Early and Accurate Prediction of Heart Disease Using Machine Learning Model

B Padmaja[a], Chintala Srinidhi[b], Kotha Sindhu[c], Kalali Vanaja[d], N M Deepika[e], E Krishna Rao Patro[f]

b.padmaja@gmail.com, srinidhireddy000@gmail.com, sindhureddy962@gmail.com, vanajakalali793@gmail.com, nm.deepika@iare.ac.in, patrosir@gmail.com

[b,c, d]Student of Department of Computer Science and Engineering, Institute of Aeronautical Engineering
[a, e, f]Assistant Professor, Institute of Aeronautical Engineering, Hyderabad, India

Abstract: Heart disease is one of the critical health issues and many people across the world are suffering with this disease. It is important to identify this disease in early stages to save many lives. The purpose of this article is to design a model to predict the heart diseases using machine learning techniques. This model is developed using classification algorithms, as they play important role in prediction. The model is developed using different classification algorithms which include Logistic Regression, Random Forest, Support vector machine, Gaussian Naïve Bayes, Gradient boosting, K-nearest neighbours, Multinomial Naïve bayes and Decision trees. Cleveland data repository is used to train and test the classifiers. In addition to this, feature selection algorithm named chi square is used to select key features from the input data set, which will decrease the execution time and increases the performance of the classifiers. Out of all the classifiers evaluated using performance metrics, Random forest is giving good accuracy. So, the model built using Random forest is efficient and feasible solution in identifying heart diseases and it can be implemented in healthcare which plays key role in the stream of cardiology.

Keywords: Heart disease, Classification, Prediction, feature selection, Random forest.

1. Introduction

Heart disease is considered as one of the fatal diseases across the globe. From the past few years, millions of cases are increasing and many people died due to heart related issues. According to a report conducted by the World Health Organization (WHO), heart disease is responsible for 17 million deaths worldwide. As heart is important organ in the human body, any issue related to it highly affects human health. The main symptoms of heart disease are chest pain, bloating, swollen legs, breathing issues, fatigue and irregular heart beat rhythm. The factors that cause heart disease are age, overweight, stress, unhealthy diet and smoking.

The main goal of this paper is to create a model to predict heart disease using machine learning algorithms, which will aid doctors in detecting the disease early on with less medical tests and providing appropriate care, potentially saving many lives.There is a traditional approach to identify heart disease in hospitals then why machine learning? In hospitals, large amount of data related to patients suffering from heart diseases and other diseases is generated each day, it is difficult for doctors to use or handle the patient's data efficiently to make decisions without datamining techniques. Data mining is highly recommended for the prediction of heart diseases as it extracts more accurate and useful data from large amount of data which makes prediction easy. It is primary foundation of machine learning that helps to handles large amount of data, the processing speed of machine learning is high and it makes predictions in early stages.

There are different datamining techniques can be used such as classification, prediction and recognizing patterns for diagnosing heart disease. In this paper, classification models which are part of machine learning are used for identifying cardio vascular diseases. Classification algorithms uses input data to predict and classify them to which class or category that the data belongs. Some of the classification techniques are Logistic regression, Decision trees, Random Forest, Gradient Boosting, Support vector machine, Naïve Bayes and K-Nearest Neighbour. In this paper, all classification models are trained to predict the heart disease and compare the performance of them using evaluation metrics such as sensitivity, accuracy and so on, which gives the best classification model for prediction of occurrence of heart disease.

1.1.Machine learning model

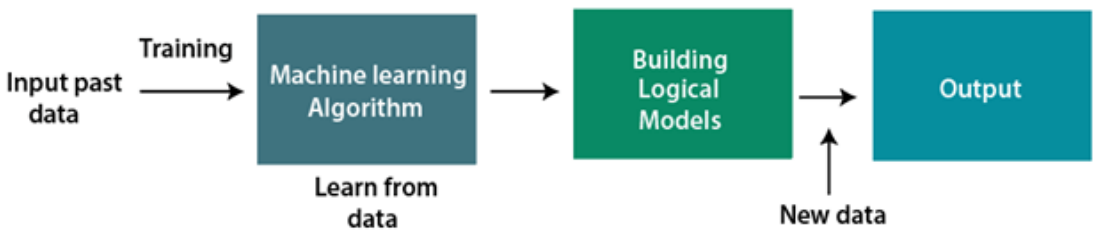

Figure 1. Phases in Machine Learning Model

Machine learning is a part of artificial intelligence which can learn by itself and improve from past experiences, makes decisions and predictions. According to Fig 1In this model, training and testing is conducted by using Cleveland data set. Initially, we train the classification algorithm using the input dataset features, then the model learns from the input data and able to find or recognize the patterns from dataset, then by testing with new data it makes the prediction that represents to which class the data belongs to. The model built using this technique learn from the data and classify the patients into normal and heart disease category and predicts the heart disease.

2. Significance of the Study

The modern lifestyle or fast forward life has significant impact on lives of people. Many people across the globe are suffering from heart related diseases due to stress, lifestyle habits and some from genes irrespective of their age. The objective of this article is to predict the occurrence of heart disease in early stages, so that it will help doctors to take pro-active measures to control many lives of the people. Although there is traditional approach in hospitals for dealing with heart issues but most of them are re-active, they can only know after its occurrence and hospitals cannot handle huge amount of data that generates each day related to patient's data, so it becomes difficult for doctors to make accurate predictions. In this paper, machine learning techniques are used for prediction as they can handle huge data and their performance is high. Classification models that are part of machine learning techniques are used, as they play key role in prediction. In this project, Random Forest algorithm along with some other classification algorithms are used to predict the heart disease. The purpose of this project is to reduce the deaths caused due to heart diseases and to predict its occurrence efficiently.

3. Review of Related Studies

Jayshril S. Sonawane et al. [1] proposed prediction of heart disease using multilayer perception neural network(2014). The accuracy by using this technique is 80%. Since the time complexity will be more due to usage of complex Boolean functions while using limited data sets for training, independently trained subnetworks scale quite well. Ketut Agung Enrico et.al [2] a system was proposed by him for heart disease prediction using KNN algorithm with simplified parameters(2016). The accuracy of this algorithm is 81.85%. Using KNN , with increase of number of parameters the performance decreases and it considers 90% of data for training which is computationally expensive and does nothing during. M.Akhil jabbar et al [3] studies about heart disease prediction using Lazy Associative classification. The vast amount of space used to store the entire data collection. Since no abstraction is made during the training phases, especially noisy training data increases the case base unnecessarily.Jaymin Patel et.al [4] proposed prediction of heart disease using data mining techniques(2015). The accuracy is 56.76%. The drawback of J48 is that the tree increases linearly with large data. LMT is slower and takes long time for implementation and accuracy is low. Rifki wijaya et al [5] studies about preliminary design of estimating heart disease by using machine learning ANN within one year(2013). The accuracy is 81.85%. To function, the neural network needs to be trained. Big neural networks necessitate a long processing time. Microprocessor design and history necessitate their emulation. Carlos Ordonez et al [6] proposed association rules to implement this prediction system. The accuracy is 70%. It uses too many parameters from patients record and produce irrelevant rules. So, this technique is computationally expensive and performance is low. Jyoti soni et al [7] did evaluation of Weighted Associative Classifier (WAC).The accuracy is 81.51%. All attributes are not equally important in predicting the class mark in the prediction model. As a result, various weights may be allocated to different attributes based on their ability to predict. Idticeme sedjelmaci et al [8] proposed detection of some heart diseases using fractal dimension and chaos theory(2013). The accuacy of this technique is 80%. Fractal analysis was created to analyse complex irregular objects. The drawbacks of applying Chaos Theory are primarily due to the input parameters chosen. The methods used to compute these parameters are determined by the underlying dynamics of the data as well as the type of analysis being performed, which is usually complex and not always precise. D.R.Patilet al [9] proposed prediction of heart disease using learning vector quantization algorithm. The accuracy of this algorithm is 85.55%. There is no limit to the number of prototypes that can be used per class; the only requirement is that each class have at least one prototype. AH Chen et al. [10] presented a heart disease prediction system that can help doctors predict heart disease status using patient clinical data. The C language is used an artificial neural networks for classification and prediction of heart disease.The C and C# programming languages are used to develop system.The proposed method accuracy is 80%. m. Anbarasi et al [11] proposed prediction of heart disease with feature subset selection using genetic algorithm(2010). The accuracy is 70%. The language used to specify candidate solution must be robust. Manpreet Singh et.al [12] proposed cardiovascular disease prediction system based on structural equation modelling (SEM) and Fuzzy cognitive map (FCM) (2016). The accuracy of SEM and FCM 74%. It doesn't work well with large data and accuracy is low. Kathleen j. Miao et al [13] studies about coronary heart disease diagnosis using Deep Neutral Network(2015). The accuracy is 83.67%. It is difficult to be adopted by people who are less experienced. It is difficult to comprehend performance based solely on understanding, and this necessitates the use of classifiers.. Jae Kwon Kim et al [14] proposed neural network based coronary heart disease risk prediction using feature correlation analysis(2017). The accuracy of this is 81.163%. A correlational analysis can only be used when the variables are two measurable on scale. It's difficult to tell which variables cause which effects, and a high correlation between variables can be misleading. Sairabi H. Mujawar et.al [15] proposed a model for prediction of heart disease using modified K-means and Naïve Bayes(2015). Naïve Bayes assumes that all predictors are independent and it also have zero frequency problem.

3.1.Tabular analysis of existing work

Ref no	Author	Technique	Dataset	Accuracy	Limitations
[1]	Jayshril s. sonawane	Multilayer Perception Neural Networks	Cleveland heart disease database	97.5%	Independently trained subnetworks scale well because the learning time of multilayer perceptron networks with backpropagation scales exponentially for complex boolean functions by using minimal training sets.
[2]	Ketut Agung Enrico	K-Nearest Neighbours	Hungarian dataset	81.85%	Using KNN , with increase of number of parameters the performance decreases and it considers 90% of data for training which is computationally expensive and does nothing during training phase.
[3]	M.Akhil jabbar	Lazy association classification	UCI repository	80%	The vast amount of space used to store the entire data collection. Since no abstraction is made during the training phases, especially noisy training data increases the case base unnecessarily.
[4]	Jaymin Patel	Data mining technique-Decision Tree model (J48, Logistic model tree-LMT, Random forest)	Cleveland dataset	56.76%	The drawback of J48 is that the tree increases linearly with large data. LMT is slower and takes long time for implementation and accuracy is low
[5]	Rifki wijaya	Artificial Neural Network	Diff tools or database	100%	To function, the neural network needs to be trained. Big neural networks necessitate a long

					processing time. Microprocessor design and history necessitate their emulation.
[6]	Carlos Ordonez	Association rules	Medical dataset from hospital	70%	It uses too many parameters from patients record and produce irrelevant rules. So, this technique is computationally expensive and performance is low.
[7]	Jyoti soni	Weighted Associative classifier	UCI machine learning dataset	81.51%	All attributes are not equally important in predicting the class mark in the prediction model. As a result, different weights can be assigned to different attributes depending on their predictive performance.
[8]	Idticeme sedjelmaci	Fractal dimension and chaos theory	Hospital database	80%	The drawbacks of applying Chaos Theory are primarily due to the input parameters chosen. The underlying dynamics of the data, as well as the type of analysis being done, which is usually complex and not always accurate, determine the methods used to measure these parameters.
[9]	D.R.Patil	Learning Vector Quantization Algorithm	Cleveland heart disease database	85.55%	There is no limitation on how many prototype can be used per class, the only requirement being that there is at least 1 for each class.
[10]	AH CHEN	Artificial neural network algorithm	ML UCI repository	80%	It exhibits black box nature, which doesn't give information about how much time required for

					prediction, amount of data required. It is computationally expensive.
[11]	m. Anbarasi	Feature subset selection using genetic algorithm	Hospital database	70%	The language used to specify candidate solution must be robust. Population size, mutation rate, and crossover rate are all parameters of a Genetic Algorithm that must be carefully selected. A bad fitness function option can cause serious issues, such as being unable to solve a problem or, even worse, returning an incorrect answer to a problem the solution to the issue
[12]	Manpreet Singh	Structural equation modelling and Fuzzy cognitive mapping	Canadian community health survey (CCHS)dataset	74%	It doesn't work well with large data and accuracy is low.
[13]	Kathleen j. Miao	Deep Neural Network	Cleveland heart disease database	83.67%	It is difficult to be adopted by people who are less experienced. It is difficult to comprehend performance based solely on understanding, and this necessitates the use of classifiers.
[14]	Jae Kwon Kim	Feature correlation Analysis	KNHANES-VI dataset	81.163%	A correlational analysis can only be used when the variables are two measurable on scale. Cannot conclude cause and effect, strong association between variables can be misleading.
[15]	Sairabi H. Mujawar	Modified K-means and Naïve Bayes	Cleveland dataset	93%for presence of HD. 89%for	Naïve Bayes assumes that all predictors are independent and it

				absence of HD.	also have zero frequency problem.

4.Methodology

Complete machine learning methodology is mentioned below that is how input dataset is loaded into program and how it will processed is mentioned clearly in text format and also in block diagram format.

4.1Block diagram

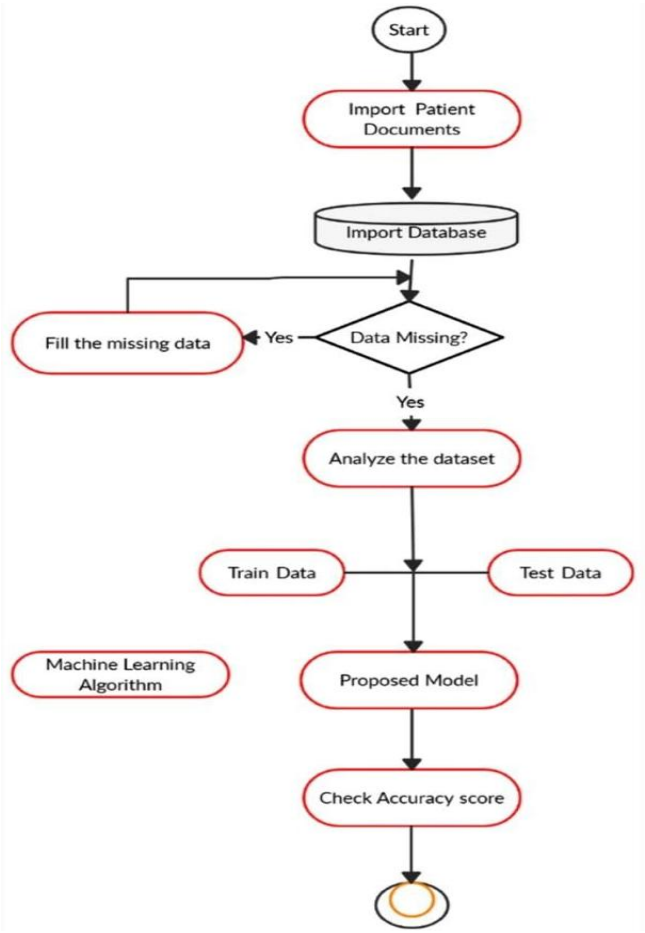

Figure.2 Block Diagram for Proposed model

The system's output is measured using the Cleveland heart disease database, which was obtained from UCI's data-set repository. There are 303 records in this database, each with 13 clinical attributes such as age, sex, form of chest pain, resting blood pressure, cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise induced angina, old peak, slope, number of vessels coloured, and thal. Out of the 303 records in this report, 164 belong to the stable category and 139 to the heart disease category.

Huge amounts of incomplete and noisy data are present in real-life data. Noise and missing values are common when cleaning collected data. The data must be cleaned of noise and missing values must be filled in order to obtain an accurate and efficient result. Transformation is the process of changing the format of data from one type to another in order to make it more understandable.

In this scheme, the database is randomly divided into two sets: training and research. When a data set is split into a training and testing set, the bulk of the data is used for training and only a small part ofthe data is for testing. To ensure that the training and testing sets are equivalent, Analysis Services are selectes the data randomly. By using similar data for

training and research, you can reduce the effect of data inconsistencies and gain a better understanding of the model's characteristics. The test set is a set of observations that are used to evaluate the model's results using a performance metric.In last step it check accuracy of the algorithm which is used to predict the heart disease is exist or not.

5. Dataset for Heart Disease Prediction

Heart disease predicting model is trained by using Cleveland dataset which take age, sex, chest pain, resting blood pressure, rest ECG, maximum heart rate, exercise include angina, ST depression, ST slope, number of major blood vessels, types of thalassemia as inputs and produce output stating whether patient is suffering from heart disease or not. Random forest classifier is used to classify the data and for prediction. Feature selection algorithms are also used which helps to improve accuracy of the model. Basically random forest classifier is used for both classification and regression but we use this for classification purpose and can overcome missing values. Accuracy obtained for random forest classifier is 93.44%.Univariate data analysis is done in that process categorical variables are dropped.

5.1 Random Forest Classifier

The random forest is a classification algorithm that uses several decision trees to classify data. When constructing each individual tree, it employs bagging and feature randomness in order to establish an uncorrelated forest of trees whose committee prediction is more reliable than that of any single tree. It also aims to reduce the difficulties associated with high blood pressure. By averaging, you can establish a natural equilibrium between high variation and high bias between two extremities. This approach can be implemented in R and Python using robust libraries.

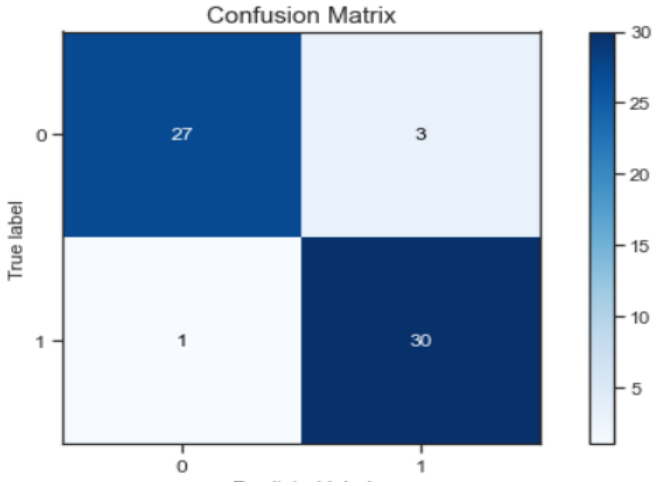

Figure.3 Confusion Matrix of Random Forest Classifier

Random Forest Classifier Accuracy --> 93.44262295081968

5.2 Logistic Regression Classifier

Logistic regression is supervised classification technique. It's used to forecast a categorical dependent variable using a variety of independent variables. The output of a categorical dependent variable is predicted using logistic regression. As a result, the outcome must be a discrete or categorical attribute. It's easier to put into practice, interpret, and train with. Logistic regression should not be employed if the number of observations is less than the number of features in input data set; otherwise, overfitting may occur.

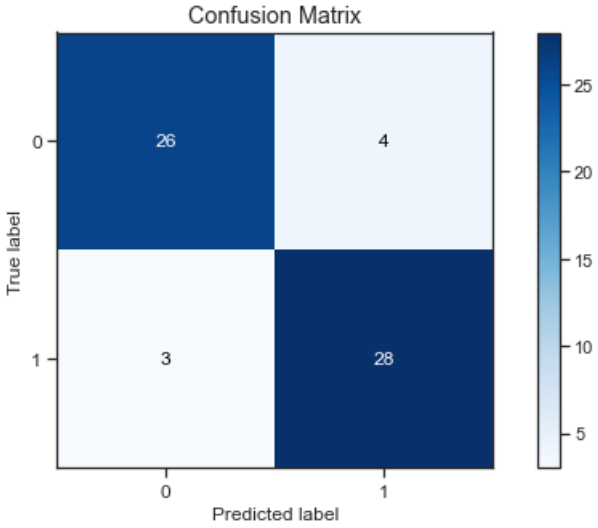

Figure.4 Confusion Matrix of Logistic Regression

Logistic Regression Accuracy --> 88.52459016393442

5.3 KNN Classifier

KNN is a simple classification model, which act as non-parametric algorithm which doesn't make any pre assumptions about data distribution during analysis. KNN considers 90% of data for training which is computationally expensive and does nothing during training phase. K-nearest neighbour is a classification method that uses an imaginary border to classify data. When fresh data points are received, the algorithm will attempt to anticipate them as closely as possible to the boundary line. It is imported from scikit-learn package

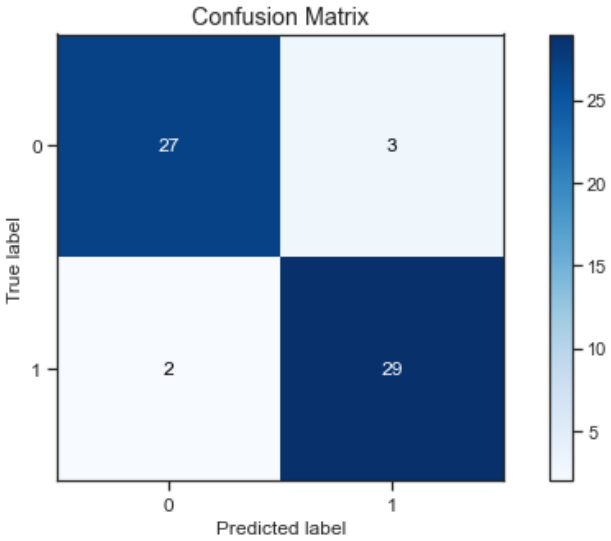

Figure.5 Confusion Matrix of K-Nearest Neighbour

K-nearest neighbour Classifier Accuracy --> 91.80327868852459

5.4Support Vector Machine Classifier

A support vector machine (SVM) is a supervised machine learning model that solves two-group classification problems with classification algorithms. SVM models will categorize new text after being fed sets of named training data for each group. In this, maximizing the hyperplane margin will help to overcome the problems of misclassification. Some of well-known support vector machine implementations are Scikit-learn, MATLAB, and LIBSVM.

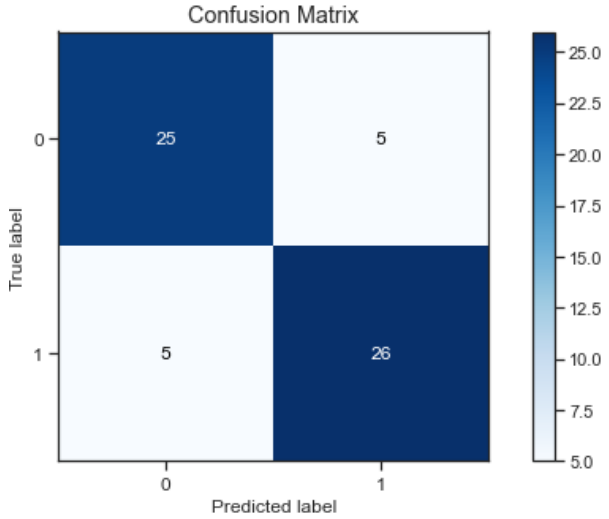

Figure.6 Confusion Matrix of Support Vector Machine

SVM Accuracy --> 83.60655737704919

5.5 Decision Tree Classifier

Decision trees are similar to tree structures which are mainly used for decision making in machine learning. For classification and regression, Decision Trees (DTs) are a non-parametric supervised learning process. The aim is to learn basic decision rules from data features to build a model that predicts the value of a target variable. In decision tree, each internal node represents an attribute query, each branch a test result, and each a label class.

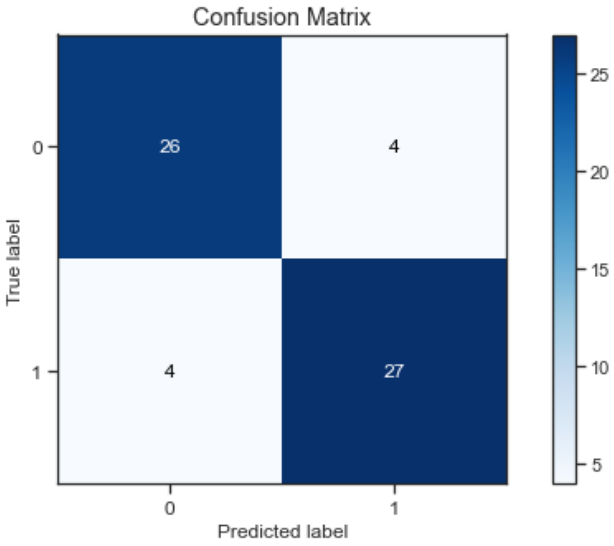

Figure.7 Confusion Matrix of Decision Tree

Decision Tree Classifier Accuracy --> 86.88524590163934

5.6Gradient Boosting

Gradient Boosting is a well-known boosting method. Each prediction corrects the error of the one before (weak predictors called decision trees) it in gradient boosting. Unlike Adaboost, the training instance weights are not adjusted; instead, each predictor is trained using the predecessor's residual errors as labels. It also concentrates on optimization of loss function and

as it is a greedy approach, it overfits the training data easily. Regularization methods that penalize various parts of the algorithm and minimize overfitting would aid in its efficiency.

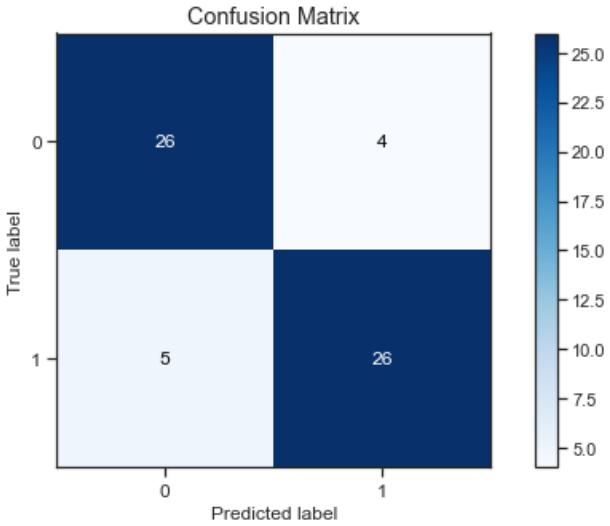

Figure.8 Confusion Matrix of Gradient Boosting

Gradient BoostingClassifier Accuracy --> 85.24590163934425

5.7Naive Bayes

Naive Bayes methods are a class of supervised learning algorithms based on Bayes' theorem.Given the class variable, a naive Bayes classifier assumes that the existence (or absence) of one aspect of a class is unrelated to the existence (or absence) of any other feature. It's "naive" in the sense that it makes assumptions that may or may not be right. It assumes that each feature being classified is independent on other features.

high blood pressure. By averaging, you can establish a natural equilibrium between high variation and high bias between two extremities. This approach can be implemented in R and Python using robust libraries.

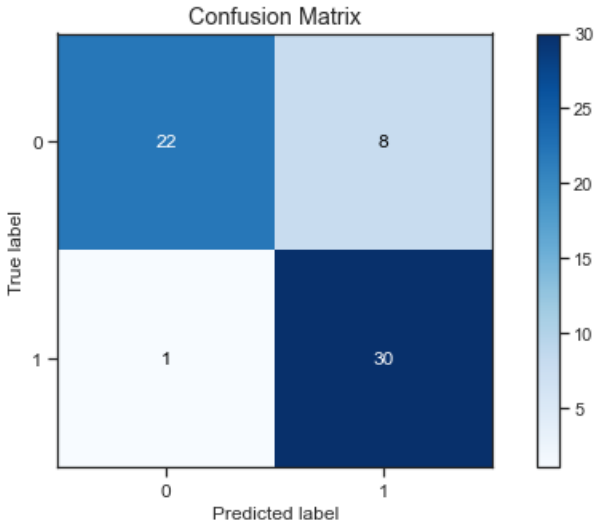

Figure.9 Confusion Matrix of Naïve Bayes

Naïve Bayes Classifier Accuracy --> 85.24590163934425

6.Implementation and Results

6.1 Dataset Evaluation

Cleveland dataset which is used for training data and classified by using datamining classifier algorithms. From 303 values of the dataset we reduced input values to 14 main data inputs taken from user and univariate, bivariate analysis is done on the data. We split the dataset randomly into training and testing datasets then we preprocess them through this we get analytical results and based on that heart disease heart disease prediction will be done by the machine learning model.

6.2 Output generation

Inputs taken through interface made by sing flask app will be classified by random forest classifier and it will make the predictions. All inputs that are taken from the user are based on the Cleveland dataset and output is predicted based on this inputs.

Web application is made by using python flask application. As shown in the below figures fig.10 is the execution snap shot of the application and fig.11 is the user interface that is home page of the application through which users give input. Those input data which is given by users are considered as dataset values. Splitting, preprocessing, training, testing are done on that data input given by users. After submitting user inputs internal processing is done by machine learning model and in fig.12 that is result prediction page final prediction will be displayed.

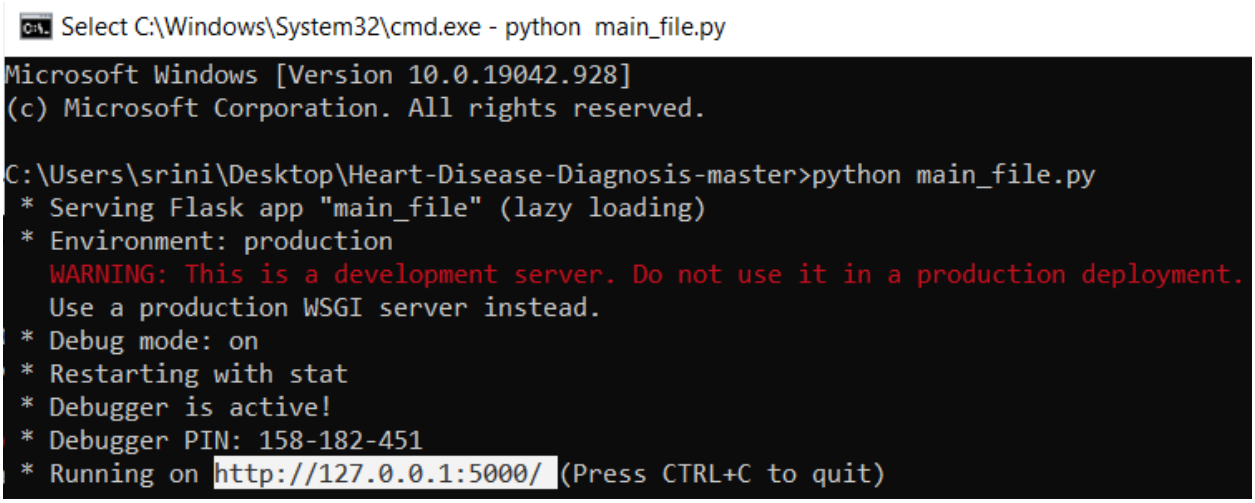

Figure.10 Execution

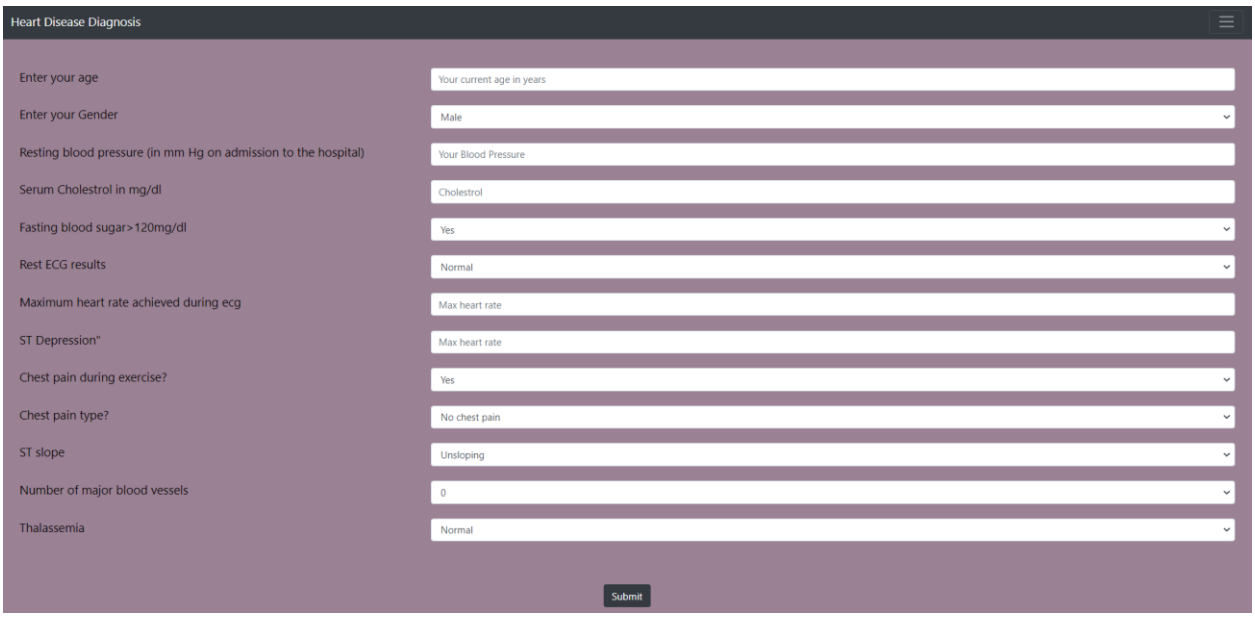

Figure.11 Home page

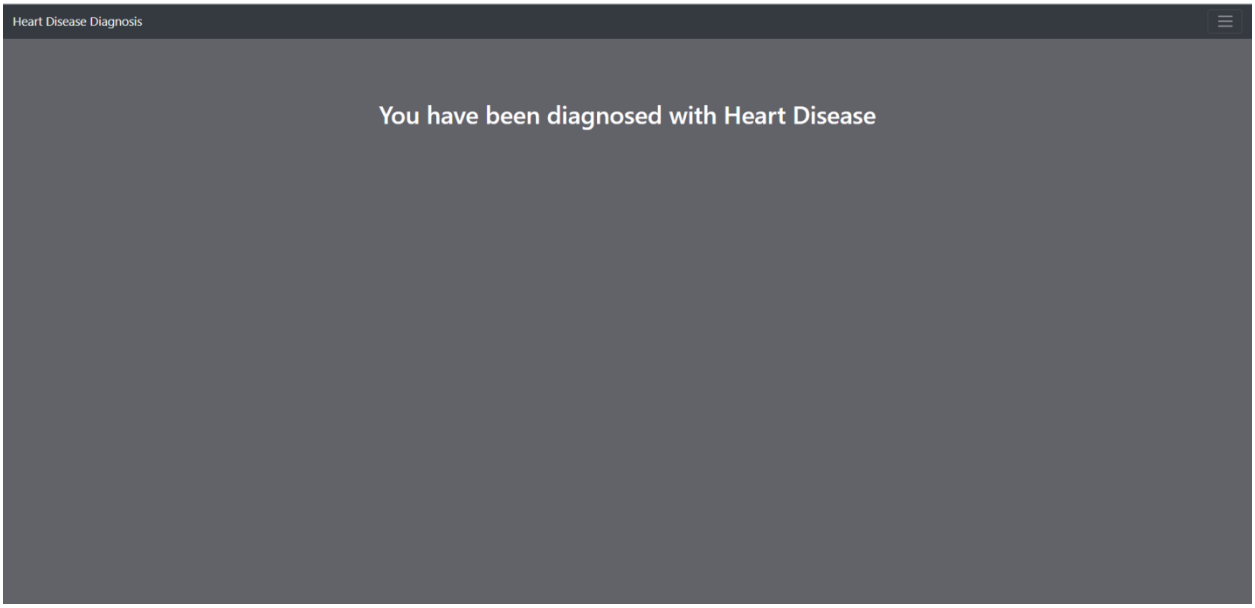

Figure.12 Result prediction

This way by classifying inputs which are given by user from home page random forest classifier will predict the result that is whether that user is suffering from heart disease or not will predicted.

7.Conclusion and future scope

Different datamining techniques are used for classification purpose and minimum number of required attributes that is 14 required attributes are taken and random forest data mining classification algorithm is used which gave very good results and future scope is extend the application like by using same attributes or by adding some more attributes we can predict different other diseases also like kidney related lungs related diseases. This application can be made as common platform for predicting all kind of diseases. We are excited to continue our project and add some more attributes and we try to this application as disease prediction application for all kind of diseases.

References (APA)

[1] Jayshril S. Sonawane, D.R Patil, 2014, " Prediction Of Heart Disease Using Multilayer Perceptron Neural Network ", IEEE International Conference on Information Communication and Embedded Systems (ICICES2014).

[2]I Ketut Agung Enriko, Muhammad Suryanegara,Dinda Agnes Gunawan, " Heart disease prediction system using k-Nearest neighbor algorithm with simplified patient's health parameters", 2016.

[3] M. Akhil Jabbar; B. L Deekshatulu; Priti Chandra, " Heart disease prediction using lazy associative classification", : 2013, IEEE International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s).

[4] Jaymin Patel, Prof. Tejal Upadhyay, and Dr. Samir Patel, Sep 2015-Mar 2016, "Heart Disease Prediction using Machine Learning and Data Mining Technique", Vol. 7, No.1, pp. 129-137.

[5] Rifki Wijaya, ArySetijadiPrihatmanto, Kuspriyanto, " Preliminary design of estimation heart disease by using machine learning ANN within one year", 2013, IEEE Joint International Conference on Rural Information & Communication Technology and Electric-Vehicle Technology (rICT&ICeV-T).

[6] Carlos Ordonez,2006, "Association Rule Discovery with the Train and Test Approach for Heart Disease Prediction", IEEE Transactions on Information Technology in Biomedicine (TITB), pp. 334-343, vol. 10, no. 2.

[7] Jyoti Soni, Uzma Ansari, Dipesh Sharma, and SunitaSoni,June 2011, "Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers", International Journal on Computer Science and Engineering (IJCSE), Vol. 3, No. 6, pp. 2385-2392.

[8] IbticemeSedielmaci; F. BereksiReguig, " Detection of some heart diseases using fractal dimension and chaos theory", 2013, IEEE 8th International Workshop on Systems, Signal Processing and their Applications (WoSSPA).

[9] Jayshril S. Sonawane; D. R. Patil, " Prediction of Heart Disease Using Learning Vector Quantization Algorithm ", 2014, IEEE Conference on IT in Business, Industry and Government (CSIBIG).

[10] AH Chen, SY Huang, PS Hong, CH Cheng, and EJ Lin,2011, "HDPS: Heart Disease Prediction System", Computing in Cardiology, ISSN: 0276-6574, pp.557- 560.

[11] Anbarasi Masilamani, ANUPRIYA, N Ch Sriman Narayana Iyenger, "Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm", October 2010, International Journal of Engineering Science and Technology 2(10).

[12] Manpreet Singh, Levi Monteiro Martins, Patrick Joanis, and Vijay K. Mago,2016, "Building a Cardiovascular Disease Predictive Model using Structural Equation Model & Fuzzy Cognitive Map", IEEE International Conference on Fuzzy Systems (FUZZ), pp. 1377-1382.

[13] Kathleen H. Miao, Julia H. Miao, "Coronary Heart Disease Diagnosis using Deep Neural Networks ",(IJACSA)International Journal of Advanced Computer Science and Applications, Vol. 9, No. 10, 2018 .

[14] Jae Kwon Kim and Sanggil, "Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis", 2017.

[15] Sairabi H. Mujawar, and P. R. Devale, October 2015,"Prediction of Heart Disease using Modified k-means and by using Naive Bayes", International Journal of Innovative Research in Computer and Communication Engineering(An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 10, pp. 10265-10273.

[16] B Padmaja, V V Rama Prasad, K V N Sunitha (2016), "TreeNet analysis of human stress behavior using socio-mobile data," in Journal of Big Data, 3(1), pp. 1-15, 2016, Springer.

[17] B Padmaja, Myneni Madhu Bala, E Krishna Rao Patro (2020), "A Comparison on visual prediction models for MAMO(multi activity multi-object) recognition using Deep Learning," in Journal of Big Data, 7(24), pp. 1-15, Springer.

[18] B Padmaja, V V Rama Prasad, K V N Sunitha (2020), " A Novel Random Split Point Procedure using Extremely Randomized Trees Ensemble Method for Human Activity Recognition," in EAI Endorsed transactions on Pervasive Health and Technology, 6(22), PP. 1-10.

[19] B Padmaja, V V Rama Prasad, K V N Sunitha (2018), "Machine Learning Approach for Stress Detection using Wireless Physical Activity Tracker," in International Journal of Machine Learning and Computing, 8(1), pp. 33-38.