IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, December 2022
Impact Factor: 6.252

Heart Disease Prediction Using Machine Learning Algorithms A Survey Paper

Dhiraj Chatur, Hanmant Karhale, Neel Lihitkar, Ranjit Kale
UG Students, Department of Computer Science
G. H. Raisoni Institute of Engineering and Technology, Pune, Maharashtra, India

Abstract: *Increasing Heart diseases in urban areas is major concern. It is not possible for common man to frequently undergo test like ECG and so there is need of system which can predict heart diseases which is reliable and budget friendly. Data mining techniques can be used to identify whether a patient is normal or having heart disease. We can predict the vulnerability on the basic symptoms like age, sex, pulse rate, etc. Machine learning algorithms can be used to precisely predict heart diseases. This paper presents a survey of various Machine learning algorithms like Naive Bayes, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine.*

Keywords: Heart Disease, Naive Bayes, Machine Learning, Support Vector Machine, Decision Tree

I. INTRODUCTION

Heart is one of the vital organ of human body, so the care of heart is essential. It pumps blood to every part of human body. If it fails to function properly then brain and other organs will stop working, within a few minutes person will die. Normal person heart beat is 72 per minute and normal blood pressure is less than 120/80mmHg. In humans, the heart is approximately the size of closed fist.

According to World Health Organization 31% deaths are caused due to heart diseases. Cardiovascular diseases(CVD) is a class that involves heart or blood vessels. CVD includes heart failure, stroke, abnormal heart rhythm, valvular diseases, venous thrombosis, carditis, etc. These diseases can be caused due to smoking, high blood pressure, pure diet, lack of exercise, obesity, high blood cholesterol, diabetes milletus, etc.

Various machine learning algorithms under various conditions can be used to predict heart diseases. Algorithms such as Naive Bayes, Logistic Regression, Random Forest, Support Vector Machine. Naive Bayes is a classification technique based on Bayes' theorem with an assumption of independence among predictor. Logistic Regression computes the probability of even occurrence. Random Forest is a classifier that contains a number of decision tree on various subsets of the given data sets and takes the average to improve the predictive accuracy of the that data set. Support Vector Machine is one of the most popular Supervised Learning Algorithm which is used for classification as well as Regression problems.

II. RELATED WORK

Archana Singh and Rakesh Kumar used methodologies like Data Collection, Attribute Selection, Preprocessing of data, Data Balancing and Histogram of Attribute. They used algorithms like Linear Regression, Decision Tree, Support Vector Machine. Logistic Regression gives the accuracy of 87.1%, Support Vector Machine gives the accuracy of 85.71%, Ada-boot Classifier gives the accuracy up to 98.57% which is good in prediction point of view. He described attributes of data set like Age, Sex , Cp(Chest Pain), Trest BPS, Cholesterol, FBS(Fasting Blood Sugar), Resting.

Sayali Ambekar and Rashmi Palnikar, for the heart disease prediction they used following algorithms like Naive Bayes, KNN Algorithm, CNN-UDRP Algorithm. They successfully derived that time required for execution of two algorithms Naive Bayes and KNN. Naives Bayes takes 30ms, 90ms, 20ms, 120ms on different training data set 80%, 70%, 60%, 50% respectively. KNN takes 60ms, 175ms, 55ms, 208ms on training data set 80%, 70%, 60%, 50%. According to these Naives Bayes takes less time than KNN. They performed heart diseases prediction using Naive Bayes algorithm

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, December 2022
Impact Factor: 6.252

and KNN algorithm. They compared the results between KNN and Naive Bayes algorithm and the acuuracy of Naive Bayes is 82% and KNN algorithm is 65%. According to them Naive Bayes algorithm accuracy is more than KNN

Senthil Kumar Mohan, Chndrasegar Thirumalai and Gautam Srivastava used classification models like Decision Tree, Language Model, Support Vector Machine and Random Forest . They have used an R-Studio rattle to perform heart disease classification of the cleavland UCI repository. R-Studio rattle provides an easy to use visual representation of the data set working environment and building the predictive analytics. They used attributes like Age, Sex, Cp, Trestbps, Chol, FBS, Resting, Thali(Accomplishment of the maximum rate of heart), Exaing, Old peak, Slope, Ca, Thal, Num, etc.According to them HRFLM (proposed) accuracy is 88.4% which is highest accuracy than other models.

Jaymam Patel and Sameer Patel performed Decision Tree Classification using J48 algorithm and Logistic Model Tree algorithm , Random Forest Algorithm on UCI repository. They found that j48 algorithm has train error of 0.1423221 and test error of 0.1666667. Logistic Model tree on UCI repository they found train error 0.1156716 and test error of 0.137931.Random Forest Tree on UCI Repository they found train error 0 and test error 0.2. On comparing these methodology they found that J48 has best overall performance. When j48 was used on UCI data has the highest accuracy and total time to build model is 0.04 second while LMT has lowest accuracy and time to build model was is 0.39 second. They concluded that combination and more complex models to increase the accuracy of predicting heart diseases.

Sanathan Krishanan.J and Geetha.s assistant professor used different algorithms like Decision Tree, Naive Bayes to predict heart diseases. They found that on training dataset 70%instance of dataset displayed possibility of having heart disease and on testing dataset shows 30%instance possibility of having heart disease. Naive Bayes on train dataset shows 70% instance of possibility of having heart diseases and on test dataset 30% instance of data set shows possibility of having heart disease. Accuracy of Decision Tree was 91% and Naive Bayes was 87%. They concluded that Decision Tree Classification Algorithm is better for handling medical dataset.

III. ALGORITHM AND TECHNIQUE

Machine Learning is one of efficient technology which on based two terms namely testing and training. There are are three types of machine learning algorithm Supervised Unsupervised and Reinforced. Supervised algorithm are type of machine learning algorithm which machines are trained using well “labeled” training data, and on bases of that data, machine predict the output. There are two types of supervised algorithm Regression and Classification.

Regression algorithm are used if there is a relation ship between the input variable and output variable. Types of regression algorithm which come under supervised algorithm are :

- Leaner regression
- Regression trees
- Non leaner regression
- Bayesian leaner regression
- Polynomial regression

Classification algorthem are used when output variable is categorical that means there are two classes like Yes-No, Male-Female, True -False. Types of classification algorithm which comes under supervised algorithm are:

- Random Forest
- Support Vector Machine
- Decision tree
- Naive Bayes
- Logistic Regression

Unsupervised Algorithm are those algorithm where machine are trained with unlabeled data. Types of unsupervised algorithm are Clustering and Association.

Reinforced Algorithm are those algorithm where an intelligent agent interacts with the environment and learns to act within that.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, December 2022
Impact Factor: 6.252

Algorithm which we are going to discuses are Decision tree, Naive Bayes, Logistic Regression, Random Forest, Support Vector Machine.

3.1 Decision Tree

Decision tree is the type of supervised machine learning algorithm where there is graphical representation of data. It start with root node which expands on further branches and constructs a tree like structure, it represent enter data set. Leaf nodes gives output where as decision node are used to make decision. Branch tree is formed by spiting the main tree. Entropy is matric to measure the impurity of the given attribute. Entropy can be calculated by :

Entropy(s)= -P(yes)log2 P(yes) - P(no)

3.2 Naive Bayes

Nave Bayes algorithm helps us in building fast machine learning model. These algorithm assumes that occurrence of certain feature is independent of the occurrence of other feature. These algorithm is based on Bayes algorithm.
It is used to determine the probability of a hypotheses with prior knowledge. It is also known as Bayes Rule or Bayes Law. It is use in medical data classification.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

3.3 Logistic Regression

Logistic Regression predicts the output of categorically dependent variable. It has ability to provide probability and classify new data using continues and discrete data set. It uses the concept of predictive modeling as regression.
Logistic regression is commonly used in Natural Language Processing (NLP) . Types of logistic Regression :

- Binomial : There can be two possible types of dependent variables such as 0 or 1 , true or false.
- Multinomial : There can be three or more possible unordered dependent variable.
- Ordinal : There can be three or more types of order dependent variable.

3.4 Random Forest

Random forest combines multiple trees to predict the class of data set. It is based on the concept of ensemble learning. It enhance the accuracy of the model and prevents the over fitting issues. Random Forest randomly selected observations, builds a decision tree and then the result is obtained based on the majority voting. No Formulas are required here. Random forest is mostly used in Banking Sector, Medicine, Marketing and Land use.There are many advantages of Random forest :

- It takes less training time as compared to other algorithm.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

3.5 Support Vector Machine

It works on the concept of Hyper plane means it classify the data by creating hyper plane between them. A simple SVM works by making a straight line between two classes. That means all the data points on one side of line will represent a category and the data points on the other side of the line will be put into different category. SVM Algorithm better because it chooses the best line to classify your data points. SVM are used in application like Handwriting recognition, Face Detection, Email Classification, Gene Classification and in Web Pages. There are advantages of SVM like :

- It is Effective on Datasets with multiple features like financial or medical data.
- It is Effective in cases where no of features is greeter than no of data points.
- It uses Subset of training points in the decision function called Support Vectors which makes it memory efficient.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, December 2022
Impact Factor: 6.252

IV. COMPARISON BETWEEN MAJOR MACHINE LEARNING ALGORITHMS

Algorithm	Accuracy	Classification Error	Precision
Decision Tree	85	15.0	86
Naive Bayes	75.8	24.2	90.5
Logistic Regression	82.9	17.1	89.6
Random Forest	86.7	13.9	87.1
SVM	86.1	13.9	86.1

V. CONCLUSION

Heart disease is crucial problem in human society. Prediction about heart diseases is also important concern so the accuracy of algorithm is one of the parameter for analysis . Accuracy depends on dataset that is used for training and testing purpose. There is huge scope for machine learning algorithm in predicting heart disease. When naive byes and Decision tree were applied on same dataset Decision tree accuracy is 91% were naive bayes has 87%. Decision tree algorithm is better for handling medical dataset . All algorithm mentioned have perform extremely well in some cases but poorly in some other cases. Decision tree with PCA performed well but Decision tree performed extremely poor in some other cases due to over fitting. Random forest in ensembles model they performed well because the solve overfittting problem by implying multiple algorithm. Models like naive bayes were computationally very fast and have also performed well. SVM performed extremly well for most of the cases. A lot of research is required to handle high dimensional data and over fitting problem.

REFERENCES

[1]. Archana Singh, Rakesh kumar. "Heart Disease Prediction using Machine Learning Algorithm" 2020. International Conference on Electrical and ELectronics Engineering (ICE3-2020).

[2]. Mr Santhana Krishnan,Dr Geetha. s. "Prediction of Heart disease using Machine Learning Algorithm ".

[3]. Himanshu Sharma, M A Rizvi. "Prediction of Heart Disease using Machine Learning Algorithm. International Journal on Recent and Innovation trends in computing and communicate".

[4]. Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava."Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques".Received May 13,2019,accepted june9,2019,date of publication June 19, 2019,date of current version July 3, 2019.

[5]. Prof. Tejal Upadhyay, Dr. Samir Patel, Jaymin Patel."Heart Disease Predictio using Machine Learning and Mining technique."VOlume 7 Sept 2015 - March 2016.

[6]. D. Tain, J. Zhou, Y. Wang, Y. Lu, H. Xia, and Z. Yi, "A dyanamic and self -adaptive network selection method for multimode communication in hetrogenous vahicular telematics," IEEE Transaction on Intelligent Transportation Systems, vol. 16, no. 6, pp. 3033-3049, 2015.

[7]. M. Chen, Y. Ma, Y. Li, D.Wu, Y. Zhang, C. Youn, "Wearable 2.0:En-ableHuman-Cloud Integration in Next Generation Healthcare System, "IEEE Communication, Vol, 55, No. 1, pp. 54-61, jan. 2017.

[8]. M. Chen , Y. Ma , J. Song, C. Lai, B. HU, "Smart Clothing: Connection Human With Clouds and Big Data for Systainable Health Monitoring , "ACM/SpringerMObile Networks and Appliction , Vol. 21, No. 5, pp.825c845,2016

[9]. M. Chen, P. Zhou, G. Fortino, "Emotion Communication System,"IEEE Access , DOI:10.1109/ACCESS.2016.2641480,2016.

[10]. J. Wang, M. Qui, and B. Guo, "Enabling real-time information service on telehealth system over cloud-based big data platform", Journal of System Architecture, vol. 72, pp. 69-79, 2017.

[11]. Y. Zhang, M. Qui, C.-W. Tsai, M.M.Hassan, and A. Alamri, "Healthcps:Healthcare cyber-physical system assistance by cloud and big data ",IEEE Sys-tems Journal, 2015.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, December 2022
Impact Factor: 6.252

[12]. K. Lin, J. Luo, L. Hu, M.S. Hossain, and A. Ghoneim , "Localization based on social big data analysis in the vahicular networks,"IEEE Transaction on Industriallnformatics,2016.

[13]. K Lin, M. Chen, J. Deng, M. M. Hassan, and G.fortino, "Enhanced finger- printing and trajectory prediction for iot localization in smart buildings ," IEEE Transaction on Automation Science and Engineering, vol. 13, pp. 1294-1307,2016.

[14]. B. Quin. X, Wang, N. Cao, H. Li, and Y. G.Jaing, "A relatively similarity Basedmethod for intractive patient risk prediction , " data miming and knowladge Discovery ,vol. 29, no.4, pp. 1070-1093, 2015.

[15]. S. Bandyopadhayay, J. Wwolfson, D.M. Vock, G. Vazquez-Benitez, G.Ado-mavicius, M. Elidrisi,P. E.Johnson, and P. J. O'Connor,"Data mining for censored time -to-event data: a bayesian network model for predicting cardio-vascular risk from electronic health record data,"Data mining and Knowladge Discovery , vol.29,no.4,pp.1033-1069,2015

[16]. J. Wan, S. Tang, D. Li, S. Wang, C. Liu, H. Abbas and A. Vasilakos, "A manu-facturing Big Data Solution for Active Preventive Maintanance ", IEEE Trans-action on industrial Informatics, DOI: 10.1109/TII.2017.2670505,2017.