A NEW AND COMPREHENSIVE ANALYSIS FOR PREDICTING HEART DISEASE USING DATA MINING TECHNIQUES

Ms R.A.Taley[1]
Assistant Professor C.O.E&T , Akola

Abstract: *Data mining is a technique that is performed on large databases for extracting hidden patterns by using combinational strategy from statistical analysis, machine learning and database technology. Further, the medical data mining is an extremely important research field due to its importance in the development of various applications in flourishing healthcare domain. While summarizing the deaths occurring worldwide, the heart disease appears to be the leading cause. The identification of the possibility of heart disease in a person is complicated task for medical practitioners because it requires years of experience and intense medical tests to be conducted. In this work, three data mining classification algorithms like Random Forest, Decision Tree and Naïve Bayes are addressed and used to develop a prediction system in order to analyse and predict the possibility of heart disease. The main objective of this significant research work is to identify the best classification algorithm suitable for providing maximum accuracy when classification of normal and abnormal person is carried out. Thus prevention of the loss of lives at an earlier stage is possible. The experimental setup has been made for the evaluation of the performance of algorithms with the help of heart disease benchmark dataset retrieved from UCI machine learning repository. It is found that Random Forest algorithm performs best with 89% precision when compared to other algorithms for heart disease prediction*

Keywords*: Coronary heart disease, Decision tree, K nearest neighbour, Machine Learning, Naïve Bayes, Support vector Machine.*

I. INTRODUCTION

In day to day life many factors that affect a human heart. Many problems are occurring at a rapid pace and new heart diseases are rapidly being identified. 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 essential and its health is to be conserved for a healthy living. The health of a human heart is based on the experiences in a person's life and is completely dependent on professional and personal behaviours of a person. There may also be several genetic factors through which a type of heart disease is passed down from generations. According to the World Health Organization, every year more than 12 million deaths are occurring worldwide due to the various types of heart diseases which is also known by the term cardiovascular disease. The term Heart disease includes many diseases that are diverse and specifically affect the heart and the arteries of a human being. Even young aged people around their 20-30 years of lifespan are getting affected by heart diseases. The increase in the possibility of heart disease among young may be due to the bad eating habits, lack of sleep, restless nature, depression and numerous other factors such as obesity, poor diet, family history, high blood pressure, high blood cholesterol, idle behaviour, family history, smoking and hypertension.

Data Mining is a task of extracting the vital decision making information from a collective of past records for future analysis or prediction. The information may be hidden and is not identifiable without the use of data mining. The classification is one data mining technique through which the future outcome or predictions can be made based on the historical data that is available.

In this research work, the supervised machine learning concept is utilized for making the predictions. A comparative analysis of the three data mining classification algorithms namely Random Forest, Decision Tree and Naïve Bayes are used to make predictions. The analysis is done at several levels of cross validation and several percentage of percentage split evaluation methods respectively. The StatLog dataset from UCI machine learning repository is utilized for making heart disease predictions in this research work. The predictions are made using the classification model

that is built from the classification algorithms when the heart disease dataset is used for training. This final model can be used for prediction of any types of heart diseases.

The following are the type of heart disease: Heart means "cardio". Hence all heart diseases concern to category of cardiovascular diseases. The different kinds of heart disease are:

- Coronary heart diseases.
- Angina pectoris
- Congestive heart failure.
- Cardiomyopathy
- Congenital heart diseases. [9]

Coronary heart disease or coronary artery disease is the narrowing of the coronary arteries. The coronary arteries supply oxygen and blood to the heart. It causes a large number of people to become ill or to face death. It is one of the popular type of heart disease. High blood glucose from diabetes can damage blood vessels and nerves that control heart and blood vessels. If a person has diabetes for a longer time, there are high chances for that person to have heart disease in future. With diabetes, there are other reasons which contribute to heart disease. They are smoking which raises the risk of developing heart disease, high blood pressure makes the heart work harder to pump blood and it can strain heart and damage blood vessels, abnormal cholesterol levels also contribute to heart disease and obesity. Also, family history of heart disease can be a cause of having heart disease. But this history is not considered in this paper for prediction of heart disease.

The other risk factors include age, gender, stress and unhealthy diet. Chance of having a heart disease increases when a person is getting older. Men have a greater risk of heart disease. However, women also have the same risk after menopause. Leading a stressed life can also damage the arteries and increase the chance of coronary heart disease.

So, in this paper based on the factors mentioned above we try to predict the risk of heart disease. A large amount of work has been done related to heart prediction system by using various techniques and algorithms by many authors. These techniques may be based on deep-learning, machine-learning, data mining and so on. The aim of all those papers is to achieve better accuracy and to make the system more efficient so that it can predict the chances of heart attack.

II. LITERATURE REVIEW

Monika Gandhi et.al, [1] used Naïve Bayes, Decision tree and neural network algorithms and analysed the medical dataset. There are a huge number of features involved. So, there is a need to reduce the number of features. This can be done by feature selection. On doing this, they say that time is reduced. They made use of decision tree and neural networks.

J Thomas, R Theresa Princy [2] made use of K nearest neighbour algorithm, neural network, naïve Bayes and decision tree for heart disease prediction. They made use of data mining techniques to detect the heart disease risk rate.

Sana Bharti, Shailendra Narayan Singh [3] made use of Particle Swarm Optimization, Artificial neural network, Genetic algorithm for prediction. Associative classification is a new and efficient technique which integrates association rule mining and classification to a model for prediction and achieved good accuracy.

Purushottam et.al, [4] proposed "An automated system in medical diagnosis would enhance medical care and it can also reduce costs. In this study, we have designed a system that can efficiently discover the rules to predict the risk level of patients based on the given parameter about their health. The rules can be prioritized based on the user's requirement. The performance of the system is evaluated in terms of classification accuracy and the results shows that the system has great potential in predicting the heart disease risk level more accurately".

Sellappan Palaniyappan, Rafiah Awang [5] made use of decision tree Naïve Bayes, Decision tree,

Artificial Neural Networks to build Intelligent Heart Disease Prediction Systems (IHDPS).To enhance visualization and ease of interpretation, it displays the results both in tabular and graphical forms. By providing effective treatments, it also helps to reduce treatment costs. Discovery of hidden patterns and relationships often has gone unexploited. Advanced data mining techniques helped remedy this situation.

Himanshu Sharma,M A Rizvi [6] made use of Decision tree, support vector machine, deep learning, K nearest neighbour algorithms. Since the datasets contain noise, they tried to reduce the noise by cleaning and pre-processing the dataset and also tried to reduce the dimensionality of the dataset. They found that good accuracy can be achieved with neural networks.

Animesh Hazra et.al, [7] discussed in detail the cardiovascular disease and different symptoms of heart attack. The different types of classification and clustering algorithms and tools were used.
V.Krishnaiah, G.Narsimha, N.Subhash Chandra [8] presented an analysis using data mining. The analysis showed that using different techniques and taking different number of attributes gives different accuracies for predicting heart diseases.

Ramandeep Kaur, Er.Prabhsharn Kaur [9] have showed that the heart disease data contains unnecessary, duplicate information. This has to be pre processed. Also, they say that feature selection has to be done on the dataset for achieving better results.

J.Vijayashree and N.Ch.SrimanNarayanaIyengar [10] used data mining. A huge amount of data is produced on a daily basis. As such, it cannot be interpreted manually. Data mining can be effectively used to predict diseases from these datasets. In this paper, different data mining techniques are analysed on heart disease database. In conclusion, this paper analyses and compares how different classification algorithms work on a heart disease database.

Benjamin EJ et.al [11] says that there are seven key factors for heart disease such as smoking, physical inactivity, nutrition, obesity, cholesterol, diabetes and high blood pressure. They also discussed the statistics of heart disease including stroke and cardio vascular disease.

Abhay Kishore et.al [12] on their experimentation showed that recurrent neural network gives good accuracy when compared to other algorithms like CNN, Naïve Bayes and SVM. Hence, neural networks perform well in heart disease prediction. They also achieved a system that could predict silent heart attacks and inform the user as earliest possible.

M.Nikhil Kumar et.al [13] used various algorithms – Decision tree, random forest, Naïve Bayes, KNN, Support vector machine, logistic model tree algorithm. Naïve Bayes algorithm gave good results when compared to other algorithms. They made use of UCI repository of heart disease dataset. Also, J48 algorithm took less time to build and gave good results.

Amandeep Kaur et.al [14] compared various algorithms such as artificial neural network, K – nearest neighbour, Naïve Bayes, Support vector machine on heart disease prediction.

Stephen F Weng et.al [15] used four machine learning algorithms such as logistic regression, random forest, gradient boosting machines and neural networks. They showed that machine learning algorithms perform well at predicting the heart disease cases correctly. They say that this is the first experimentation using machine learning techniques to routine patient data in electronic records. The source of the dataset is the Clinical Practice Research Datalink (CPRD). These are the electronic medical records which contains all the medical related data such as statistics of human population, medical history, specialists. It also contains details of medicine intake, outcomes and details of hospital admissions.

Sahaya Arthy et.al [16] analyse the existing works on heart disease prediction which uses data mining. The data mining techniques are commonly used in heart disease prediction. They also discuss the databases used such as the heart disease dataset

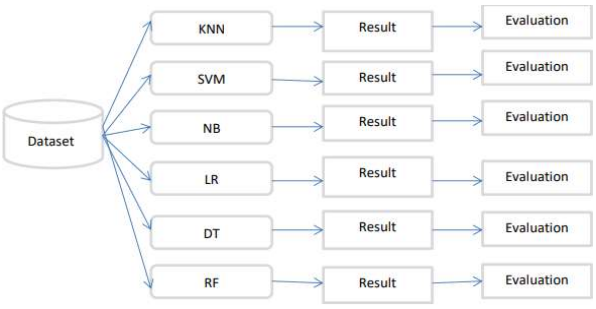
from UCI repository, tools used such as Weka, Rapid Miner, Data melt, Apache Mahout, Rattle, KEEL, R data mining and so on. They conclude that use of single algorithm results in better accuracy in prediction. But use of hybridization of two or more algorithms can enhance and improve the heart disease prediction with good accuracy.

A.Sudha et.al [18] discusses the data mining technology. They also propose an architecture diagram which includes the steps – dataset collection, normalization and pre-processing, dimensionality reduction using Principal Component analysis, feature subset selection, classification algorithm and result analysis. They made use of three classifiers decision tree, Naïve Bayes and neural networks. They conclude that neural networks perform well than other classifiers.

III. PROPOSED METHODOLOGY

In this paper, comparison of various machine learning methods is done for predicting the 10 year risk of coronary heart disease of the patients from their medical data. The following is the flowchart for proposed methodology:

FIGURE 1: PROPOSED WORK



The heart disease data set is taken as input. It is then pre-processed by replacing non-available values with column means.

Four different methods were used in this paper. The different methods used are depicted in figure 3.The output is the accuracy metrics of the machine learning models. The model can then be used in prediction.

K-Nearest Neighbours (KNN)

KNN is a non-parametric machine learning algorithm. The KNN algorithm is a supervised learning method. This means that all the data is labelled and the algorithm learns to predict the output from the input data. It performs well even if the training data is large and contains noisy values.

The data is divided into training and test sets. The train set is used for model building and training. A k- value is decided which is often the square root of the number of observations. Now the test data is predicted on the model built. There are different distance measures. For continuous variables, Euclidean distance, Manhattan distance and Minkowski distance measures can be used.

However, the commonly used measure is Euclidean distance. The formula for Euclidean distance is as follows:

$$d = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

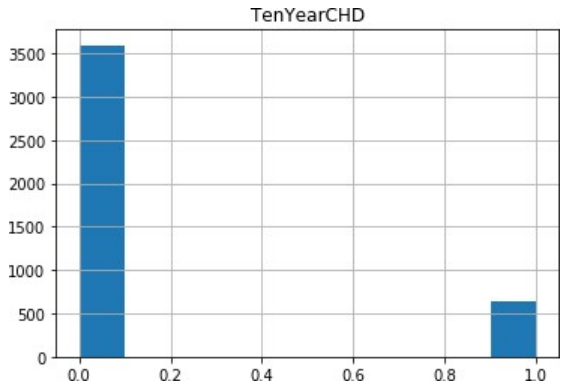
The ROC curve for k-nearest neighbour is depicted in figure 5.

Support Vector Machine (SVM)

The SVM algorithm is used to predict this disease by plotting the train dataset where a hyper plane classifies the points into two - presence and absence of heart disease. SVM works by identifying the hyper plane which maximises the margin between two classes. The ROC curve is depicted in figure 7.

Here, penalized SVM is used to handle class imbalance. Class imbalance is a problem in machine learning when total number of positive and negative class is not the same. If the class imbalance is not handled then the classifier will not perform well. The following plot shows the class imbalance.

FIGURE 2: PLOT SHOWING CLASS IMBALANCE



SVM algorithms uses a set of mathematical functions called kernel. In this proposed methodology, linear kernel is used.

$$K(x,x') = \exp((-\|x-x'\|^2)/2\sigma^2)$$

The performance of the SVM classifier can be increased by fine-tuning the hyper parameters. This can be done by using Grid Search CV. Different values of C can be given as input to this method. It builds different SVM models with given values and then finds the best value of c for which the model performs well.

Random Forest Method (RFM)

The algorithm for random forest is given below:
Step 1: Randomly select k features from entire m features, where $k << m$.
Step 2: Surrounded by the k features, calculate the node "d" using the best split point.
Step 3: Split the node into daughter nodes using the best split.
Step 4: Repeat 1 to 3 steps until l number of nodes has been reached.
Step 5: Construct forest by repeating steps 1 to 4 for n number times to create n number of trees.

Logistic regression

Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

Below are the steps:
1. Data Pre-processing step.
2. Fitting **Logistic Regression** to the Training set.
3. Predicting the test result.
4. Test accuracy of the result(Creation of Confusion matrix)
5. Visualizing the test set result

Naive Bayes algorithm (NB)

This is a classification algorithm which is used when the dimensionality of the input is very high. A Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. It is based on Bayes theorem. The Bayes theorem is as follows:

$$P(Y/X) = P(X/Y)\ P(X)$$

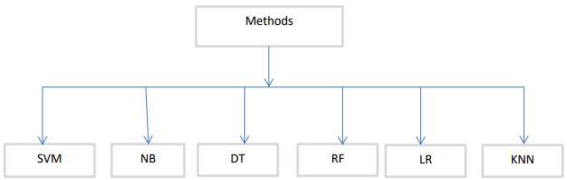
This calculates the probability of Y given X where X is the prior event and Y is the dependence event. The ROC curve is depicted in figure 6.

It needs less training data. It can be used for binary classification problems and is very simple.

Decision trees

Decision trees is one of the ways to display an algorithm. It is a classic machine learning algorithm. In heart disease, there are several factors such as cigarette, BP, Hypertension, age etc. The challenge of the decision tree lies in the selection of the root node. This factor used in root node must clearly classify the data. We make use of age as the root node. The ROC curve is depicted in figure 4. The decision tree is easy to interpret. They are non-parametric and they implicitly do feature selection.

FIGURE 3: METHODS USED



IV. RESULTS AND DISCUSSION

Data source

Cleveland Clinic Foundation Heart disease dataset has been collected at the University of California, Irvine. The Dataset has 76 raw attributes. However, all the published experiments only refer to 13 of them, because these features are considered the key attributes based on experienced cardiac clinicians and other features have so many missing values. The dataset of Cleveland contains 303 rows, which 297 instances of them are complete. Six instances contain missing values and they are removed from the experiment. This dataset can be downloaded from this address: https://archive.ics.uci.edu/ml/datasets/Heart+Disease. The 13 attributes are used as input. The dataset has 1 attribute as class of heart disease with two output classes denoted as one (heart failure presence) and zero (heart failure absence)

The machine learning models is evaluated using the AUC-ROC metric. This can be used to understand the model performance.

The ROC curve of the algorithms is as follows:

FIGURE 4: ROC CURVE FOR DECISION TREE

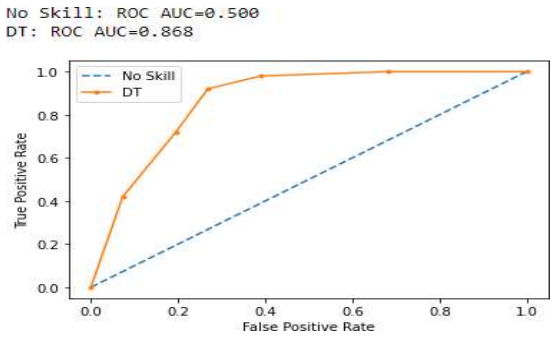


FIGURE 5: ROC CURVE FOR KNN

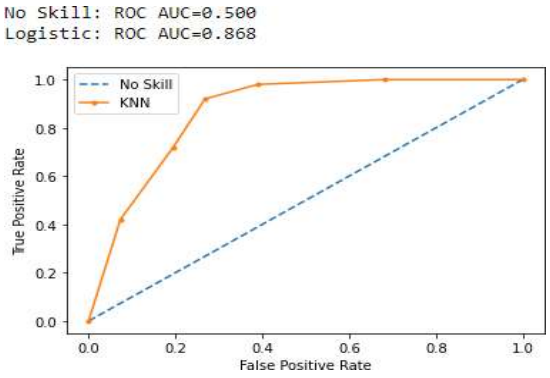


FIGURE 6: ROC CURVE FOR NAÏVE BAYES

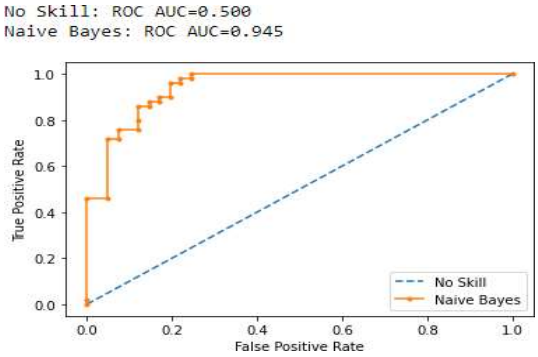


FIGURE 7: ROC CURVE FOR SVM

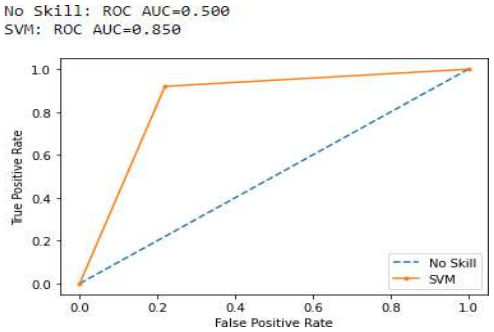


FIGURE 8: ROC CURVE FOR RF

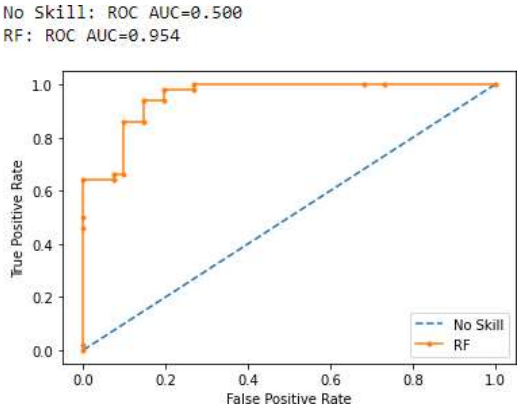
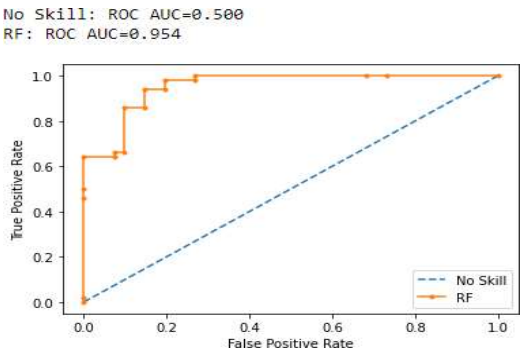


FIGURE 8: ROC CURVE FOR RF



The ROC curve is the Receiver Operating Characteristic curve. The AUC is the area under the ROC curve. If the AUC score is high, the model performance is high and vice versa. The figures 4, 5, 6 and 7 gives the ROC curve of the machine learning algorithms. The comparison of AUC score of the some algorithms is as follows:

Algorithm	AUC score
SVM	0.70
NB	0.68
KNN	0.56
Decision tree	0.53

The accuracy of the algorithms is calculated. The accuracy results are tabulated as follows:

Method	Accuracy
KNN	83.81%
NB	85.71%
Decision tree	83.51%
SVM	83.82%
LR	86.81%
RF	89.01%

The accuracy of K-nearest neighbor algorithm is good when compared to other algorithms.

V. CONCLUSION AND FUTURE WORK

This paper discusses the various machine learning algorithms such as support vector machine, Naïve Bayes, decision tree and k- nearest neighbour which were applied to the data set. It utilizes the data such as blood pressure, cholesterol, diabetes and then tries to predict the possible coronary heart disease patient in next 10 years.

Family history of heart disease can also be a reason for developing a heart disease as mentioned earlier. So, this data of the patient can also be included for further increasing the accuracy of the model.

This work will be useful in identifying the possible patients who may suffer from heart disease in the next 10 years. This may help in taking preventive measures and hence try to avoid the possibility of heart disease for the patient. So when a patient is predicted as positive for heart disease, then the medical data for the patient can be closely analysed by the doctors. An example would be - suppose the patient has diabetes which may be the cause for heart disease in future and then the patient can be given treatment to have diabetes in control which in turn may prevent the heart disease.

The heart disease prediction can be done using other machine learning algorithms. Logistic regression can also perform well in case of binary classification problems such as heart disease prediction. Random forests can perform well than decision trees. Also, the ensemble methods and artificial neural networks can be applied to the data set. The results can be compared and improvised.

REFERENCES

[1] Giarratano, J. and G. Riley, Expert Systems Principles and Programming. 2 ed. Vol. 1. 1994, Boston: PWS Publishing Company.

[2] Kraft, M.R., K.C. Desouza, and I. Androwich. Data mining in healthcare information systems: case study of a veterans' administration spinal cord injury population. in System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on. 2003. IEEE.

[3] Anbarasi, M., E. Anupriya, and N. Iyengar, Enhanced prediction of heart disease with feature subset selection using genetic algorithm. International Journal of Engineering Science and Technology, 2010. 2(10): p. 5370-5376.

[4] Stilou, S., et al., Mining association rules from clinical databases: an intelligent diagnostic process in healthcare. Studies in health technology and informatics, 2001(2): p. 1399-1403.ISO/DIS 16000-6.2 (2002) Indoor Air - Part 6: Determination of Volatile Organic Compounds in

Indoor and Chamber Air by Active Sampling on TENAX TA Sorbent, Thermal Desorption and Gas Chromatography using MSD/FID. Geneva, International Organization for Standardization.

[5] Lemke, F., and J.-A. Mueller, Medical data analysis using self-organizing data mining technologies. Systems Analysis Modelling Simulation, 2003. 43(10): p. 1399-1408.

[6] Lin, F.-r., et al., Mining time dependency patterns in clinical pathways. International Journal of Medical Informatics, 2001. 62(1): p. 11-25.

[7] Yazdani A, Ebrahimi T, Hoffmann U. Classification of EEG signals using Dempster Shafer theory and a K-nearest neighbor classifier. IEEE. In: Proc of the 4th int EMBS conf on neural engineering, 2009: 327-30.

[8] Sumit B, Praveen P, G.N. Pillai. SVM Based Decision Support System for Heart Disease Classification with IntegerCoded Genetic Algorithm to Select Critical Features. WCECS. Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA, October 22 - 24, 2008.

[9] Vapnik, V. N. The nature of statistical learning theory. New York:Springer, 1995. [10] L. I. Kuncheva, Combining pattern classifiers: methods and algorithms: John Wiley & Sons, 2004.

[11] Larose, D., Discovering Knowledge in Data: An Introduction to Data Mining. 2005, New Jersey: John Wiley & Sons, Inc