DATA MINING

ASSIGNMENT 1-WEKA REPORT

Submitted by

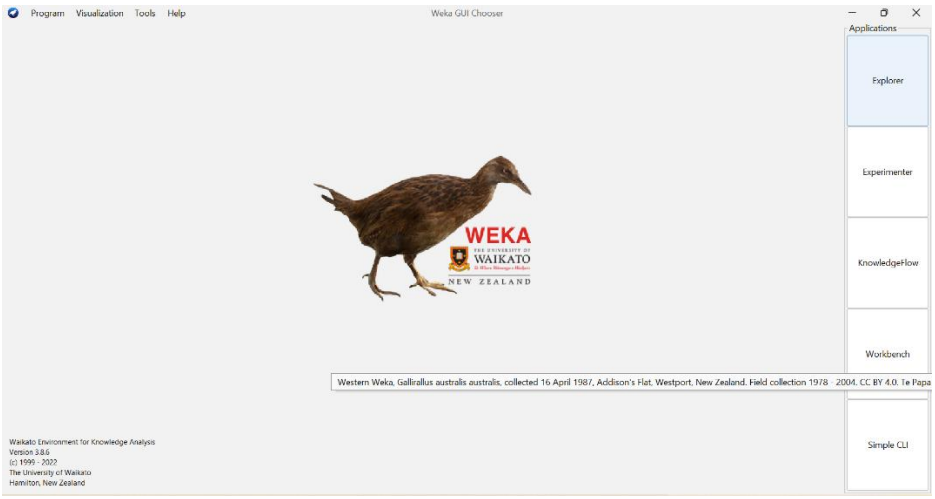
Abhishek Joshi (1002050821)

Amanul Rahiman Attar (1002071319)

Anushka Chandrakant Pawar (1002071263)

Introduction

Weka is an open-source tool used for performing various data preprocessing, Data mining, and visualization tasks. It includes a number of visualization tools, algorithms, and a graphical user interface for data analysis and predictive modeling. Java was used to write the software. Weka supports a number of common data mining operations, including feature selection, data pre-processing, clustering, classification, and regressing.

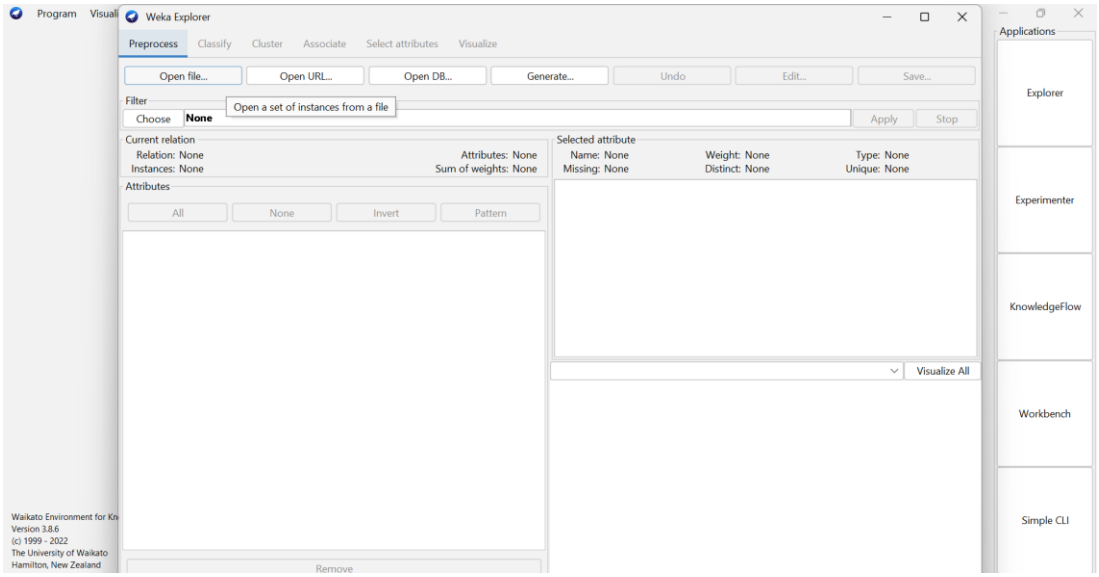


The above image is of the workspace of Weka.

To start the visualization, we need to load our data to Weka.

Pre-Processing

To load our Data on Weka, we need to select the ‘Explore’ option top right-hand side, and an empty window appears.

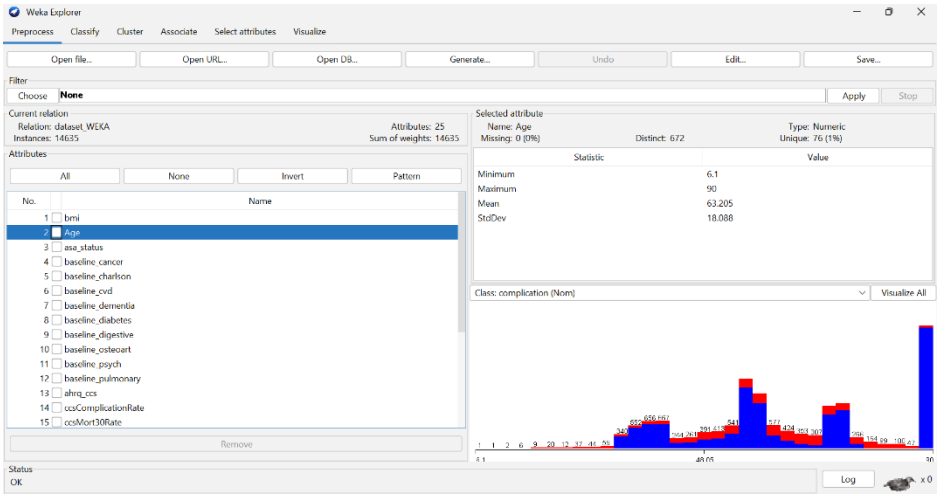


Select "Open" from the "File" menu by clicking. Change the "Files to type" option to "csv file(.csv)" after navigating to the current working directory, and then click "Open."

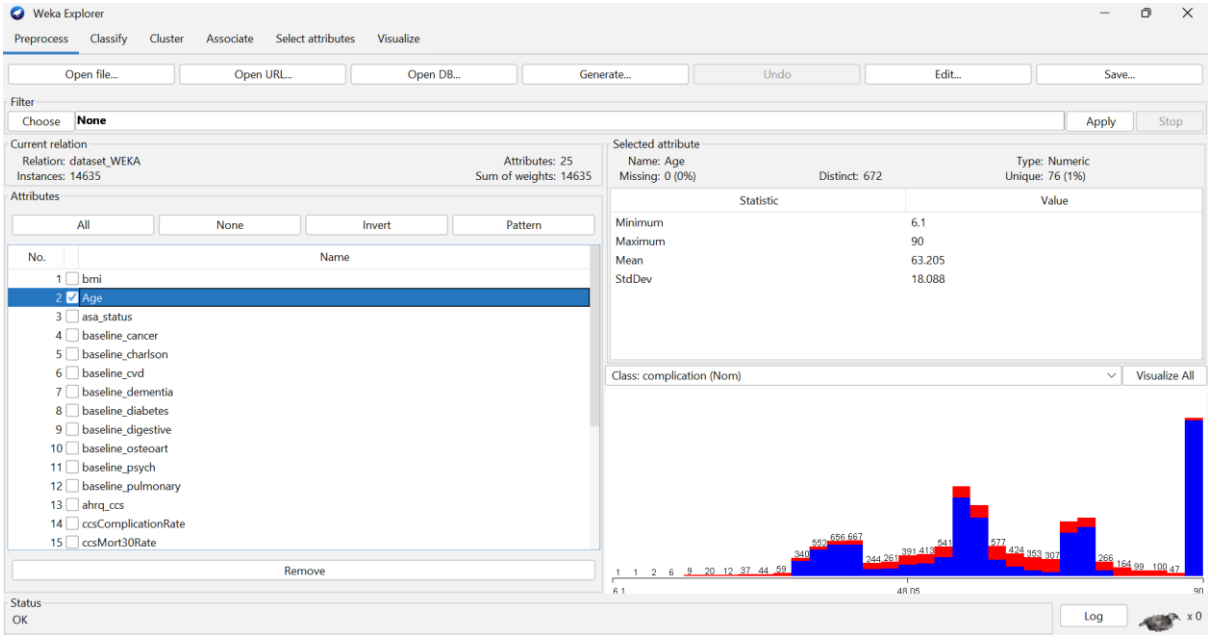
Weka Explorer's Preprocess tab allows us to view each attribute's descriptive information and visual representation.

The number of values in the range and the numeric attribute type are displayed.

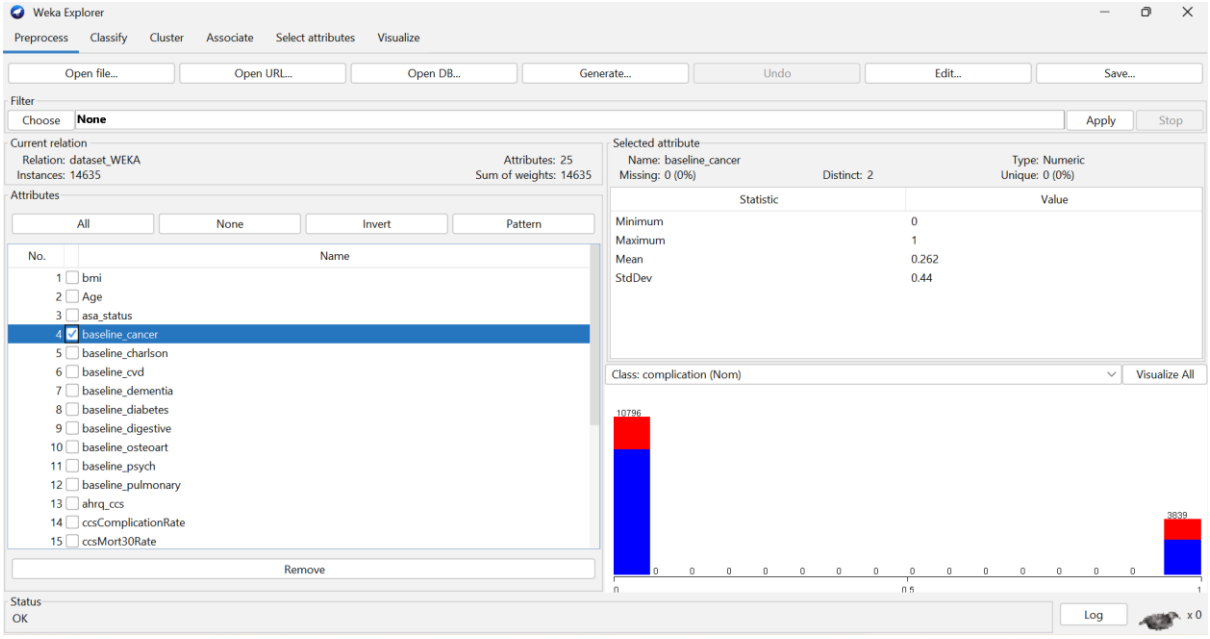
In the given Weka Dataset, we have 25 attributes among which all attributes except complications are numeric attributes. We can determine the attribute's range of values and the number of instances in that range by choosing each attribute.



A total of 25 attributes and 14635 instances were identified by WEKA. Basic statistics were also provided for each attribute. As the Age attribute is selected in the image, the minimum value is displayed as 6.1, the maximum value as 90, the mean as 63,205, and the standard deviation as 18.088.

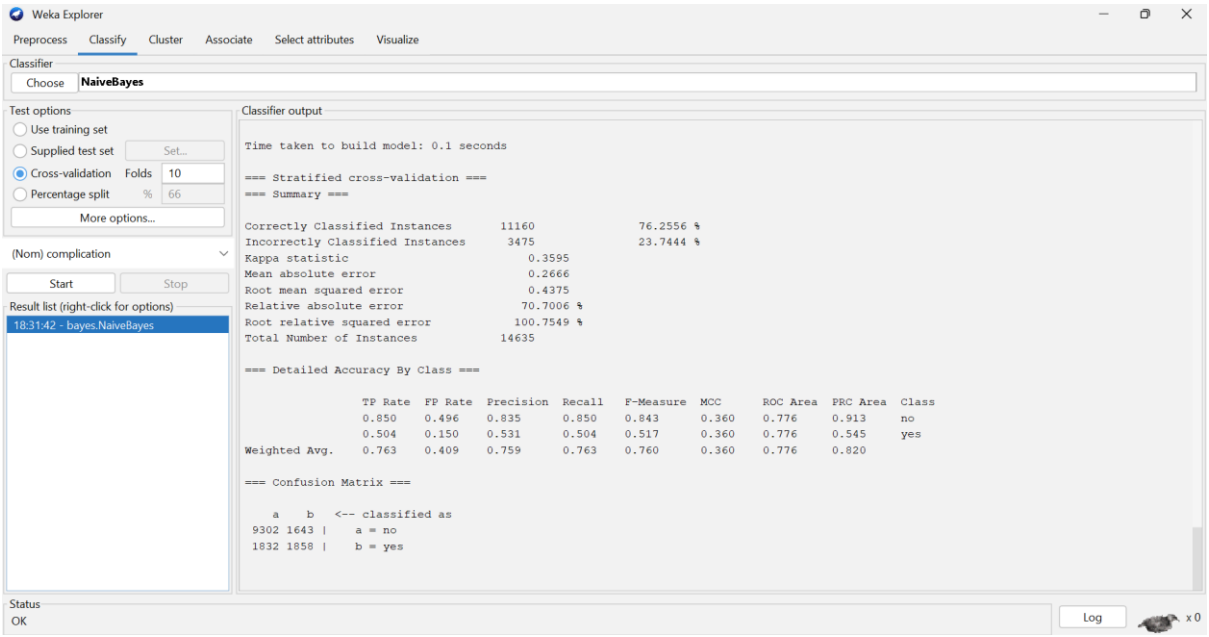


Similarly, As the Age attribute is selected in the image, the minimum value is displayed as 0, the maximum value as 1, the mean as 0.262, and the standard deviation as 0.44.



This is the preprocessing of the Dataset.

Classification of Data



We would attempt to use cross-validation or percentage split to separate the current data into training and testing data as we don't have a separate training set. Trees should be the first classifier taken into account. Classification and regression trees are other names for decision trees (CART). They function by learning the answers to a series of if/else questions that lead to a choice. The name comes from the fact that these questions take the form of a tree. We will attempt

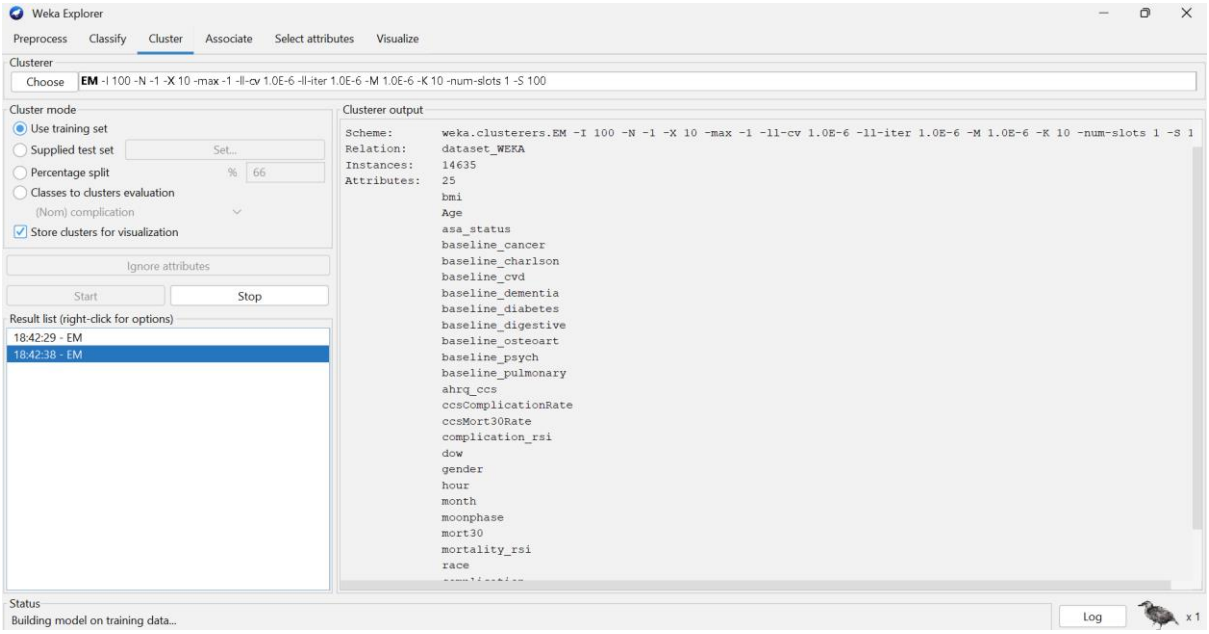
to categorize the Dataset using WEKA's Naïve Bayes method since it is widely regarded as the best classifier. We tried categorizing using default criteria.

Right-click on the results generated on the result list and choose "visualize tree" to see the classification as a decision tree.

Clustering

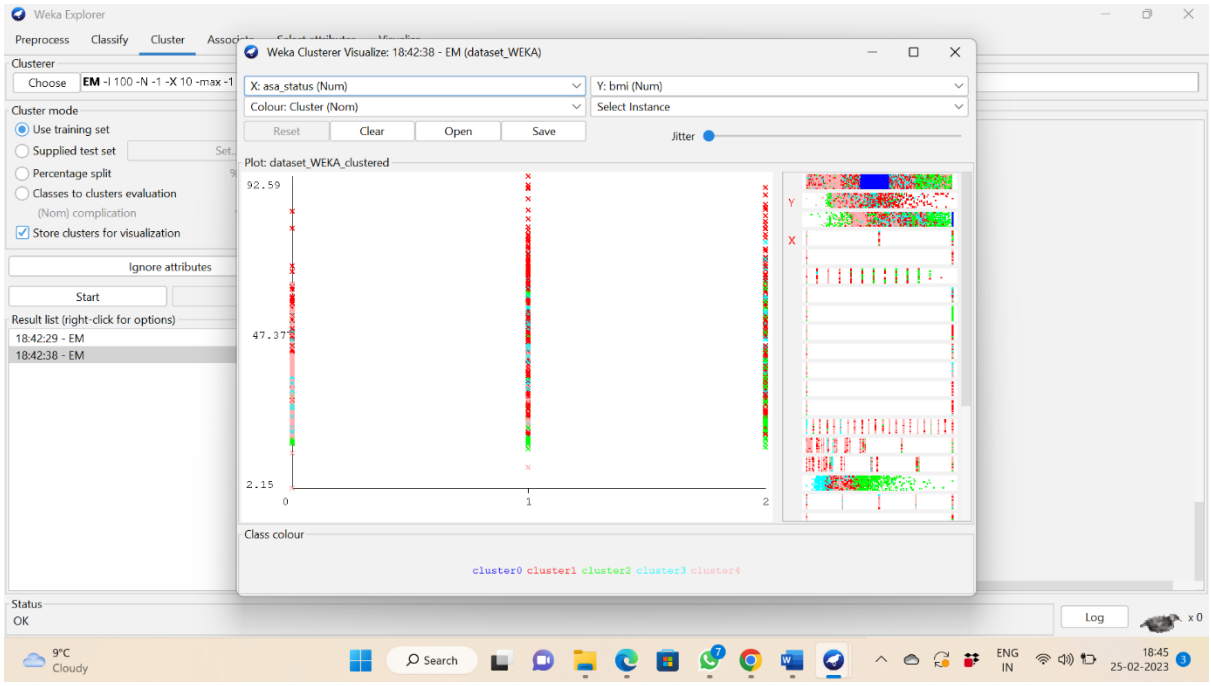
By grouping together sets of related types of data, clustering aids in our knowledge of the dataset.

We chose the EM (Expectation Maximization) clustering method, which assigns each instance a probability distribution indicating the likelihood that it will belong to each cluster from the group.



On the provided dataset, the EM clustering method was able to identify 5 groups. Additionally, it offered the means and standard deviations of each attribute within a certain cluster.

The clustered data can also be represented graphically as a graph. Based on the various colors utilized for each cluster group, we discovered the following result after choosing the asa_status vs. BMI for graphing:

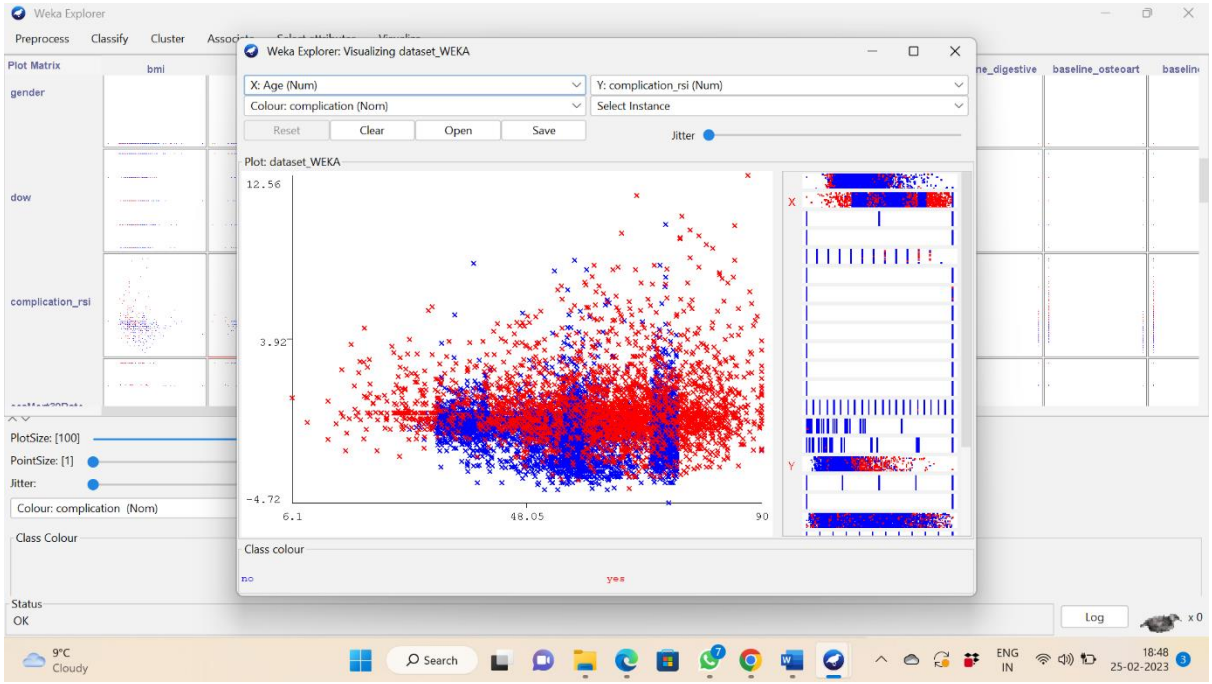


Visualization

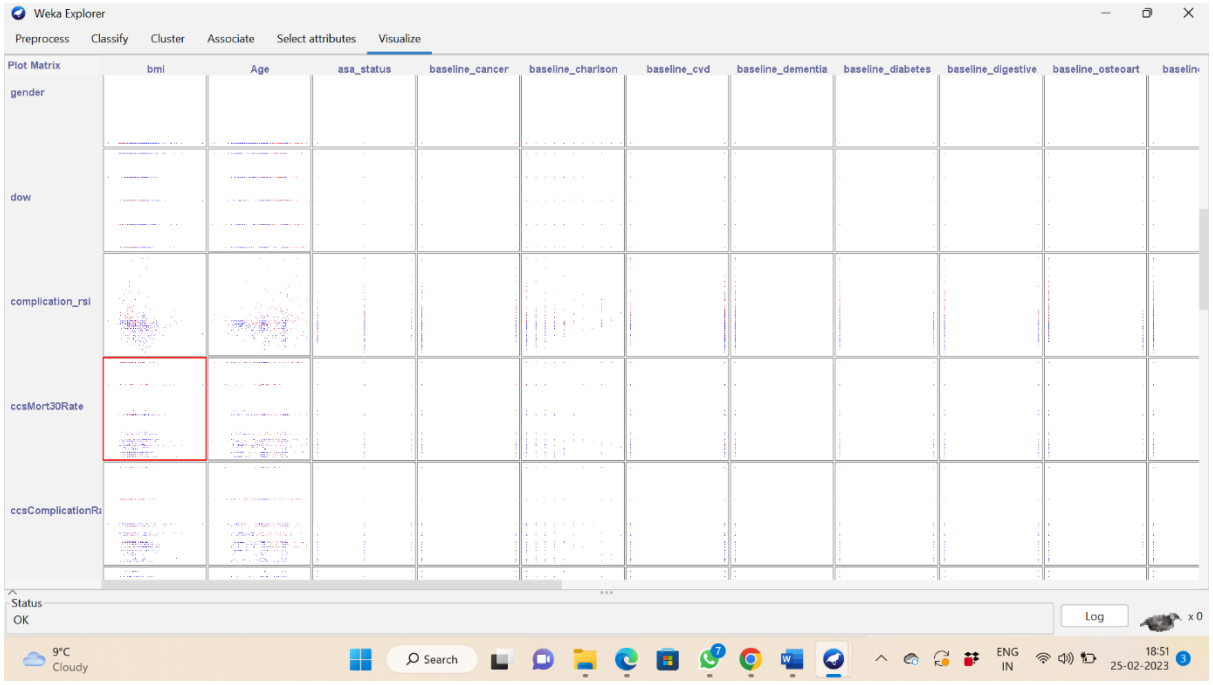
WEKA offers a visualization tool that allows users to obtain graphic insights into the provided Dataset. Complex problems can be resolved with its help. The jitter feature spreads out the data points by adding random noise, which could help find buried data points.

Based on any attribute combinations for the X and Y axes, we may view the data.

Here is Age Vs Complication graph:



Following are all the scatter graphs of all attributes of the data set given:



Conclusion:

- WEKA offers sophisticated methods for processing and visualizing datasets.
- The data can be generalized and grouped into sets of related data kinds.
- By putting the necessary properties on the X-axis and Y-axis of a graphical graph, one can gain clear insights into how linked attributes connect to one another.
- One of the common characteristics of those with problems is a high BMI. Age itself is a significant determinant of these difficulties.
- Preprocessing makes it simple to gain a general picture of each attribute.

Contribution by team members

Python – Amanul Rahiman Attar

R language – Abhishek Joshi

Weka – Anushka Chandrakant Pawar

References

- https://www.futurelearn.com/info/courses/data-mining-with-weka/0/steps/25377#:~:text=Weka's%20Visualize%20panel%20lets%20you,on%20the%20points%20inside%20it.
- https://www.softwaretestinghelp.com/weka-explorer-tutorial/
- https://www.analyticsvidhya.com/blog/2020/03/decisiontree-weka-no-cod