A Machine Learning based framework for Heart Disease Prediction

[1] Mogali .V.P. P. Akshita, [2] Dr.S Rao Chintalapudi

[1]Student, [2]Associate Professor,

Department of Computer Science and Engineering, Pragati Engineering College (Autonomous),

Surampalem, East Godavari District, Andhra Pradesh, India.

Abstract : In today's era deaths due to heart disease has become a major issue approximately one person dies per minute due to heart disease. Accurate and on time diagnosis of heart disease is important for coronary failure prevention and treatment. In order to lower the number of deaths from heart diseases, there has to be a fast and efficient detection technique. This paper presents a model for detecting heart disease using machine learning algorithms. The methodology adopted is such a way that a Heart Dataset was trained using four different machine learning algorithms K-Nearest Neighbor, Support Vector Machine, Decision Tree and Random Forest. Hence, it is a binary classification problem, these models will predict the heart disease based on the attributes provided. The performance of these four models is measured in terms of accuracy. The K-Nearest Neighbor model is producing good results than the three models.

IndexTerms **- Machine Learning, Cardiovascular disease, K-Nearest Neighbors, Support Vector machine, Decision Tree, Random Forest.**

I. INTRODUCTION

In India, heart attack is one of the most common disease. Heart plays an eminent role in pumping the blood through the circulatory system of the body. Oxygen is circulated by the circulatory system of the body to all the body parts and also entire human blood system gets collapsed if the heart does not function properly. As a result it will cause serious health condition, sometimes even lead to death. The term "heart disease" is analogues to the term "cardiovascular disease". Cardiovascular disease generally refers to a condition that constitutes narrowed or blocked blood vessels that can lead to a heart failure, chest pain (angina) or stroke. Heart disease is one of the extensive sources of morbidity and mortality among the human population. Due to wide range of contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate, and other factors, it has become a strenuous task to predict the heart disease. Due to such constraints, scientists have become dependent of modern approaches like Data Mining and Machine Learning in order to predict the disease. So there is a necessity for an approach on machine-learning-based diagnosis system to predict heart disease through heart disease dataset. In our proposed we used four machine learning algorithms like –K Nearest Neighbor, Support Vector Machine, Decision Tree and Random Forest. The concept of our model lays out following objectives:

- Provides new approach to concealed patterns within the data.
- Helps avoid human biasness.
- Implementation of ML Classifiers that classifies the disease as per the input of the user.
- Reduces the cost of medical tests.

II.RELATED WORK

Heart Disease has been identified as one of the largest causes of death even in developed countries [2]. Implementation of artificial intelligence in the cardiac disease system detection improvises the performance of widely used models like models provided by American College of Cardiology/American Heart Association (ACC/AHA) models in CVD detection and prediction [1].Based on data mining, a research on heart disease prediction was carried out by [3]. The paper highlighted the utilization of data mining in discovering trends in patient data through pattern generation. They produced an affiliation guideline on a real informational index with the patients' history with regard to coronary illness to yield high exactness rate.. Kernel F-score Feature Selection was introduced to perform determination as a prepreparing venture in the characterization of therapeutic database [3]. In the paper titled "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques" [4], data mining classification techniques, namely Decision Trees, Naive Bayes, and Neural Networks were analyzed on Heart disease dataset. The performance of various techniques used in the research was compared, based on accuracy. The accuracy of Neural Networks, Decision Tree, and Naive Bayes were 100%, 99.61%, and 90.73%. As a result obesity and smoking were included in order get more accurate results in prediction of heart disease. In the paper "Intelligent Heart Disease Prediction System Using Data Mining Techniques", various data mining techniques such as Decision Trees, Naïve Bayes and Neural Network were used and a Intelligent Heart Disease Prediction System (IHDPS) model was developed. The complex "what if" queries were solved by IHDPS, contrary to traditional decision support systems which are incapable. Based on attributes such as age, sex, blood pressure and glucose it can predict the probability of patients getting a heart disease. Important knowledge is also permitted, for example patterns, relationships between medical characteristics related to heart disease, to be established [5]-[9]. In order to improve the prediction accuracy in heart diseases, an analytical approach was performed and with the help of the ensemble classification an increase of 7% accuracy for weak classifiers was attained. Enhancement of performance of the process using feature selection, and the results contributed to drastic improvement in prediction accuracy [10]. Amandeep Kaur made contrast among algorithms such as artificial neural network, K-nearest neighbor, Naïve Bayes and Support vector machine on cardiovascular disease prediction [11]. J Thomas, R Theresa Princy [12] for prediction of heart disease used K nearest neighbor algorithm, naïve Bayes and decision tree and also in order to detect the rate of heart disease risk, data mining techniques were used. Monika Gandhi et. Al, [13] used Naïve Bayes, Decision Tree and neural network algorithms for further analysis of dataset and found that large

number of features must be added and there is need to reduce the count of features. On doing feature selection, time was reduced and problems were solved. RamandeepKaur, Er. PrabhsharnKaur [14] have found that the heart disease data has to be pre-processed as it contains redundant information. A paper titled as Prediction of Heart Disease Using Machine Learning Algorithms using decision tree classifier and naïve bayes was contributed by SonamNikharet Al. Mr. Santhana Krishnan. J and Dr. Geetha. S, [15] wrote a paper that predicts heart disease for only male patient using classification techniques.

III. Heart Disease Prediction System

1. Data Acquisition:

The foremost step is to gather the data from the public repository. Heart Data set is downloaded from Kaggle website which consists of 303 patient attributes. The dataset consists of 13 attributes and therefore the 14th attribute being '1' and '0' which predicts whether the patient is suffering from heart disease or not respectively.

2. Data Pre-Processing:

The described data set is preprocessed by changing the missing values into their means. For understanding the data, Correlation Matrix is used to get the degree of association between the features. Conclusion is drawn ushers there is need for scaling as each feature and label is distributed along several ranges. By using target classes, feature scaling is done accordingly through normalizing the data by placing the values in range. Now dataset is ready to train the model.

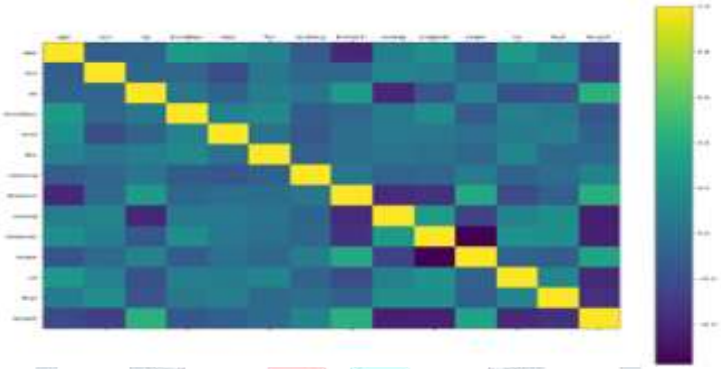

Fig 7.Correlation matrix showing degree of association

3. Feature Scaling:

Feature scaling is a method that normalizes the range of independent variables or features of data. In some machine learning algorithms as the range of values of data varies, so target functions will work improperly without normalization. It normalizes the dataset accordingly but placing the values in range for proper evaluation using the StandardScaler to scale the dataset. Data is scaled and we improve the columns using fit_transform() method. It is most important part as many classifiers govern the distance broad range of values of scaled data.

4. Algorithms

Classification is a predictive modeling problem where a class label is predicted for a given sample of input. In terms of modeling perspective, classification needs a training dataset along with many samples of inputs and outputs through which it learns. In this paper, we train and build the model using four machine learning algorithms namely K Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest Classifier.

Algorithm 1: K-Neighbors Classifier

This classifier depicts the classes of K nearest neighbors for a given data point and it assigns a class to this data point based on the important class. Anyhow, the number of neighbors can be varied. For eg, we can vary them from 1 to 20 neighbors and calculate the test 23 score in each case. After obtaining the number of neighbors and the test score in each case, a line graph is plotted.

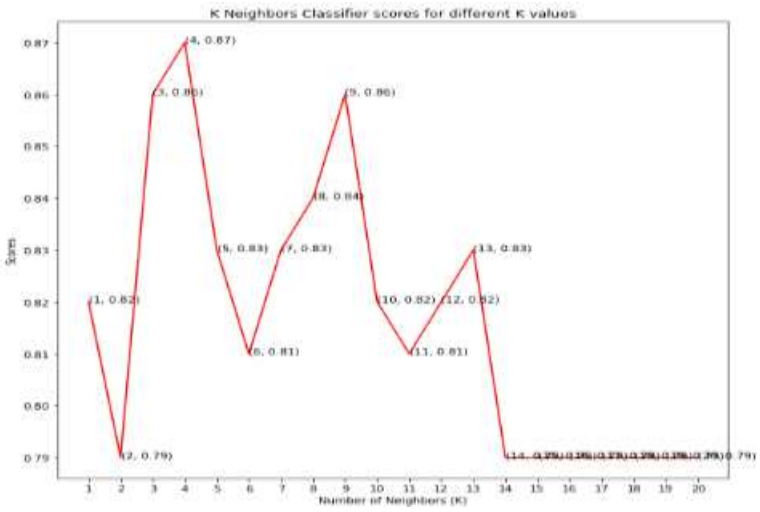

Fig 1. Accuracy versus Number of Neighbors in K-Nearest Neighbors algorithm

Algorithm 2: Support Vector Machine

This classifier separates the classes as much as possible by adjusting the distance between the data points and thus creates a hyperplane where there are several kernels that rely on which the hyperplane is characterized such as linear, sigmoid, poly and r rbf.

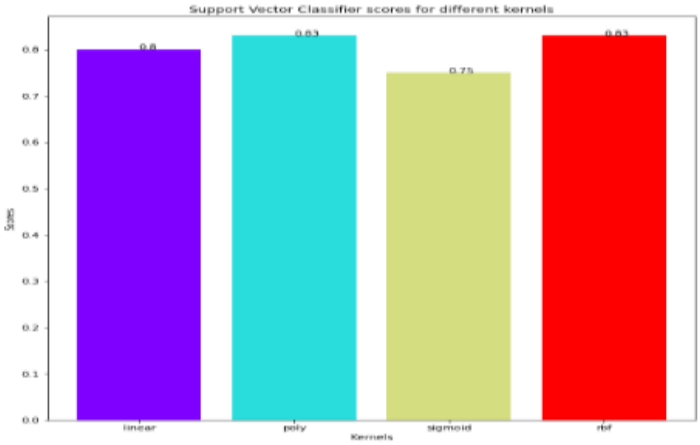

Fig 2. Accuracy Versus Kernels in Support vector machine

Algorithm 3: Decision Tree

This classifier assigns the class values to each data point with the help of decision tree formed. While creating the model, maximum number of features that were taken, are varied accordingly . In this paper, we used features ranging from 1 to 30 (the total features in the dataset after dummy columns were added).

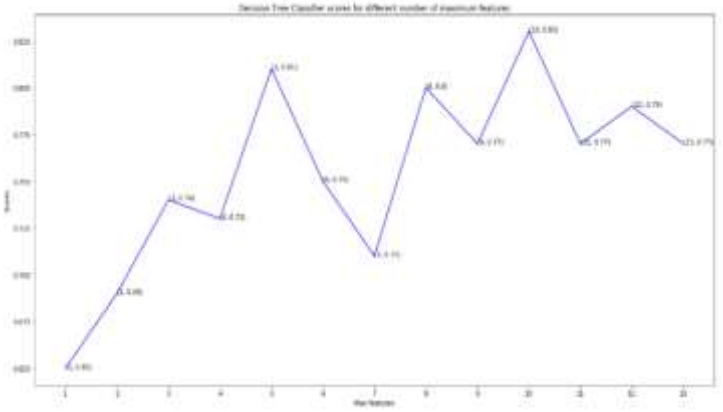

Fig 3. Accuracy versus number of features in Decision tree

Algorithm 4: Random Forest

This classifier uses the concept of decision trees to a successive level in which it creates a forest of trees where each tree is formed by a random selection of features from the total features. Fig. 4 describes accuracy versus number estimators in Random Forest classifier.

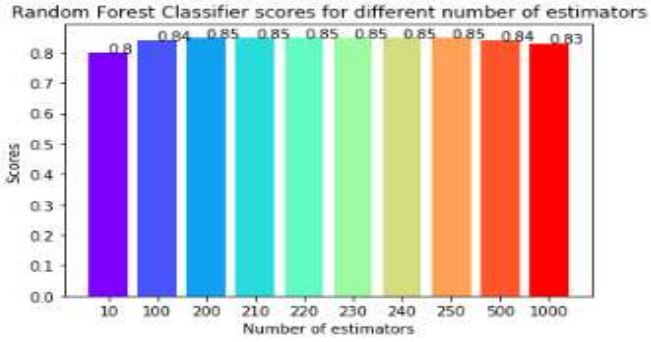

Fig 4.Demonstration of accurate scores of Random forest classifier

IV.EXPERIMENTAL SETUP

This paper implemented in python using Google colab – a cloud based jupyter notebook. The dependencies used for these models are NumPy (scientific computing with python),Pandas(imports csv, excel files),Matplotlib(plotting library), OS, Tkinter(GUI library).

Training Phase:
Dataset was split into x variable and y variable. Where the x variable consists of 13 attributes which showed various test results and the y variable contains of output. The x and y variables were further split into x_train, x_test, y_train and y_test respectively. These x_train and y_train where being fitted or trained using four machine algorithms which are K-Nearest Neighbors, Support Vector Machine, Decision Tree and Random Forest. The four algorithms were utilized in checking the percentage of accurate results using different numbers of n values.

Testing Phase:
The algorithm with highest accurate results will be saved and loaded , where a GUI will be created through which users will enter their different test result and give the patient's input to the model to predict if they are suffering from a Heart disease or not.

V. RESULTS AND DISCUSSION

The model is trained using four machine learning algorithms whose various parameters were differed and were compared to the final models. Dataset is split into 67% of training data and 33% of testing data. Then, 4 models were trained and tested and the obtained accuracy scores for K Nearest Neighbors, Support Vector Machine, Decision Tree, and Random Forest are 0.87, 0.83, 0.79 and 0.84. After the testing of accuracy, we used K-Nearest Neighbors Classifier which has one of the highest accurate results in making prediction. The K-Nearest Neighbors model was being saved and loaded into the GUI and inputs were passed to the model to detect if they have a Heart Disease or not. The result of the model will also be displayed on the web to the user.

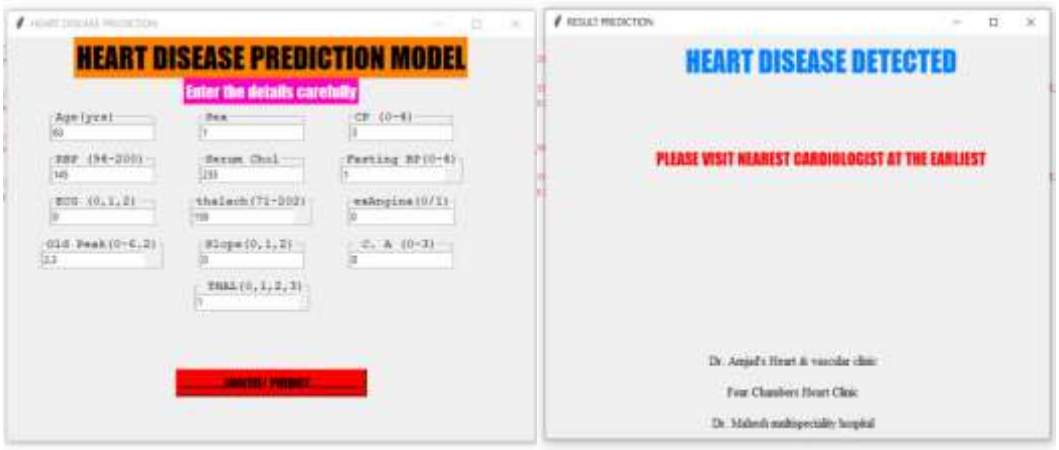

Fig 5.Demonstration of user input form to which patients will have to input some test results to the GUI (by selecting the predict button) and displays the result of the input as presence of heart disease after analysis for the heart disease prediction system

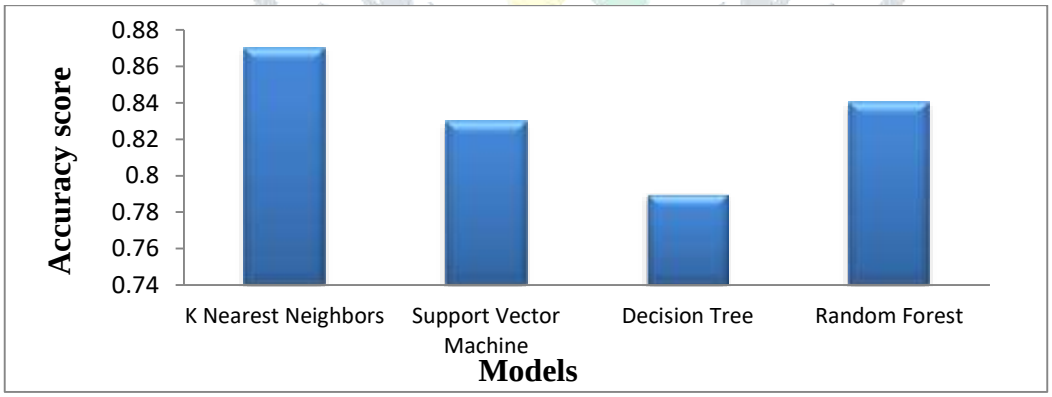

Fig 6.Demonstration of accuracy of 4 models namely; K Neighbors, Support vector machine, Decision tree, Random forest.

VI. CONCLUSION AND FUTURE SCOPE

In this paper, a diagnosis system is proposed to predict the heart disease by using ML algorithms and therefore the prediction outcomes are relied on the heart disease dataset. In this machine learning approach four algorithms were used to train and analyze the dataset which contains the test results of different patients. After testing for accuracy, for K Nearest Neighbors, Decision Tree, Random Forest and Support Vector Machine the highest accuracy score observed is for K Nearest Neighbors. K Nearest Neighbors was also integrated in the web through GUI, and it predicted good results when tested. This research can be widened to a real-time system using Deep Learning methodology, where users can upload images of their test results.

VII. ACKNOWLEDGMENT

First and foremost, warmest thanks to Pragati Engineering College, Department of Computer Science and Engineering and second author Dr. S Rao Chintalapudi, Associate Professor, Computer Science and Engineering Department. A special word of gratitude to Dr. M. Radhika Mani, Head of Department, Computer Science and Engineering Department, Pragati Engineering College, for her continued guidance and support for our project work.

REFERENCES

[1]. K. Vanisree, S. Jyothi, "Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks", International Journal of Computer Applications vol.19, issue.6, pp.6 – 12, 2011.
[2]. S.F. Weng, J. Reps, J. Kai, J.M. Garibaldi, N. Qureshi, "Can Machine-Learning Improve Cardiovascular Risk Prediction Using Routine Clinical Data", vol.1, issue.12, pp. e0174944, 2017.
[3]. M. Thiyagaraj, G. Suseendran, "Survey on heart disease prediction system based on data mining techniques", Indian Journal of Innovations and Developments vol.6 issue.1, pp.1-9, 2017.
[4]. C.S. Dangare, S.S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications vol.47, issue.10, pp. 44-48, 2012.
[5]. S. Palaniappan, R. Awang, "Intelligent heart disease prediction system using data mining techniques", In 2008 IEEE/ACS international conference on computer systems and applications, pp. 108-115, 2008.
[6]. C.B.C. Latha, S.C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques", Informatics in Medicine, Unlocked 16, pp.100203, 2019.
[7] AvinashGolande, Pavan Kumar T, (June 2019): Heart Disease Prediction Using Effective Machine Learning Techniques, International Journal of Recent Technology and Engineering (IRTE), ISN: 2277-3878, Volume-8, Issue-1S4.
[8] A. SahayaArthy, G.Murugeshwari, (April 2018): A survey on heart disease prediction using data mining techniques.
[9] AmitaMalav, KalyaniKadam, (2018): "A Hybrid Approach for Heart Disease Prediction Using Artificial Neural Network and K – Means", International Journal of Pure and Applied Mathematics.
[10] DhafarHamed, JwanK.Alwan, Mohamed Ibrahim, Mohammad B.Naeem, (march – 2017): "The Utilization of Machine Learning Approaches for Medical Data Classification" in Annual Conference on New Trends in Information & Communications Technology Applications.
[11] I KetutAgungEnriko, Muhammad Suryanegara, DadangGunawan al, (June 2018): "Heart Disease Diagnosis System with k – Nearest Neighbors Method Using Real Clinical Medical Records", 4th International Conference.
[12] Monika Gandhi, Shailendra Narayanan Singh, (2015): Predictions in heart diseases using techniques of data mining.
[13] M. S. Amin, Y. K. Chiam, K. D. Varathan, (Mar.2019): Identication of significant features and data mining techniques in predicting heart disease, Telematics Inform., vol. 36, pp. 8293.
[14] Senthilkumar Mohan, ChandrasegarThirumalai, GautamSrivastava, (2019): Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques, Digital Object Identifier 10.1109/ACCESS.2019.2923707, IEEE Access, VOLUME 7.
[15] V. V. Ramalingam, AyantanDandapath, M Karthik Raja, (2018): heart disease prediction using machine learning techniques: a survey, International Journal of Engineering & Technology (IJET), 7 (2.8) 684-687.