youtu.be/U63ExiTJMic