BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

“EARLY PREDICTION OF HEART DISEASES USING DATA MINING TECHNIQUES”

GROUP NUMBER-9:

VARUN GUPTA	2016A7PS0087P
AVICHAL JAIN	2016A7PS0046P
ADITI AGARWAL	2016A7PS0095P
SAHIL RANADIVE	2016A7PS0097P
RIGVITA SHARMA	2016A7PS0067P
AASTHA KATARIA	2016A7PS0062P

Date of Submission: Submitted To:

22/04/18 Dr. Yashvardhan Sharma(Instructor-In-Charge)

Ms. Geetika Arora(Mentor)

TABLE OF CONTENTS:

INTRODUCTION:

1.1. Background:

Largest-ever study of deaths shows that heart diseases have emerged as the largest killer in world. About 25 percent of deaths in the age group of 25- 69 years occur because of heart diseases. If all age groups are included, heart diseases account for about 19 percent of all deaths. It is the leading cause for death among males as well as females. It is also the leading cause of death in all regions though the numbers vary. The proportion of deaths caused by heart disease is the highest in south India (25 per cent) and lowest - 12 per cent - in the central region of India.

The prediction of heart disease survivability has been a challenging research problem for many researchers. Since the early dates of the related research, much advancement has been recorded in several related fields. Therefore, the main objective of this manuscript is to report on a research project where we took advantage of those available technological advancements to develop prediction models for heart disease survivability.

1.2. Motivation:

Over the past few years data mining techniques have become strong enough to predict medical conditions with accuracy. Data mining could thus help in data analysis to find the symptoms and pre effects of heart attacks successfully and predict them earlier which would in turn save lives. Using data mining techniques such as CART and Iterative Dichotomized 3 (ID3), we can accurately predict onset of heart diseases.

1.3. Objective:

1. To use data mining techniques mentioned in this document to accurately predict heart attacks using 14 attributes that have been considered.

2. CART and ID3 were used first to classify the outcome and predict the accuracy and then ensemble was applied on these two techniques to find an optimum solution (which was not a part of the paper to be implemented).

3. To check which the best classification method is to predict heart attacks.

2. Related Work:

Data mining methods like ANN, Clustering(K-nearest neighbour), Association Rules, soft computing approaches, Decision Trees, Naive Bayes, etc have been used in effective heart attack prediction systems. Techniques like Multi-layer Perceptron Neural Network with Backpropagation have also been used for training.
In previous works it has been shown that the performance of decision trees is better and Bayesian classification; gives similar accuracy as that of DT but other predictive methods like K-Nearest Neighbor, Neural Networks, Classification based on clustering do not perform well. When Genetic Algorithm is applied, the accuracy of the Decision Tree and Bayesian Classification is improved by reducing the actual data size.

3. Proposed/Used Techniques and Algorithms:

i). Pre-processing Techniques

It includes zero-mean normalization, PCA, replacing missing values with mode of data and constructing separate input and target lists. Reading data and making it ready to be passed to algorithms.

ii). CART (Classification and Regression Tree)

The algorithm is based on Classification and Regression Trees by Breiman et al (1984). A CART tree is a binary decision tree that is constructed by splitting a node into two child nodes repeatedly, beginning with the root node that contains the whole learning sample(x) . The leaf nodes of the tree contain an output variable (y) which is used to make a prediction. CART uses Gini index to measure the impurity of a partition or set of training tuples. It can handle high dimensional categorical data. Decision Trees can also handle continuous data (as in regression) but they must be converted to categorical data.

iii).Iterative Dichotomized 3(ID3)

ID3 is a decision tree learning algorithm which is used for the classification of the objects with the iterative inductive approach. In this algorithm the top to down, greedy search approach is used. It's traversing from root node to leaf nodes. Each node requires some test on the attributes which decide the level of the leaf nodes.
It uses information gain to help it decide which attribute goes into a decision node.It Does not handle numeric attributes and missing values.

iv).Random Forest Ensemble Classifier

Random forests take a multitude of decision trees and output a result by taking the mean of all the predictions or the mode of the classes. It is used here to predict heart attacks after taking models built from the previous two algorithms i.e. CART and ID3, as arguments.

4. DATASET USED:

Source of the dataset: V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.
The data used in this study is from the Cleveland Clinic Foundation. Heart disease data set is available at

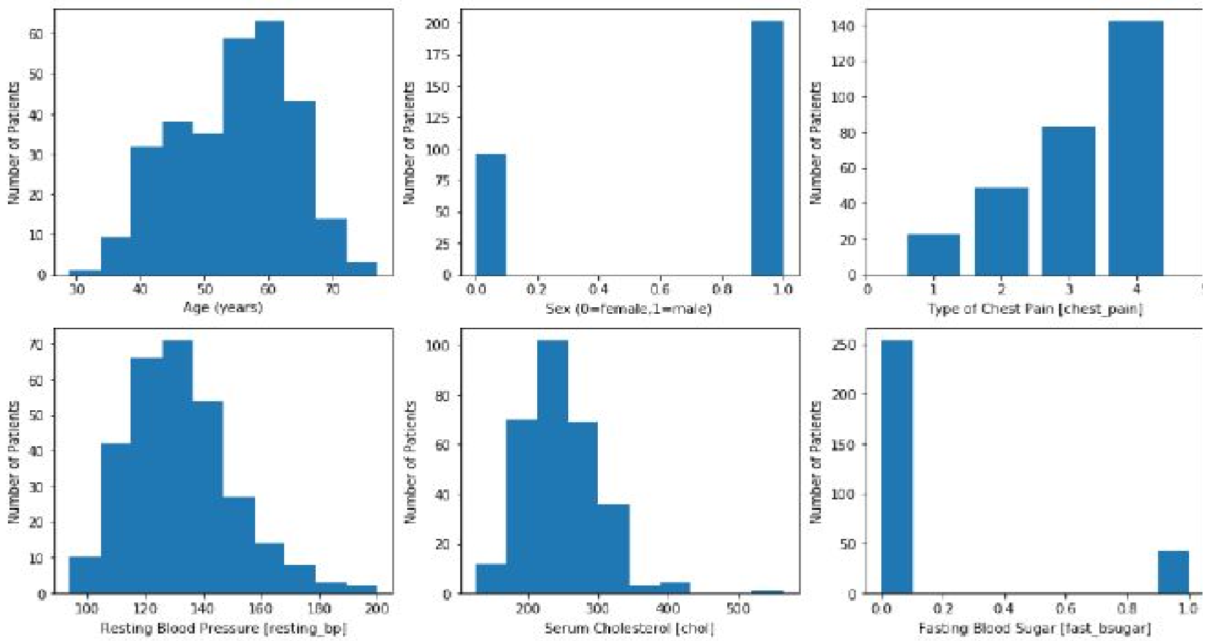
http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data

The data set has 76 raw attributes. However, all of the published experiments only refer to 14 of them. Although the paper-Early Prediction to allow comparison with the literature, we have restricted testing to these same attributes.

The data set contains 303 rows.
The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

Attribute Information:
The 14 attributes used in the dataset are:
1. age : age in years
2. sex : 4 sex: sex (1 = male; 0 = female)
3. Cp: chest pain type
-- Value 1: typical angina
-- Value 2: atypical angina
-- Value 3: non-anginal pain
-- Value 4: asymptomatic
4. trestbps : resting blood pressure (in mm Hg on admission to the hospital)
5. Chol: serum cholesterol in mg/dl
6. Fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
7. Restecg: resting electrocardiographic results
-- Value 0: normal
-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
-- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
8. thalach: maximum heart rate achieved
9. exang: exercise induced angina (1 = yes; 0 = no)
10. Oldpeak: ST depression induced by exercise relative to rest
11. slope: the slope of the peak exercise ST segment
-- Value 1: upsloping
-- Value 2: flat

-- Value 3: downsloping
12. ca: number of major vessels (0-3) colored by fluoroscopy
13. thal: : 3 = normal; 6 = fixed defect; 7 = reversible defect
14. num (the predicted attribute): diagnosis of heart disease (angiographic disease status)
-- Value 0: < 50% diameter narrowing
-- Value 1: > 50% diameter narrowing



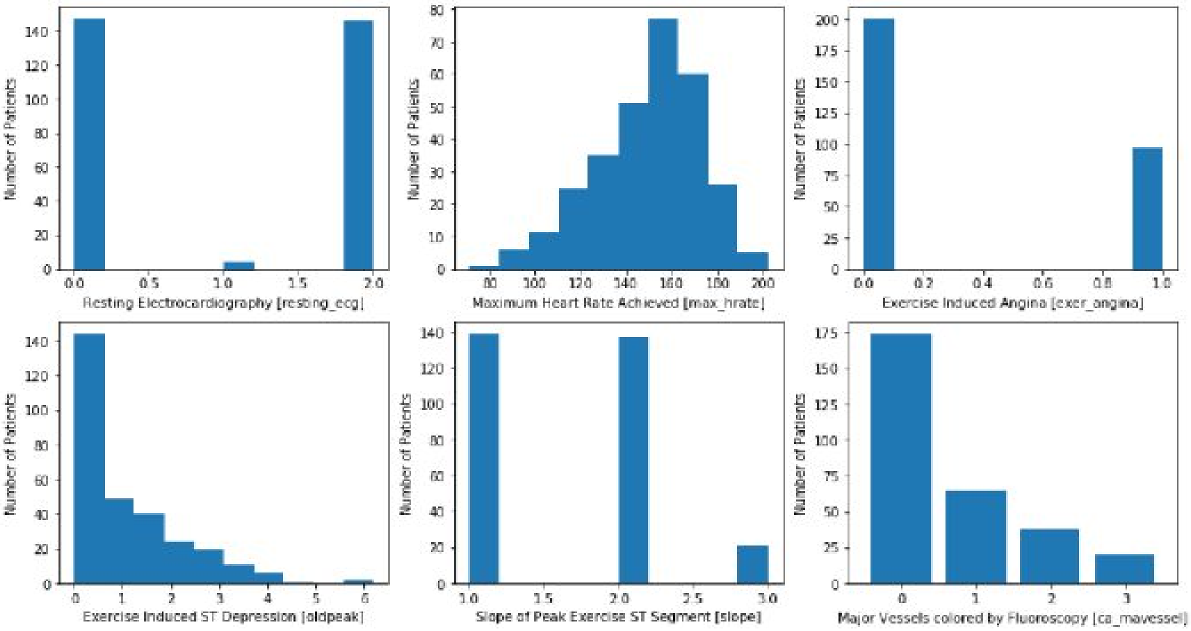


FIGURE 1: Distribution of data of heart disease patients

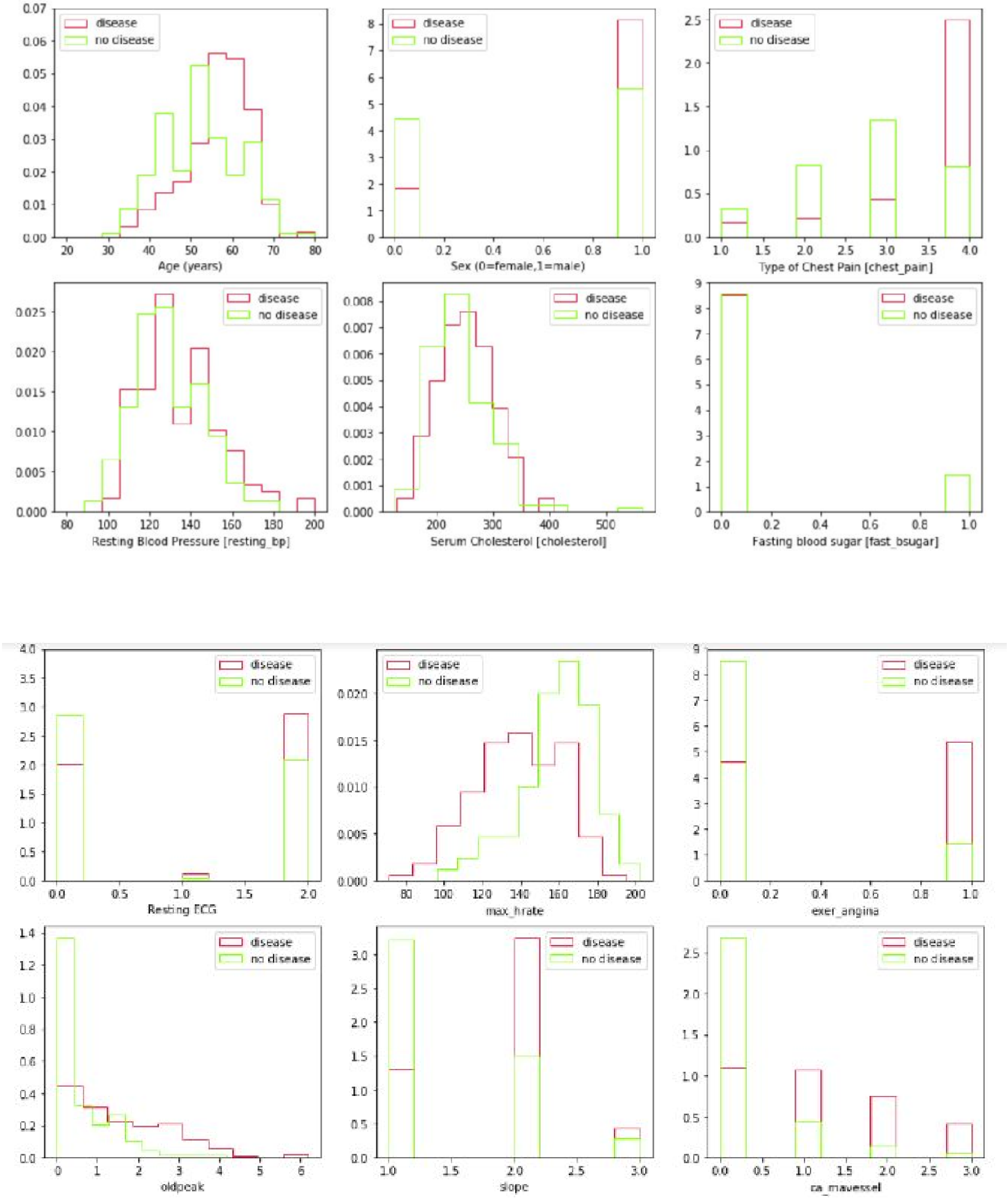
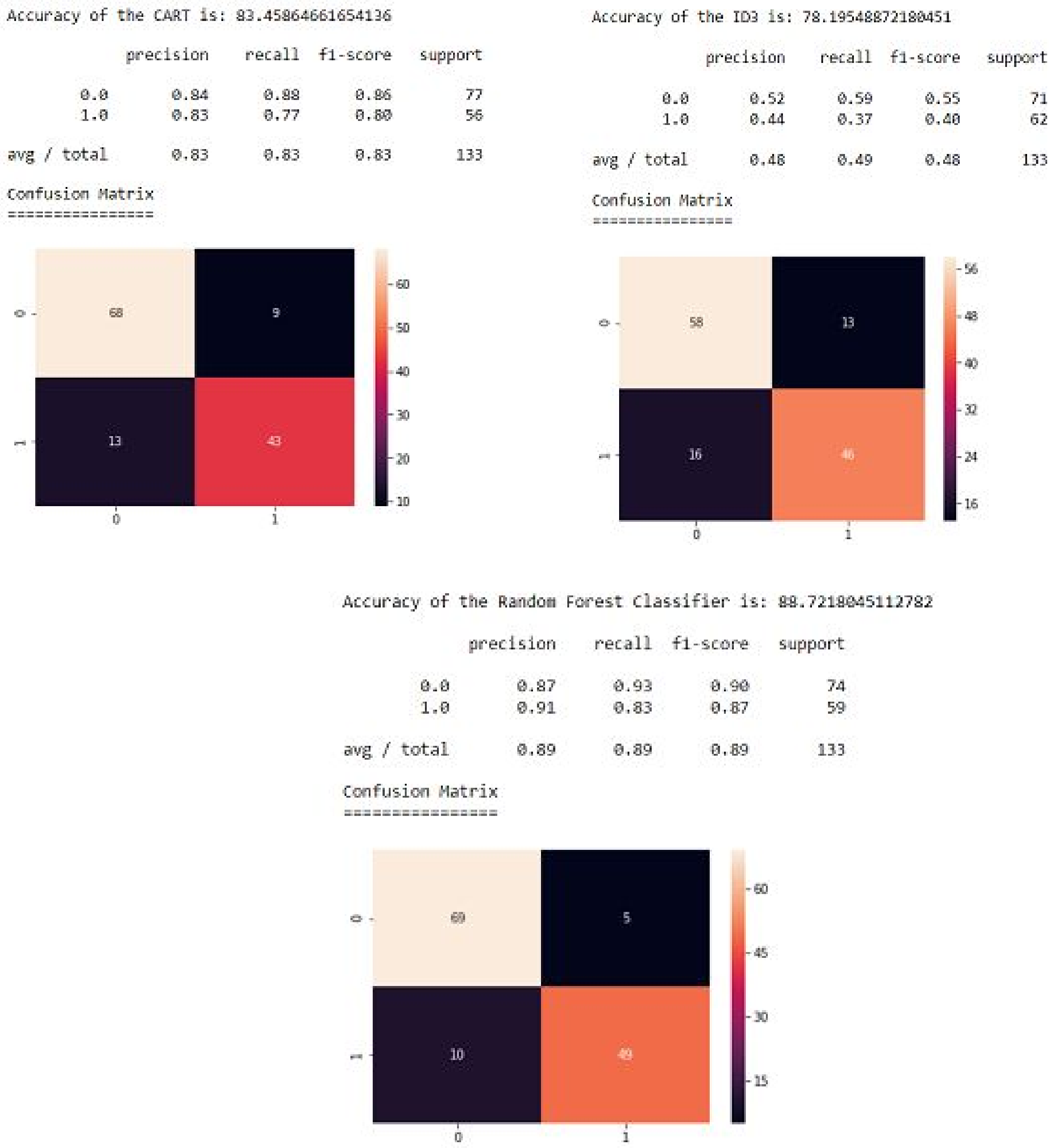


FIGURE 2: Visualization of data of heart disease patients

5. EXPERIMENT AND RESULTS:

5.1 Evaluation and Results:



```
Accuracy of the CART is: 83.45864661654136

             precision    recall  f1-score   support

        0.0       0.84      0.88      0.86        77
        1.0       0.83      0.77      0.80        56

avg / total       0.83      0.83      0.83       133

Confusion Matrix
================
```

```
Accuracy of the ID3 is: 78.19548872180451

             precision    recall  f1-score   support

        0.0       0.52      0.59      0.55        71
        1.0       0.44      0.37      0.40        62

avg / total       0.48      0.49      0.48       133

Confusion Matrix
================
```

```
Accuracy of the Random Forest Classifier is: 88.7218045112782

             precision    recall  f1-score   support

        0.0       0.87      0.93      0.90        74
        1.0       0.91      0.83      0.87        59

avg / total       0.89      0.89      0.89       133

Confusion Matrix
================
```

FIGURE 3: Accuracy of CART, ID3 and Random Forest

We have carried out some experiments in order to evaluate the performance and usefulness of different classification algorithms for predicting heart patients. Figure 3 shows the accuracy of different classification algorithms used in our experiment. It also shows the precision, recall, f1-score, support and confusion matrix of each of the algorithms.

5.2 Error Analysis:

TABLE : Training and Simulation Error

Evaluation Criteria	CART	ID3	Random Forests
Cohen Kappa Score	0.657	0.560	0.769
Mean Absolute Error	0.165	0.218	0.112
Root Mean Squared Error	0.406	0.466	0.335

TABLE 1: Parameters for evaluating algorithm performance

In the above given table we have noted the Cohen Kappa Score, Mean Absolute Error and Root Mean Squared Error to compare the three classification techniques used. From the values we conclude that Random Forests have the highest Cohen Kappa Score, the lowest Mean Absolute Error and the lowest Root mean squared error, thus making it the best choice algorithm.

6. CONCLUSION AND FUTURE WORK:

6.1 Conclusion:

In this paper, different classifiers are studied and the experiments conducted to find the best classifier for predicting the patient with heart disease. We have proposed three approaches to predict the heart diseases using data mining techniques. Two classifiers i.e. ID3, CART and and an ensemble classifier- Random Forest are used for diagnosis of

patients with heart diseases. According to our experiment Random Forests' performance has more accuracy, when compared with other two classification methods. The best algorithm based on the patient's data is Random Forest Classification with accuracy of 88.72% and lowest average error at 0.112 compared to others. Average error for ID3 and CART is 0.218 and 0.165 respectively. Out of the two classifiers CART(83.45%) has more accuracy than ID3(78.14%).
The empirical results show that we can produce short but accurate prediction list for the heart patients by applying the predictive models to the records of new patients. This study will also work to identify those patients who needed special attention.

6.2 Future Work

There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system. Due to time limitation, the following research work needs to be performed in the future.

1. To make use of testing different discretization techniques, multiple classifiers Voting technique and different Decision tree types like information gain,gain ratio and Gini index. Eg. Experiment need to perform on use of Equal Frequency Discretization Gain Ratio Decision Trees by applying nine Voting scheme in order to enhance the accuracy and performance of diagnosis of heart disease.
2. This report proposes a framework using combinations of CART , ID and Random Forest to arrive at an accurate prediction of heart disease. Further work involves development of system using the mentioned methodology to be use for checking the imbalance with other data mining models.
3. To explore different rules such as Association, Clustering, K-means etc for better efficiency and ease of simplicity.
4. Continuous data instead of categorical data can also be analysed in further study for better results. Also, text mining, image processing and other such techniques can also be used to include unstructured healthcare data in this study and hence enhance the dataset used.

7. REFERENCES:

1. Dataset-https://archive.ics.uci.edu/ml/datasets/Heart+Disease
2. http://scikit-learn.org/stable/modules/tree.html
3. https://en.wikipedia.org/wiki/ID3_algorithm
4. http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomFores tClassifier.html
5. https://en.wikipedia.org/wiki/Decision_tree_l earning
6. https://en.wikipedia.org/wiki/Ensemble_learning
7. https://pdfs.semanticscholar.org/b343/df5ffd7534f15c00b357fe2d0d1e86243648.pdf
8. https://en.wikipedia.org/wiki/Random_forest