PREDICTION OF HEART DISEASE USING DATA MINING

GROUP MEMBERS

1. P19101064 – SHAHID IQBAL
2. P19101063 – SANA ASHFAQ
3. P19101023 – IQRA MEHDI
4. P19101008 – ANUM ZEHRA (GROUP LEAD)

SUPERVISOR

DR. TEHSEEN A. JILANI

PROJECT DETAILS

MCS MORNING (FINAL)

DATA MINING AND DATA WAREHOUSING

WORKING ON WEKA AND SPSS

ABSTRACT

Our project is to predict the heart disease by examining the various attributes. In this project diverse strategies have been utilized to detect heart disease such as Decision tree, K-Means, Confusion Matrix. And among all these calculations the final result gives us the finest precision of 91.8%.

TABLE OF CONTENT

1. INTRODUCTION

In daily life many factors influence a human heart. Many problems are happening at a fast pace and new heart diseases are rapidly being recognized. In today's world of stress Heart, being an essential organ in a human body which pumps blood through the body for the blood circulation is fundamental and its health is to be conserved for a sound living. The health of a human heart is based on the encounters in a person's life and is completely dependent on proficient and personal behaviors of a person. There may also be a few genetic factors through which a sort of heart illness is passed down from eras. Concurring to the World Health Organization, every year more than 12 million deaths are happening around the world due to the different sorts of heart diseases which is additionally known by the term cardiovascular disease. The term heart disease includes numerous diseases that are diverse and particularly affect the heart and the arteries of a human being. Even youthful matured individuals around their 20-30 a long time of life expectancy are getting influenced by heart diseases. The increment within the possibility of heart disease among young may be due to the bad eating habits, lack of rest, anxious nature, depression, discouragement and various other factors such as obesity, poor diet, family history, high blood pressure, high blood cholesterol, idle behavior, smoking and hypertension.

The diagnosis of the heart diseases could be an exceptionally important and is itself the most complicated task in the medical field. All the mentioned components are taken into consideration when analyzing and understanding the patients by the specialist through manual check-ups at regular intervals of time. The symptoms of heart disease significantly depend upon which of the distress felt by a person. A few side effects are not usually identified by the common people. However, common symptoms include chest pain, breathlessness, and heart palpitations. The chest pain common to many types of heart disease is known as angina, or angina pectoris, and happens when a portion of the heart does not get sufficient oxygen. Angina may be activated by stressful events or physical effort and normally lasts under 10 minutes. Heart attacks can also happen as a result of different types of heart disease.

Data Mining is an important decision-making process information from past collections for future analysis or forecast. Information may be anonymous and may not be identified without using a data mine. The section says a single data mining process where the future result or predictions can be made based on historical data i.e., available. Digging for medical data has created a possible solution

combine classification techniques and deliver by computer database training that leads continuously to hidden tests patterns in medical data sets used for prediction of the patient's future status. So, using medical data to dig it is able to provide information about patient history and is capable provided clinical support through analysis. Clinical analysis in patients, these patterns are very important. In English, medical data mining uses classification algorithms that is an important part of diagnosing the possibility of a heart attack before it happened. Separation algorithms can be trained and tested to make decisive predictions a person's condition of heart attack.

2. DECISION TREE

2.1. INTRODUCTION

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

For instance, in the example below, decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model.

2.2. ADVANTAGES

Some advantages of decision trees are:

- Simple to understand and to interpret. Trees can be visualized.
- Requires little data preparation. Other techniques often require data normalization, dummy variables need to be created and blank values to be removed. Note however that this module does not support missing values.
- The cost of using the tree (i.e., predicting data) is logarithmic in the number of data points used to train the tree.
- Able to handle both numerical and categorical data. However, scikit-learn implementation does not support categorical variables for now. Other techniques are usually specialized in analyzing datasets that have only one type of variable. See algorithms for more information.
- Able to handle multi-output problems.
- to account for the reliability of the model.
- Performs well even if its assumptions are somewhat violated by the true model from which the data were generated.

2.3. DISADVANTAGES

- Decision-tree learners can create over-complex trees that do not generalize the data well. This is called overfitting. Mechanisms such as pruning, setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree are necessary to avoid this problem.
- Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. This problem is mitigated by using decision trees within an ensemble.
- Predictions of decision trees are neither smooth nor continuous, but piecewise constant approximations as seen in the above figure. Therefore, they are not good at extrapolation.

- The problem of learning an optimal decision tree is known to be NP-complete under several aspects of optimality and even for simple concepts. Consequently, practical decision-tree learning algorithms are based on heuristic algorithms such as the greedy algorithm where locally optimal decisions are made at each node. Such algorithms cannot guarantee to return the globally optimal decision tree. This can be mitigated by training multiple trees in an ensemble learner, where the features and samples are randomly sampled with replacement.

- There are concepts that are hard to learn because decision trees do not express them easily, such as XOR, parity or multiplexer problems.

- Decision tree learners create biased trees if some classes dominate. It is therefore recommended to balance the dataset prior to fitting with the decision tree.

3. TREE ALGORITHMS: ID3, C4.5

What are all the various decision tree algorithms and how do they differ from each other? Which one is implemented in scikit-learn?

3.1. ID3 (ITERATIVE DICHOTOMISER 3)

It was developed in 1986 by Ross Quinlan. The algorithm creates a multiway tree, finding for each node (i.e., in a greedy manner) the categorical feature that will yield the largest information gain for categorical targets. Trees are grown to their maximum size and then a pruning step is usually applied to improve the ability of the tree to generalize to unseen data.

3.2. C4.5

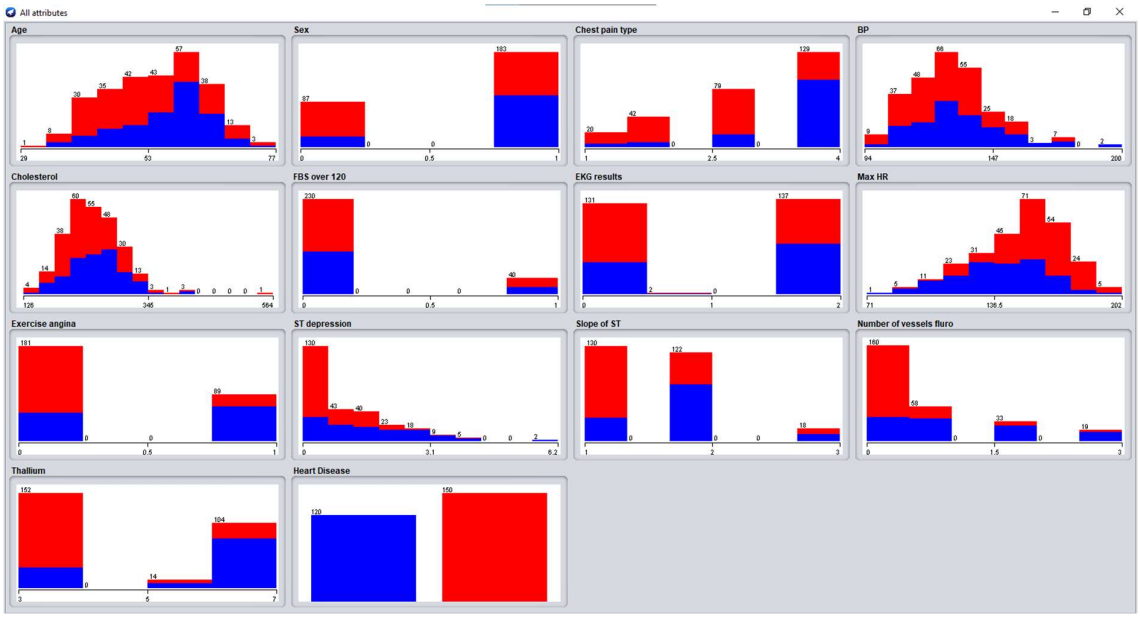
It is the successor to ID3 and removed the restriction that features must be categorical by dynamically defining a discrete attribute (based on numerical variables) that partitions the continuous attribute value into a discrete set of intervals. C4.5 converts the trained trees (i.e., the output of the ID3 algorithm) into sets of if-then rules. This accuracy of each rule is then evaluated to determine the order in which they should be applied. Pruning is done by removing a rule's precondition if the accuracy of the rule improves without it.

4. LOGISTIC REGRESSION

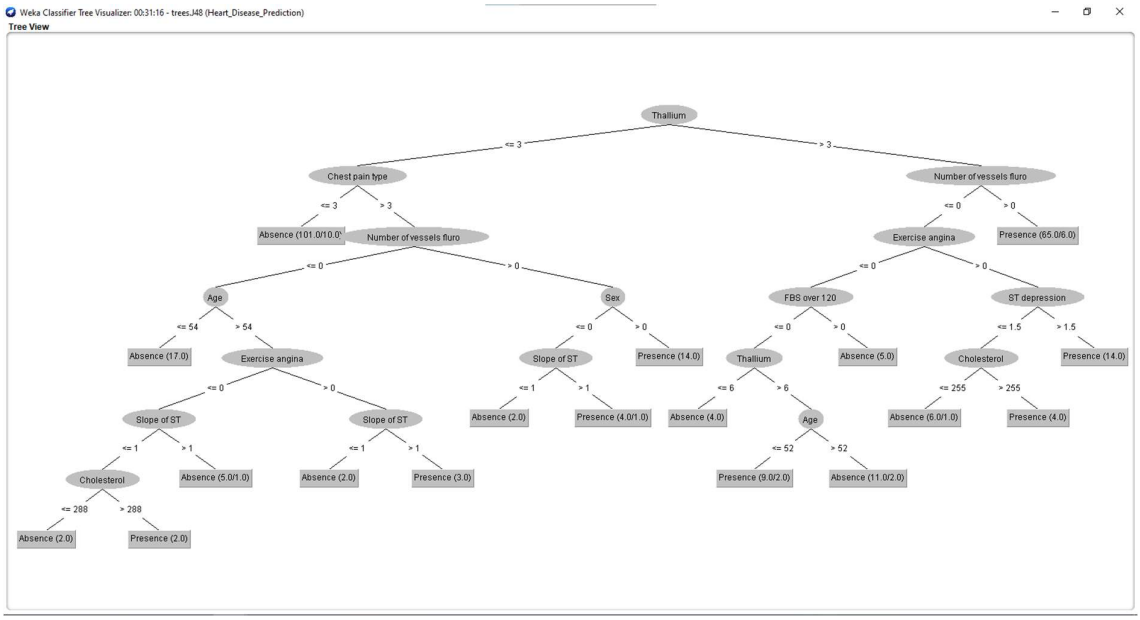
In statistics, multinomial logistic regression is a classification method that generalizes logistic regression to multiclass problems, i.e., with more than two possible discrete outcomes. That is, it is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables (which may be real-valued, binary-valued, categorical-valued, etc.).

5. DATA VISUALIZATION

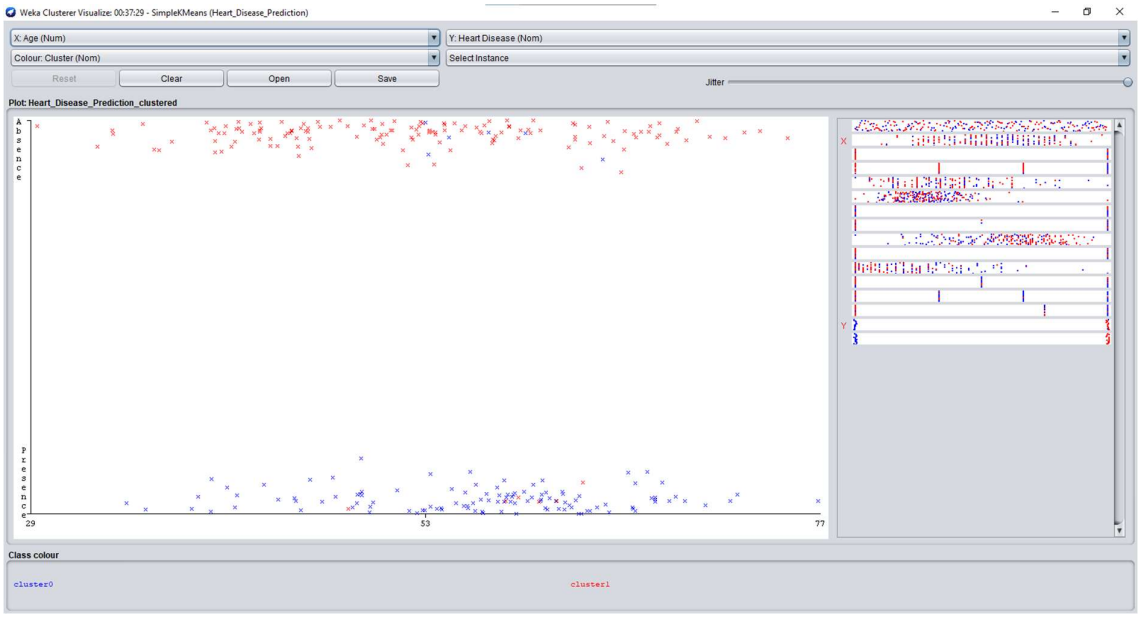
5.1. ALL ATTRIBUTES WRT HEART DISEASE



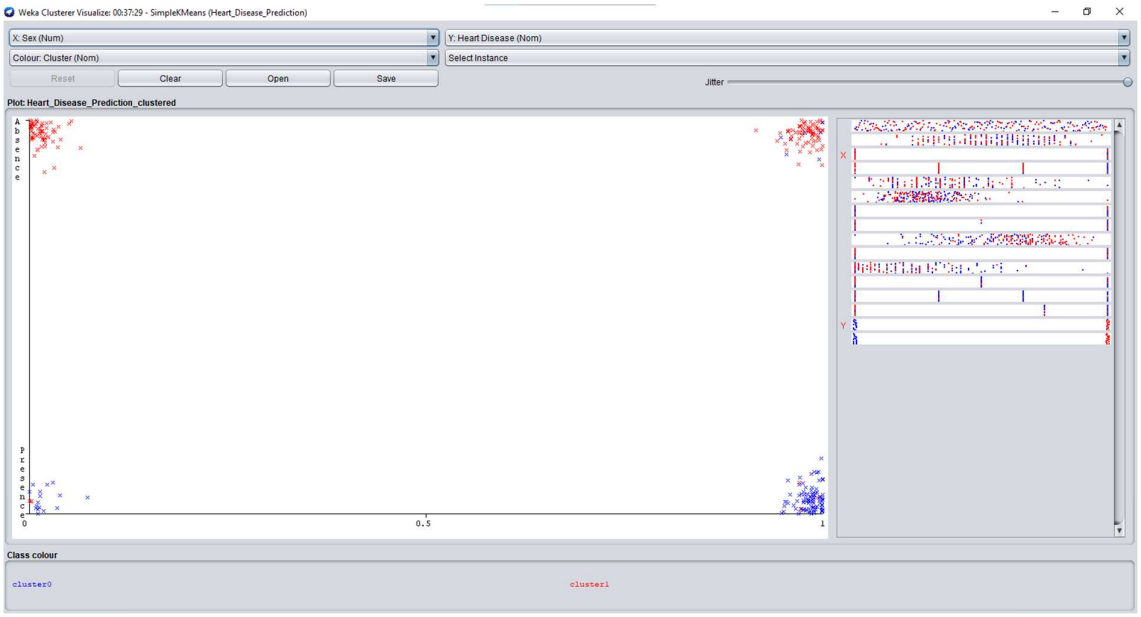
5.2. C4.5 IMPLEMENTATION IN WEKA USING JAVA LIBRARY J48



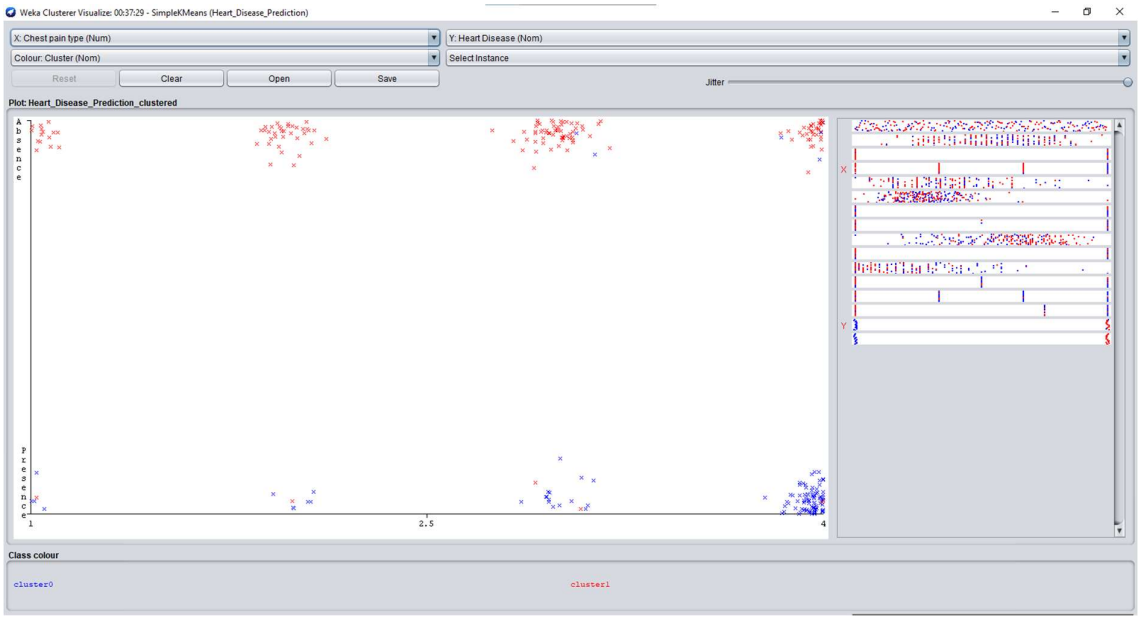
5.3. AGE VS HEART DISEASE



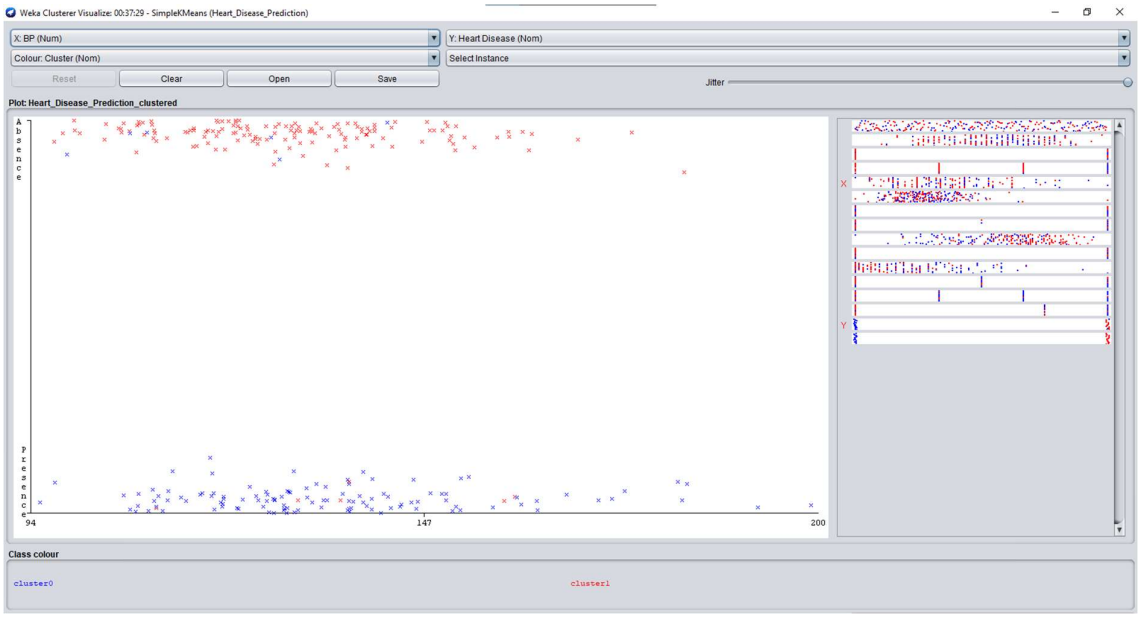
5.4. SEX VS HEART DISEASE



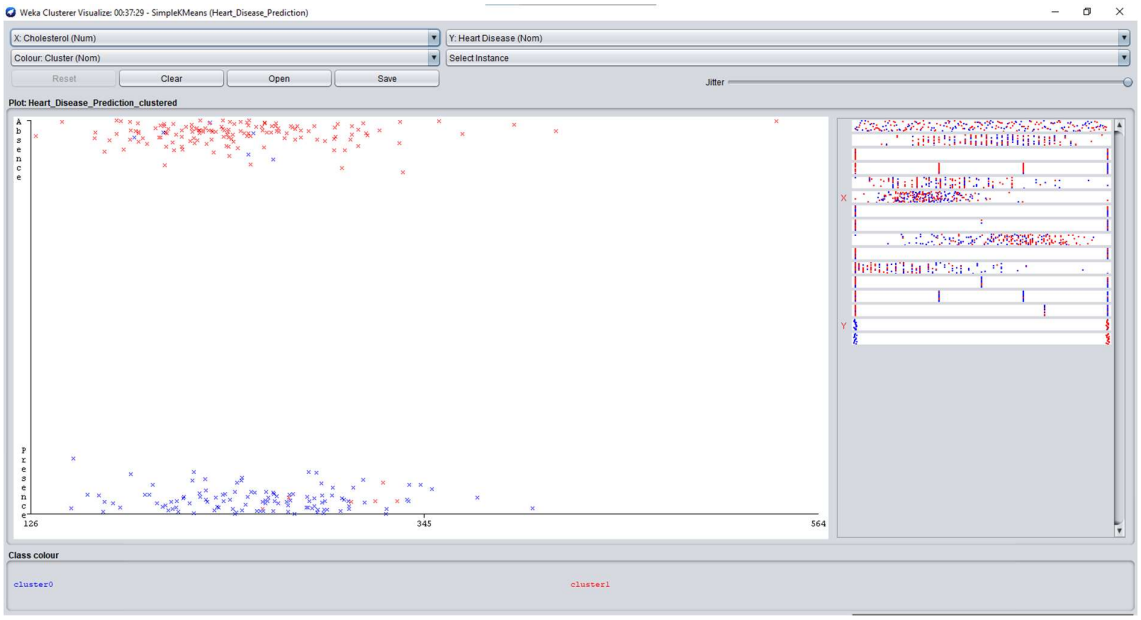
5.5. CHEST PAINT TYPE VS HEART DISEASE



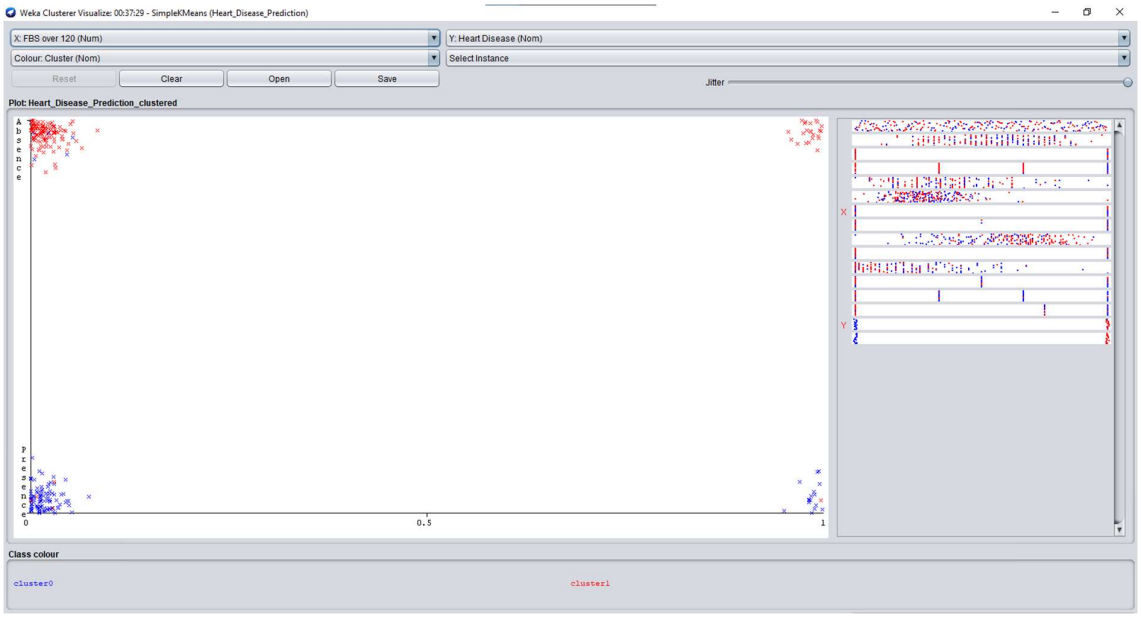
5.6. BP VS HEART DISEASE



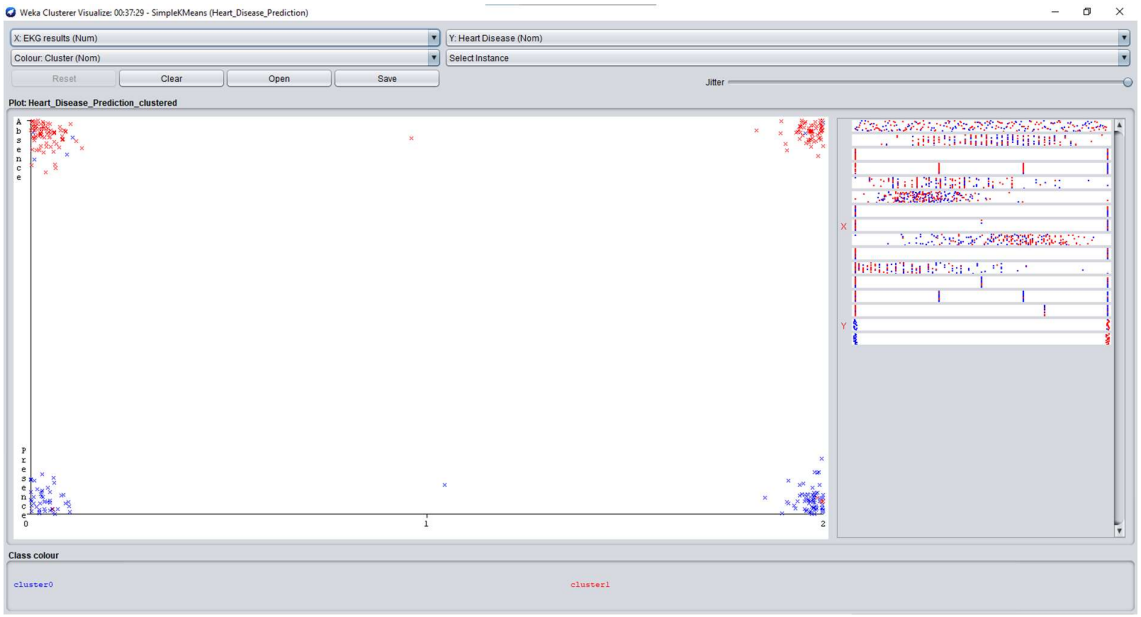
5.7. CHOLESTROL VS HEART DISEASE



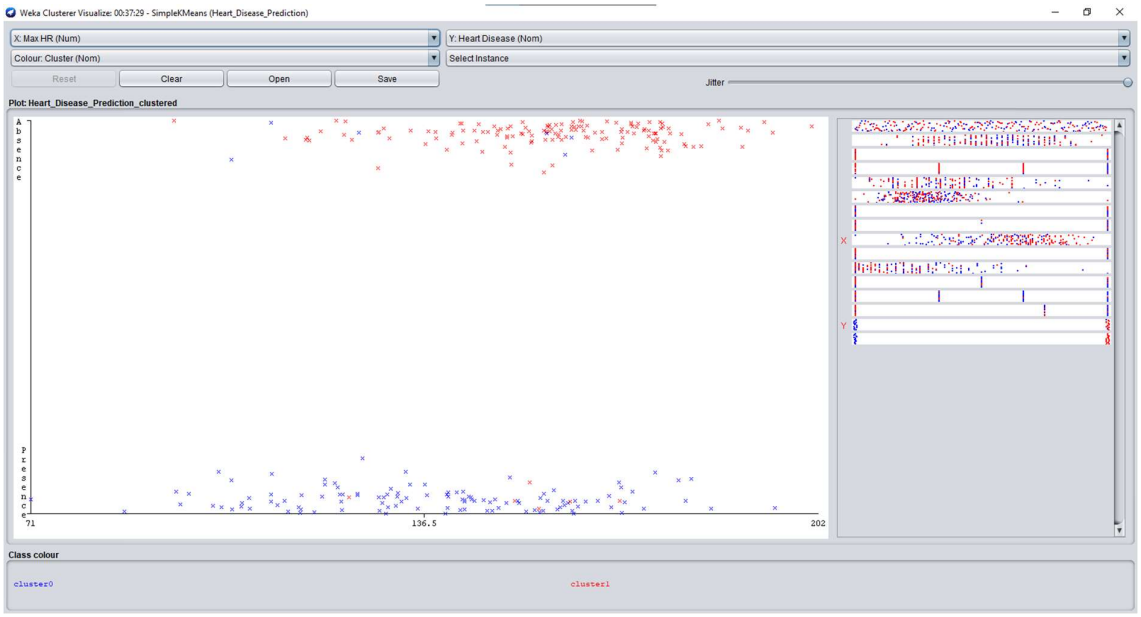
5.8. FBS OVER 120 VS HEART DISEASE



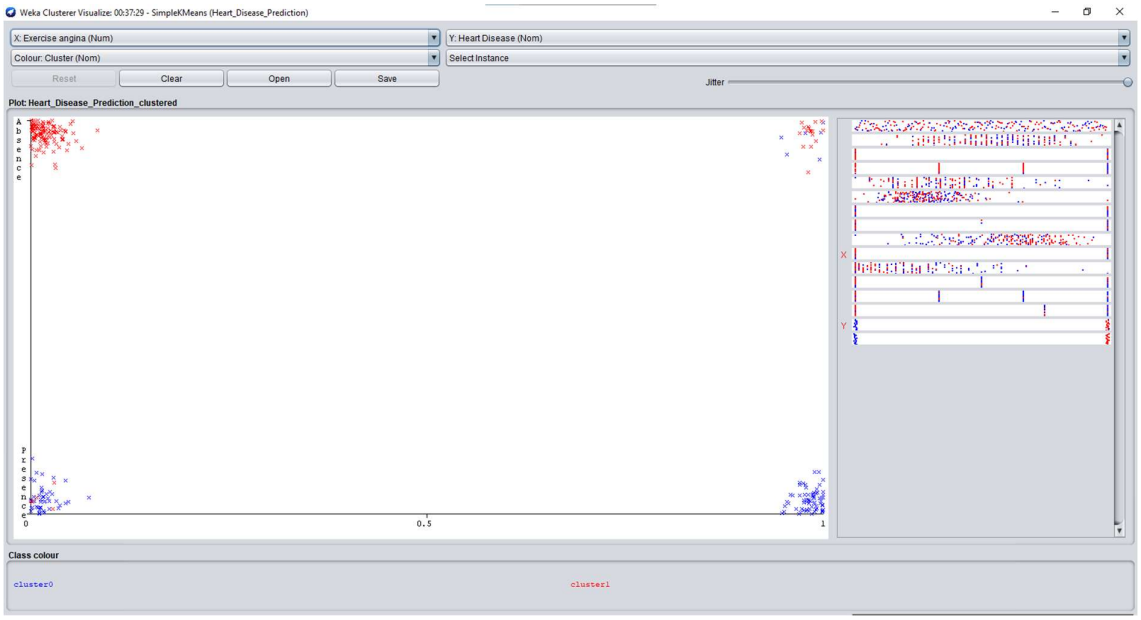
5.9. EKG RESULTS VS HEART DISEASE



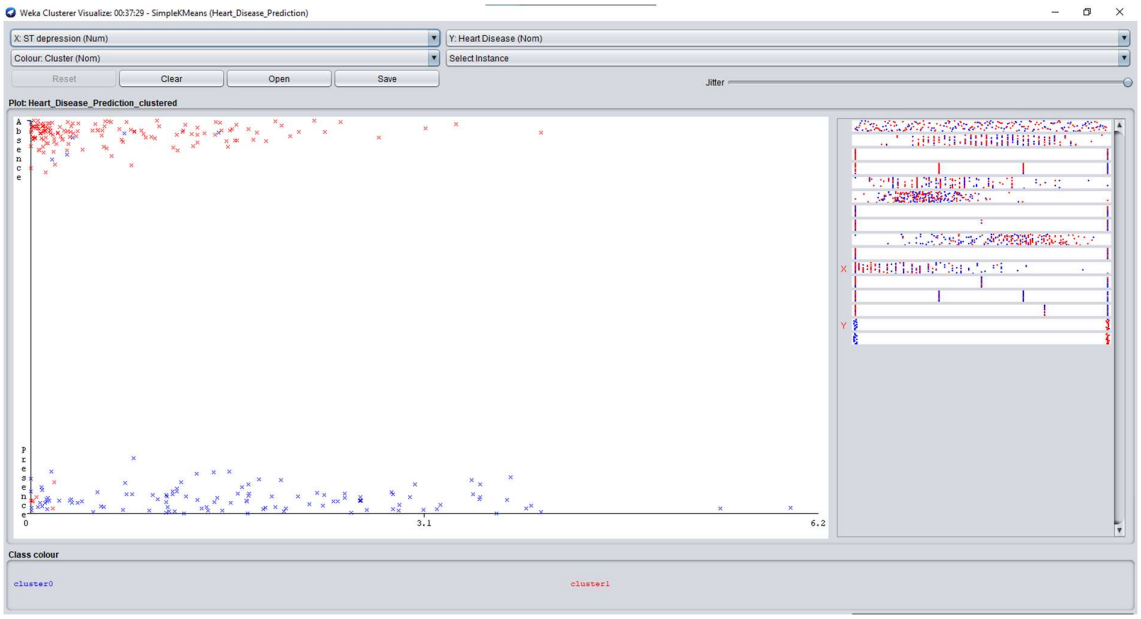
5.10. MAX HR VS HEART DISEASE



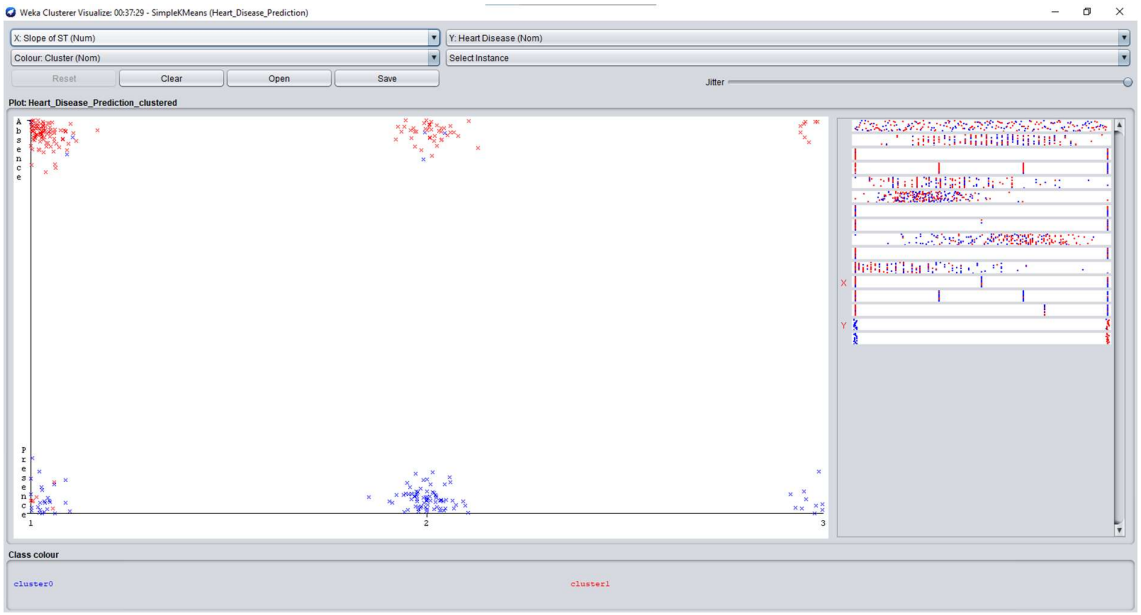
5.11. EXERCISE ANGINA VS HEART DISEASE



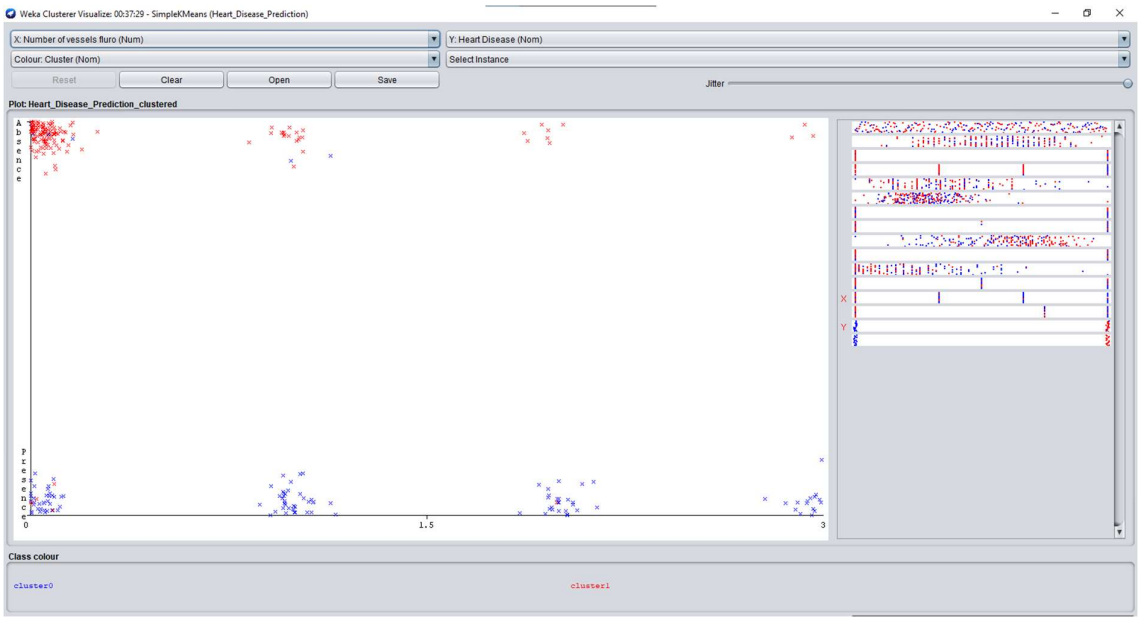
5.12. ST DEPRESSION VS HEART DISEASE



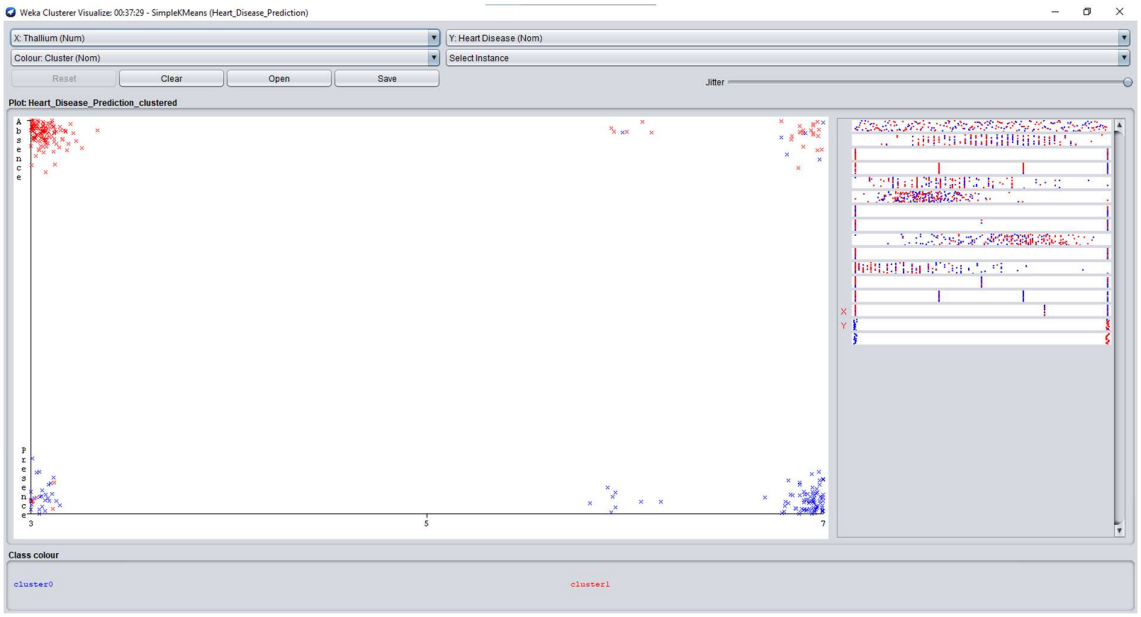
5.13. SLOPE OF ST VS HEART DISEASE



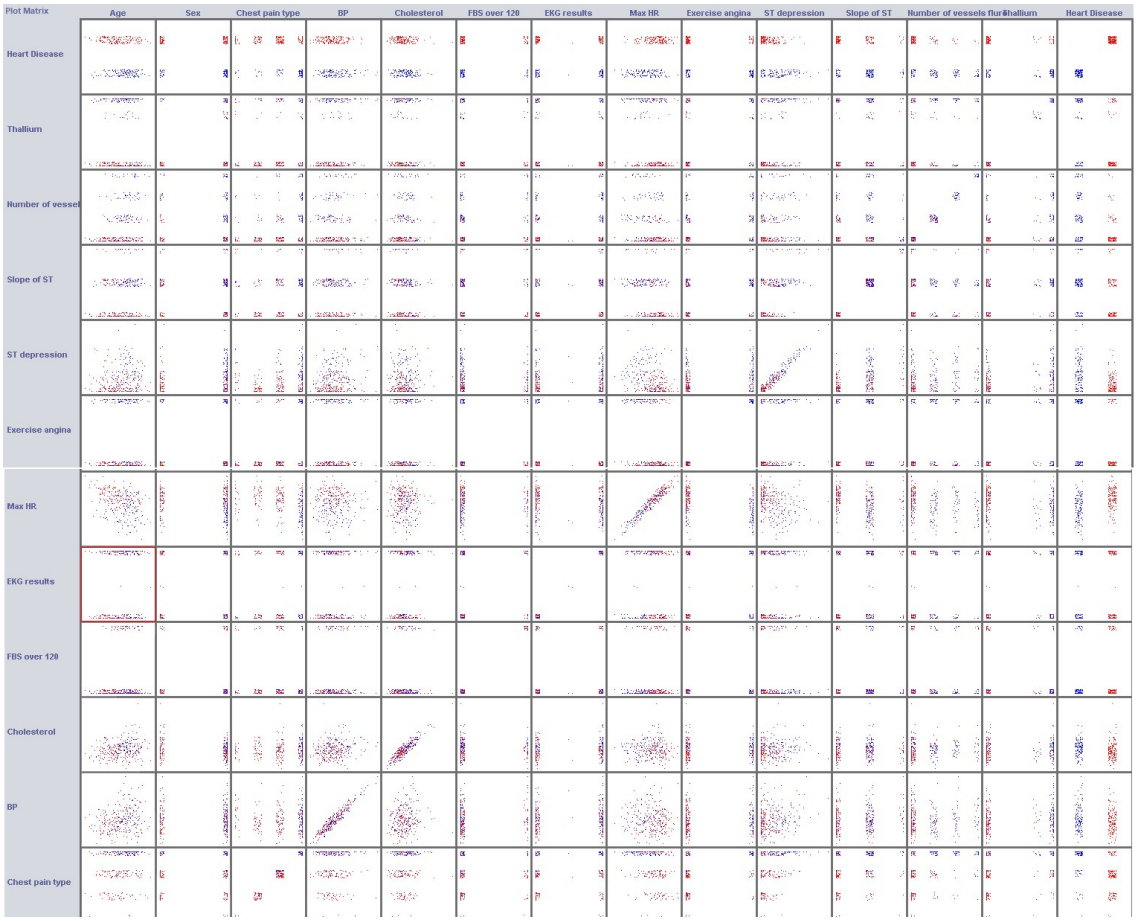
5.14. NO. OF VESSELS FLURO VS HEART DISEASE



5.15. THALLIUM VS HEART DISEASE



5.16. PLOT MATRIX



6. CONCLUSION

Following the algorithm and technique, we can deduce the presence of heart disease on the basis of some attributes. In light of this model, we can predict but prediction can't be 100% true. Therefore we will increase the set of attributes for more accurate prediction in near future .