DOI: http://dx.doi.org/10.26483/ijarcs.v11i3.6532

ISSN No. 0976-5697

Volume 11, No. 3, May-June 2020

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

ANALYSIS OF MACHINE LEARNING ALGORITHM FOR PREDICTION OF HEART DISEASE

Parthasarathy G
Associate Professor
School of C&IT REVA University
Bangalore, India

Lakhan S Kesaragopp
Student
School of C&IT, REVA University
Bangalore, India

M M Vishal
Student
School of C&IT, REVA University
Bangalore, India

Manigandan S
Student,
School of C&IT, REVA University
Bangalore, India

Kundan Kulkarni
Student
School of C&IT, REVA University
Bangalore, India

Abstract: In today's era Heart disease has become one of the vital causes of death in the world. The age group which has the major death rate because of heart disease is from 30-69. Therefore, the prediction of heart disease at the early stage is a prime challenge nowadays, because of many risk factors. Predicting and Analyzing heart disease using a single data mining approach does not provide us the best accuracy with precision. So in this project we are using various Machine Learning Algorithm and Data Mining techniques like, Random Forest, Naive Bayes algorithm, Decision Tree, Support Vector Machine in order to get the best accuracy with precision and to reduce the number of tests required to be carried out to find the result, and provide the results quickly. This project includes collecting the best attributes to analyze and predict the heart disease, and use various machine learning algorithms to successfully generate the accurate result. And ensures that heart disease is predicted at the early stage.

Keywords: Machine Learning, Data Mining, Random Forest, Decision Tree, Naive Bayes, Support Vector Machine.

I. INTRODUCTION

Cardiovascular disease or heart disease generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain, or stroke. In India fatality rate due to heart disease is 24.8%. Diagnosing heart disease or cardiovascular disease is very much abstruse because of dominant risk factors such as unhealthy diet, high blood pressure, high cholesterol level, diabetes, smoking, etc. Diagnosing heart disease before an emergency, will increase the chance of survival. Heart disease is very much Complex. it requires continuous monitoring and treatment based on severity heart disease. In medical and health care areas, a huge amount of data is available for that instance hospital being one of the huge storage of data (big data) which provides the required data in building a medical diagnosis system. Recent advancement of technology in data mining helps in extracting big data (i.e. Patient data from hospital repository) which provides us the information in building a heart disease prediction system used to diagnose heart disease with precision and accuracy. In this project, various machine learning algorithms and Data mining techniques have been used which provide us the accurate result in diagnosing heart disease.

Data mining plays a dominant role in extracting and analyzing huge data and providing accurate results. Techniques like classification, prediction, clustering make heart disease diagnosis easier and faster. Data mining, classification techniques like Nave Bayes, Decision Tree are used to analyze the dataset based on the attribute.

In machine learning the training of the model is done by allowing the model to analyze and interpret the outcome of the data. Therefore, based on the new and different data the model can make a better prediction in diagnosing heart disease.

I. LITERATURE SURVEY

There is abundant related work in the fields which is directly related to this paper. Machine Learning and Data mining has been introduced to build the highest accuracy prediction in the medical field. The obtained results are compared with the results of existing models within the same domain and found to be improved. The data of patients who has heart disease which are collected from the UCI laboratory is used to determine the patterns with Random Forest Decision Tree, Support Vector Machines, and Naive Bayes. The results are compared to study the performance and accuracy with these algorithms. The proposed method returns results of 79.47%, competing with the other existing methods. A huge amount of data generated by the medical industry has not been used effectively previously. The new approaches presented here decrease the cost and improve the prediction of heart disease in an easy and effective way. The various different research techniques considered in this work for prediction and classification of heart disease using Machine Learning and Data Mining techniques are highly accurate in establishing the efficacy of these methods.

K. Polaraju, proposed Prediction of Heart Disease using Multiple Regression Model. This paper proves that Multiple Linear Regression yields best results in predicting heart disease. This model is trained by using training data set which consists of 3000 instances with 13 different attributes. This paper proves that multiple linear regression yields best result in prediction of heart disease.

Niti Guru, Anil Dahiya, Navin Rajpal has given paper in Decision Support System for Heart Disease Diagnosis using Neural Network.This paper mainly focused on incorporating Neural Network where only 78 instances were used to train neural network which resulted to be less effective.

Dilip Roy Chowdhury, Mridula Chatterjee, R.K. Samanta has proposed a paper in An Artificial Neural Network Model for Neonatal Disease Diagnosis. This paper focused on using Multi-Layered Perceptron where backpropagation is used which is a supervised learning technique to train the model. They have used about 94 attributes which is very huge data to train neural network which gives a result of about 75% accuracy.

Mohammed Abdul Khaleel has given paper in the Survey of Techniques for mining of data on Medical Data for Finding Frequent Diseases locally. This paper focus on dissect information mining procedures which are required for mining medicinal information. In this, the algorithm used was Naive Bayes algorithm in which Bayes theorem was used.. This model used more than 500 dataset. The tool used is Weka and classification is achieved by using 70% of Percentage Split and the accuracy offered by Naive Bayes is 86.419%.

For the detection of stroke disease, Sudha shetal studied classification algorithms like Naive Bayes, Decision Tree, and Neural Network. This study incorporates classification techniques such as Decision Tree, Bayesian classifiers, and Back-Propagation Neural Networks. This study states that before the mining process, records with irrelevant data were detected from the data warehouse.

S. Seema focuses on techniques that can predict chronic disease by mining the data containing in historical health records using Nave 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 Nave Bayes gives the highest accuracy.

R. Sharmila proposed to use non- linear classification algorithm for heart disease prediction. This paper uses tools such as Hadoop Distributed File System, Map-reduce ,and Support Vector Machine for prediction of heart disease using minimum attributes. In this paper HDFS is used for storing large data in different nodes and executing SVM in more than one node simultaneously. SVM is used parallelly which yielded better computation time than sequential SVM.

II. Methodology

In this study, we have used Random Forest Classifier to perform heart disease classification of the Cleveland UCI repository. Machine Learning process starts from a Pre-Processing data phase followed by feature selection based on Decision Tree entropy, classification of modeling performance evaluation, and the results with improved accuracy. The feature selection and modeling keep on iterating for different configuration of attributes. The performance of each model is generated based on 9 features from 13 features and Machine Learning techniques are used for each progression and each performance is recorded. Section A describes the method for pre-processing of data, Section B summarizes the feature selection and reduction, and Section C explains classification modeling. Figure 1 repersents the data flow in the model.

A. Data Pre-Processing

The data from Cleveland UCI repository is collected and the data is pre-processed to remove the missing attribute from the data-set. The dataset contains a total of 302 patient data, where 6 data are having missing values and these 6 data have been removed from the data-set and the remaining 296 data have been used in pre-processing. The multiclass classification variable is used to check the presence or absence of heart disease. In the instance if the patient is having heart disease, the value is set to 1, and if the patient is not having heart disease then the value is set to 0. The pre-processing of data is carried to transmute medical records into diagnosis values. The results of data pre-processing for 296 Patient data indicate that 139 records show the value of 1 establishing the presence of heart disease while the remaining 163 reflected the value of 0 indicating the absence of heart disease.

B. Feature Selection and Reduction

From among the 13 attributes of the data set only 9 attributes are used in which two attributes pertaining to age and sex are used to identify the personal information of the patient and the remaining 11 attributes are considered important as they contain vital clinical records. Clinical records are vital to diagnosis and learning the severity of heart disease. As previously mentioned in this experiment, several ML techniques are used namely Random Forest, Nave Bayes and SVM. The experiment will be repeated with all the ML techniques using all 13 attributes.

C. Classification Modeling

The clustering of datasets is done based on the variables and criteria of Decision Tree features. Then, the classifiers are applied to each clustered dataset in order to estimate its performance. The accurate performing models are identified from the above results based on their low rate of error. The performance is further optimized by choosing the Decision Tree cluster with a high rate of error and extraction of its corresponding classifier features. The performance of the classifier is evaluated for error optimization on this data set.

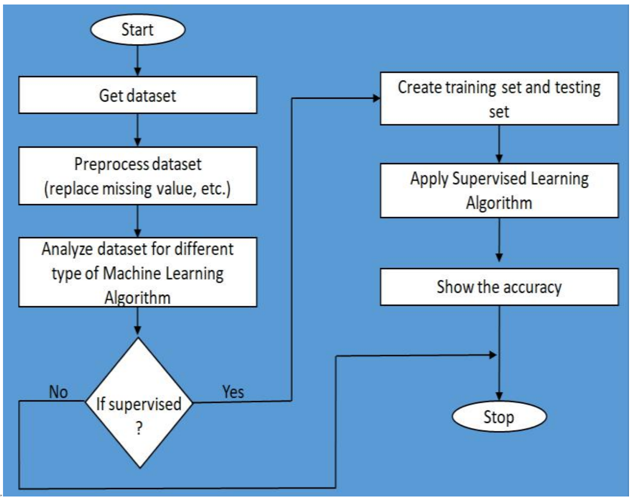

Start
Get dataset
Preprocess dataset
(replace missing value, etc.)
Analyze dataset for different type of Machine Learning Algorithm
No
If supervised ?
Yes
Create training set and testing set
Apply Supervised Learning Algorithm
Show the accuracy
Stop

Fig. 1. Data Flow Diagram

III. Proposed System

In this study, we used UCI repository which contains 9 attributes such as age, gender, resting blood pressure(in mm Hg on admission to the hospital), serum cholesterol in mg/dl, fasting blood sugar, Rest ECG results, maximum heart rate achieved during ECG, chest pain during exercise, chest pain type. 75% of the dataset is used to train the model using the classifier and 25% is used for testing. A GUI is constructed using Visual Studio Code to extract the details of the user and then is subjected to the model which classifies the patient into four stages depending on the exigency of the disease or as if the patient is out of danger. From arbitrarily selected subset of training data set, random forest classifier creates few data set of decision trees. It then aggregates the results from distinct decision trees to determine the final class of the test object. In this model we used sklearn classifiers which requires data cleaning. Hence the data is pre-processed before it was trained. As random forest classifier is a group of algorithm which gives more accurate result because of the cause of principle, Number of weak estimators when combined forms a strong estimator.

a) *Decision Tree*: Decision Tree is a form of the inductive learning task, which uses an objective like, using a training dataset of the patient data to create a hypothesis that gives compared results. The trees are constructed based on high entropy inputs. These trees are simple and fast constructed in a top-down recursive divide and conquer (DAC) approach. In the dataset, Tree pruning is performed to remove the irrelevant samples.

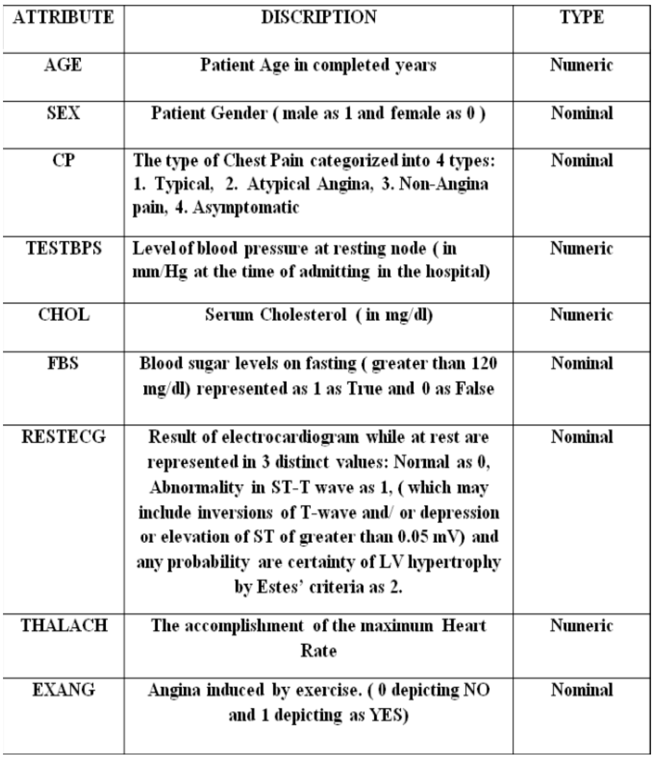
b) *Random Forest*: This uses an ensemble learning method, which amalgamates multiple classifiers to solve a complex dataset and improve the performance of the model. This algorithm contains several decision trees on a various subset of the given dataset and takes the average to improve accuracy. This algorithm tends to take less training time as compared to other algorithms and gives higher accuracy even for large datasets.

c) *Naive Bayes*: The Naive Bayes classifier depends upon Bayesian hypothesizes with autonomy which suspects among the attributes. The output is not hard to run, with no entrapped repetitive parameter estima- tion, which makes it particularly supportive for broad datasets despite its effortlessness, the Naive Bayesian classifier generally completes its job with precision and is broadly used in consideration that it frequently vanquish high order techniques which are complex.

d) *Support Vector Machine*: SVM is fundamentally used for classification problems. The objective is to create an accurate line or decision boundary that dissociates n-dimensional space into classes so that the model can easily search for the correct class and store the new data point in that class after execution. The best and accurate decision boundary is known as hyper-plane which is created by using extreme point called support vectors.

IV. Data Sets

Healthcare databases have collected a significant amount of patients records. The data is obtained from the Cleveland heart disease database (UCI Repository). Datasets segregate the patterns related to the disease. The records classify into two datasets: 75% training dataset and 25% testing dataset. Figure 2 represents data type and figure 3 represents the dataset range used for training the model.



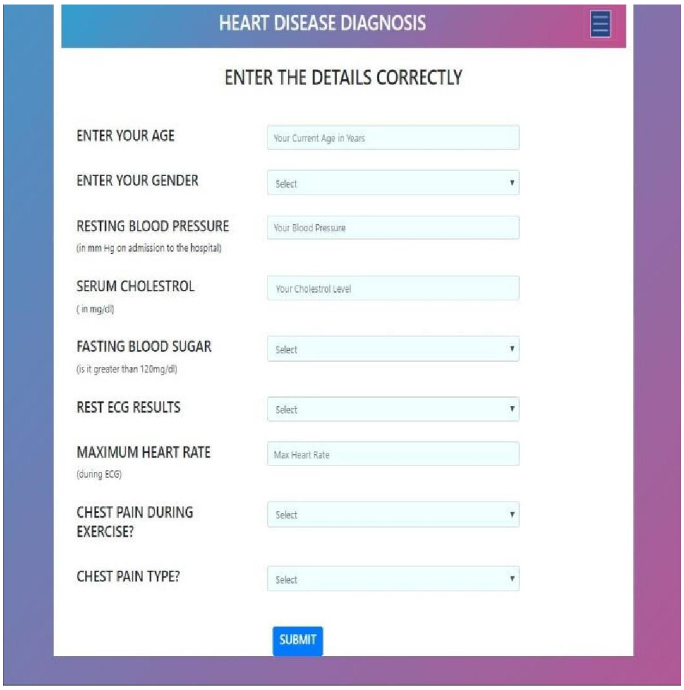
ATTRIBUTE	DISCRIPTION	TYPE
AGE	Patient Age in completed years	Numeric
SEX	Patient Gender (male as 1 and female as 0)	Nominal
CP	The type of Chest Pain categorized into 4 types: 1. Typical, 2. Atypical Angina, 3. Non-Angina pain, 4. Asymptomatic	Nominal
TESTBPS	Level of blood pressure at resting node (in mm/Hg at the time of admitting in the hospital)	Numeric
CHOL	Serum Cholesterol (in mg/dl)	Numeric
FBS	Blood sugar levels on fasting (greater than 120 mg/dl) represented as 1 as True and 0 as False	Nominal
RESTECG	Result of electrocardiogram while at rest are represented in 3 distinct values: Normal as 0, Abnormality in ST-T wave as 1, (which may include inversions of T-wave and/ or depression or elevation of ST of greater than 0.05 mV) and any probability are certainty of LV hypertrophy by Estes' criteria as 2.	Nominal
THALACH	The accomplishment of the maximum Heart Rate	Numeric
EXANG	Angina induced by exercise. (0 depicting NO and 1 depicting as YES)	Nominal

Fig. 2. Data Type

AGE	Numeric[29 to 77]	Unique = 41 ; Mean = 54.4 ; Median = 56
SEX	Numeric[0 to 1]	Unique = 2 ; Mean = 0.68 ; Median = 1
CP	Numeric[1 to 4]	Unique = 4 ; Mean = 3.16 ; Median = 3
TESTBPS	Numeric[94 to 200]	Unique = 50 ; Mean = 131.69 ; Median = 130
CHOL	Numeric[126 to 564]	Unique = 152 ; Mean = 246.69 ; Median = 241
FBS	Numeric[0 to 1]	Unique = 2 ; Mean = 0.15 ; Median = 0
RESTECG	Numeric[0 to 2]	Unique = 3 ; Mean = 0.99 ; Median = 1
THALACH	Numeric[71 to 202]	Unique = 91 ; Mean = 149.61 ; Median = 153
EXANG	Numeric[0 to 1]	Unique = 2 ; Mean = 0.33 ; Median = 0

Fig. 3. Dataset Range

V. RESULT AND ANALYSIS

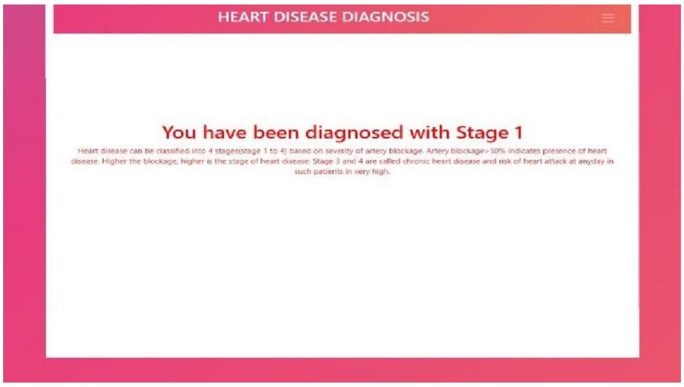


HEART DISEASE DIAGNOSIS

ENTER THE DETAILS CORRECTLY

ENTER YOUR AGE	Your Current Age in Years
ENTER YOUR GENDER	Select
RESTING BLOOD PRESSURE (in mm Hg on admission to the hospital)	Your Blood Pressure
SERUM CHOLESTROL (in mg/dl)	Your Cholestrol Level
FASTING BLOOD SUGAR (is it greater than 120mg/dl)	Select
REST ECG RESULTS	Select
MAXIMUM HEART RATE (during ECG)	Max Heart Rate
CHEST PAIN DURING EXERCISE?	Select
CHEST PAIN TYPE?	Select

Fig. 4. Input

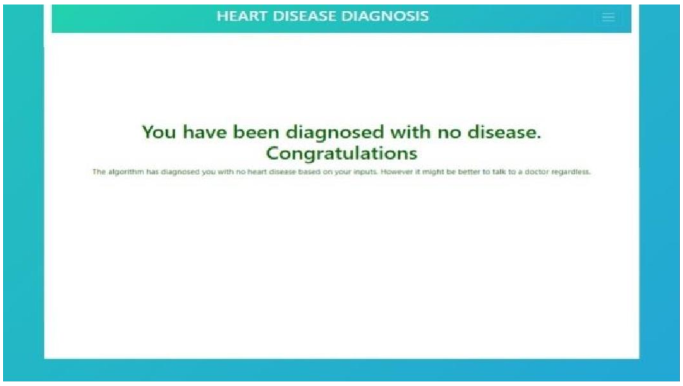


HEART DISEASE DIAGNOSIS

You have been diagnosed with Stage 1

Heart disease can be classified into 4 stages(stage 1 to 4) based on severity of artery blockage. Artery blockage>50% indicates presence of heart disease. Higher the blockage, higher is the stage of heart disease. Stage 3 and 4 are called chronic heart disease and risk of heart attack at anyday in such patients in very high.

Fig. 5. Output 1



HEART DISEASE DIAGNOSIS

You have been diagnosed with no disease. Congratulations

The algorithm has diagnosed you with no heart disease based on your inputs. However it might be better to talk to a doctor regardless.

Fig. 6. Output 2

VI. CONCLUSION

Identifying the processing of raw healthcare data of heart information will help in the long term saving of human lives and early detection of abnormalities in heart conditions. However, the mortality rate can be significantly supervised if the disease is recognized from the outset and preventative measures are primly embraced as soon as possible. Machine Learning and Data Mining techniques made our work easy in processing raw data and provide a new and novel discernment towards heart disease. Heart disease diagnosis is challenging and very important in the medical field. The data collected from the repository is very large, therefore the decision tree model usually works best for categorizing data and storing the correct data into the new data points. The different attributes in our dataset represent the symptoms most favourable leading to heart disease in patients. Training the model with selected attributes is one of the important aspects, therefore using a random forest classifier the model yields results with the best precision and accuracy. Thus, our work has successfully contrived machine learning techniques and data mining in diagnosing heart disease which enhances accuracy and reduces cost factors. Our model is predicting heart disease with an accuracy of about 79.47%.

VIII. REFERENCES

[1] Algorithms in the International Journal of Advanced Engineering, Management and Science (IJAEMS) June-2016 vol-2.

[2] Deeanna Kelley Heart Sonam Nikhar, A.M. Karandikar Prediction of Heart Disease Using Machine Learning Disease: Causes, Prevention, and Current Research in JCCC Honors Journal.

[3] Nabil Alshurafa, Costas Sideris, Mohammad Pourhomayoun, Haik Kalantarian, Majid Sarrafzadeh ”Remote Health Monitoring Outcome Success Prediction using Baseline and First Month Intervention Data” in IEEE Journal of Biomedical and Health Informatics.

[4] Ponrathi Athilingam, Bradlee Jenkins, Marcia Johansson, Miguel Labrador ”A Mobile Health Intervention to Improve Self-Care in Patients With Heart Failure: Pilot Randomized Control Trial” in JMIR Cardio 2017, vol. 1, issue 2, pg no:1.

[5] DhafarHamed, Jwan K. Alwan, Mohamed Ibrahim, Mohammad B. Naeem ”The Utilisation of Machine Learning Approaches for Med- ical Data Classification” in Annual Conference on New Trends in Information and Communications Technology Applications - march-

2017.

[6] Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients Mai Shouman, Tim Turner, and Rob Stocker International Journal of Information and Education Technology, Vol. 2, No. 3,June 2012.

[7] Amudhavel, J., Padmapriya, S., Nandhini, R., Kavipriya, G., Dha- vachelvan, P., Venkatachalapathy, V.S.K., ”Recursive ant colony op- timization routing in wireless mesh network”, (2016) Advances in Intelligent Systems and Computing, 381, pp. 341-351.

[8] Alapatt, B.P., Kavitha, A., Amudhavel, J., ”A novel encryption al- gorithm for end to end secured fiber optic communication”, (2017) International Journal of Pure and Applied Mathematics, 117 (19 Special Issue), pp. 269-275.

[9] Amudhavel, J., Inbavalli, P., Bhuvaneswari, B., Anandaraj, B., Vengat- taraman, T., Premkumar, K., ”An effective analysis on harmony search optimization approaches”, (2015) International Journal of Applied Engineering Research, 10 (3), pp. 2035-2038.

[10] Amudhavel, J., Kathavate, P., Reddy, L.S.S., Bhuvaneswari Aad- harshini, A., ”Assessment on authentication mechanisms in distributed system: A case study”, (2017) Journal of Advanced Re- search in Dynamical and Control Systems, 9 (Special Issue 12), pp. 1437-1448.

[11] Amudhavel, J., Kodeeshwari, C., Premkumar, K., Jaiganesh, S., Ra- jaguru, D., Vengattatraman, T., Haripriya, R., ”Comprehensive analysis on information dissemination protocols in vehicular ad hoc networks”, (2015) International Journal of Applied Engineering Research, 10 (3), pp. 2058-2061.

[12] Amudhavel, J., Kathavate, P., Reddy, L.S.S., Satyanarayana, K.V.V., ”Effects, challenges, opportunities and analysis on security based cloud resource virtualization”, (2017) Journal of Advanced Research in Dynamical and Control Systems, 9 (Special Issue 12), pp. 1458- 1463.

[13] N.Al-milli, ”Backpropogation neural network for prediction of heart disease”, J.Theor. Appl.Inf. Technol., vol. 56, pp. 131-135, 2013.