Comparative Study of Heart Disease Prediction Using Machine Learning Algorithms

PushkalaV[1] ,Agalya T[2], S A Angayarkanni[3]
[1,2]Student, Department of Information Technology,R.M.K. Engineering College, Kavaraipettai, India
[3]Assistant Professor, Department of Information Technology,R.M.K. Engineering College, Kavaraipettai, India

Abstract—The heart disease puts itself in the category of modern plague. The need to predict the heart diseases brings machine learning technology into picture. This paper aims to predict the heart disease using various machine learning algorithms such as Random forest, Support vector classifier, Naive Bayes, K-nearest neighbors and the accuracy of decision tree with and without using the Application Programming Interface (API) and compare their accuracy. One way to implement the algorithms as machine learning model is through a APIs and the another way is to implement it without any APIs. The dataset used to train and test the prediction model is obtained from the UCI repository. The algorithm with highest accuracy is used in the web application as a final product.

Keywords—Decision Tree, Application Programming Interface (API), Heart Disease, Classification, Support Vector machine Classifier, K-Nearest Neighbors Classifier, Random Forest, Gaussian Naive Bayes, Machine Learning

I. INTRODUCTION

Heart diseases are most common in the modern world with increasing stress, pressure and irregular lifestyle. Heart is one of the vital device of human body responsible for transporting the blood and oxygen throughout the body. If any irregularity in heart functioning results in different functional disorders and it will paralyze the normal flow of life. According to the Centers for disease control and prevention, about 6,10,000 people die in united states due to heart diseases and that is 1 in every 4 deaths that occur. Heart diseases are responsible for killing over 3,70,000 people annually. Due to uncontrolled high blood pressure and low-density lipoprotein cholesterol or current smoking increases the risk of heart disease in about 47% adults [11].

Machine learning has become one of the most powerful technology in giving predictive results. As formally defined, machine learning is gives the ability for the machines to learn from their past experiences. The past experiences here denotes the dataset that is used to train the model. The machine learning are broadly classified as supervised learning and unsupervised learning. Supervised learning refers to the algorithm in which the there's a correct prediction that is defined and on the other hand, the unsupervised learning refers to the technique in which the algorithm detects some pattern or clusters them into group. Heart disease classification, comes under the supervised learning. In this paper we discuss about the algorithms that can be used for the prediction of the heart disease. The prediction model gets the data from the patients health check reports and provide a suggestion to the user about the health of their heart. It will help to predict the occurence of the disease accurately, eliminating the chances for human errors.

II. LITERATURE REVIEW

K.S.Shalet et al, used a dataset with 240 instances which is splitted into train data and test data with 120 data each. Decision tree and SVM algorithm is used. The accuracy of 77.91% is obtained from the decision tree algorithm [1].

The MudasirManzoorKirmani et al, discretization techniques of Decision Trees algorithm is used. They used a dataset with 76 attributes they have choose 13 attributes.The accuracy of 79.1% is obtained from the decision tree algorithm [2].

Atul Kumar Pandey et al, used a dataset with 303 instances of which 139 instances belonged to the heart disease where training instances 200 and testing instances 103 using split test mode and they used Pruned J48 Decision Tree with Reduced Error Pruning Approach. The accuracy of 75.73% is obtained from the decision tree algorithm [3].

SankariKarthiga et al, used a dataset consists of total 573 records in which training consists of 303 records and testing consists of 270 records. Decision Tree Algorithm and Naive Bayes Algorithm. They have achieved 89% accuracy by applying Naive Bayes algorithm [4].

Mai Shouman et al, Decision tree algorithm is used to obtain high accuracy. 79.1% accuracy is obtained from the decision tree algorithm. They have used dataset containing 303 instances of which 297 are complete and they have choose 13 attributes [5].

SeyedaminPouriyeh et al, used a dataset with 303 instances. The machine learning algorithms used are Decision Tree, Naive Bayes, Multilayer Perceptron, K-Nearest Neighbor, Single Conjunctive Rule Learner, Radial Basis Function and support vector Machine. 77.55% accuracy is obtained by decision tree algorithm [6].

AnimeshHazra et al, used 75% of data as training data and 25% as testing data, they have used some of the tools likeWEKA, RapidMiner, TANAGRA, MATLAB etc for prediction. Naive Bayes, Decision list and K-NN algorithms are used. The accuracy obtained is 52% is obtained from the decision tree algorithm [9].

Jaymin Patel et al, chosen 13 attributes for prediction from the dataset of 303 instances and 76 attributes. Decision tree algorithm is used. 56.76% accuracy is obtained from this algorithm[10].

Lakshmishree et al, the Decision tree algorithm is used in the dataset consisting of 100 patient details, of which 51 male and 49 female patients the chances of getting heart disease totally is 26 patients and 74 are predicted as no heart disease [7].

M.A.Jabbar et al, used a 75% of data set to train the classifier and to build the classifier. Remaining 25% data set is used for testing. In 10-fold cross validation all the instances of the data set are used and are divided into 10 disjoint groups. They have used Random forest, Decision tree, Feature selection, Chi square, Genetic algorithms. The accuracy of 63.3% is obtained from the Decision tree algorithm [12].

Stephen R. Alty et al, used a Support Vector Machines, Digital Volume Pulse, Pulse Wave Velocity for predicting the Heart Disease. The SVM method yields a high degree of classification accuracy, with a significantly high proportion, 93%, of true positives achieved (i.e. the sensitivity).There was a slightly lower result of only 78% true negatives (i.e. the specificity) [13].

SonamNikhar et al, used a dataset total of 303 records with 76 medical attributes only 19 attributes were used for the prediction model.Naïve Bayes Classifier, Decision tree Classifier algorithm are used. They have analyzed that the decision tree has better accuracy as compared to naïve Bayes classifier. The future work of this paper is to improve the performance of the Naïve Bayesian classifier by removing unnecessary and irrelevant attributes from the dataset [14].

JyotiSoni et al, used a dataset of 909 records with 15 medical attributes.The records were splitted equally into two datasets 455 records for training and 454 records for testing the dataset. k-nearest neighbor, Neural Networks, Bayesian classification, Classification based on clustering, Decision Tree Algorithm are used in the prediction model. The accuracy of 89% is obtained for the decision tree algorithm [15].

Himanshu Sharma et al, used a Machine learning techniques with Naïve Bayes, Decision Tree, Neural Network, SVM and Deep Learning algorithms are used for the Prediction of Heart Disease.Decision tree algorithm as used few approaches for building tree such as ID3, CART, CYT, C5.0 and J48. This paper showed some high accuracy with promising result in other field of medical diagnose [16].

RavindraYadav et al,used a Machine learning techniques with Decision Tree algorithm, Naive Bayes, Neural Network, Deep Learning and SVM are used for the Survey on Heart Disease Prediction. CART, ID3, CYT, C5.0 and J48 are used to build the decision tree result [17].

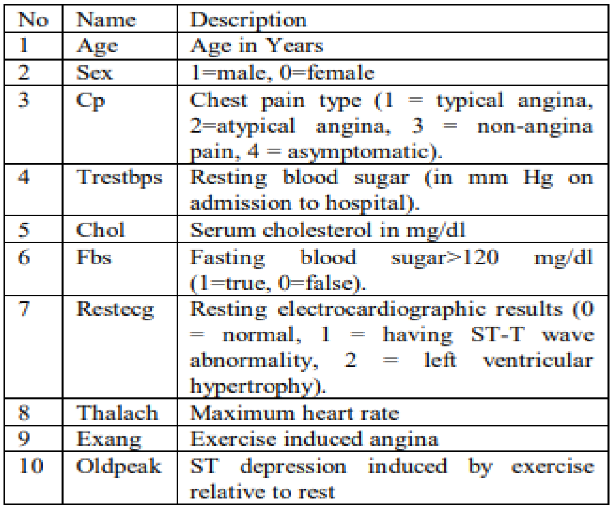
SellappanPalaniappan et al,used a model Intelligent Heart Disease Prediction System built with the Neural Network, Naïve Bayes, and Decision Tree algorithm. Results show that each technique has its infrequent strength in realizing the objectives of the defined goals. IHDPS can answer complex “what if” queries which conventional decision support systems cannot be proposed [21].

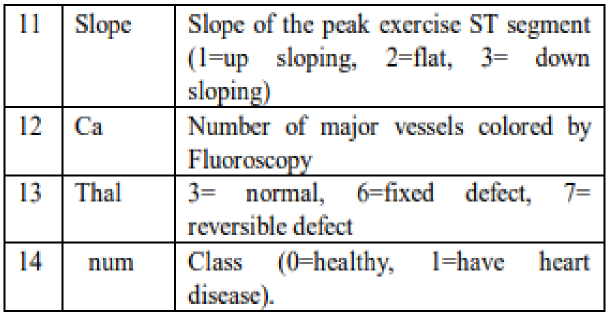
T.John Peter et al,used the preprocessed data is clustered using clustering algorithms as K-means to gather relevant data in a database. Maximal Frequent Item set Algorithm (MAFIA) is applied for maximal frequent model in heart disease database. The regular patterns can be classified into different classes using the C4.5 algorithm as training algorithm using the concept of information entropy. The result demonstrates that the designed prediction system is capable of predicting the heart attack successfully [22].

III. DATASET

The Heart disease dataset is obtained from the UCI repository. The processed Cleveland data is used for training and testing the prediction model. The dataset possess 75 attributes and 303 instances. From the 75 attributes, only 14 attributes are used, shown in Table 1. The dataset is preprocessed and splitted as training and testing data. The prediction model is trained with the training data and tested with the testing data. 80% of the data in the dataset is used as training data and remaining 20% as test data to obtain the highest accuracy. 13 attributes is used as input and the ‘num’ class is predicted by the machine learning model.

Table 1: Used attributes from the dataset





No	Name	Description
1	Age	Age in Years
2	Sex	1=male, 0=female
3	Cp	Chest pain type (1 = typical angina, 2=atypical angina, 3 = non-angina pain, 4 = asymptomatic).
4	Trestbps	Resting blood sugar (in mm Hg on admission to hospital).
5	Chol	Serum cholesterol in mg/dl
6	Fbs	Fasting blood sugar>120 mg/dl (1=true, 0=false).
7	Restecg	Resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality, 2 = left ventricular hypertrophy).
8	Thalach	Maximum heart rate
9	Exang	Exercise induced angina
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	Slope of the peak exercise ST segment (1=up sloping, 2=flat, 3= down sloping)
12	Ca	Number of major vessels colored by Fluoroscopy
13	Thal	3= normal, 6=fixed defect, 7= reversible defect
14	num	Class (0=healthy, 1=have heart disease).

IV. ALGORITHMS USED

4.1 Decision Tree

Decision tree is a tree-like structure or a flowchart-like structure used as a decision support tool in both classification and regression problems which helps to build automated predictive models. It is a non-parametric supervised learning algorithm. It creates a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The decision tree model is developed by placing the best attribute of the dataset at the root of the tree. The training set is splitted up into several subsets. Each subset contains data with the same value for an attribute. Until all the leaf nodes are found i.e., expected result is found the algorithm continues. The number of hyper-parameters to be tuned is almost null. The decision tree is associated with decision rules where the outcome is the contents of the leaf node, and the conditions along the path form a if else rules to make the algorithm decide based on the situation. When the algorithm continues to be deeper and deeper to reduce the training set error, it may result in increased test set error. This overfitting problem may occur frequently while building a decision tree model. The implementation of decision tree helps to visualize the logic. It generates every possible outcomes of a decision.The main aim of decision tree algorithm is to prioritize the attribute which can give highest accuracy. It creates a model to predict the 'num' variable from the inferred dataset.

4.1.1 Application programming interface used

API for decision tree implementation is used from Scikit-learn. Decision tree classifier prediction model, evaluation, test data split is done by different APIs available in Scikit-learn. It gives the accuracy of 78%

4.1.2 Implementation Without API

Decision tree is implemented manually without using any APIs. The test data split, construction of decision tree, evaluation of algorithm and prediction of heart disease are done without any API. The accuracy obtained is 79.59% which is high compared to the decision tree implementation using the API.

4.2 Support Vector Machine - Classifier

The Support vector machine is used in both classification and regression analysis. It is a supervised learning model to find the points closest to the line from both the classes called support vectors. The computation of the distance between the line and the support vectors is called the margin. The main goal is to maximize the margin. The optimal hyperplane is the hyperplane for which the margin is maximum. Thus SVM algorithm try to make a decision boundary with a wide possible separation between the two classes. This data is clearly linearly separable. Thus the data can be classified linearly. SVM classifier learning model with associated learning algorithms is used to analyse the data and recognize patterns. The accuracy obtained by applying the SVM classifier is 61%.

4.3 Random Forest

Random forest builds a forest in a random fashion. The forest that is build is the ensemble of decision tree classifiers. Bootstrap Aggregation or Bagging is the is the base of random forest algorithm which combine the predictions of multiple decision trees and put it together to construct a forest to make more accurate predictions. It fits a sub-samples of dataset using a decision tree classifier. It is a supervised learning method and this algorithm that has high variance are like Classification and Regression Trees (CART). They are sensitive to the training data. Bagging of CART algorithm creates many random sub-samples from the dataset, each of those samples are used to train the CART model and finally using the test data, the average prediction from each model is calculated. These trees have low-bias and high variance. This process might take long time to prepare but the overfitting problem will not occur. The problem with CART is they are greedy. Random forest redesign the CART algorithm in such a way that the sub-trees are learned. The resultant prediction from these sub-trees are less correlated. The samples that are left behind from the bootstrap samples taken from the training data is called Out-Of-Bag samples.The estimated accuracy is provided by the performance on each model on its left out samples. The features that are needed for prediction process can be selectively picked from the dataset which is called feature importance. The less important feature that has less contribution in prediction process can be left out while constructing the random forest. The accuracy of 86.67% is obtained.

4.4 K-Nearest Neighbors Classifier

The K nearest neighbors algorithm is a instance based learning and widely used in the real-life scenarios. The K Nearest Neighbors algorithm can be used for both the classification and the regression problems also. The K Nearest Neighbor algorithm is also called lazy learning. The steps followed in the K Nearest Neighbor are preprocessing the dataset, training the model and testing the model. The preprocessing of the dataset generally includes the cleaning, removing erroneous and outlier values in the dataset.This is the most important step in the algorithmic process. Also, the analysing the quality of the dataset is very crucial before running the algorithmic tests on them. The missing values and the outliers should be taken care.

The K Nearest Neighbor uses the curve and plots the new test data. The K factor refers to the how many neighbours it considers for the classification. Usually the K value should be a single digit number. After the test data point is plotted, the K nearest neighbors are determined by using the distance formula. The K neighbors can only be in odd numbers. The accuracy of 68% is obtained.

4.5 Gaussian Naive Bayes

Naive Bayes classifier is derived from Bayes' Theorem. It can be extended to real-valued attributes by assuming Gaussian distribution. Gaussian Naive Bayes is this extension of Naive Bayes. It is a supervised learning classification algorithm used for both binary and multi-class classification problems. This algorithm is also called idiot Bayes because the hypothesis calculation probabilities are tractable. Training the data is fast, no optimal coefficient fitting is needed. By using this classifier in order to make a prediction, independency between attributes of the data set is the main assumption. The probabilities of each class, mean and standard deviation for each input variables in each class is calculated.

The Mean and Standard deviation is calculated by the equations,
$\text{mean}(x) = 1/n * \text{sum}(x)$
$\text{standard deviation}(x) = \text{sqrt}(1/n * \text{sum}(xi-\text{mean}(x)\hat{}2))$
where x is the input variable and xi is a specific value of the x variable for the i'th instance.
When making predictions, the parameters are plugged into Gaussian Probability Density Function and is calculated using
the equation, $\text{pdf}(x, \text{mean}, sd) = (1 / (\text{sqrt}(2 * PI) * sd)) * \exp(-((x-\text{mean}\hat{}2)/(2*sd\hat{}2)))$
One of the main advantages of Naive Bayes classifier is that it requires small amount of data for parameters estimation. 91% accuracy is obtained which is the highest one when compared to all other algorithms.

V. RESULTS

The proposed work consists of the applying various machine learning algorithms including SVM- classifier, KNN classifier, Random forest, Naive Bayes, decision tree implementation using the API from Scikit-learn and decision tree algorithm without using APIs. The prediction model is implemented as a web application to put to use by all type of end users. The angiographic heart disease status is predicted using this algorithms. If the ‘num’ class is 0, then the person has no heart disease. If it is 1, then the person has heart disease. When it is implemented as web application, the result is conveyed to the end-user as a web page presenting the content, describing whether heart disease is present or not.

The accuracy between these six different algorithm implementations are analysed and the highest accuracy of 91% is obtained from the Naive Bayes algorithm in comparison with all other implemented algorithms. The accuracy obtained from the manual implementation of decision tree algorithm without using any APIs is 79.59% which is high when compared to the DT implementation using the API.

Machine Learning Algorithms	Accuracy
Decision Tree with API	78%
Decision Tree without API	79.59%
Support Vector Classifier	61%
Random Forest	86.67%
K-Nearest Neighbors Classifier	68%
Naive Bayes	91%

VI. CONCLUSION AND FUTURE WORK

Heart disease prediction system is implemented as a web application which can be put into use by various end-users specifically doctors, lab-technicians and patients. Five machine learning algorithms such as Decision Tree, SVM-Classifier, Random Forest, KNN-Classifier and Naive Bayes is used to predict the heart disease. The Naive Bayes algorithm gave the highest accuracy. These algorithms predict whether the person has the heart disease or not.

Our future work is to improve the accuracy of algorithms using different other methods and techniques. The prediction model has to be tested with a large volume of real-time medical data. So that the accuracy will get improved and any necessary changes that has to be made into the prediction model to improve its working and efficiency. The other additional work is automating the way that the web application is using to read the data from the user. This will make all set of users use the system without any knowledge about its working and all complexity will be eliminated making it a perfect end product.

VII. REFERENCES

[1] K.S.Shalet, V. Sabarinathan, V.Sugumaran and V.J.Sarath Kumar, “Diagnosis of Heart Disease Using decision Tree and SVM Classifier”, International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 10 No.68, 2015.

[2] MudasirManzoorKirmani and Syed ImmamulAnsarullah, “Prediction of Heart Disease using Decision Tree”, IJCSN International Journal of Computer Science and Network, Volume 5, Issue 6, December 2016

[3] Atul Kumar Pandey, PrabhatPandey, K.L. Jaiswal and Ashish Kumar Sen, “Predicting heart disease status based on the clinical data of patients using the decision tree algorithm”, IOSR Journal of Computer Engineering Volume 12, Issue 6, 2013

[4] SankariKarthiga, M.Safish Mar andM.Yogasini, “Early Prediction of HeartDisease Using Decision Tree Algorithm”, International Journal of AdvancedResearch in Basic Engineering Sciences and Technology Vol.3,Issue.3, March 2017.

[5] Mai Shouman, Tim Turner and Rob Stocker, “Using Decision Tree for DiagnosingHeart Disease Patients”, Proceedings of the 9-th Australasian Data MiningConference, 2011.

[6] SeyedaminPouriyeh, Sara Vahid, Hamid Reza Arabni and Giovanna Sannino, “Acomprehensive investigation on comparison of Machine Learning Techniques onheart disease domain”, IEEE Symposium on Computers and Communication, 2017.

[7] Lakshmishree J and K Paramesha, “Prediction of Heart Disease Based on DecisionTrees”, International Journal for Research in Applied Science EngineeringTechnology (IJRASET) Volume 5 Issue V, May 2017.

[8] V.V. Ramalingam , AyantanDandapath and M Karthik Raja, Heart disease prediction using machine learning techniques International Journal of Engineering Technology, 2018.

[9] AnimeshHazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee and Asmita Mukherjee, “Heart Disease Diagnosis and Prediction Using Machine Learning”, Advances in Computational Sciences and Technology ISSN 0973-6107 Volume 10, 2017.

[10] Jaymin Patel, Prof. TejalUpadhyay, and Dr. Samir Patel, “Predict the presence of heart disease in patients Using Machine learning”, IJCSC conference paper, 2016.

[11] Cheryl D. Fryar, Te-Ching Chen, and Xianfen Li, “Prevalence of Uncontrolled Risk Factors for Cardiovascular Disease: United States, 1999–2010”, NCHS Data Brief, No. 103, August 2012

[12] M.A.Jabbar, B.L.Deekshatulu and Priti Chandra ,Intelligent heart disease prediction system using random forest and evolutionary approach, Journal of Network and Innovative Computing ISSN 2160-2174 Volume 4 (2016) pp. 175-184, 2016

[13] Stephen R. Alty, Sandrine C. Millasseau ,Philip J. Chowienczyk and Andreas Jakobsson, Cardiovascular Disease Prediction Using Support Vector Machines, Circuits and Systems, 2003 IEEE 46th Midwest Symposium on, Volume: 1, 2004

[14] SonamNikhar, A.M. Karandikar, “Prediction of Heart Disease Using Machine Learning Algorithms”, International Journal of Advanced Engineering, Management and Science (IJAEMS) [Vol-2, Issue-6, June- 2016] Infogain Publication (Infogainpublication.com) ISSN : 2454-1311, 2016

[15] JyotiSoni, Ujma Ansari, Dipesh Sharma, SunitaSoni, Predictive techniques for Medical Diagnosis: An Overview of Heart Disease Prediction, International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011.

[16] Himanshu Sharma and M A Rizvi, “Prediction of Heart Disease using Machine Learning Algorithms: A Survey”, International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 5 Issue: 8

[17] RavindraYadav, Upendrasingh,Survey on Heart Disease Prediction by using Machine Learning Technique, International Journal of Creative Research Thoughts (IJCRT) www.ijcrt.org, Volume 7, Issue 1 March 2019

[18] M. Sultana, A. Haider, and M. S. Uddin, “Analysis of data mining techniques for heart disease prediction,” 2016 3rd Int. Conf. Electr. Eng. Inf. Commun. Technol. iCEEiCT 2016, 2017.

[19] V. Manikantan and S. Latha, “Predicting the analysis of heart disease symptoms using medicinal data mining methods”, International Journal of Advanced Computer Theory and Engineering, vol. 2, pp.46-51, 2013.

[20] M. Akhil, B. L. Deekshatulu, and P. Chandra, “Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm,” Procedia Technol., vol. 10, pp. 85–94, 2013.

[21] SellappanPalaniappan and RafiahAwang, “Intelligent heart disease prediction system using data mining techniques”, International Journal of Computer Science and Network Security, vol.8, no.8, pp. 343-350, 2008.

[22] T.John Peter , K. Somasundaram, “An Empirical Study on Prediction of Heart Disease using classification data mining technique” IEEE International Conference On Advances In Engineering, Science And Management (ICAESM - 2012) March 2012.

[23] AbhishekTaneja, “Heart Disease Prediction System Using Data Mining Techniques”, Taneja, Orient. J. Comp. Sci. &amp; Technol., Vol. 6(4), 457-466 (2013).

[24] Hlaudi Daniel Masethe, Mosima Anna Masethe, “Prediction of Heart Disease using Classification Algorithms”, Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II WCECS 2014, 22-24 October, 2014.