Predicting Presence of Heart Disease in Patients using Decision Tree ClassificationTechniques

Rajiv Gupta
Master of Science in IT
Hood College
Frederick, MD, USA

Abstract – **Heart disease is the leading cause of death in the world over the past 10 years. Almost one person dies of heart disease about every minute in the United States alone. Researchers have been using several data mining techniques to help health care professionals in the diagnosis of heart disease. However using data mining technique can reduce the number of test that are required. In order to reduce number of deaths from heart diseases there have to be a quick and efficient detection technique. Decision Tree is one of the successful data mining techniques used. This research compares different algorithms of Decision Tree classification seeking better performance in heart disease diagnosis using WEKA. The algorithms which are tested is J48 algorithm, Logistic model tree algorithm and Random Forest algorithm. The existing datasets of heart disease patients from Cleveland database of UCI repository is used to test and justify the performance of decision tree algorithms. This datasets consists of 303 instances and 76 attributes. Subsequently, the classification algorithm that has optimal potential will be suggested for use in sizeable data. The goal of this study is to extract hidden patterns by applying data mining techniques, which are noteworthy to heart diseases and to predict the presence of heart disease in patients where this presence is valued from "no presence" to "likely presence".**

Keywords: **Data Mining, Decision Tree, Heart Disease, J48, Logistic Model Tree, Random Forest, WEKA**

I. Introduction

Heart disease is the leading cause of death in the world over the past 10 years (World Health Organization 2007). The European Public Health Alliance reported that heart attacks, strokes and other circulatory diseases account for 41% of all deaths (European Public Health Alliance 2010).

Several different symptoms are associated with heart disease, which makes it difficult to diagnose it quicker and better. Working on heart disease patients databases can be compared to real-life application. Doctor's knowledge to assign the weight to each attribute. More weight is assigned to the attribute having high impact on disease prediction. Therefore it appears reasonable to try utilizing the knowledge and experience of several specialists collected in databases towards assisting the diagnosis process. It also provides healthcare professionals an extra source of knowledge for making decisions.

The healthcare industry collects large amounts of healthcare data and that need to be mined to discover hidden information for effective decision making. Motivated by the world-wide increasing mortality of heart disease patients each year and the availability of huge amount of patients' data from which to extract useful knowledge, researchers have been using data mining techniques to help health care professionals in the diagnosis of heart disease (Helma, Gottmann et al. 2000).

Data mining is the exploration of large datasets to extract hidden and previously unknown patterns, relationships and knowledge that are difficult to detect with traditional statistical methods (Lee, Liao et al. 2000). Thus data mining refers to mining or extracting knowledge from large amounts of data. Data mining applications will be used for better health policy-making and prevention of hospital errors, early detection, prevention of diseases and preventable hospital deaths (Ruben 2009).

Heart disease prediction system can assist medical professionals in predicting heart disease based on the clinical data of patients [3]. Hence by implementing a heart disease prediction system using Data Mining techniques and doing some sort of data mining on various heart disease attributes, it can able to predict more probabilistically that the patients will be diagnosed with heart disease.

This paper presents a new model that enhances the Decision Tree accuracy in identifying heart disease patients. It uses the different algorithm of Decision Trees.

Literature Review

Prediction of heart disease using data mining techniques has been an ongoing effort for the past two decades. Most of the papers have implemented several data mining techniques for diagnosis of heart disease such as Decision Tree, Naive Bayes, neural network, kernel density, automatically defined groups, bagging algorithm and support vector machine showing different levels of accuracies (Yan, Zheng et al. 2003; Andreeva 2006; Das, Turkoglu et al. 2009; Sitar-Taut, Zdrenghea et al. 2009; Raj Kumar and Reena 2010; Srinivas, Rani et al. 2010) on multiple databases of patients from around the world.

One of the bases on which the papers differ are the selection of parameters on which the methods have been used. Many authors have specified different parameters and databases for testing the accuracies. In particular, researchers have been investigating the application of the Decision Tree technique in the diagnosis of heart disease with considerable success. Sitair-Taut et al. used the weka tool to investigate applying Naive Bayes and J4.8 Decision Trees for the detection of coronary heart disease. Tu et al. used the bagging algorithm in the weka tool and compared it with J4.8 Decision Tree in the diagnosis of heart disease. In [9], the decision making process of heart disease is effectively diagnosed by Random forest algorithm. In [10] based on the probability of decision support, the heart disease is predicted. As a result the author concluded that decision tree performs well and sometimes the accuracy is similar in Bayesian classification.

In year 2013, S. Vijiyarani et. al. [4] performed a work, “An Efficient Classification Tree Technique for Heart Disease Prediction**”**. This paper analyzes the classification tree techniques in data mining. The classification tree algorithms used and tested in this paper are Decision Stump, Random Forest and LMT Tree algorithm. The objective of this research was to compare the outcomes of the performance of different classification techniques for a heart disease dataset.

II. BACKGROUND

Millions of people are getting some sort of heart disease every year and heart disease is the biggest killer of both men and women in the United States and around the world. The World Health Organization (WHO) analyzed that twelve million deaths occurs worldwide due to Heart diseases. In almost every 34 seconds the heart disease kills one person in world.

Medical diagnosis plays vital role and yet complicated task that needs to be executed efficiently and accurately. To reduce cost for achieving clinical tests an appropriate computer based information and decision support should be aided. Data mining is the use of software techniques for finding patterns and consistency in sets of data. Also, with the advent of data mining in the last two decades, there is a big opportunity to allow computers to directly construct and classify the different attributes or classes.

Knowledge of the risk factors associated with heart disease helps health care professionals to identify patients at high risk of having heart disease. Statistical analysis has identified risk factors associated with heart disease to be age, blood pressure, total cholesterol, diabetes, hyper tension, family history of heart disease, obesity and lack of physical exercise, fasting blood sugar etc [7].

Researchers have been applying different data mining Techniques to help health care professionals with improved accuracy in the diagnosis of heart disease. Neural network, Naive Bayes, Decision Tree etc. are some techniques used in the diagnosis of heart disease.

Applying Decision Tree techniques has shown useful accuracy in the diagnosis of heart disease. But assisting health care professionals in the diagnosis of the world’s biggest killer demands higher accuracy. Our research seeks to improve diagnosis accuracy to improve health outcomes.

Decision Tree is one of the data mining techniques that cannot handle continuous variables directly so the continuous attributes must be converted to discrete attributes. Couple of Decision Tree use binary discretization for continuous-valued features. Other important accuracy improving is applying reduced error pruning to Decision Tree in the diagnosis of heart disease patients. Intuitively, more complex models might be expected to produce more accurate results, but which techniques is best? Seeking to thoroughly investigate options for accuracy improvements in heart disease diagnosis this paper systematically investigates comparing multiple classifiers decision tree technique.

This research uses Waikato Environment for Knowledge Analysis (WEKA). The data of UCI repository often presented in a database or spreadsheet. In order to use this data for WEKA tool, the data sets need to be in the ARFF format (attribute-relation file format). WEKA tool is used for to preprocess the dataset. After reviewing all these 76 different attributes, the unimportant attributes is dropped and only the important attributes (i.e. 14 attributes in this case) is considered for analysis to yield more accurate and better results. The 14th one is basically a predicted attribute, which is referred as Class. With thorough comparison between different decision tree algorithms within WEKA tool and deriving the decisions out of it, would help the system to predict the likely presence of heart disease in the patient and will definitely help to diagnose heart disease well in advance and able to cure it in right time.

III. APPROACH AND METHODOLOGY

The following objectives are set for this heart prediction system.

1) The prediction system should not assume any prior knowledge about the patient records it is comparing.

2) The chosen system must be scalable to run against large database with thousands of data.

This chosen approach is implemented using WEKA tool. WEKA is an open source software tool which consists of a collection of machine learning algorithms for Data Mining tasks. It contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization [9]. For testing, the classification tools and explorer mode of WEKA are used. Decision Tree classifiers with Cross Validation 10-fold in Test mode is considered for this study.

The following steps are performed in WEKA.

- Start the WEKA Explorer.
- Open .CSV dataset file & save in .ARFF format.
- Click on Classify tab & select J48 etc (from Trees) from choose button.
- Select appropriate Test mode option.
- Click on Start button & result will be displayed.

DATA

For comparing various Decision Tree classification techniques, Cleveland dataset from UCI repository is used, which is available at http://archive.ics.uci.edu/ml/datasets/ Heart+Disease. The dataset has 76 attributes and 303 records. However, only 13 attributes are used for this study & testing as shown in Table 1.

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	1 = male 0 = female
Cp	Discrete	Chest pain type: 1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 =asymptomatic
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar > 120 mg/dl: 1 = true 0 = false
Restecg	Discrete	Resting electrocardiographic results: = normal = having ST-T wave abnormality =showing probable or define left ventricular hypertrophy
Thalach	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise induced angina: 1 = yes 0 = no
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment : 1 = up sloping 2 = flat 3= down sloping
Ca	Discrete	Number of major vessels colored by fluoroscopy that ranged between 0 and 3.
Thal	Discrete	3 = normal 6= fixed defect 7= reversible defect
Class	Discrete	Diagnosis classes: 0 = No Presence 1=Least likely to have heart disease 2= >1 3= >2 4=More likely have heart disease

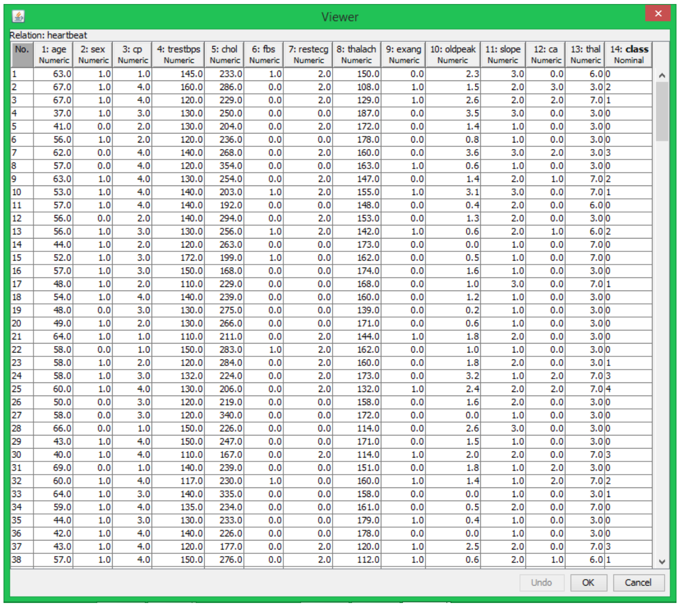
Table 1: Selected Heart Disease Attributes

This paper has emphasized specifically on decision tree classifiers for heart beat prediction within WEKA. Decision tree was considered here among all types of Data mining techniques due to these below reasons. Decision tree filters are easy to implement and easy to understand. It is a method commonly used in data mining. Decision tree is one of the data mining techniques showing considerable success when compared to other data mining techniques. It is a decision support system that uses a tree-like graph decisions. Decision trees are the most powerful approaches in knowledge discovery and data mining. Decision trees are highly effective tools in many areas such as data and text mining, information extraction, machine learning, and pattern recognition. It can handle input data like Nominal, Numeric & Text. It is able to process erroneous datasets or missing values.

A Decision Tree is used to learn a classification function which concludes the value of a dependent attribute (variable) given the values of the independent (input) attributes. This verifies a problem known as supervised classification because the dependent attribute and the counting of classes (values) are given [4]. Tree complexity has its effect on its accuracy. Usually the tree complexity can be measured by a metrics that contains: the total number of nodes, total number of leaves, depth of tree and number of attributes used in tree construction. Tree size should be relatively small that can be controlled by using a technique called **pruning** [13].

Univariate decision tree approach will be used here. In this technique, splitting is performed by using one attribute at internal nodes. This study can able to distinguish the dominant attributes and provides different labels of LIKELY PRESENCE for heart disease.

In this paper, three decision tree algorithms namely J48 algorithm, logistic model tree algorithm and Random Forest decision tree algorithm are used for comparison. The proposed methodology involves reduced error pruning, confident factor and seed parameters to be considered in the diagnosis of heart disease patients. Reduced error pruning has shown to drastically improve decision tree performance. These three decision tree algorithms are then tested to identify which combination will provide the best performance in diagnosing heart disease patients.



Relation: heartbeat

No.	1: age Numeric	2: sex Numeric	3: cp Numeric	4: trestbps Numeric	5: chol Numeric	6: fbs Numeric	7: restecg Numeric	8: thalach Numeric	9: exang Numeric	10: oldpeak Numeric	11: slope Numeric	12: ca Numeric	13: thal Numeric	14: **class** Nominal
1	63.0	1.0	1.0	145.0	233.0	1.0	2.0	150.0	0.0	2.3	3.0	0.0	6.0	0
2	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	2.0	3.0	3.0	2
3	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	2.0	2.0	7.0	1
4	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	3.0	0.0	3.0	0
5	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	1.0	0.0	3.0	0
6	56.0	1.0	2.0	120.0	236.0	0.0	0.0	178.0	0.0	0.8	1.0	0.0	3.0	0
7	62.0	0.0	4.0	140.0	268.0	0.0	2.0	160.0	0.0	3.6	3.0	2.0	3.0	3
8	57.0	0.0	4.0	120.0	354.0	0.0	0.0	163.0	1.0	0.6	1.0	0.0	3.0	0
9	63.0	1.0	4.0	130.0	254.0	0.0	2.0	147.0	0.0	1.4	2.0	1.0	7.0	2
10	53.0	1.0	4.0	140.0	203.0	1.0	2.0	155.0	1.0	3.1	3.0	0.0	7.0	1
11	57.0	1.0	4.0	140.0	192.0	0.0	0.0	148.0	0.0	0.4	2.0	0.0	6.0	0
12	56.0	0.0	2.0	140.0	294.0	0.0	2.0	153.0	0.0	1.3	2.0	0.0	3.0	0
13	56.0	1.0	3.0	130.0	256.0	1.0	2.0	142.0	1.0	0.6	2.0	1.0	6.0	2
14	44.0	1.0	2.0	120.0	263.0	0.0	0.0	173.0	0.0	0.0	1.0	0.0	7.0	0
15	52.0	1.0	3.0	172.0	199.0	1.0	0.0	162.0	0.0	0.5	1.0	0.0	7.0	0
16	57.0	1.0	3.0	150.0	168.0	0.0	0.0	174.0	0.0	1.6	1.0	0.0	3.0	0
17	48.0	1.0	2.0	110.0	229.0	0.0	0.0	168.0	0.0	1.0	3.0	0.0	7.0	1
18	54.0	1.0	4.0	140.0	239.0	0.0	0.0	160.0	0.0	1.2	1.0	0.0	3.0	0
19	48.0	0.0	3.0	130.0	275.0	0.0	0.0	139.0	0.0	0.2	1.0	0.0	3.0	0
20	49.0	1.0	2.0	130.0	266.0	0.0	0.0	171.0	0.0	0.6	1.0	0.0	3.0	0
21	64.0	1.0	1.0	110.0	211.0	0.0	2.0	144.0	1.0	1.8	2.0	0.0	3.0	0
22	58.0	0.0	1.0	150.0	283.0	1.0	2.0	162.0	0.0	1.0	1.0	0.0	3.0	0
23	58.0	1.0	2.0	120.0	284.0	0.0	2.0	160.0	0.0	1.8	2.0	0.0	3.0	1
24	58.0	1.0	3.0	132.0	224.0	0.0	2.0	173.0	0.0	3.2	1.0	2.0	7.0	3
25	60.0	1.0	4.0	130.0	206.0	0.0	2.0	132.0	1.0	2.4	2.0	2.0	7.0	4
26	50.0	0.0	3.0	120.0	219.0	0.0	0.0	158.0	0.0	1.6	2.0	0.0	3.0	0
27	58.0	0.0	3.0	120.0	340.0	0.0	0.0	172.0	0.0	0.0	1.0	0.0	3.0	0
28	66.0	0.0	1.0	150.0	226.0	0.0	0.0	114.0	0.0	2.6	3.0	0.0	3.0	0
29	43.0	1.0	4.0	150.0	247.0	0.0	0.0	171.0	0.0	1.5	1.0	0.0	3.0	0
30	40.0	1.0	4.0	110.0	167.0	0.0	2.0	114.0	1.0	2.0	2.0	0.0	7.0	3
31	69.0	0.0	1.0	140.0	239.0	0.0	0.0	151.0	0.0	1.8	1.0	2.0	3.0	0
32	60.0	1.0	4.0	117.0	230.0	1.0	0.0	160.0	1.0	1.4	1.0	2.0	7.0	2
33	64.0	1.0	3.0	140.0	335.0	0.0	0.0	158.0	0.0	0.0	1.0	0.0	3.0	1
34	59.0	1.0	4.0	135.0	234.0	0.0	0.0	161.0	0.0	0.5	2.0	0.0	7.0	0
35	44.0	1.0	3.0	130.0	233.0	0.0	0.0	179.0	1.0	0.4	1.0	0.0	3.0	0
36	42.0	1.0	4.0	140.0	226.0	0.0	0.0	178.0	0.0	0.0	1.0	0.0	3.0	0
37	43.0	1.0	4.0	120.0	177.0	0.0	2.0	120.0	1.0	2.5	2.0	0.0	7.0	3
38	57.0	1.0	4.0	150.0	276.0	0.0	2.0	112.0	1.0	0.6	2.0	1.0	6.0	1

Fig. 1 Data sample for Creating Decision Tree

A comparison is based on sensitivity, specificity and accuracy by true positive and false positive in confusion matrix. To have a fair comparison between these algorithms, training time in seconds and tree size ratio for each technique is considered with 10-fold stratified cross validation. The overall methodology followed for Decision Tree classification for fulfilling the goal is

Training=>Algorithm=>Model=>Testing=>Evaluation

Classification Tree Algorithms Used

J48 algorithm:

J48 is an open source Java implementation of the C4.5 algorithm in the WEKA tool. This algorithm uses a greedy technique to create decision trees for classification and uses reduced-error pruning [8]. Decision tree is built by analyzing data nodes, which are used to evaluate significance of existing features. J48 algorithm is an extension of ID3 algorithm and possibly creates a small tree. It uses divide and conquers approach to growing decision trees. [5]. At each node of the tree, the algorithm chooses an attribute that can further split the samples into subsets. Each leaf node represents a class or decision.

Basic steps to construct tree are

- Check whether all cases belongs to same class, then the tree is a leaf and is labeled with that class.
- For each attribute, calculate the information and information gain.
- Find the best splitting attribute (depending upon current selection criterion).

J48 with Reduced error Pruning:

Pruning is very important technique to be used in tree creation because of outliers. It also addresses overfitting. Datasets may contain little subsets of instances that are not well defined. To classify them correctly, pruning can be used. Separate and Conquer rule learning algorithm is basis to prune any tree. This rule learning scheme starts with an empty set of rules and the full set of training instances. Reduced-error pruning is one of such separate and conquer rule learning scheme. There are two types of pruning i.e.

- Post pruning (performed after creation of tree)
- Online pruning (performed during creation of tree).

After extracting the decision tree rules, reduced error pruning was used to prune the extracted decision rules. Reduced error pruning is one of the fastest pruning methods and known to produce both accurate and small decision rules (Esposito, Malerba et al. 1997). Applying reduced error pruning provides more compact decision rules and reduces the number of extracted rules.

The run-time complexity of J48 algorithm matches to the tree depth which is linked to tree size and number of examples. So their greatest disadvantage is size of J48 trees, which increases linearly with the number of examples. J48 rules slow for large and noisy datasets. Space complexity is very large as we have to store the values repeatedly in arrays.

Logistic Model Tree Algorithm:

Logistic Model Tree is the classifier for building ‘logistic model trees’, which consist of a decision tree structure with logistic regression function at the leaves. The algorithm can deal with binary and multi-class target variables, numeric and nominal attributes and missing values [11]. A combination of learners that rely on simple regression models if only little and/or noisy data is available and add a more complex tree structure if there is enough data to warrant such a structure. LMT uses cost-complexity pruning. This algorithm is significantly slower than the other algorithms.

As in decision tree, the tested attributes is associated with every inner node. The attributes with k values, the node has k child nodes for nominal attributes and depending on the value of the attribute, the instances are sorted down. For the attributes of numeric, the node has two child nodes and comparing the attributes of tested value to a threshold (the instances are sorted down based on threshold [12].

Logistic Model Trees have been shown to be very accurate and compact classifiers in different research areas. Their greatest disadvantage is the computational complexity of inducing the logistic regression models in the tree. But the prediction of a model is obtained by sorting it down to a leaf and using the logistic prediction model associated with that leaf. A single logistic model is easier to interpret than J48 trees. However building LMTs takes longer time. It can also be shown that trees generated by LMT are much smaller than those generated by J48.

To build a 'logistic model tree' by growing a standard classification tree, building logistic regression models for all nodes, pruning some of the sub-trees using a pruning criterion, and combining the logistic models along a path into a single model in some fashion is performed.

The pruning scheme uses cross-validation to obtain more stable pruning results. Although this increased the computational complexity, it resulted in smaller and generally more accurate trees. These ideas lead to the following algorithm for constructing logistic model trees:

Tree growing starts by building a logistic model at the root using the LogitBoost algorithm. The number of iterations (and simple regression functions f_{mj} to add to F_j) is determined using 10 fold cross-validation. In this process the data is split into training and test set 10 times, for every training set LogitBoost is run to a maximum number of iterations and the error rates on the test set are logged for every iteration and summed up over the different folds. The number of iterations that has the lowest sum of errors is used to train the LogitBoost algorithm on all the data. This gives the logistic regression model at the root of the tree

Like other tree induction methods, LMT does not require any tuning of parameters. LMT produces a single tree containing binary splits on numeric attributes, multi-way splits on nominal ones and logistic regression models at the leaves, and the algorithm ensures that only relevant attributes are included in the latter.

Random Forest Algorithm:

Random forest is an ensemble classifier that consists of many decision trees. The output of the classes is represented by individual trees. It is derived from random decision a forest that was proposed by Tin Kam Ho of Bell Labs in 1995 [12]. This method combines with random selection of features to construct a decision trees with controlled variations. The tree is constructed using algorithm as discussed.

i) Let N be the number of training classes and M be the number of variables in classifier.

ii) The input variable m is used to determine the node of the tree. Note that m<M

iii) Choosing n times of training sets with the replacement of all available training cases N by predicting the classes, estimate the error of the tree.

iv) Choose m variable randomly for each node of the tree and calculate the best split.

v) At last the tree is fully grown and it is not pruned. The tree is pushed down for predicting a new sample. When the terminal node ends up, the label is assigned the training sample. This procedure is iterated over all trees and it is reported as random forest prediction.

Multi-classifiers are the result of combining several individual classifiers. Ensembles of classifiers towards increasing the performance have been introduced [13]. Random Forest (RF) is one of the example of such techniques. RF as a multi-classifier composed by decision trees where every tree h_t had been generated from the set of data training and a vector θ_t of random numbers identically distributed and independent from the vectors . Vectors θ_1, θ_2 ,.., θ_{t-1} used to generate the classifiers h_1, h_2 , .., h_{t-1} . Each decision tree is built from a random subset of the training dataset. It used a random vector that is generated from some fixed probability distribution, where the probability distribution is varied to focus examples that are hard to classify. A random vector can be incorporated into the tree-growing process in many ways. The leaf nodes of each tree are labelled by estimates of the posterior distribution over the data class labels. Each internal node contains a test that best splits the space of data to be classified. A new, unseen instance is classified by sending it down every tree and aggregating the reached leaf distributions.

There are three approaches for Random Forest such as Forest-RI (Random Input selection) and Forest-RC (Random combination) and mixed of Forest-RI and Forest-RC.

The Random Forest technique has some desirable characteristics such as

- It is easy to use, simple and easily parallelized.
- It does not require models or parameters to select except for the number of predictors to choose at random at each node.
- It runs efficiently on large databases; it is relatively robust to outliers and noise
- It can handle thousands of input variables without variable deletion; it gives estimates of what variables are important in the classification
- It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing, it has methods for balancing error in class population unbalanced data sets.

EVALUATION OF CLASSIFICATION ALGORITHMS

The performance of classification algorithms is usually examined by evaluating the sensitivity, specificity, and accuracy of the classification. The sensitivity is proportion of positive instances that are correctly classified as positive (i.e. the proportion of patients known to have the disease, who test positive for it). The specificity is the proportion of negative instances that are correctly classified as negative (i.e. the proportion of patients known not to have the disease, who test negative for it). The accuracy is the proportion of instances that are correctly classified. To measure the stability of the performance of proposed model, the data is divided into training and testing data with 10-fold stratified cross validation. These values are defined as,

Sensitivity = True Positive/(True Positive + False Negative)

Specificity = True Negative/(True Negative +False Positive)

Accuracy = (True Positive + True Negative) / (True Positive + True Negative+ False Negative+ False Positive)

All measures can be calculated based on four values namely True Positive, False Positive, False Negative, and False Positive where,

- True Positive (TP) is a number of correctly classified that an instances positive.
- False Positive (FP) is a number of incorrectly classified that an instance is positive.
- False Negative (FN) is a number of incorrectly classified that an instance is negative.
- True Negative (TN) is a number of correctly classified that an instance is negative.

- F-Measure is a way of combining recall and precision scores into a single measure of performance.
- Recall is the ratio of relevant instances found in the search result to the total of all relevant instances.
- Precision is the proportion of relevant instances in the results returned.
- Receiver Operating Characteristics (ROC) Area is a traditional to plot this same information in a normalized form with 1-false negative rate plotted against the false positive rate.
- For each algorithm, the test option *cross-validation* were used. Instead of reserving a part for testing, the cross-validation repeats the training and testing process several times with different random samples. The standard for this is 10-fold cross-validation. The data is divided randomly into 10 parts in which the classes are represented in the same proportions as in the full dataset (stratification). Each part is held out in turn and the algorithm is trained on the nine remaining parts; then its error rate is calculated on the holdout set. Finally, the 10 error estimates are averaged to yield an overall error estimate. For J48 and Random Forest, all the tests were run with ten different random seeds. Choosing the different random seeds is done to average out statistical variations.

IV. RESULTS

The decision tree classification was performed using J48 algorithm, logistic model trees algorithm and Random Forest algorithm on UCI repository. The experimental results is under the framework of WEKA 3.6.10. All experiment were performed on Quad Core with 2.4GHz CPU and 12GB RAM. The experimental results are partitioned into several sub item for easier analysis and evaluation. The first part is sensitivity (SE), specificity (SP), accuracy (AC), Kappa statistics (KS) and time taken to build model (TTBM) is partitioned in one table while the second part has the relative mean absolute error (RMAE), root mean squared error (RMSE), relative absolute error (RAE) and root relative squared error (RRSE) for reference and evaluation.

J48 with ReducedErrorpruning Algorithm:

The example of J48 algorithm is applied on UCI repository and the confusion matrix is generated for class having 5 possible values are shown in Fig 2. The confusion matrix is important aspect to be considered. From this matrix, classifications can be made. The results of the J48 algorithm are shown in Table 2.

```
-----------Confusion Matrix-----------
  a    b   c    d   e   <-- classified as
 146   8   4    6    0 |   a = 0
  31    9   9   6    0  |   b = 1
   9    5   13   8   1 |   c = 2
  11    7  10   4   3 |   d = 3
  2    5    3    3    0 |   e = 4
```

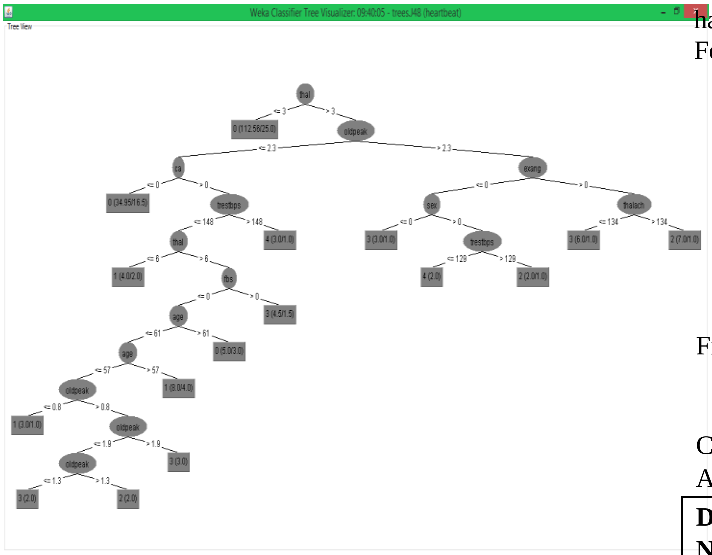
TABLE II
CLASSIFICATION RESULT FOR J48 WITH PRUNING ALGORITHM

Data Name	**SE**	**SP**	**AC (%)**	**KS**	**TTBM (s)**
Heart Disease	0.568	0.752	56.76	0.277	0.04

ERROR RESULT FOR J48 WITH PRUNING ALGORITHM

Data Name	**MAE**	**RMSE**	**RAE (%)**	**RRSE (%)**
Heart Disease	0.20	0.34	79.94	97.22

Fig 3. Visualize Tree of J48 with Reduced Error Pruning:



Logistic Model Tree Algorithm

The example of logistic model trees algorithm is applied on UCI repository and the confusion matrix is generated for class gender having two possible values are shown in Fig 4. The results of LMT algorithm are shown in Table 3.

```
---------Confusion Matrix-----------
  a     b   c   d   e   <-- classified as
148  12   2   1   1 |   a = 0
 31   10   6   8   0 |   b = 1
  8   12   4  10   2 |   c = 2
  4   11  11   7   2 |   d = 3
  0    5   2    6   0 |   e = 4
```

Fig. 4 The confusion matrix of Logistic model trees algorithm.

TABLE III
CLASSIFICATION RESULT FOR LOGISTIC MODEL TREES ALGORITHM

Data Name	SE	SP	AC (%)	KS	TTBM (s)
Heart Disease	0.558	0.781	55.77	0.27	0.39

ERROR RESULT FOR LOGISTIC MODEL TREES ALGORITHM

Data Name	MAE	RMSE	RAE (%)	RRSE (%)
Heart Disease	0.19	0.32	76.2	91.24

Random Forest algorithm

The example of Random Forest algorithm is applied on UCI repository and the confusion matrix is generated for class having 5 values are shown in Fig 5. The results of the Random Forest algorithm are shown in Table 4.

```
-----------Confusion Matrix-----------
  a    b   c   d   e   <-- classified as
152   7   2   3   0 |   a = 0
 34   4  10   5   2 |   b = 1
 10  11   7   7   1 |   c = 2
  5  11  12   5   2 |   d = 3
  1   5   2   3     2 |   e = 4
```

Fig. 5 The confusion matrix of Random Forest algorithm

TABLE V
CLASSIFICATION RESULT FOR RANDOM FOREST ALGORITHM

Data Name	SE	SP	AC (%)	KS	TTBM (s)
Heart Disease	0.561	0.760	56.1	0.26	0.05

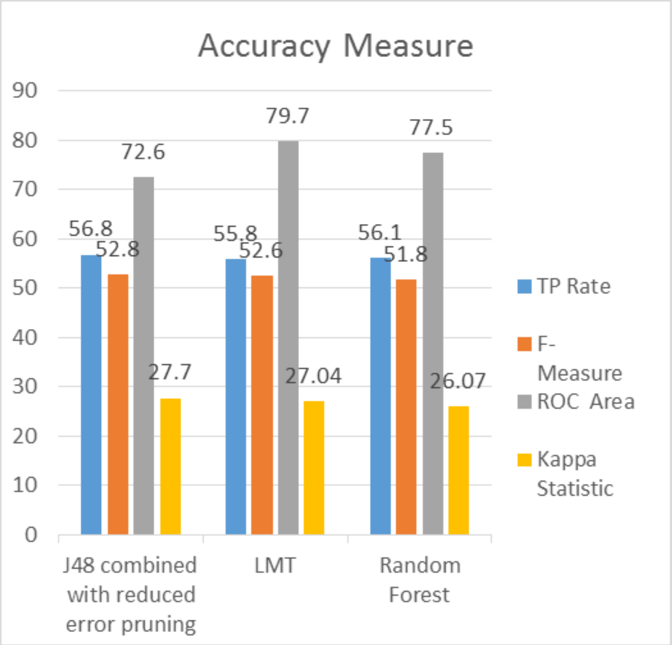
ERROR RESULT FOR RANDOM FOREST ALGORITHM

Data Name	MAE	RMSE	RAE (%)	RRSE (%)
Heart Disease	0.20	0.33	77.66	93.31

Accuracy Measure

The following table shows the accuracy measure of classification techniques. They are the TP rate, F-measure, ROC Area, Kappa Statistics and FP rate.

Fig. 6



From the graph, we analyzed that, TP rate accuracy of J48 performs better when compared to two algorithms. When compared to F-Measure accuracy both LMT and J48 have produced better results than Random Forest. The ROC Area of Random Forest is good, but it attains the highest accuracy in LMT algorithm. At last the accuracy measure of Kappa statistics performs better in J48 algorithm compared to LMT and Random Forest. So in terms of final accuracy among classification tree techniques, J48 outperforms well when compared to LMT and Random forest algorithm.

Error Rate

The table 6 shows the Error rate of classification techniques. The mean absolute error (MAE) is defined as the quantity used to measure how close predictions or forecasts are to the eventual outcomes. The root mean square error (RMSE) is defined as frequently used measure of differences between values predicted by a model and the values actually observed. It is a good measure of accuracy. The relative error (RAE) is a measure of the uncertainty of measurement compared to the size of the measurement. The relative squared error (RRSR) manipulates by taking the total squared error and normalizes it by dividing by the total squared error of the simple predictor. From all error rate analysis, it is concluded that LMT algorithm performs well because it contains least error rate when compared to other two algorithms.

TABLE VI

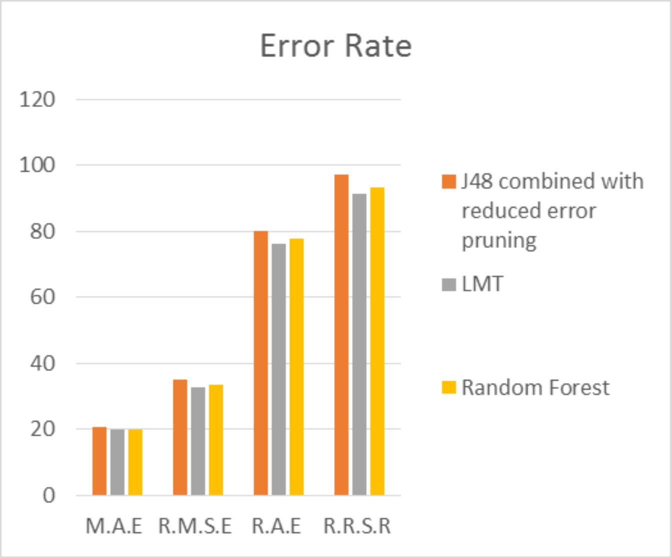
Dataset	**J48 with pruning** tree size	**Random Forest** tree size	**LMT** tree size
Heart Disease	31	10	1

Algorithm	**TP Rate %**	**F-Measure %**	**ROC Area %**	**Kappa Statistic %**
J48 with Reduced error pruning	56.8	52.8	72.6	27.7
LMT	55.8	52.6	79.7	27.04
Random Forest	56.1	51.8	77.5	26.07

TABLE VII

Algorithm	**M.A.E %**	**R.M.S.E %**	**R.A.E**	**R.R.S.R**
J48 with reduced error pruning	20.71	34.93	79.94	97.22
LMT	19.75	32.78	76.20	91.24
Random Forest	20.12	33.52	77.66	93.31

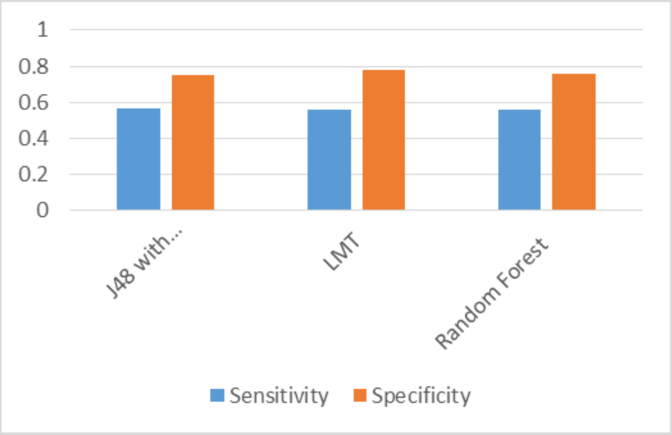
Fig. 7



Algorithm	**Sensitivity %**	**Specificity %**	**Accuracy %**
J48 with ReducedErrorPruning	0.568	0.752	56.76
LMT	0.558	0.781	55.77
Random Forest	0.561	0.760	56.1

Table VIII: Comparison of Different Algorithm Results

Fig. 8



When comparing the results with LMT and Random Forest algorithm, J48 algorithm achieved higher sensitivity and accuracy while LMT achieved higher specificity than J48 and Random Forest algorithm. So overall from Table 10 and Table 11, it is concluded that J48 (with ReducedErrorPruning) has got the best overall performance.

Also, J48 algorithms use reduced-error pruning build less number of trees. The LMT algorithm builds the smallest trees. This could indicate that cost-complexity pruning prunes down to smaller trees than reduced-error pruning, but it also indicate that the LMT algorithm does not need to build large trees to classify the data. The LMT algorithm seems to perform better on data sets with many numerical attributes, whereas for good performance for 3 algorithms, the data sets with few numerical attributes gave a better performance. We can see from the results that J48 is the best classification tree algorithm among the three with pruning method.

V. 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. Like 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 paper proposes a framework using combinations of support vector machines, logistic regression and decision trees 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. Like to explore different rules such as Association, Clustering, K-means etc for better efficiency and ease of simplicity.
4. To make use of Multivariate Decision Tree approach on smaller and larger amount of data.

VI. CONCLUSION

By analyzing the experimental results, it is concluded that J48 tree technique turned out to be best classifier for heart disease prediction because it contains more accuracy and least total time to build. We can clearly see that highest accuracy belongs to J48 algorithm with reduced error pruning followed by LMT and Random Forest algorithm respectively. Also observed that applying reduced error pruning to J48 results in higher performance while without pruning, it results in lower performance. The best algorithm J48 based on UCI data has the highest accuracy i.e. 56.76% and the total time to build model is 0.04 seconds while LMT algorithm has the lowest accuracy i.e 55.77% and the total time to build model is 0.39 seconds

In conclusion, as identified through the literature review, we believe only a marginal success is achieved in the creation of predictive model for heart disease patients and hence there is a need for combinational and more complex models to increase the accuracy of predicting the early onset of heart disease.

VII. REFERENCES

[1] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, “Introduction To Data Mining” Addison Wesley 2006

[2] Ian Witten, Eibe Frank and Mark Hall, “Data Mining ….Practical Machine Learning Tools and Techniques”, Morgan Kaufmann, 2011

[3] Arthur Asuncion and David Newman, “UCI Machine Learning Repository”, 2007 http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[4] Taylor Kulp “book-parser-recommender paper”, 2013 https://code.google.com/p/book-parser-recommender/source/browse/#svn%2Ftrunk%2FPaper

[5] Ian Witten “Data Mining with Weka”,Weka MOOC 2013 https://www.youtube.com/playlist?list=PLm4W7_iX_v4NqPUjceOGd-OKNVO4c_cPD

[6]. INTRODUCTION OF DATA MINING, “Data Mining: What is Data Mining” http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm.

[7] ESCAP (2010) Retrieved 7-February-2011 http://www.unescap.org/stat/data/syb2009/9.Healthrisks-causes-of-death.asp.

[8] Esposito, F., D. Malerba, et al. (1997). "A Comparative Analysis of Methods for Pruning Decision Trees." IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE VOL. 19

[9] Utgoff, Paul E. "Linear Machine Decision Tree." (1991): 6 Feb. 2013.

10] C.L. Blake and C.J. Merz. UCI repository of machine learningdatabases,1998.www.ics.uci.edu/∼mlearn/MLRepository.html

[11]WEKA available at http://www.cs.waikato.ac.nz/ml/weka/ 2006

[12] J.Quinlan C4.5: Programs for Machine Learning. Morgan Kaufmann, 1992.

[13]Ruben D Canlas Jr,”DATA MINING IN HEALTHCARE CURRENT APPLICATIONS AND ISSUES” August 2009