HEART DISEASE PREDICTION USING ML

Vaibhav gupta, Dr Pallavi Murghai Goel

Vaibhavgupta290498@gmail.com, Pallavi30nov@gmail.com

ABSTRACT-

Heart disease is a major cause of death throughout the world. It is difficult to predict by medical practitioners as it requires expertise and higher knowledge of prediction. The environment in healthcare sector is information rich but lacks knowledge. A lot of data is available in healthcare systems over the internet but there is a lack of effective analysis tool to discover hidden patterns in data. An automated system will enhance medical efficiency and reduce cost and time. This software intends to predict the occurrence of a disease based on the data which is gathered from kaggle. The objective is to extract the hidden patterns by applying data mining techniques on the dataset and to predict the presence value on a scale. The prediction of heart disease requires a huge size of data which is too massive and complex to process and analyse by conventional technique. Our aim is to find out an suitable technique that is efficient and accurate for prediction of cardiac disease.

Keywords- prediction, heart disease, machine learning, algorithms, analysis.

INRODUCTION-

Data processing process involves mining/extracting of very significant, hidden and valuable information from large databases [1]. Usually the Healthcare sector involves abundant of knowledge regarding patients, various diagnosis of the diseases etc… [2]. Nowadays the hospitals are adopting the culture of hospital IMS (information management systems) so on handle their or patients data systematically and effectively. [3]. great quantity of knowledge is produced by such systems that's represented using charts, numbers, text and pictures. Though such quite data is hardly employed for creating any clinical decisions[4]. This research emphasizes on heart condition diagnosis. Various techniques of knowledge mining are incorporated for diagnosing the disease thereby obtaining several probabilities [5]. Concerning the middle disease prediction numerous systems are being recommended which are being deployed by the means of varied techniques and algorithms. Gaining quality service at affordable price remains the prime and challenging concern for the healthcare establishments. For offering quality services at par, there must be accurate diagnosis of the patients in conjunction with effective dosage of medicines. Inferiority clinical diagnosis and treatment can yield in undesired and inadequate results. One solution for cut by Healthcare establishments are often utilization of computer generated data or use of DSS (decision support systems). Usually the Healthcare sector involves abundant of knowledge regarding patients, various diagnosis of the diseases, resource management etc. This information or data must be further weakened by the Human services. Using computerized system, patients treatments records are often stored and using mining methods one can acquire significant information and queries concerning the hospital. Supervised and unsupervised learning are the two processing methods. Supervised learning involves usage of coaching for learning model parameters where else no training set is required in unsupervised learning. Classification and prediction are the essential approach of knowledge mining. The Classification models helps in classifying distinct, disorganized data values on the opposite hand prediction model anticipated values that are continuous.Following are the stages within the proposed approach: user registration and login supported Application, dataset collection, classification via Navies Bayesian, prediction and secure data transfer by the means of AES (Advanced Encryption Standard) and lastly output in PDF format. AES helps in transmitting user data to the database during a secured manner. From the safety point of view, patient’s personalized data is replaced with some mock values. The study considers and employs medicine datasets performances for predicting heart condition in contrast to other Machine Learning techniques. The proposed technique assures to be extremely significant and effective in handling classification, resembling ML (Machine Learning) with reference to Naive Bayesian model. Following represent journal classification: Section 2 illustrate work of previous author. Section 3 put forth the proposed system of heart condition classification and prediction and overview of varied levels. Section 4, presents the experimental outcome. Lastly, Section 5 presents the conclusion and proposes research work for future.

LITERATURE REVIEW -

Monika Gandhi et.al, [1] used Naïve Bayes, Decision tree and neural network algorithms and analysed the medical dataset. There are an enormous number of features involved. So, there's a requirement to scale back the amount of features. this will be done by feature selection. On doing this, they assert that point 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 condition prediction. They made use of knowledge mining techniques to detect the guts disease risk rate.

Sana Bharti, Shailendra Narayan Singh [3] made use of Particle Swarm Optimization, Artificial neural network, Genetic algorithm for prediction. Associative classification may be 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 diagnosis would enhance medical aid and it also can reduce costs. during this study, we've designed a system which will efficiently discover the principles to predict the danger level of patients supported the given parameter about their health. the principles are often prioritized supported the user's requirement. The performance of the system is evaluated in terms of classification accuracy and therefore the results shows that the system has great potential in predicting the guts disease risk level more accurately".

Sellappan Palaniyappan, Rafiah Awang [5] made use of decision tree Naïve Bayes, Decision tree, Artificial Neural Networks to create Intelligent heart condition Prediction Systems (IHDPS).To enhance visualization and simple interpretation, it displays the results both in tabular and graphical forms. By providing effective treatments, it also helps to scale back treatment costs. Discovery of hidden patterns and relationships often has gone unexploited. Advanced data processing techniques helped remedy this example.

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 scale back the noise by cleaning and pre-processing the dataset and also tried to scale back the dimensionality of the dataset. They found that good accuracy are often achieved with neural networks.

Animesh Hazra et.al, [7] discussed intimately the disorder and different symptoms of attack. the various sorts of classification and clustering algorithms and tools were used.

V.Krishnaiah, G.Narsimha, N.Subhash Chandra [8] presented an analysis using data processing. 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 guts disease data contains unnecessary, duplicate information. This has got to be pre processed. Also, they assert that feature selection has got to be done on the dataset for achieving better results.

J.Vijayashree and N.Ch.SrimanNarayanaIyengar [10] used data processing. an enormous amount of knowledge is produced on a day to day. As such, it can't be interpreted manually. data processing are often effectively wont to predict diseases from these datasets. during this paper, different data processing techniques are analysed on heart condition database. last, this paper analyses and compares how different classification algorithms work on a heart condition database.

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

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

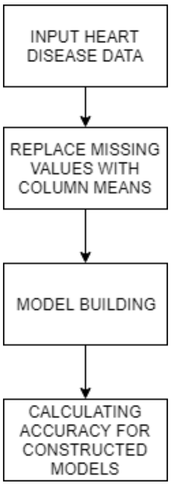
M.Nikhil Kumar et.al [13] used various algorithms – Decision tree, random forest, Naïve Bayes, KNN, Support vector machine, logistic model tree algorithm. Naïve Bayes algorithm gave good results in comparison to other algorithms. They made use of UCI repository of heart condition dataset. Also, J48 algorithm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

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

PROPOSED METHODOLOGY-

In this paper, comparison of varied machine learning methods is completed for predicting the ten year risk of coronary heart condition of the patients from their medical data. the subsequent is that the flowchart for proposed methodology.

FIGURE 1: PROPOSED WORK


INPUT HEART DISEASE DATA
REPLACE MISSING VALUES WITH COLUMN MEANS
MODEL BUILDING
CALCULATING ACCURACY FOR CONSTRUCTED MODELS

The cardiac disease data set is taken as input. it's then pre-processed by replacing non-available values with column means.

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

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

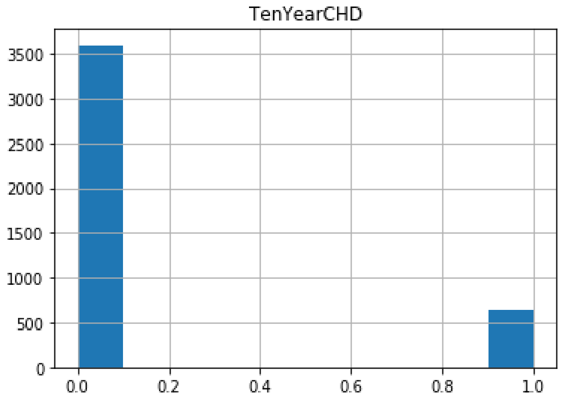
The ROC curve for k-nearest neighbour is depicted in figure 5.

Support Vector Machine (SVM)

The SVM algorithm is employed to predict this disease by plotting the train dataset where a hyper plane classifies the points into two - presence and absence of heart condition.

Here, penalized SVM is employed to handle class imbalance. Class imbalance may be a problem in machine learning when total number of positive and negative class isn't an equivalent. If the category imbalance isn't handled then the classifier won't perform well. the subsequent plot shows the category imbalance.

FIGURE 2: PLOT SHOWING CLASS IMBALANCE



SVM algorithms uses a group of mathematical functions called kernel. during this proposed methodology, linear kernel is employed.

$K(x,x') = \exp((-||x-x'||2)/2\sigma2)$

The performance of the SVM classifier are often increased by fine-tuning the hyper parameters. this will be done by using Grid Search CV. Different values of C are often given as input to the present method. It builds different SVM models with given values then finds the simplest value of c that the model performs well.

Naive Bayes algorithm (NB)

This is a classification algorithm which is used when the dimensionality of the input is extremely high. A Naive Bayes classifier assumes that the presence of a selected feature during a category is unrelated to the presence of the opposite feature. it's supported 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 that the prior event and Y is that the dependence event. The ROC curve is given in figure 6.

It needs less training data. It are often 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 like cigarette, BP, Hypertension, age etc. The challenge of the selection tree lies within the choice of the idea node. This factor utilized in root node must clearly classify the data. We make use aged because the basis node. The ROC curve is given in figure 4.

The decision tree is simple to interpret. they're non-parametric which they implicitly do feature selection.

RESULT ANALYSIS AND COMPARISON-

The dataset used is Framingham taken from Kaggle [17].
There were 16 attributes were as follows: Male – gender 0 for female and 1 for male, Age – age of the patient, Education – values 1-5, education of the patient. Current smoker – 1 if current smoker and 0 otherwise, Cigarette per day – if current smoker then number of cigarette per day, BP Meds – vital sign, Prevalent BP – prevalent vital sign, Prevalent Hyp – prevalent hyper tension, Diabetes – 1 if diabetes 0 otherwise, Total cholesterol – cholesterol level, Sys BP – systolic vital sign, Dia BP – diastolic vital sign, BMI – body mass index, pulse – pulse or pulse of the patient, Glucose – glucose level, Ten Year CHD – has chronic heart condition or not.

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

The ROC curve of the algorithms is as follows:

FIGURE 3: ROC CURVE FOR DECISION TREE

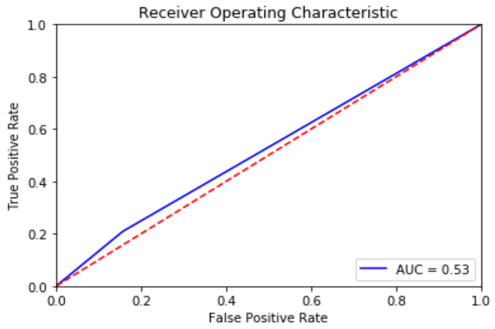


FIGURE 4: ROC CURVE FOR KNN

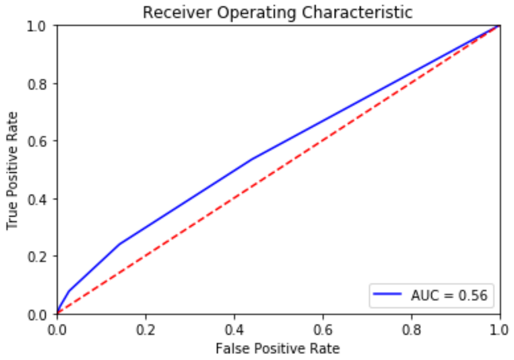


FIGURE 5: ROC CURVE FOR NAIVE BAYES

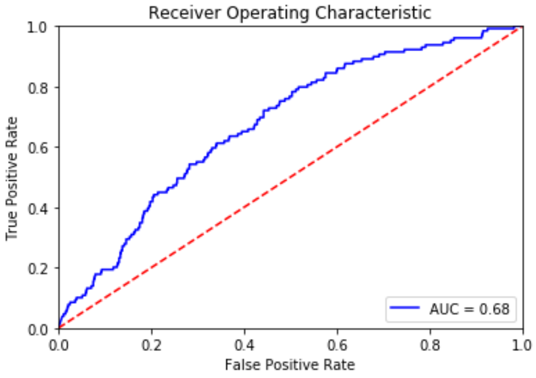
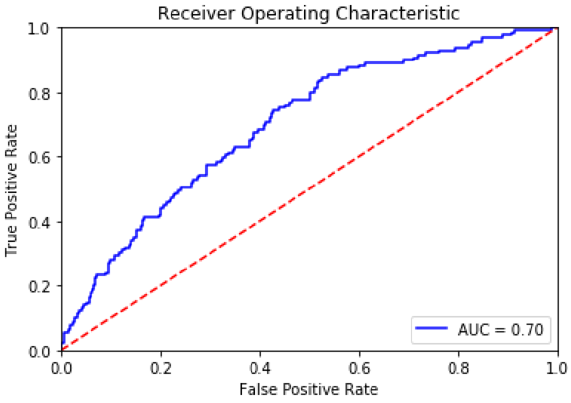


FIGURE 6: ROC CURVE FOR SVM



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 various 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	86.30%
NB	79.66%
Decision tree	77.58%
SVM	63.56%

CONCLUSION-

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]Gandhi, Monika & Singh, Shailendra. (2015). Predictions in heart disease using techniques of data mining. 2015 1st International Conference on Futuristic Trends in Computational Analysis and Knowledge Management, ABLAZE 2015. 520-525. 10.1109/ABLAZE.2015.7154917.

[2] Thomas, J. & Princy, R. (2016). Human heart disease prediction system using data mining techniques. 1-5. 10.1109/ICCPCT.2016.7530265.

[3] S. Bharti and S. N. Singh, "Analytical study of heart disease prediction comparing with different algorithms," International Conference on Computing, Communication & Automation, Noida, 2015, pp. 78-82.

[4] Purushottam, & Saxena, Kanak & Sharma, Richa. (2015). Efficient heart disease prediction system using decision tree. International Conference on Computing, Communication and Automation, ICCCA 2015. 72-77. 10.1109/CCAA.2015.7148346.

[5] Mat Ghani, Mohd & Awang, Raflah. (2008). Intelligent heart disease prediction system using data mining techniques. 8. 108 - 115. 10.1109/AICCSA.2008.4493524.

[6] Sharma, Himanshu. “Prediction of Heart Disease using Machine Learning Algorithms: A Survey.” (2017).

[7] Hazra, Animesh & Mandal, Subrata & Gupta, Amit & Mukherjee, Arkomita & Mukherjee, Asmita. (2017). Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review. Advances in Computational Sciences and Technology. 10. 2137-2159.

[8] V Krishnaiah, G Narsimha and Subhash N Chandra. Article: Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review. *International Journal of Computer Applications* 136(2):43-51, February 2016. Published by Foundation of Computer Science (FCS), NY, USA

[9] Kaur, Ramandeep and Er. Prabhsharn Kaur. “A Review-Heart Disease Forecasting Pattern using Various Data Mining Techniques.” (2016).

[10] Vijayashree, J. & Iyenger, N Ch Sriman Narayana. (2016). Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques: A Review. International Journal of Bio-Science and Bio-Technology. 8. 139-148. 10.14257/ijbsbt.2016.8.4.16.

[11] Benjamin EJ et.al heart condition and Stroke Statistics 2018 At-a-Glance (2018)

[12] Kishore, A. G. Ravi et al. “Heart Attack Prediction Using Deep Learning.” (2018).

[13]Mutyala, Nikhil Kumar & Koushik, K.V.s & Krishna, K.. (2018). Prediction of Heart Diseases Using Data Mining and Machine Learning Algorithms and Tools. 10.13140/RG.2.2.28488.83203.

[14] Razia, Shaik & M, Shaik. (2019). Heart Disease Prediction using Machine Learning Techniques. International Journal of Recent Technology and Engineering. 8. 10.35940/ijrte.D4537.118419.