International Journal of Grid and Distributed Computing
Vol. 11, No. 4 (2018), pp.37-48
http://dx.doi.org/10.14257/ijgdc.2018.11.4.04

Parametric Analysis of Heart Attack Prediction Using Machine Learning Techniques

Virender Ranga* and D. Rohila

Department of Computer Engineering, NIT Kurukshetra
**virender.ranga@nitkkr.ac.in*

Abstract

Heart diseases have become one of the most dreadful diseases nowadays around the world. It is a major problem in all groups of ages. Heart diseases are not easy to predictable. However, with the evolving technology, it is not really difficult to accomplish this task of prediction. The key point here is to evaluate datasets containing patient's health factors and analyze the data using some machine learning techniques after which the results can be used for prediction and prevention of this disease. Machine learning techniques are very popular nowadays due to the capability of their learning from massive amounts of data. Also, it is important to ensure that the patient's data remains private to the third parties. In this research paper, our proposed methodology considers a dataset with each row containing the patient's health monitoring factors. We test few well known machine learning techniques for the prediction of heart attacks. We also present a comparison of used machine learning algorithms over different evaluation metrics.

Keywords: *Heart disease, Machine learning, Datasets, classification, clustering, Analysis, Metrics*

1. Introduction

Cardiovascular diseases are the leading cause of death for both men and women globally. Cardiovascular diseases include coronary artery diseases (CAD) such as angina and myocardial infarction (commonly known as heart attack). Heart attack is one of the deadliest diseases existing right now. The primary reason for most heart attacks is a blockage which causes blood flow to one of the coronary arteries, vital channels through which blood travels to the heart muscle, to become reduced or obstructed. Heart attacks have become major concern to everybody because of its sudden occurrence [2]. Heart attack risk factors include age, sex, tobacco, genetics, physical inactivity etc. A number of major contributing risk factors increase the chance of developing heart disease. Some of these can be changed and some cannot. Heart rate variability plays major role in judging the role of autonomic nervous system fluctuations in normal healthy individuals and in patients with various cardiovascular and non-cardiovascular disorders. The chance of developing heart diseases is directly proportional to the number of risk factors. A medical practitioner finds it difficult to predict any chance of occurrence of heart attack since it requires more experience and knowledge. But many methods have been developed to predict the occurrence of the heart attack by considering datasets and some of these methods have very good accuracy percentages [5], [6]. Government or respective authorities may use this analysis and can take the further step to cure or prevention for the disease. Machine learning algorithms also play a major role in heart attack prediction. Machine learning is an application of artificial intelligence (AI) that provides systems to learn and improve from experience without being explicitly programmed [5]. Broadly, machine learning algorithms are classified into the following types:

Received (December 29, 2017), Review Result (February 26, 2018), Accepted (February 28, 2018)

ISSN: 2005-4262 IJGDC

- Supervised learning
- Unsupervised learning
- Reinforcement learning

Our methodology aims here to show the performance of machine learning techniques and discusses comparison between the different techniques under different parameters.

Dataset: The dataset is taken from UCI repository known as Cleveland database [1]. It consists of total 303 patient records. Each row represents one patient record. The record consists of 14 attributes out of which one is the predictable attribute called 'num' whose value indicates the presence and absence of heart attack. Even the presence of heart attack ranges from 1 to 4 indicating the severity of heart disease. The remaining 13 attributes are used in the prediction part. The data problems like missing and inconsistencies in the data are resolved [5]. All the 14 attributes are categorical attributes. The following table describes the dataset taken in the research paper.

Table 1. Dataset Description

S. No	Attribute Name	Attribute Description	Values
1.	Age	Age of the person	No particular range
2.	Sex	Gender of the person (Binary value)	Female=0, Male=1
3.	Cp	Chest pain type	Typical angina=1, Atypical angina=2, Non-angina pain=3, Asymptomatic=4.
4.	Trestbps	Resting blood pressure in mmHg	No particular range
5.	Chol	Serum cholesterol measured in mg/dl.	No particular range
6.	Fbs	Fasting blood sugar (Binary value)	Fasting blood sugar > 120 mg/dl – 1 = True and 0 = False
7.	Restecg	Electrocardiographic value	Normal=0, Having ST-T wave abnormality=1, Showing left-ventricular hypertrophy=2.
8.	Thalach	Maximum heart rate achieved	No particular value
9.	Exang	Exercise induced angine (Binary value)	Yes=1, No=0
10.	Oldpeak	ST depression induced by exercise relative to rest.	No particular range
11.	Slope	Slope of the peak exercise ST segment	Upsloping=1, Flat=2, Downsloping=3
12.	Ca	Number of major vessels colored by fluoroscopy	No particular range
13.	Thal	3 values	Nomal=3, Fixed defect=6, Reversible defect=7
14.	Num	Result (Target variable)	Absence=0, Presence=1 to 4

The aim of this research work is to perform parametric analysis of the UCI dataset using machine learning algorithms. Here we used different machine learning techniques that include classification techniques like k-nearest neighbor, support vector machine, decision tree, random forest and clustering like k-means clustering, EM and artificial neural network. We used some prominent evaluation metrics for the comparison of these algorithms. Generated confusion matrices are also shown to check the effectiveness of the evaluated techniques. The remainder of the paper is organized as follows: Section 2 presents related work. Section 3 shows our proposed work. Implementation of the proposed solutions and the results of the simulations are shown in Sections 4 and 5 respectively. Finally, Section 6 concludes the paper with future scope.

2. Related Work

The authors in [3] extracted the prominent patterns by using preprocessed dataset. After the preprocessing, the new dataset is used as input to extract patterns from k-means clustering algorithm. Using the frequent patterns a new neural network is built which is used for the prediction of heart attack. In [4] comparison of different data mining techniques is shown and weighed patterns are extracted from them. Here, authors assumed five goals and evaluated them using the trained models. In [5] the authors built 3 models using naïve bayes, neural network and decision trees. They have also compared the results and found out the best effective model to be neural network. The work done in [7], [8] used Apache spark to load the dataset and its machine learning libraries for prediction. Three layers model generation and storage layer and a subsequent data analysis layer followed by disease prediction layer are used in total. In [9], [10] automated surveillance systems is built to prove advantageous over manual ones in healthcare. These are built using data mining and machine learning algorithms. The authors in [11] used feature selection to highlight the prominent attributes and used them for prediction to train the models. Different feature selection and ranking attributes methods are used. In [13] an analysis of heart attack and prediction is done in coal mining regions. The authors used different data mining methods based on accuracy and sensitivity.

3. Proposed Work

3.1. Research Methodology:

The following steps are used in our research methodology:

a. Dataset is taken from uci repository [1] and it is preprocessed to make it suitable for algorithms.
b. Simulation is performed on popular Data Mining tool Weka.
c. The predictable attribute is converted to nominal and a merge filter is applied on it, just to the presence and absence of heart disease.
d. Data is divided into training testing parts where training part consists of 60% of dataset and testing part consists of 40% of dataset.
e. Classifiers are loaded with the processed dataset and a training model is achieved which is used to test the testing part of dataset.
f. Clustering techniques are applied on the data and clusters are evaluated with respect to class using the dataset.
g. Classification and clustering algorithms are simulated and results are tabulated for further analysis. Figure 1 shows the flow diagram of classifier.

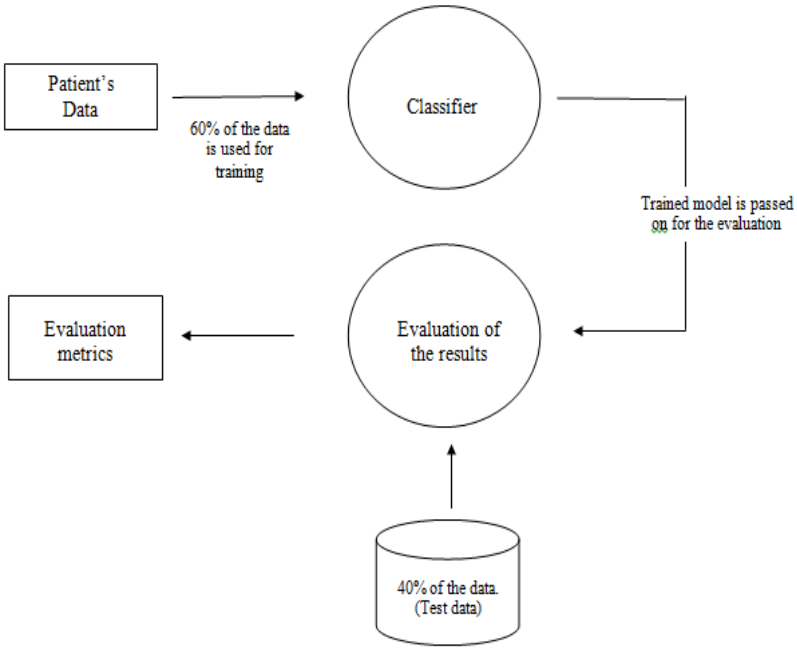

Figure 1. Flow Diagram of Classifier

3.2. Weka

Weka actually is a little bird and it finds only in the islands of New Zealand, but in this case weka is a data mining toolkit. It is a workbench is an acronym for Waikato Environment for Knowledge Analysis and it was produced in the University of Waikato. Weka is used to work on own datasets and perform own data mining without any programming. It is an open source software. Data mining is done using a group of machine learning algorithms embedded on weka. All the tools are incorporated in weka that is required for data preprocessing. It has also customizable options for classification, clustering and feature selection algorithms. Comma separated file format is supported in weka.

3.3. k-nearest Neighbor Classifier

In pattern recognition, the K Nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. It is supervised learning. The 'k' in kNN algorithm is the number of nearest neighbors taken into consideration.

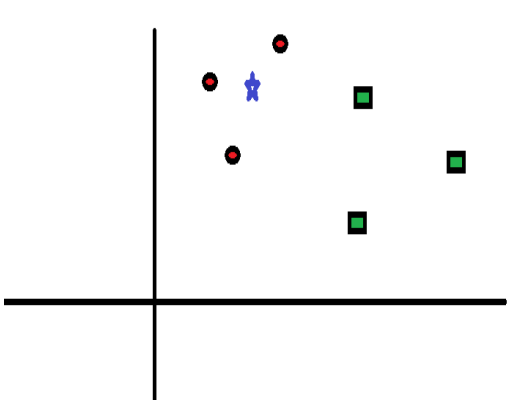

Figure 2. Before k-NN Classification

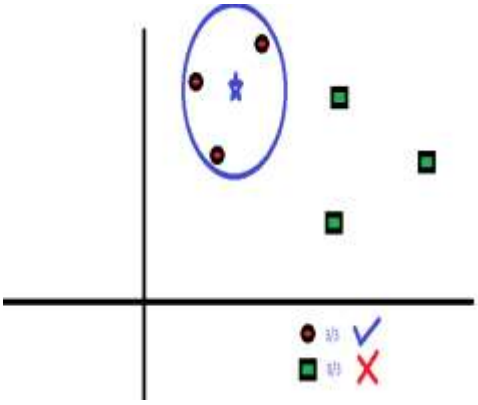

Figure 3. After k-NN Classification

Figures 2 & 3 show the scenario before k-NN classification and after k-NN classification.

Here, 'Bluestar' has to be put in one of the two classes – 'Redcirlcles' or 'Greensquares'. With star as center and radius sufficient to encircle 3 nearest neighbor (k=3), a cicle is drawn. Class of bluestar is defined based on the majority. Weights are assigned to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. According to the following example, a common weighting scheme consists of giving each neighbor a weight of 1/d, where d is the distance to the neighbor.

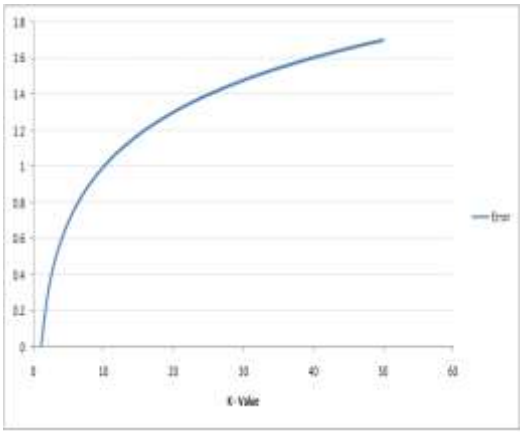

Figure 4. Training Rate vs. k-value

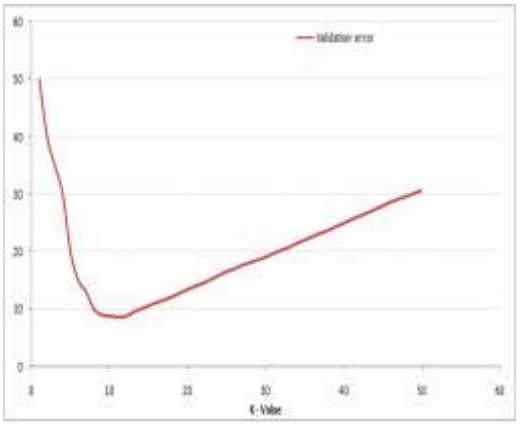

Figure 5. Validation Error vs. k-value

Figure 4 shows the graph between training error rate and the k-value. For $k = 1$, the training error is always zero. If this is the case for validation error, our choice of k would always have been 1. Figure 5 shows the graph between validation error and the k-value. It is observed that error rate initially decreases and reaches a minimal. After the minima point, it is directly proportional to K.

3.4. Artificial Neural Network (ANN)

The word neural network derives its origin from the human brain which consists of a massively large parallel interconnection of a large number of neurons. Brain is the natural neural network which is highly complex, nonlinear and parallel computer. ANN is a mimic of that network. A single perceptron can solve a linearly separable problem. So here we use a multilayer perceptron model. This model classifies instances based on back propagation. This network could be created by an algorithm or by hand or both. During training time this network is monitored and modified. The layers are classified into three categories namely input layer, hidden layer and output layer. In this model, an average of number of attributes and classes is taken and that number of hidden layers is used. The connections between neurons correspond to some weights. So training a neural network involves adjusting all those weights such as if given an input after performing all the calculations it gives the correct output. In this back propagation algorithm first we go in one way from input to output and then propagate back from output to input adjusting weights. There is also a bias unit whose weights are adjusted. And finally an activation function is used to get the correct output according to the training set. The back propagation is repeated until the error is reduced to a very small value. Multilayer perceptron is more advantageous with features like non-linear mapping and noise tolerance. It is more used in data mining because of its good behavior regarding predictive knowledge.

3.5. Decision Tree Classifier

In this algorithm at a particular node, a split of data happens. So the best attribute to split is to be identified. After the split for each value, a child node is created. For each child node if the subset is pure then it would stop otherwise a recursive split happens. This recursion is top down and goes on in dividend conquer manner. There are 3 types of nodes in a decision tree namely root node, branch node and leaf node. Leaf node represents a class. Partitioning of data is stopped when the following conditions are satisfied:

i. When all samples for a given node belong to the same class.
ii. No attributes should be remained for further partitioning and in order to classify leaf and choose majority voting is used.
iii. No data samples should be left.

Once the decision tree is built using that the testing set, it is predicted by traversing across the tree for each sample and choosing the appropriate value at each node.

3.6. Random Forest Classifier

The combination of learning models increases the classification accuracy. This technique is called bagging and the main idea of this is to average noisy and unbiased models in order to create another model with a lower variance in terms of classification. This algorithm works as a large collection of de-correlated decision trees and is based on the above mentioned bagging technique. From the sample set a lot of subsets with random values are created. Thus using the subsets a corresponding decision tree is created for each subset. With all these decision trees different variations of the main classification is achieved. All these decision trees are used to create a ranking of the classifiers. For each sample in testing data, prediction is done using all the decision trees. The majority voted class is selected as the result.

3.7. k-mean Clustering

One of the categories of machine learning is unsupervised learning that is nothing but clustering. Simplest of all the clustering techniques is *k-means*. As the name indicates it is not at all related with *k-nearest neighbor* classifier. In this algorithm k indicates the number of clusters. All the samples of dataset are partitioned into clusters. Each sample is thrown into the cluster with closest mean. The same is done in iterative manner till convergence is reached. Convergence is nothing but no movement of samples between clusters and centroids are also stabilized. This algorithm assumes Euclidean space or distance. The complexity of each iteration happens to be $O(kn)$ where k the number of clusters is and n is the number of instances.

3.8. Expectation-Maximization Clustering

This technique tries to maximize the likelihood or the power of posterior probability. If clusters are overlapped then assignment of instance to a cluster is difficult. So here clusters are modeled as Gaussians. Gaussian mixture models are the extension of k-means model in which clusters are modeled with Gaussian distributions so we have not only their mean but also their covariance that describes their ellipsoidal shape. It is like *k-means* except that it assigns data with soft probability. In the expectation step for each instance the probability that it belongs to a particular cluster is calculated and normalized as shown below.

$$r_{ic} = \frac{\pi_c N(x_i; \mu_{c,} \sum c)}{\sum c' \pi_{c'} N(x_i; \mu_{c',\sum c'})} \quad (1)$$

Where r_{ic} is the normalized probability that it belongs to cluster c, μ_c is the mean of the cluster and $\sum c$ is the variance of the cluster and π_c is the size of cluster. In the maximization step for each cluster c its parameters are updated using the computed r_{ic} values. Thus each step increases the log-likelihood of the model and iterates until convergence. There are other EM algorithms that are less smooth like hard EM, stochastic EM.

3.9. Support Vector Machine

Support Vector Machines are supervised learning models that can be used for classification or regression machine learning problems. The goal of SVM is to design a hyper-plane that classifies all training vectors in two classes. In the basic SVM algorithm, a set of input observations and associated binary outputs are taken and a model is designed that can classify new observations into either of the classes. The training observations are mapped as points in space, separating the observation input sets linearly. There can be many hyper-planes that can divide the training sets but the best choice will be the hyper-plane that leaves the maximum margin from both classes as shown in Figure 6. If there are two hyper-planes A which classifies the vectors accurately but has less margin and B which has higher margin but has small error in classification, hyper-plane A is selected. A partitioning in higher dimensional space by a linear hyper-plane corresponds to a nonlinear partition in the output space. This higher dimensional partitioning is known as the SVM kernel, and can be defined by any mathematical surface. Some of the more common kernels are linear, quadratic, polynomial and Gaussian radial basis function.

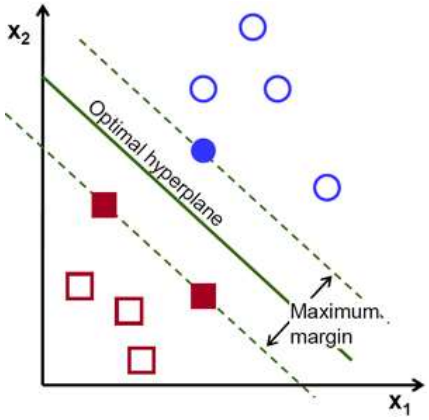

Figure 6. SVM with Hyper-plane

4. Evaluation Metrics

Parametric analysis of heart attack prediction is done using many evaluation metrics. These metrics include Recall, False Positive rate (FPR), F-measure, Accuracy and Precision. Among all these measures accuracy is the well-known measure that indicates closeness to the target result. Dataset complexity can be known using accuracy and FPR. But accuracy alone is not sufficient to decide how good a model is *i.e.*, a high accuracy model needs not be a good model. In a binary classification there might be four possible outcomes TN, TP, FN, FP where TN is actually negative and predicted to be negative similarly TP is true positive that is actually positive and predicted to be positive. Similarly, FN is false negative that is actually positive and predicted to be negative and FP is false positive that is actually negative but predicted to be positive. Accuracy is the proportion of correct classifications (true positives and negatives) from overall number of cases as given in equation 2. Recall is the proportion of correct positive classifications

(true positives) from cases that are actually positive as given in below equation 3. Precision is the proportion of correct positive classification (true positive) from cases that are predicted as positive as given in below equation 4. Recall and precision are a bit more stringent measures that indicate how good the classification system is. Harmonic mean of precision and recall gives the F-measure as given in equation 5. Below given is the confusion matrix that describes the performance of a binary classifier. True positive rate as shown in equation 6 is the recall and also indicates sensitivity of the classifier. Receiver operating characteristic (ROC) curves are used for visual comparison of classification models. This shows the tradeoff between true positive rate and false positive rate as shown below:

Actual Class	Predicted Class	
	Negative	Positive
Negative	**TN**	**FP**
Positive	**FN**	**TP**

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$F - measure = \frac{2TP}{2TP + FP + FN} \tag{5}$$

$$False\ Positive\ rate = \frac{FP}{FP + TN} \tag{6}$$

5. Experimental Results and Discussion

The experiment has been conducted to compare the obtained results by using different classification and clustering algorithms on the same dataset *i.e.*, Cleveland database. Two clustering techniques are used namely *k-means* clustering and *EM* (Expectation-Maximization) clustering. Five classifiers are used namely kNN, SVM, decision tree, random forest and artificial neural network classifiers. Weka of version 3.8.1 is used to simulate the models on Intel i5 system having clock speed of 3.60GHz processor with 4 GB memory. The dataset is first transformed into either csv or arff file format. Then it is preprocessed in weka and the result attribute is converted to nominal. The merge filter is applied on result attribute for all the algorithms. This processed dataset is split into 60% to use it for training the model and the remaining 40% is used to test the trained model.

5.1. Analysis with Classifiers

First kNN classification is used with *k* value as 1. The training model is built with 60% of the dataset. Performance of the kNN classifier is tabulated. For each instance class is predicted and checked with the actual value. Thus the obtained accuracy is 80.17%. Next Artificial neural network model is built using multilayer perceptron and it also used 60/40 split of data. The performance is tabulated. The observed accuracy is 81.82%. Next Decision Tree algorithm is used to build the classifier and the decision tree is also visualized as shown in Figure 7. The achieved accuracy is 81.82%. Next random forest algorithm is used to build the classifier and the accuracy achieved through this being the highest among all the classifiers is 85.95%. Finally, a classifier is built using John Platt's

sequential algorithm and a support vector classifier is trained. This achieved an accuracy of 82.64%.

Table 2. Performance of Classifiers

Name of the classifier	Mean absolute error	FP Rate	Precision	Recall	F-Measure	Area under ROC	Accuracy (in %)
kNN	0.2016	0.198	0.802	0.802	0.802	0.802	80.1653
Artificial neural network	0.1909	0.182	0.818	0.818	0.818	0.889	81.8182
Decision Tree	0.2052	0.183	0.818	0.818	0.818	0.829	81.8182
Random Forest	0.2724	0.142	0.860	0.860	0.859	0.887	85.9504
Support Vector	0.1405	0.143	0.862	0.860	0.859	0.858	82.6400

Table 3. Confusion Metrics

*k*NN Classifier		Predicted Class	
		Absence	Presence
Actual Class	Absence	49	13
	Presence	11	48

Ann Classifier		Predicted Class	
		Absence	Presence
Actual Class	Absence	51	11
	Presence	11	48

Decision tree		Predicted Class	
		Absence	Presence
Actual Class	Absence	52	10
	Presence	12	47

Random forest		Predicted Class	
		Absence	Presence
Actual Class	Absence	55	7
	Presence	10	49

SVM Classifier		Predicted Class	
		Absence	Presence
Actual Class	Absence	56	6
	Presence	11	48

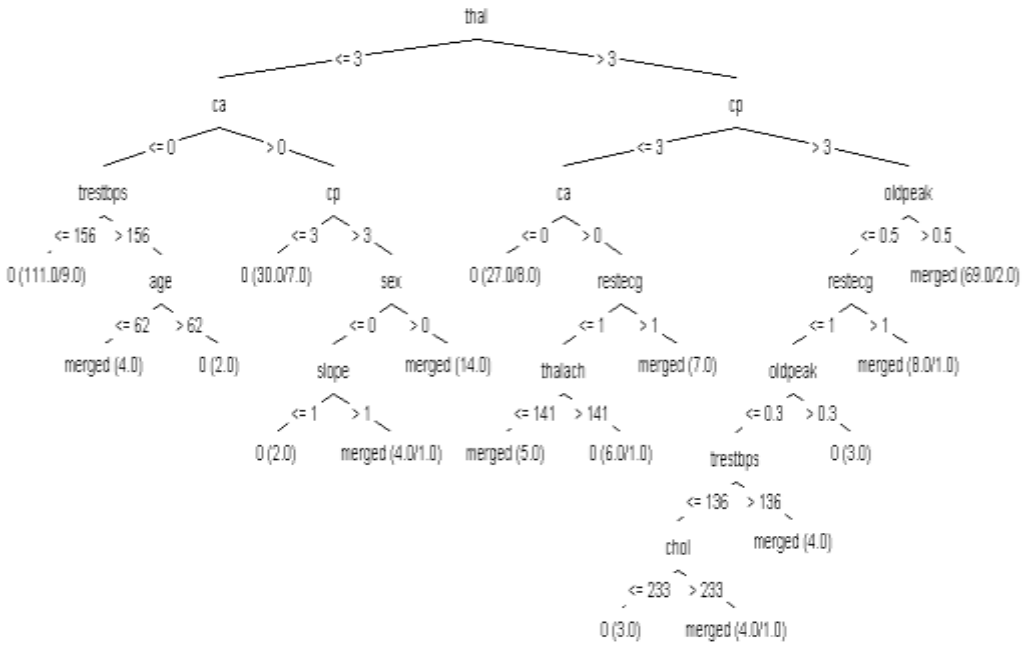

Figure 7. Decision Tree obtained after Classification

5.2. Analysis with Clustering

First *k-means* clustering algorithm is applied on the dataset and each instance is classified into nearby clusters. Clusters are evaluated with respect to class and confusion matrix is visualized. Every instance in the dataset is thrown into one cluster and verified with the actual class. Here the incorrectly clustered instances are 20.13%. Next *EM* clustering is used and the data is classified into 2 clusters in which one is considered presence of heart attack class and the other absence of heart attack class. Here the incorrectly clustered instances are 21.45%.

Table 4. Confusion Metrics of Clustering Techniques

k-means clustering		**Predicted Class**	
		Absence	**Presence**
Actual Class	Absence	129	35
	Presence	26	113
EM clustering		**Predicted Class**	
		Absence	**Presence**
Actual Class	Absence	49	13
	Presence	30	109

6. Conclusion and Future Scope

In this research paper, we have discussed the parametric analysis of heart attack prediction using the UCI dataset. A total of five classification techniques and two clustering techniques are described and their valuation metrics are calculated. According to the results Random forest classifier works better among all the other classifiers for this dataset. It can be observed that due to the less number of instances in the dataset the accuracy percentages and other prominent metrics are not that high. Therefore, in the future we have planned to do the analysis with more instances while taking categorical data not continuous data. Also we will try to perform analysis of continuous or non-numerical data and then study the complexity of datasets.

References

[1] C. L. Blake and C. J. Mertz, "UCI Machine Learning Databases", **(2004)** http://mlearn.ics.uci.edu/databases/heait-disease/.
[2] A. Miller, B. Blott and T. Hames, "Review of neural network applications in medical imaging and signal processing", Med. Biol. Engg. Comp., vol. 30, **(1992)**, pp. 449-464.
[3] S. B. Patil and Y. S. Kumaraswamy, "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network", in European Journal of Scientific Research, EuroJournals Publishing, Inc., ISSN 1450-216X, vol. 31, no. 4, **(2009)**, pp. 642-656.
[4] K. Srinivas, B. Kavihta Rani and A. Govrdhan, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks", IJCSE, vol. 02, no. 02, **(2010)**, pp. 250-255.
[5] S. Palaniappan and R. Awang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", Computer Systems and Applications, 2008. IEEE/ACS International Conference on, Doha, Qatar, **(2008)**, pp. 108-115.
[6] P. Sellappan and S. L. Chua, "Model-based Healthcare Decision Support System", Proc. Of Int. Conf. on Information Technology in Asia CITA'05, 45-50, Kuching, Sarawak, Malaysia, **(2005)**, pp. 45-50.
[7] R. Mehta, "Heart Disease Prediction Using Machine Learning and Big Data Stack - DZone AI", Dzone.com, dzone.com/articles/a-tutorial-on-using-the-big-data-stack-and-machine, **(2017)**.
[8] J. Han and M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, **(2006)**.
[9] M. K. Obenshain, "Application of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, vol. 25, no. 8, **(2004)**, pp. 690-695.
[10] M. Anbarasi, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", International Journal of Engineering Science and Technology, vol. 2, no. 10, **(2010)**, pp. 5370-5376.
[11] H. Decell and J. Quirein, "An Iterative Approach to the Feature Selection Problem", Proc. Purdue Univ. Conf. Machine Processing of Remotely Sensed Data, vol. 1, **(1972)**, pp. 1-3.
[12] Y. Yang, S. Slattery and R. Ghani, "A Study of Approaches to Hypertext Categorization", J. Intelligent Information Systems, vol. 18, no. 2-3, **(2002)**, pp. 219-241**.**
[13] K. Srinivas, "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques", IEEE Transaction on Computer Science and Education (ICCSE), **(2010),** pp. 1344-1349.

