IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-2, Feb 2019 E-ISSN: 2347-2693

Predictive Analysis on Heart Disease Using Different Machine Learning Techniques

Niraj Kalantri[1*], Kumar R[2]

[1,2]School of Computer Science and Engineering, VIT University, Chennai, India

[*]*Corresponding Author: nirajkalantri99@gmail.com, Mob.: +91-74052-73327*

DOI: https://doi.org/10.26438/ijcse/v7i2.97101 | Available online at: www.ijcseonline.org

Accepted: 22/Feb//2019, Published: 28/Feb/2019

***Abstract*—** Heart Disease is the one of the major cause of death especially in developed countries. Some of its types include Arrhythmia, Stroke, High Blood pressure, Cardiac Arrest etc. Thus to help clinicians for early diagnose disease related conditions, some medical decision support system are also designed. Data mining plays an essential role in analyzing huge amount of data. These quick predicting techniques helps medical practitioners to analyze the same. Classification is the most common Machine Learning algorithm used to classify the disease/non-disease patient. In this paper we will analyze and predict the occurrence of heart disease by applying some of the machine learning algorithms like K-Nearest Neighbor, Decision Trees, Random Forest, Adaptive boosting, SVM and Logistic Regression. It will help physicians to estimate the risk in different age groups. The dataset used is taken from Heart Disease database of UCI Machine Learning Datasets. Factors like blood pressure, heart rate, sugar level, cholesterol, age, gender etc. highly affects the result of the algorithm. The accuracy has been improved by working on high-contributing attributes found using feature importance technique.

***Keywords*—** Heart Disease, Predictive Analysis, Data Mining, SVM, Classification, Decision Tree

I. INTRODUCTION

Heart Disease remains the primary reason of death around the world. According to the data provided by WHO, one-third or the world's population died from heart disease. It is even worse in developing countries where lifestyle is very poor. Most people are aware of the fact that this type of disease is sometimes occurs due to a family history. High Blood pressure increases the risk of heart disease because of stretching of walls of blood vessels. Obesity and smoking are also the factor which turns out to be a major affecting parameter for the result. Sometimes a person might not even know about multiple disease due to poor health and lifestyle. So it is difficult for a physician to find out the real reason and factors for the disease. There is an urgent need for suggesting a suitable approach to control the rate of mortality due to heart disease. Precise prediction at an initial stage and reduce the death rate due to a specified disease. In the past years, data mining have played a pivotal role in heart disease research. The major reason for choosing data mining and analyzing over the data is the wide availability of the data and turning the same data into some useful piece of information. It will be helpful for finding interesting patterns and insights about the dataset. There will be plenty of factors affecting our human health in different ways but this paper mainly focuses on predicting the heart disease and helps for reducing the efforts of doctors since the main challenge for any healthcare organization is quality service at affordable costs.

The paper follows the following structure. Related works have been described which shows the previous works done by various authors and their contribution towards similar topic. Then the work methodology for this project is briefly explained. It includes various sections about the dataset used, workflow, and models used. It is followed by results where the performance outputs from various algorithms are discussed. The final section focuses upon the conclusion of the work done so far and about the work to be done in this project as a part of continuation of this project.

II. RELATED WORK

Various research works have been done for the proposal of prediction of heart disease and different algorithms have been conducted on each of them. Palaniappan and Rafiah [1] have developed a prototype called Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques. The techniques used are Decision Trees, Neural networks and Naïve Bayes. Naïve Bayes appears to be the most effective among all three. Results shows that in both models,

the most significant factor influencing heart disease is "Chest Pain Type". Other significant factors include Thal, CA and Exang. Decision Trees model shows 'Trest Blood Pressure" as the weakest factor while Naïve Bayes model shows 'Fasting Blood Sugar' as the weakest factor. Naïve Bayes appears to fare better than Decision Trees as it shows the significance of all input attributes. Doctors can use this information to further analyse the strengths and weaknesses of the medical attributes associated with heart disease.

Sultana, Afrin and ShorifUddin [3] have focus on approaches like such as K-Star, J48, SMO, Bayes Net and Multilayer Perceptron. The research is conducted on Weka software as training and then testing.

Theresa and Thomas [4] have shown a survey about different classification techniques used for predicting the health of a person based on different parameters. The patient risk level is classified using data mining classification techniques such as Naïve Bayes, KNN, Decision Tree Algorithm, and Neural Network. Etc.

Radhimeenakshi [6] have given a comparative study about Artificial Neural Network (ANNs) and Support Vector Machine (SVM). The study shows that precision and accuracy of SVM is beter then ANN. But in case of Specificity ANN has pretty good results compared to the other one. Sana Bharti [7] have shown the analytical study of different machine learning algorithms and proposed a comparative study among them. The author have used Neural Network, Genetic algorithms and Particle Swarm optimization to conduct the study.

III. METHODOLOGY

This paper mainly focuses on study of different machine learning algorithms on heart disease dataset and improving the accuracy with only selected attributes. It has been noticed that not all the attributes are equally important for the output (dependent) variable.

3.1. Introduction

Related works have shown the study of algorithms on whole dataset. However, if certain attributes are removed it gives greater accuracy compared to the scenario mentioned above. So in this system we are going to demonstrate the attributes and algorithms which provides decent results compared to other previous works.

3.2. Datasets

The total of more than 580 instances with fourteen different attributes were collected from the Cleveland Heart Disease database and Hungarian Database from the UCI Machine Learning Repository [10]. The attribute "diagnosis" described as the measurable field with value 0 mean person with no heart diseases and 1 mean person with heart diseases. Table 1 shows the attributes, description, values of heart disease dataset.

Predictable attribute

1. Diagnosis (value 0: < 50% diameter narrowing (no heart disease); value 1: > 50% diameter narrowing (has heart disease))

Key attribute

1. PatientID – Patient's identification number

Input attributes

1. Sex (value 1: Male; value 0 : Female)
2. Chest Pain Type (value 1: typical type 1 angina, value 2: typical type angina, value 3: non-angina pain; value 4: asymptomatic)
3. Fasting Blood Sugar (value 1: > 120 mg/dl; value 0: < 120 mg/dl)
4. Restecg – resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
5. Exang – exercise induced angina (value 1: yes; value 0: no)
6. Slope – the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: downsloping)
7. CA – number of major vessels colored by floursopy (value 0 – 3)
8. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
9. Trest Blood Pressure (mm Hg on admission to the hospital)
10. Serum Cholesterol (mg/dl)
11. Thalach – maximum heart rate achieved
12. Oldpeak – ST depression induced by exercise relative to rest
13. Age in Year

Fig.1 Dataset Attributes

3.3 Work Flow

3.3.1 Selection

There is need to select an appropriate data set for prediction of heart diseases for data analysis and to get effective knowledge from that. The sufficient quantity of data is required to perform data techniques on selected heart diseases dataset.

3.3.2 Pre-processing and Transformation

The dataset is prepared in CSV (Comma Separated Values) file format standard of heart dataset. The data is already Pre-processed with '?" values in place of missing values. Since the data is very sensitive, data imputation will not be very useful. Unnecessary fields where empty values are in abundance are removed. Since the final dataset contains databases of two different regions, normalization has been done on the data as a part of transformation.

3.3.3 Training and Testing

Classification of data can be done once the data is trained. So here only 20% of the data is kept for testing and remaining all is sent for training on data. Cross-validation and K-fold are the methods available for achieving better accuracy for the data. When the data is trained we can apply classifiers on testing data in order to get the results.

3.3.4 Applying Appropriate Classification Algorithms
The classification algorithms like Decision Tree, Random Forest, SVM, Logistic regression and k- nearest neighbour are implemented on training dataset and the output of each algorithm is evaluated of basis of corrected classified instances.

IV. RESULTS AND DISCUSSION

After creating models on training dataset and testing using the testing dataset, we get the results showing the corrected classified records for patients and the accuracy of it. Logistic regression shows the highest accuracy of predicting the disease i.e. 87.7 %. It is followed by Random forest (83 %) and then KNN (81%) respectively.

The data in table given below are results which are classified after applying various algorithms. **They are shown by True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN).**

Table 1: DECISION TREE

	0	1
0	48	9
1	15	41

Decision Tree Accuracy: 76 %

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Although random forest is used which is collection of decision trees. Still this technique is used to check the stability of model and determine the efficiency for RF.

TABLE 2: RANDOM FOREST

	0	1
0	51	7
1	12	43

Random Forest Accuracy: 83 %
Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over-fitting to their training set.

TABLE 3: SVM

	0	1
0	52	11
1	11	39

Support Vector Machine Accuracy: 80 %

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well.

TABLE 4: K-NEAREST NEIGHBOUR

	0	1
0	54	12
1	9	38

K-Nearest Neighbor Accuracy: 86 %

In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

TABLE 5: LOGISTIC REGRESSION

	0	1
0	55	6
1	8	44

Logistic Regression Accuracy: 87 %

Logistic regression is a statistical method for analysing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

These were the various algorithms which were applied for our model creation and their results described above. A ROC curve is shown as a summarized result showing the performance of each model in TPR (true positive rate) vs FPR (false positive rate) plot.

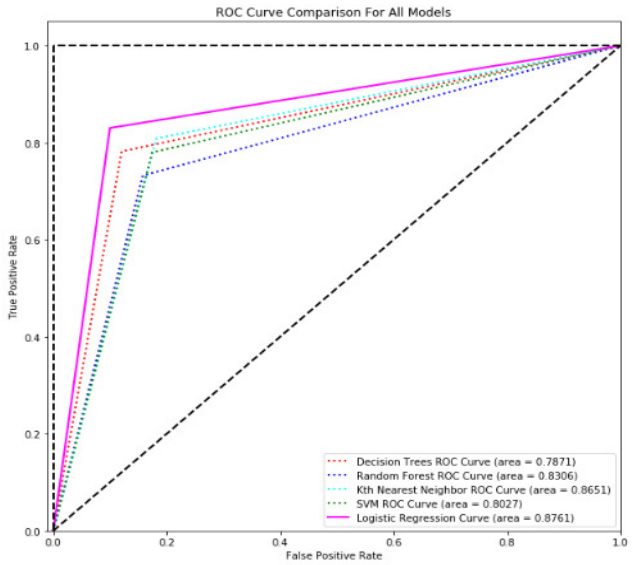

Fig.2. ROC Curve

V. CONCLUSION AND FUTURE SCOPE

Logistic regression gives the maximum accuracy out of all the 5 algorithms. Moreover it has been noticed that the accuracy of same algorithm differs according to the scenarios of training them. Training the dataset with simple train test split is better than training with cross validation. Using K-fold increases the accuracy of the algorithm. The future work will be to increase the accuracy of remaining algorithm using boosting techniques. The work will also focus on integrating different datasets and discovering other useful information.

ACKNOWLEDGMENT

I would like to express my heartfelt gratitude and sincere thanks to my research faculty Dr. Kumar R for his support and guidance throughout the course of my research process. His valuable comments and sharing of knowledge of related work are the added advantage for the research work. Any of the errors in this project are on my own and should not tarnish the reputation of the esteemed person.

REFERENCES

[1] Sellappan Palaniappan, Rafiah Awang, "*Intelligent Heart Disease Prediction System Using Data Mining Techniques*", (IJCSNS), Vol.8 No.8, August 2008.

[2] Amit Kamra, Jagdeep Singh, Harbhag Singh, "*Prediction of Heart Diseases Using Associative Classification*", International Conference on Wireless Networks and Embedded Systems, pp.1-7, 2016.

[3] M. Sultana, A. Haider and M. S. Uddin, "Analysis of data mining techniques for heart disease prediction," *2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, Dhaka, 2016, pp. 1-5

[4] R, J. Thomas, Theresa Princy, "*Human Heart Disease Prediction System using Data Mining Techniques*", In the Proceedings of the 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT) ,pp.1-5

[5] Prof. Dhomse Kanchan B. ,Mr. Mahale Kishor M. *"Analysis of Supervised Machine Learning Algorithms for Heart Disease Prediction with Reduced Number of Attributes using Principal Component Analysis*", International Journal of Computer Applications (0975 – 8887) ,Volume 140 – No.2, April 2016

[6] Mrs.S.Radhimeenakshi, "*Classification and Prediction of Heart Disease Risk Using Data Mining Techniques of Support Vector Machine and Artificial Neural Network*" IEEE-2016.

[7] T. J. Peter and K. Somasundaram, "An empirical study on prediction of heart disease using classification data mining techniques," *IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM -2012)*, Nagapattinam, Tamil Nadu, 2012, pp. 514-518.

[8] Sana Bharti Dr.Shaliendra Narayan Singh, "*Analytical Study of heart Disease Prediction Comparing with different algorithms*", Proceedings published by IEEE Xplore Digital Library 1st International Conference on Computing, Communication and Automation (ICCA-2015), Greater Noida, India,pp-1-12.

[9] Monika Gandhi, Dr. Shailendra Narayan Singh "*Predictions in Heart Disease Using Techniques of Data Mining*" 2015 IEEE.

[10] AH Chen,SY Huang,PS Hong,CH Cheng,EJ Lin, "*HDPS:Heart Disease Prediction System*",Computing in Cardiology 2011, Department of Medical Informatics,Tzu Chi University,Hualien City,Taiwan.

[11] Wu R, Peters W, Morgan MW. The next generation clinical decision support: linking evidence to best practice. J Healthc Inf Manag, 2002;16:50-5

[12] Ho, T. J.: "Data Mining and Data Warehousing", Prentice Hall, 2005.

[13] Giudici, P.: "Applied Data Mining: Statistical Methods for Business and Industry", New York: John Wiley, 2003.

[14] M.C. Tu and D. Shin, "Effective diagnosis of heart disease through bagging approach," *IEEE 2nd Int. Conf. on Biomedical Engineering and Informatics '09,* Tianjin, China, 2009, pp. 1-4.

[15] W. Li, J. Han, J. Pei, "CMAR: Accurate and efficient classification based on multiple class association rules," *in Proc. of IEEE Int. Conf. on Data Mining*, Washington, DC, USA, 2001, pp. 369–376.

[16] Beant Kaur h, Williamjeet Singh, "Review on Heart Disease Prediction System using Data Mining Techniques", *International Journal on Recent and Innovation Trends in Computing and Communication*, Volume: 2 Issue: 10, pp.3003-08,October 2014.

Authors Profile

Mr. Niraj Kalantri is currently pursuing his second year of M.tech from Vellore Institute of Technology, chennai campus in computer science and engineering with specialization in Big Data Analytics. He completed his bachelor in computer science and engineering in year 2017. His major area of interests are Machine Learning , Deep Learning , Data Mining , Data Visualizations and Internet of things.

Dr. R. Kumar pursed Bachelor of Computer Science and Engineering at University of Madras, Chennai in 1998 and Master of Computer Science and Engineering at Jadavpur University in 2003. He has completed Doctorate of Philosophy (Computer Science and Engineering) at Madras Institute of Technology Campus, Anna University, Chennai. He is currently working as an Associate Professor at Vellore Institute of Technology, Chennai campus. He has 8 years of Research and Development experience as Senior Research Associate at Madras Institute of Technology, Anna University, Chennai and Seven years of work experience in Teaching. He has in-depth knowledge and expertise in Cluster and Grid computing and Virtualization. Experience in Scheduling and Semantic Discovery and its integration with Grid Meta-Scheduler. Possesses proven research capability with decent number of publications in renowned international Journals and conferences in the field of Distributed Systems, Grid Computing and Multicore architecture.

