Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7766225

A SURVEY ON HEART DISEASE PREDICTION USING MACHINE LEARNING TECHNIQUES

Mr. Mohit Patel[1], Dr. Dushyantsinh B. Rathod[2]
PhD Scholar[1] ,Professor & HOD[2]
[1,2,]Computer Engineering Department ,
Swarnim Startup & Innovation University,Kalol,Gujarat[1],
Ahmedabad Institute of Technology,Ahmedabad,Gujarat[2]

ABSTRACT

Heart plays significant role in living organisms. Diagnosis and prediction of heart related diseases requires more precision, perfection and correctness because a little mistake can cause fatigue problem or death of the person, there are numerous death cases related to heart and their counting is increasing exponentially day by day. To deal with the problem there is essential need of prediction system for awareness about diseases. Machine learning is the branch of Artificial Intelligence(AI), it provides prestigious support in predicting any kind of event which take training from natural events. In this paper, we calculate accuracy of machine learning algorithms for predicting heart disease, for this algorithms are logistic regression , random forest classifier and support vector machine(SVM) by using kaggle dataset for training and testing. For implementation of Python programming Anaconda(jupyter) notebook is best tool, which have many type of library, header file, that make the work more accurate and precise.

Keywords: Supervised; Unsupervised; Reinforced; Logistic Regression; SVM; Random Forest Classifier Python Programming; Jupyter Notebook.

I. INTRODUCTION

Heart is one of the most extensive and vital organ of human body so the care of heart is essential. Most of diseases are related to heart so the prediction about heart diseases is necessary and for this purpose comparative study needed in this field, today most of patient are died because their diseases are recognized at last stage due to lack of accuracy of instrument so there is need to know about the more efficient algorithms for diseases prediction. Machine Learning is one of the efficient technology for the testing, which is based on training and testing. It is the branch of Artificial Intelligence(AI) which is one of broad area of learning where machines emulating human abilities, machine learning is a specific branch of AI. On the other hand machines learning systems are trained to learn how to process and make use of data hence the combination of both technology is also called as Machine Intelligence.

As the definition of machine learning, it learns from the natural phenomenon, natural things so in this project we uses the biological parameter as testing data such as cholesterol, Blood pressure, sex, age, etc. and on the basis of these, comparison is done in the terms of accuracy of algorithms such as in this project we have used three algorithms which are logistic regression, random forest , SVM. In this paper, we calculate the accuracy of three different machine learning approaches and on the basis of calculation we conclude that which one is best among them.

The Heart is an indispensable organ in Human beings. Heart disease is the main reason for the deaths of many people in the world. As per WHO, every year 12 million deaths are caused due to cardiovascular disease. Heart Disease is like a silent killer which results in the death of a person without obvious symptoms. Early identification of disease leads to prevention of disease and which in turn reduces the complications. As we say, prevention is better than cure, preventing heart disease can be able to prevent many premature deaths and reduce the mortality rate. Doctors may not be able to monitor the patient for 24 hours. Although there are lots of instruments in the market, they are not capable of detecting heart disease accurately and some of the instruments are very expensive and would also require expertise in the field. Machine learning is a trending technology, which is a subclass of artificial intelligence. Machine learning allows machines to enhance at tasks with experience. Machine learning enables a system to identify patterns by itself and make predictions.

In this Research, we use machine learning to predict whether a person is having heart disease or not. We consider various attributes of patients like gender, chest pain type, fasting blood pressure, serum cholesterol, exang, etc. We utilize various algorithms like SVM, Random Forest, Logistic Regression. Based on attributes, we perform a comparative analysis of algorithms regarding the accuracy, and whichever algorithm is giving better accuracy, is considered for heart disease prediction.

II. LITERATURE SURVEY

This paper describes the prediction of heart disease in the medical field through the use of data science. Because a lot of research carries out research related to that problem, the accuracy of the forecast has yet to be Improved. Therefore, this research focuses on features Selection techniques and algorithms in which multiple data sets on heart disease are used for experimental analysis and to show greater accuracy. we propose a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction Model is introduced with different combinations of Features and several known classification techniques.

In this paper, they analyze the commonly used classification algorithms in the medical data set that helps predict heart diseases that are the main ones Cause of death throughout the world. It is complex for doctors Professionals to anticipate the heart attack as required experience and knowledge The healthcare sector today contains hidden but meaningful information to create decisions The experiments carried out reveal this algorithm As expected

Table 1: Literature Survey 1

AUTHOR	TITLE	PURPOSE	ALGORITHMS USED AND ACCURACY

1) Mr.Santhana Krishnan.J, Dr.Geetha.S	Prediction of Heart Disease Using Machine Learning Algorithms Year-2018	In this system, a heart disease data set is used. The main aim of this system is to predict the possibilities of 91% occurring heart disease of the patients in terms of percentage. This is performed through data mining classification technique	Decision tree 91% Naive Bayes 87%
2) M. Marimuthu, S.Deivarani, Gayathri.R	Analysis of Heart Disease Prediction using Various Machine Learning Techniques year-2018	To achieve better accuracy and to make the system more efficient so that it can predict the chances of heart attack.	KNN NB 83.60% 80.66% Decision tree 75.58% SVM 65.56%
3) Sanchayita Dhar Pritha Datta, Ankur Biswas,Tanusree Dey. Krishna Roy	A Hybrid Machine Learning Approach for Prediction of Heart Diseases Year-2018	In this paper, to develop a prediction system that be capable to envisage heart diseases based on measurements, are extracted from The ERIC laboratory consisting of 209 test cases..	Naive Bayes Decision tree Random Forest
4) Rajesh N, T Maneesha, Shaik Hafeez, Hari Krishna	Year-2018 Prediction of Heart Disease Using Machine Learning Algorithms	In this paper, processing patient's dataset and a data of patients to whom we need to predict the chance of occurrence of a heart disease.	Naive Bayes Decision tree (ID3 Algorithm)

Table 2: Literature Survey 2

	Paper 1 (Base)	Paper 2	Paper 3	Paper 4
Title	Effective heart disease prediction using hybrid machine learning techniques	Prediction of Heart Disease using machine learning	Prediction of Heart Disease using Machine Learning Algorithms (Decision tree and naive Bayes)	Prediction of Heart Disease and early stage using data mining and big data Analytics: A survay
Year	2019	2018	2019	2019
Methods used	HRFLM(Hybrid random forest)	Neural netrworks	Decision Tree, Naive Bayes	Decision Tree, Naive Bayes, Neural networks, KNN, support vector machine
Parameters used	Age, cp, fbs, restecg, exang, old peak, slope, sex, ca,thal	Age, Sex, Blood Pressure, Heart Rate, Diabetes, Hyper cholesterol, body mass indesiopacity),	Tree, Naive Bayes Cholesterol, blood pressure, Diabetics, Smoking, alcohol, overweight or obese, history of coronary illness	Chestpain, blood pressure, cholestrol and blood sugar

Best method	HRLFM	Neural network	Decision Tree	Neural network
Strength	Accuracy is high among all. 2) Classification error rateis low	MLP algorithm gives reliable op terms of accuracy, efficiency	Accuracy rate higher than naïve bayes andother.	Accuracy and MLP of NNis higher than DT, NB, SVM, KNN.

Table 3: Literature Survey 3

Sr.no	Title	Authors	Year	Algorithm used
1	Heart Disease Prediction Using Effective Machine Learning Techniques	Avinash Golande, Pavan Kumar T	2019	Decision Tree, KNN ,k-mean, addboost
2	Prediction of Heart Disease Using Machine Learning Algorithms	Mr.Santhana Krishnan.J Dr.Geetha S	2018	Decision tree, naive bayes
3	A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms	Amin Ul Haq,1 Jan Ping Li,1 Muhammad Hammad Memon,1 Shah Nazir,2 and Ruinan Sun	2018	Logistic regression, SVM, naive bayes, Neural network

III. METHODOLOGY

This paper shows the analysis of various machine learning algorithms, the algorithms that are used are SVM, Logistic Regression and Random Forest Classifiers, which can be helpful for practitioners or medical analysts for accurately diagnosing Heart Disease. This paperwork includes examining the journals, published paper and the data of cardiovascular disease of the recent times. Methodology gives a framework for the proposed model .

The methodology is a process which includes steps that transform given data into recognized data patterns for the knowledge of the users. The proposed methodology includes steps, where first step is referred as the collection of the data than in second stage it extracts significant values than the 3rd is the processing stage where we explore the data. Data processing deals with the missing values, cleaning of data and normalization depending on algorithms used . After processing of data, classifier is used to classify the processed data the classifier used in the proposed model are SVM. Logistic Regression, Random Forest Classifier.

Finally, the proposed model is undertaken, where we evaluated our model on the basis of accuracy and performance using various performance metrics. Here in this model, an effective Heart Disease Prediction System (EHDPS) has been developed using different classifiers. This model uses 13 medical parameters such as chest pain, fasting sugar, blood pressure, cholesterol, age, sex etc. for prediction .

Table 4: Attributes of the Dataset.

Sr. no	Attribute	Description
1.	Age	Patient's age
2.	Sex	Gender of patient(male-0 female-1)
3.	Cp	Chest pain type
4.	Trestbps	Resting blood pressure(in mm Hg on admission to hospital)
5.	Chol	Serum cholesterol in mg/dl
6.	Fbs	fasting blood sugar > 120 mg/dl
7.	Resteng	resting electrocardiographic results (values 0,1,2)
8.	Thal	maximum heart rate achieved
9.	Exang	exercise induced angina
10.	Oldpeak	oldpeak = ST depression induced by exercise relative to rest
11.	slope	the slope of the peak exercise ST segment
12.	ca	number of major vessels (0-3) colored by flourosopy
13.	Thal	thal: 0 = normal; 1 = fixed defect; 2 = reversable defect
14.	Target	1 or 0

IV. PROPOSED SYSTEM

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is processed into the required format. The data is then divided into two parts training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data. This system is implemented using the following modules.

1. Collection of Dataset.
2. Data Processing.
3. EDA.
4. Splitting of dataset.
5. Applying different algorithms
6. checking accuracy scores of algorithms.
7. Applying different model evaluation techniques.
8. Prediction on new data values.

9. Save model using JOBLIB.
10. GUI.

V. MODELING AND ANALYSIS

We are using different machine learning algorithms like.

1. logistic regression.
2. SVM (support vector machine)
3. Random forest classifier

Among these algorithms which one gives better accuracy will be the final algorithm for our prediction system which predicts whether patient is suffering from heart disease or not.

Algorithms and Techniques Used :

1. Logistic Regression:

Logistic regression is also a supervised learning classification algorithm that is used to solve both classification and regression problems. In classification problems, the target variable may be in a binary or discrete format either 0 or 1. Logistic regression algorithm works on the sigmoid function, so the categorical variable results as 0 or 1, Yes or No, True or False, etc. It is a predictive analysis algorithm that works on mathematical functions.

Logistic regression uses a sigmoid function or logistic function which is a complex cost function. The sigmoid functions return the value between 0 and 1. If the value less than 0.5 then it is considered as 0 and greater than

0.5 it is considered as 1. Thus to build a model using logistic regression sigmoid function is required.

2. Random Forest

Random Forest classifier is a supervised learning technique in machine learning. It can be used to solve both Classification and Regression problems in machine learning. It is based on the process of combining multiple classifiers to solve a complex problem and to improve the performance of the model, which is known as ensemble learning. Random Forest consists of several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Rather than relying on a single decision tree, the random forest acquires the prediction from each tree, and based on the majority of votes for predictions, it predicts the final output. The higher number of trees in the forest leads to better accuracy and also prevents the problem of over fitting. The final output is taken by using the majority voting classifier for a classification problem while in the case of a regression problem the final output is the mean of all the outputs.

3. Support Vector Machine

Support vector machine (SVM) is a supervised learning algorithm that is used to analyze data. It is used to resolve classification and regression problems. An SVM model is a delineation of the examples as points in space, mapped so that the examples of the discrete categories are divided by a clear gap. The points are separated by a plane which is known as a hyper plane. A set of training data is given to it to mark them as belonging to either one of two categories; an SVM training algorithm then builds a model that assigns new

examples of the same space are mapped and then predicts to which category they belong, making it a non- probabilistic binary linear classifier.

VI. RESULT AND DISCUSSION

Based on the above review, it can be concluded that there is a huge scope for machine learning algorithms in predicting cardiovascular diseases or heart related diseases. Each of the above-mentioned algorithms have performed extremely well in some cases but poorly in some other cases. Random Forest performed extremely well because they solve the problem of overfitting. SVM performed well for most of the cases. Systems based on machine learning algorithms and techniques have been very accurate in predicting the heart related diseases but still there is a lot scope of research to be done on how to handle high dimensional data and overfitting.

After performing the machine learning approach for testing and training we use three algorithms as SVM , Random forest classifier , logistic regression. Accuracy should be calculated of each algorithms .Among these 3 algorithms which one gives better accuracy will be the final algorithm for our heart disease prediction system.

VII. CONCLUSION

This project predicts people with heart disease by extracting the patient medical history that leads to a fatal heart disease from a dataset that includes patients' medical history such as chest pain, sugar level, blood pressure,etc.

Heart is one of the essential and vital organ of human body and prediction about heart diseases is also important concern for the human beings so that the accuracy for algorithm is one of parameter for analysis of performance of algorithms. Accuracy of the algorithms in machine learning depends upon the dataset that used for training and testing purpose.

VIII. REFERENCES

[1] Ujma Ansari, Jyoti Soni, Dipesh Sharma, Sunita Soni. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", 258493784 Predictive Data Mining for Medical Diagnosis An Overview of Heart Disease Prediction. March 2011Data Mining in Healthcare for Heart Diseases.

[2] C. Beyene, P. Kamat, "Survey on Prediction and Analysis the Occurrence Of Heart Disease Using Data Mining Techniques" https://www.researchgate.net/publication /323277772 Survey on prediction and analysis the occurrence of heart disease using data mining techniques. 118(8):165-173 January 2018

[3] Muhammad Usama Riaz, SHAHID MEHMOOD AWAN, ABDULGHAFFAR KHAN, "PREDICTION OF HEART DISEASE USING ARTIFICIAL NEURAL NETWORK",https://www.researchgate.net/publication 1328630348 PREDICTION_OF_HEART DISEASE_USING_ARTIFICIAL_N EURAL NETWORK. October 2018

[4] Komal Kumar Napa, G.Sarika Sindhu, D.Krishna, Prashanthi, A.Shaeen Sulthana, "Analysis and Prediction of Cardio Vascular Disease using Machine Classifiers" https://www.researchgate.net/publication 1340885231 Analysis and Prediction of Cardio Vascular Disease using machine Learning Classifiers, April 2020.

[5] Prediction of Heart Disease Using Machine Learning:- Aditi Gavhane, Gouthami Kokkula, Isha Pandya, Prof. Kailas Devadkar (PhD) - 2018.

[6] Effective heart disease prediction using hybrid learning techniques:-Senthil Kumar mohan, chandrasegar thirumalai, and Gautam Srivastava (19 JUNE 2019)

[7] Prediction of Heart Disease at early stage using Data Mining and Big Data Analytics: A Survey: Salma Banu N.K. Suma Swamy.-2019.

[8] Prediction of Heart Disease Using Machine Learning Algorithms. Mr.Santhana Krishnan.J,Dr.Geetha.S- 2019.

[9] https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset