HEART DISEASE PREDICTION USING MACHINE LEARNING

Dr.M.Vasumathi Devi[1], K.P.N.Vijaya sri[2], J.Manasa Gayathri[3], B.Kusuma kumari[4], P.Akshitha[5], M.Kausthubha Ramya[6]

Department of Computer Science and Engineering, Vignan'sNirula Institute of Technology and Science for Women, Pedapalakaluru, AP, India.

Abstract

Now a days millions of people are suffering from various diseases that can cause death. Some of the diseases like cancer, heart diseases, diabetes etc. if not identified in early stages can cause a lot of problems sometimes even instant death. Diagnosis of diseases at the right time will be of great help. Therefore to help the diagnosis process many data mining and machine learning techniques can be used. A lot of information about patient's history is present in today's health industry. The huge amount of data can be scrutinized using data mining techniques, later machine learning algorithms can be used for the prediction process. Machine learning in recent years has been evolving with reliable and supporting tools in medical field and has provided the best support for predicting disease with correct cases of training and testing. This research is intended to supply an in depth description of Random forest that are applied in our research particularly within the prediction of heart condition. Some experiments have been conducted to match the execution of predictive techniques on an equivalent data set, and therefore the consequence reveals that Random forest outperforms over logistic regression and other algorithms.

Introduction

The heart is a muscular organ which pumps blood into the body and is the central a part of the body's circulatory system along with lungs. Circulatory system also comprises a network of blood vessels, veins, arteries, and capillaries. Blood is delivered to whole body by these blood vessels. Abnormalities in normal blood flow from the guts cause several sorts of heart diseases which are commonly referred to as cardiovascular diseases (CVD). Heart diseases are the major reasons for death worldwide. Heart disease term includes a number of diseases such as blood vessel diseases, such as coronary artery disease; heart rhythm problems (arrhythmias); and heart defects you're born with (congenital heart defects). Cardiovascular disease (CVD) generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack (Myocardial infarctions), chest pain (angina) or stroke. Other heart conditions, such as those that affect your heart's muscle, valves or rhythm, also are considered forms of heart disease.17.9 million People die each year from CVDs, an estimated 31% of all deaths worldwide. Nowadays health care sector produces large amount of information about patients, disease diagnosis etc. however this data is not use deficiently by the researchers and practitioners. Today a major challenge faced by Healthcare industry is quality of service (QoS). QoS implies diagnosing disease correctly &provides effective treatments to patients. Poor diagnosis can lead to disastrous consequences which are unacceptable. There are various heart disease risk factors. Family history, Increasing age, Ethnicity and being male are some risk factors that cannot be controlled. But Smoking, Diabetes, High cholesterol, High blood pressure, not being physically active, being overweight or obese are those factors that can be controlled or prevented.
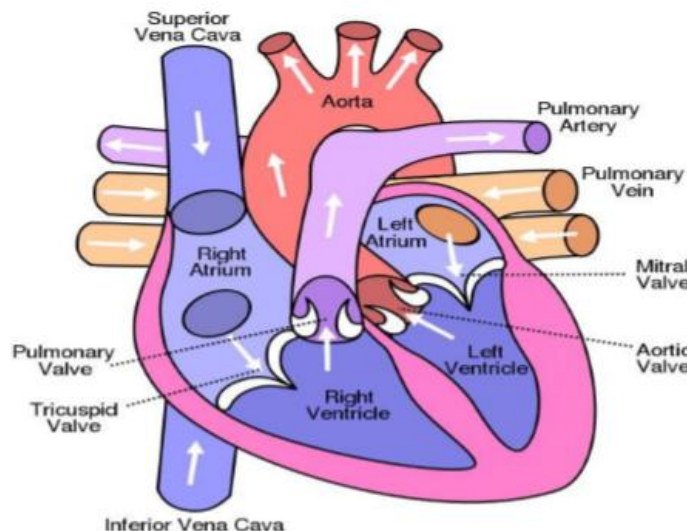
Fig: Structure of heart

Data mining is the process of discovering unknown hidden patterns (knowledge) from large pre- existing data sets with the involvement of data mining and machine learning techniques, statistics and database systems. The discovered knowledge can be used to build intelligent predictive decision systems in different fields like health care for accurate diagnosis at accurate time to provide affordable services and save precious lives. Machine learning provides computer programs the ability to learn from predetermined data and improve performance from experiences without human intervention and then apply what have learned to make an informed decision. At every successful decision machine learning program improves its performance. Given below figure depicts the knowledge discovery from data (KDD) process. Machine learning can be used to determine the automated end of diagnostic guidelines from the past descriptions, efficiently treat an affected person, as well as experts and professionals will assist and make the diagnostic system more reliable. The system makes use of machine learning algorithms to analyze the data available, train the models which are later evaluated. Algorithms used for prediction purpose are decision tree, svm, KNN, naive bayes, random forest and logistic regression. Models are built using all the four algorithms and their accuracies are compared. These models can be used to predict the type of heart disease patient is suffering from.

LITERATURE SURVEY

Machine Learning techniques are used to analyze and predict the medical data information resources. Diagnosis of heart disease is a significant and tedious task in medicine. (Sung, S.F. et al., 2015) have brought about the two Machine Learning techniques, k-nearest neighbor model and existing multi linear regression to predict the stroke severity index (SSI) of the patients. Their study show that knearest neighbor performed better than Multi Linear Regression model. (Arslan, A. K. et al., 2016) have suggested various Machine Learning techniques such as support vector machine (SVM), penalized logistic regression (PLR) to predict the heart stroke. Their results show that SVM produced the best performance in prediction when compared to other models.

Boshra Brahmi et al[7], developed different Machine Learning techniques to evaluate the prediction and diagnosis of heart disease. The main objective is to evaluate the different classification techniques such as J48, Decision Tree, KNN and Naïve Bayes. After this, evaluating some performance in measures of accuracy, precision, sensitivity, specificity are evaluated.

K. Polaraju et al[1], proposed Prediction of Heart Disease using Multiple Regression Model and it proves that Multiple Linear Regression is appropriate for predicting heart disease chance. The work is performed using training data set consists of 3000 instances with 13 different attributes which has mentioned earlier. The data set is divided into two parts that is 70% of the data are used for training and 30% used for testing. Based on the results, it is clear that the classification accuracy of Regression algorithm is better compared to other algorithms.

S. Seema et al[2], focuses on techniques that can predict chronic disease by mining the data containing in historical health records using Naïve Bayes, Decision tree, Support Vector Machine(SVM) and Artificial Neural Network(ANN). A comparative study is performed on classifiers to measure the better performance on an accurate rate. From this experiment, SVM gives highest accuracy rate, whereas for diabetes Naïve Bayes gives the highest accuracy.

Ashok Kumar Dwivedi et al[3], recommended different algorithms like Naive Bayes, Classification Tree, KNN, Logistic Regression, SVM and ANN. The Logistic Regression gives better accuracy compared to other algorithms.

MeghaShahi et al[4], suggested Heart Disease Prediction System using Data Mining Techniques. WEKA software used for automatic diagnosis of disease and to give qualities of services in healthcare centres. The paper used various algorithms like SVM, Naïve Bayes, Association rule, KNN, ANN, and Decision Tree. The paper recommended SVM is effective and provides more accuracy as compared with other data mining algorithms.

ChalaBeyene et al[5], recommended Prediction and Analysis the occurrence of Heart Disease Using Data Mining Techniques. The main objective is to predict the occurrence of heart disease for early automatic diagnosis of the disease within result in short time. The proposed methodology is also critical in healthcare organisation with experts that have no more knowledge and skill. It uses different medical attributes such as blood sugar and heart rate, age, sex are some of the attributes are included to identify if the person has heart disease or not. Analyses of dataset are computed using WEKA software.

K.Gomathi et al[6], suggested multi disease prediction using data mining techniques. Nowadays, data mining plays vital role in predicting multiple disease. By using data mining techniques the number of tests can be reduced. This paper mainly concentrates on predicting the heart disease, diabetes and breast cancer etc

Jaymin Patel et al[10]. compared different algorithms of Decision tree classification for better performance in heart disease diagnosis using WEKA. But here the greatest disadvantage is size, which increases linearly with the examples.

Proposed work

After evaluating the results from the existing methodologies, we proposed an intelligent and user friendly heart disease prediction system. We analyze the multiple machine learning algorithms and identify the better accurate model. Random forest algorithm is used to improve the accuracy of the system. In the proposed system we consider the relevant features from the dataset for building the model. The dataset is preprocessed, then trained and then tested. When the user gives the input, the model predict the chance of getting the heart disease.

ADVANTAGES
   ● Increased accuracy for effective heart disease diagnosis.
   ● Handles enormous amount of data using random forest algorithm and feature selection.
   ● Cost effective for patients.

SYSTEM ARCHITECTURE

The below figure shows the process flow diagram or proposed work. First we collected the Heart Disease Database from UCI website then pre-processed the dataset and select 16 important features. It is defined as the

process of cleaning, transforming, filling missing values and modeling the data to give us helpful information for healthcare decision making. The purpose of this is to preprocess the data and to get useful information by data and taking the decisions based upon the data analysis.
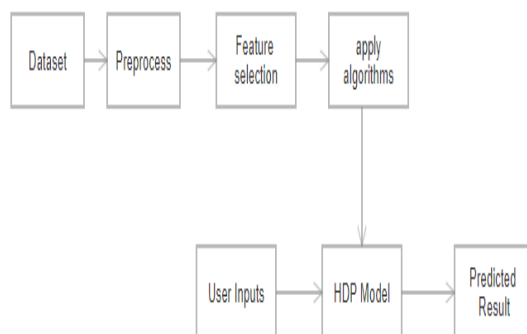
Fig: System Architecture

ALGORITHMS USED:

Random forest algorithm:

Random forest is a supervised learning algorithm which is used for both classification as well as regression .But however ,it is mainly used for classification problems .As we know that a forest is made up of trees and more trees means more robust forest . Similarly ,random forest creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting .It is ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result .

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of the given dataset.

Working of Random forest as the following steps:

- First, start with the choice of random samples from a given dataset.
- Next, the algorithm will construct a tree for each sample dataset.
- Then it'll give the prediction result for every decision tree.
- Now, voting is performed for predicted results.
- Finally, select the most voted prediction results as the final prediction result.
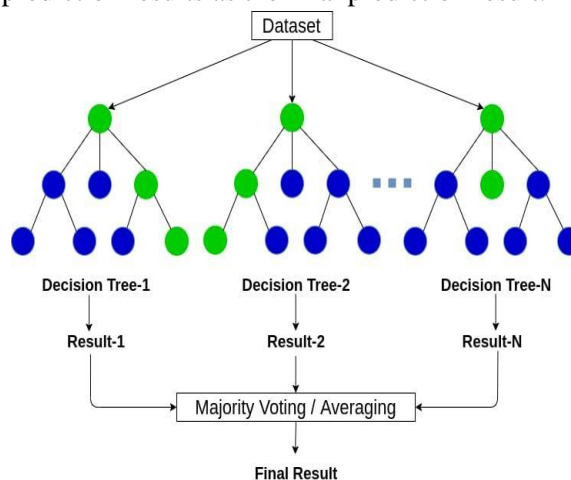
Fig: Working of Random Forest

Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, it predicts the final output.
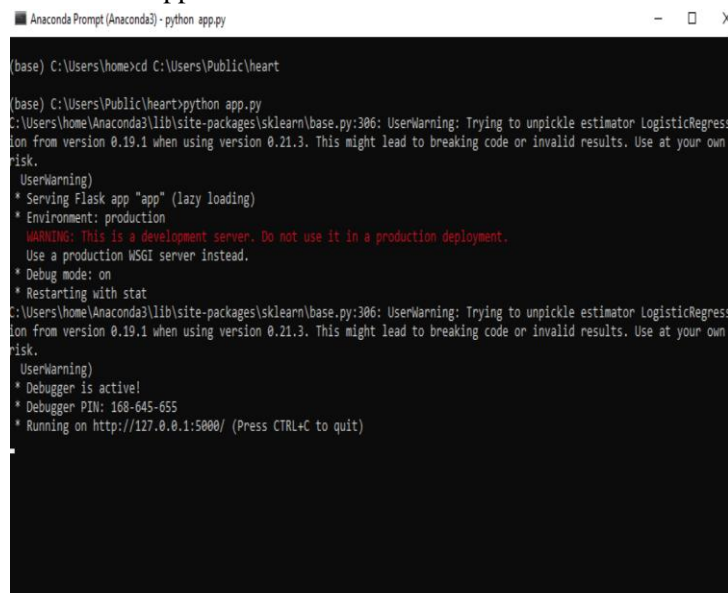
The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

The output doesn't depend on one decision tree but on various decision trees

In classification problem, the final output is taken by using majority taking classifiers and in regression problem, the final output is the mean of all outputs.

Results

For predicting the heart disease, first step is to execute the project file on the Anaconda Navigator. Launch the Anaconda prompt and run the app.py file and it displays the url where the project is running. The url should be entered on the web browser to run the application.

Fig: Anaconda prompt

The below screenshot is about the home page, that is displayed after executing the project file. The home page contains five text boxes for user to give respective inputs and one button called predict which predicts heart disease with a click.
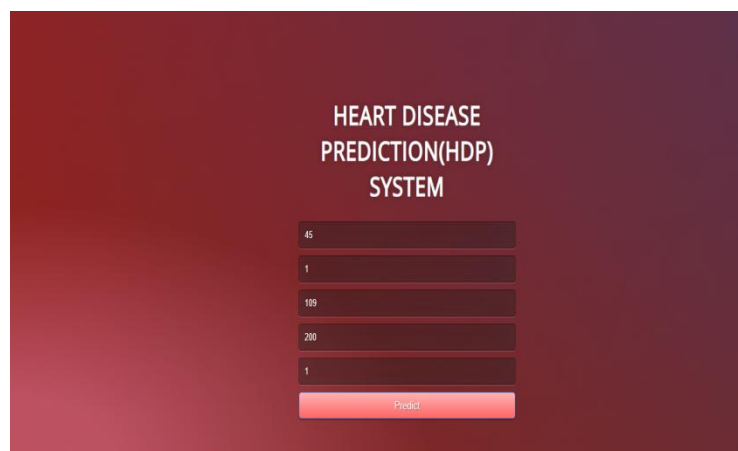
Fig: Home Screen

The below screenshot is about giving the inputs. The inputs that are to be given are age, gender, blood pressure or trestbps, cholesterol level and chest pain value. All the inputs are to be given as numerical values. Gender with values 1,0 representing male and female respectively, chest pain with values 0 or 1 or 2 or 3. The browser prompts the alert boxes on entering invalid values.

Now to predict the risk of heart disease, user can enter the values of these various parameters on the basis of which his risk factor to get disease will be calculated.

Fig: Input screen

After entering all the values, clicking on Predict button is done. The page will be reloaded and the result will be shown. If result displayed is may have heart disease, user may have a risk of heart disease. If result displayed is

may not have heart disease, the user may not have a risk of getting heart disease.
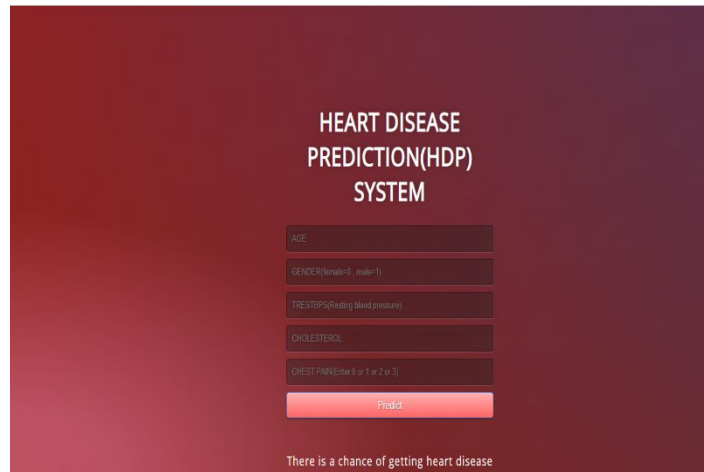

Fig: Output Screen

Conclusion

In this project, we introduce about the heart disease prediction system with different classifier techniques for the prediction of heart disease. We have analyzed that the Random Forest has better accuracy as compared to Logistic Regression. Our purpose is to improve the performance of the Random Forest by removing unnecessary and irrelevant attributes from the dataset and only picking those that are most informative for the classification task. .

The proposed model has wide area of application like grid computing, cloud computing, robotic modeling, etc. To increase the performance of our classifier in future, we will work on ensembling two algorithms called Random Forest and Adaboost. By ensembling these two algorithms we will achieve high performance.

References

[1] K. Polaraju, D. Durga Prasad, "Prediction of Heart Disease using Multiple Linear Regression Model", International Journal of Engineering Development and Research Development, ISSN:2321-9939, 2017

[2] Dr.S.SeemaShedole, Kumari Deepika, "Predictive analytics to prevent and control chronic disease", https://www.researchgate.net/punlication/316530782, January 2016.

[3] Ashok kumar Dwivedi, "Evaluate the performance of different machine learning techniques for prediction of heart disease using ten-fold cross-validation", Springer, 17 September 2016.

[4] Megha Shahi, R. Kaur Gurm, "Heart Disease Prediction System using Data Mining Techniques", Orient J. Computer Science Technology, vol.6 2017, pp.457-466.

[5] Mr. ChalaBeyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques", International Journal of Pure and Applied Mathematics, 2018.

[6] K.Gomathi, Dr.D.ShanmugaPriyaa, "Multi Disease Prediction using Data Mining Techniques", International Journal of System and Software Engineering, December 2016, pp.12-14.

[7] Boshra Brahmi, Mirsaeid Hosseini Shirvani, "Prediction and Diagnosis of Heart Disease by Data Mining Techniques", Journals of Multidisciplinary Engineering Science and Technology, vol.2, 2 February 2015, pp.164168.

[8] Marjia Sultana, Afrin Haider, "Heart Disease Prediction using WEKA tool and 10-Fold crossvalidation", The Institute of Electrical and Electronics Engineers, March 2017.

[9] 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. 334343, vol. 10, no. 2.

[10] 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.

[11] P .K. Anooj, ―Clinical decision support system: Risk level predicon of heart disease using weighted fuzzy rules‖; Journal of King Saud University – Computer and Information Sciences (2012) 24, 27–40. Computer Science & Information Technology (CS & IT) 59

[12] Nidhi Bhatla, Kiran Jyoti"An Analysis of Heart Disease Prediction using Different Data Mining Techniques".International Journal of Engineering Research & Technology.

[13] Jyoti SoniUjma Ansari Dipesh Sharma, Sunita Soni. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction".

[14] Chaitrali S. DangareSulabha S. Apte, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques" International Journal of Computer Applications (0975 – 888)

[15] Dane Bertram, Amy Voida, Saul Greenberg, Robert Walker, "Communication, Collaboration, and Bugs: The Social Nature of Issue Tracking in Small, Collocated Teams".

[16] M. Anbarasi, E. Anupriya, N.Ch.S.N.Iyengar, ―Enhanced Predicon of Heart Disease with Feature Subset Selection using Genetic Algorithm‖; International Journal of Engineering Science and Technology, Vol. 2(10), 2010.

[17] Ankita Dewan, Meghna Sharma," Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification", 2nd International Conference on Computing for Sustainable Global Development IEEE 2015 pp 704-706. [2].

[18] R. Alizadehsani, J. Habibi, B. Bahadorian, H. Mashayekhi, A. Ghandeharioun, R. Boghrati, et al., "Diagnosis of coronary arteries stenosis using data mining," J Med Signals Sens, vol. 2, pp. 153-9, Jul 2012.

[19] M Akhil Jabbar, BL Deekshatulu, Priti Chandra," Heart disease classification using nearest neighbor classifier with feature subset selection", Anale. Seria Informatica, 11, 2013.