International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-10, August 2020

Review of Techniques used to find the Possibility of Getting Heart Related Disease.

Amandeep Singh, Amit Chhabra

Abstract: ***As with the changing lifestyle and people consuming high calorie diet increases the heart disease rate among the humans. Over the last decade heart related diseases are one of the leading cause of death cases every year. It is very hard to notice the symptoms of any heart related disease at early stage and in many cases it leads to sudden death before ever knowing the first symptom of any heart related problem. With the advancement of technology there are many devices which are used to perform several tests in the medical field and with the emerging trend of Machine learning doctors can be aided to find symptoms of heart disease . There is huge amount of patients health data collected by healthcare institutes which can be used for data mining and infer relationship between data and helps in predicting heart diseases. The machine learning models trained on patients record data which shows symptoms is used to predict the probability for having a heart disease.***

Keywords: Machine learning, supervised learning, unsupervised learning, heart disease.

I.INTRODUCTION

Heart disease is a very common disease and leading reason of death for the past few years. The cases of heart disease are increasing at an alarming rate due to changing lifestyle and work routine which include very minimal physical work and consumption of very high calorie diet. According to one study done by health related research institute under the government authority of USA , 1 in 4 death cases reported in the United state is due to heart disease. Diagnosing heart disease is very difficult because term heart disease is very broad and it describes many conditions that affects a persons heart health. Like heart failure , heart attack , stroke , pulmonary embolism , cardiac arrest are some of the heart problems to which we generally refer as a heart disease.There is huge amount of patients health data collected by healthcare institutes which can be used for data mining and infer relationship between data and helps in predicting the chances of a person getting heart diseases. There are many Machine Learning algorithms which can be trained to predict heart disease. Based on the training data available to us whether we have a labelled data or unlabelled data we can choose from Supervised or Unsupervised machine learning models [9] . In Supervised machine learning algorithms we provide the algorithm with the training data and the expected output and models learns the mapping between the training data and expected output.

Revised Manuscript Received on August 30, 2020.
* Correspondence Author
Amandeep Singh, Department of Computer Science and Engineering, GNDU, Amritsar, Punjab, India.
Amit Chhabra, Assistant Professor, Department of Computer Science and Engineering, GNDU, Amritsar, Punjab, India.

When we test the performance of the Supervised machine Learning model we do not give the expected output instead model predict the expected output and we check that output against the ground truth we have. Linear regression , Decision Trees , Naive Bayes , Support vector machine are few of the examples of Supervised Machine Learning algorithms. Unlike Supervised Machine Learning algorithms Unsupervised Machine Learning algorithms learn the mapping only from training data , there is no output data from which model can learn the mapping so therefore model makes the clusters from the training data only and learns the relationship between data. Clustering , association are the examples of Unsupervised Machine learning. To predict heart disease we can use both Supervised and Unsupervised Machine learning algorithms.

II.RELATED WORK

There have been many studies done which focuses on diagnosis of heart disease. In this section we discuss the existing literature related to heart disease prediction. Carlos Ordonez (2017) [1] implements the Association rules to predict heart diseases. Generally the problem with association rules is when they are applied to large dataset like medical dataset they output very large number of rules and many of these rules are not practical and its not feasible to find these impractical rules. Carlos used search constraints and validate the rules against the test dataset which gives the set of rules with high prediction accuracy.
Shadab A. Pattekari et al. (2012) [2] they developed an web application where user inputs the symptoms or health data and their application tell user the chances of getting a heart disease. They trained the Naive Bayes classifier on the patient dataset they acquired from hospitals and based on this training dataset their algorithm make predictions. M. Akhil Jabbar et al. (2013) [3] propose a lazy associative classification for predicting the possibility for a person to get heart related disease. Associative classification approach is based on some rules which when applied to medical data give us the set of rules which can be used to make predictions. They tested their associative rule based approach on 7 test data sets and 1 real life dataset. Their algorithm showed the improvement of 5%-10% against some existing algorithms like J4.8 and naive Bayes.
Krishnan. J et al. (2019) [4] compares the two machine learning algorithms and analysed their performances for predicting whether a person may have heart related disease or not. Krishnan et al. used python programming language to implement decision tree algorithm and Naive Bayes algorithm , both these algorithms are supervised machine learning algorithms and applied to same dataset to compare the results .

The decision tree algorithm gives the prediction accuracy of 91 % and Naive Bayes algorithm gives the results with 87% accuracy. Ramesh et al. (2011) [5] developed a Decision Support in Heart Disease Prediction System (DSHDPS).
According to G. Subbalakshmi et al. healthcare industry has large volume of data but unfortunately they can't discover the meaningful relationship between this data. So they using Naive Bayes algorithm made decision support system which learns the meaningful relationship between data and make prediction for heart diseases.
Latha P. and R.Subramanian (2008) [6] presented coactive neuro-fuzzy inference system (CANFIS) to predict diseases related to heart. They combine their CANFIS model with neural network , fuzzy logic qualitative and genetic algorithm. They used the dataset of heart disease from UCI and distribute the training and testing dataset into 5 classes (No disease related to heart , Type 1 heart disease , Type 2 heart disease , Type 3 heart disease and Type 4 heart disease) . They run the simulations by using the NeuroSolution software. After training and testing the model they get the mean square error of 0.000842.
Chaitrali S. and Sulabha S. (2012) [7] introduced two new attributes to the existing 13 attributes for training the algorithm which predicts heart disease. They used dataset collected from Cleveland Heart Disease foundation which have 303 records. Existing 13 attributes in training dataset are age , sex , cp (chest pain) , resting blood pressure , serum cholesterol , resting elctrographic results , fasting blood sugar , maximum heart rate achieved , exercise induced agina , ST depression induced by exercise relative to rest , slope of the peak exercise ST segment , number of major vessels colored by floursopy and defect type , Apart from these 13 attribute they introduce two more attributes which are obesity and smoking . Three classification models they used are Neural networks , Decision Trees and Naive Bayes . According to their results Neural network gives 100 % prediction accuracy , Decision Trees give 99.62% accuracy and Naive Bayes give 90.74 % accuracy.
Sellappan P., Rafiah A. (2008) [8] developed Intelligent Heart Disease Prediction System (IHDPS). They implemented data mining techniques to predict the possibility for a person to get heart related disease. They used three classification algorithms namely Decision trees , Naive Bayes and Neural networks. The better part of IHDPS implemented by Sellappan P. and Rafiah A. is that their model is able to answer complicated "what if" questions and traditional systems are failed to do so. They developed the user friendly web interface implemented in Microsoft .Net platform where user can provide all the information and their system shows the likelihood of getting a heart disease or not. Jyoti S. et al. (2011) [10] did a survey of various existing data mining techniques and compare their performances . They have conducted number of experiments with same dataset for all the techniques. The algorithms they compare are Decision Trees , KNN and Neural Networks . According to their results Decision tree outperform all the other algorithms and KNN performs the worst.

III.GAPS IN LITERATURE

From the extensive literature , following gaps have been formulated:
A) The most of the existing researchers focussed on training the models on very small subset of dataset which is available online. Large and quality dataset can be collected if requested from good healthcare institutes.
B) Most of the existing research which do the comparison or survey of existing algorithms only compares two or utmost three algorithms but there are many algorithms which are implemented independently can be compared.
C) The existing research which compare the performance of models use same and very small dataset which cause the problem of overfitting and the trained models are not very generalised.

IV.METHODOLGY

This paper is about comparing the performance of different machine learning models used for predicting heart disease. Most of the research papers only compare the performance of two or three models but we will compare more models trained on same dataset.The image below shows the flowchart of the methodology.

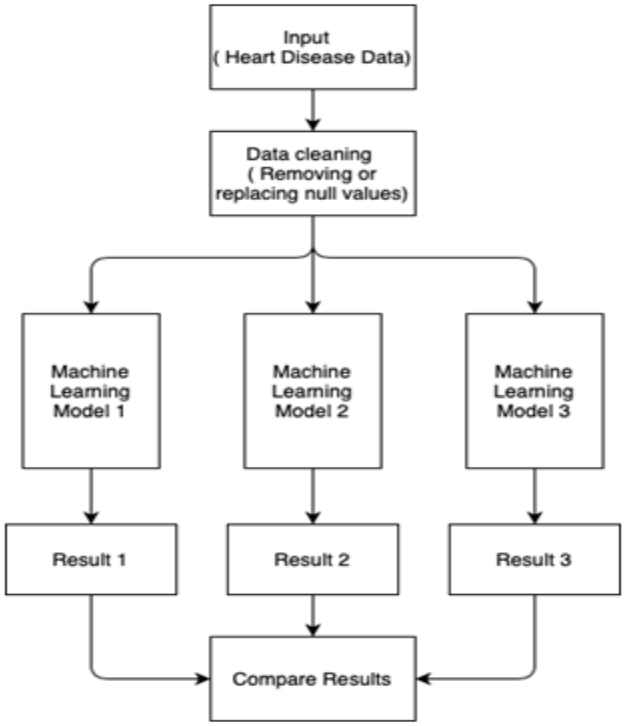


Figure 1. Flowchart of Methodology

Input Dataset: We will use heart disease dataset from UCI Machine learning repository. The data includes the patient record from Cleveland Clinic Foundation. The dataset have 14 attributes and have numeric values. The attributes of dataset are Age , Gender , Constrictive pericarditis , Resting blood pressure (trestbps) , cholesterol , Fasting blood sugar (Fbs) , rest ecg , the maximum heart rate achieved (thalach) , Exercise Induced Angina (EXANG) , old peak , slope , calcium , Thalassemia , prediction (whethere a person have heart disease or not). We split the input data into two parts one for training and other for testing. The train data is used while we train the algorithm and we test the performance on test data. 80% of the total data is used while training the algorithm and 20% of the total data is used while testing.

Data Cleaning : The input data we feed to our Machine learning algorithms must be error free and should not have any null values. The rows which contain null values for many attributes are dropped and if the row only have one or two null value that is replaced by dummy value.
Following Machine learning algorithms are trained on the input data to compare the performance.
K Nearest Neighbors Classifier: The KNN algorithm comes under supervised machine learning algorithm which is used for classification problems.

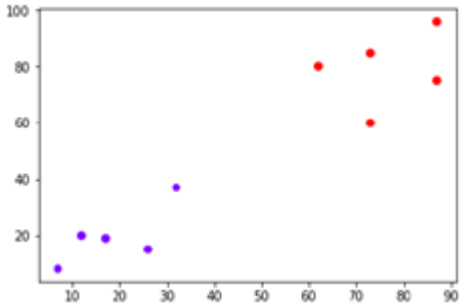


Figure 2 : Above image shows the similar datapoint exist in close proximity

The basic idea behind KNN algorithm is that similar data points are close to each other. K means the number of training sample that are needed to classify. General approach to select K is K = n^(1/2).
Support Vector Classifier : This algorithm forms a hyperplane which divides the two classes. In our case two classes are person have a possibility or does not have a possibility of getting heart related disease. The algorithm is fine tuned by tweaking the hyper parameters which changes the distance between the data points and the hyperplane.

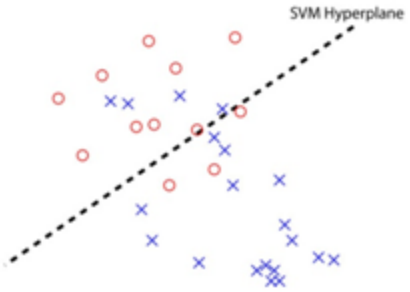
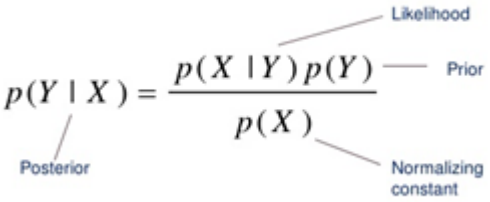


Figure 3 : Above image shows the hyperplane between two classes.

Decision Tree Classifier : Decision Tree classifier creates a decision tree and assign the class value to each data point based on the feature. In our case class values are 0 (person have no possibility of getting heart related disease) and 1 (person have possibility of getting heart related disease) is assigned based on the 14 features in our dataset.
Random Forest Classifier: The Random Forest algorithm is based on many individual decision trees. As in single decision tree we keep on splitting the data based on some condition and each path we get after split is our node. In random forest algorithm there are large number of decision trees which make predictions and our final result is the prediction made by most of the decision trees.
Naive Bayes Classifier: Naive Bayes classifier is very fast and easy to implement. The image below shows the Bayes theorem based on which Naive Bayes works :



$$p(Y \mid X) = \frac{p(X \mid Y)\, p(Y)}{p(X)}$$

Figure 4 : Bayes Theorem

In our case variable y is the class variable which is whether a person have heart disease or not.
X = (x1 , x2 , …… x14) where x1 , x2 … are the 14 features in our training dataset.

V. CONCLUSION AND FUTURE DIRECTIONS

There are many data mining techniques which can be used to find out the possibility for a person to get a heart related disease. Many researchers run different models and compare the results. But most of the researchers used same dataset and also which is very small in size which cause the problem of overfitting. Also the comparison done by many researchers is between same two or three techniques whereas there are many algorithms which can be compared and performance benchmark can be made. In this work we have analysed the existing heart disease prediction techniques and our proposed methodology is to compare the results of more algorithms for predicting the heart related disease.

REFERENCES

1. Carlos Ordonez. “Association Rule Discovery With the Train and Test
2. Approach for Heart Disease Prediction” IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 10, NO. 2, 2006
3. Shadab A. Pattekari, Asma A. Parveen, Enginering Khaja, Banda Nawaz. “PREDICTION SYSTEM FOR HEART DISEASE USING NAIVE BAYES” Semantic scholar 2012
4. Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. “Heart disease prediction using lazy associative classification.” International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s) 2013
5. Krishnan. J, S., & S, G. . “Prediction of Heart Disease Using Machine Learning Algorithms.”1st International Conference on Innovations in Information and Communication Technology (ICIICT). 2019
6. G. Subbalakshmi, Montana Tech, Edgar Dist, Ramesh M. Tech. “Decision Support in Heart Disease Prediction System using Naive Bayes.” Semantic scholar 2011
7. Latha Parthiban and R.Subramanian. “Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm.” International Journal of Biological and Medical Sciences 3:3 2008
8. Chaitrali S. Dangare , Sulabha S. Apte . “Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques.” International Journal of Computer Applications (0975 – 888) 2012
9. Palaniappan, S., & Awang, R. “Intelligent heart disease prediction system using data mining techniques.” 2008 IEEE/ACS International Conference on Computer Systems and Applications. doi:10.1109/aiccsa.2008.4493524 2008
10. Alloghani, Mohamed & Al-Jumeily, Dhiya & Mustafina, Jamila & Hussain, Abir & Aljaaf, Ahmed. “A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science.” 10.1007/978-3-030-22475-2_1. 2020

11. Jyoti Soni , Ujma Ansari , Dipesh Sharma. “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction.” International Journal of Computer Applications (0975 – 8887) Volume 17– No.8. 2011

AUTHORS PROFILE

Amandeep Singh (B.Tech in Computer Science and Engineering from GNDU main campus Amritsar and M.Tech in CSE from GNDU, Amritsar)

Amit Chhabra (Assistant Professor at GNDU, Amritsar. Areas of specialisation in Parallel and Distributed Computing)