Early Prediction of Cardiac Arrest Using Data Mining Classification Techniques

Chandi Priya N[1], Anjan Babu G[2]

Dept of Computer Science, SV University, Tirupati

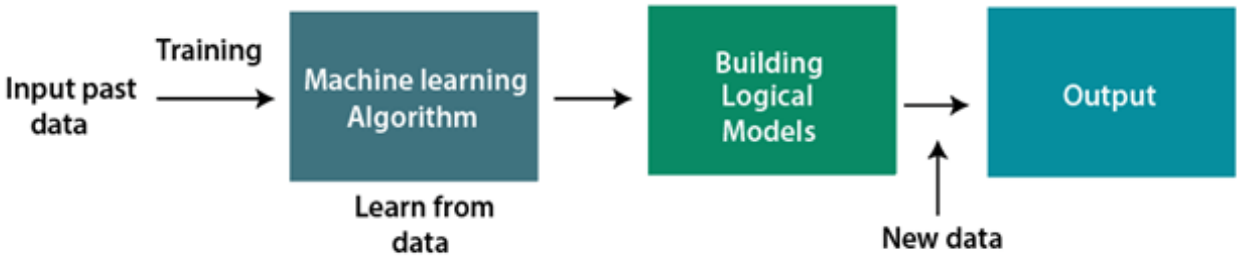
***Abstract*—** *This research work which initiated at an early detection of all the probable symptoms and signs which might further lead to detection of heart diseases using data collected from previous patients as well as data input received from the user at that particular time. Current scenario of health-care data used for surveillance are no longer simply a time building series of aggregate daily counts. Instead, a wealth of proposed spatial as well as temporal demographic, and symptom information is available at the data presented during the time of execution. Our proposed method incorporates all such information that is being used as a classification approach that compares recent healthcare data against data from that particular baseline distribution. In addition, the data sample data used is first train and test the system using machine Learning approaches. The proposed system trained to be Logistic Regression, K-Nearest Neighbours, Random Forest, XGBoost algorithm are used. Then proposed test scores have been evaluated. Classifier is further chosen to make predictions.*

I. INTRODUCTION

Cardiac arrest is one of the fatal attacks in the world that results in the supremacy of death. A heart attack is caused by a sudden occurrence of coronary thrombosis, typically which results in the death of a particular heart muscle and sometimes can be fatal. A heart attack happens if the flow of oxygen-rich blood to a section of heart muscle suddenly becomes blocked and the heart can't get oxygen. When plaque builds up in the arteries, the condition is called atherosclerosis. The professional build-up of plaques presents in the arteries occur over many years. Eventually, an area of plaque can rupture (break open) inside of an artery. This can cause a blood clot that can be formed on the surface of the plaque. The flow of blood through the clot becomes large. If any blockage present isn't treated quickly, the portion of heart muscle that can be fed by the artery can lead to death of that particular artery.

Healthy heart tissue is replaced with scar tissue. This heart damage may not be obvious, which further caused long-lasting and severe problems. A majority of the heart attacks occur as a result proportion to coronary heart disease. Coronary heart disease is a condition in which a wax like substance that can be termed as plaque builds up inside of the coronary arteries. Only early prediction could help to better diagnose the cardiac problems at the benign stage to save a person's life. A less initiated common cause of heart attack is that of severe spasm or rather tightening of a coronary artery. The spasm can cut off the flow of the blood through the artery. Atherosclerosis does not show effect on the spasms present in the coronary arteries. Heart attacks that can be associated with or can lead to severe problems that can diminish the health of an individual, such as heart failure and also can lead to life-threatening arrhythmias. Heart failure is a condition in which the heart can't pump enough blood to meet the body's needs. Irregular heartbeats are called Arrhythmias. Ventricular fibrillation present, is connected to a life-threatening arrhythmia that led to death if not treated the right away.

There are several factors that could affect a person's predisposition for Cardiac disease. Education is an important indicator of socioeconomic status that is associated with the occupation and also among the other factors affecting an individual's life style. A number of studies in developed countries have shown that Cardiac disease incidence varies between people with different levels of education.

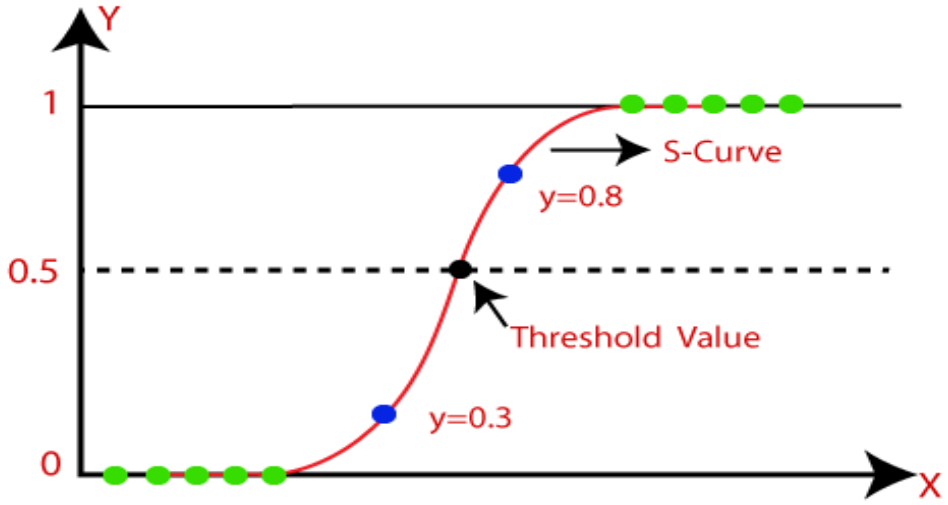


1.1 Logistic Regression Algorithm

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.

Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function.



1.1.1 Logistic Function (Sigmoid Function):

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1.
- The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.
- In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

1.2 Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model.* As the name suggests, **"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."** Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. **The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.**

Below are some points that explain why we should use the Random Forest algorithm:

- It takes less training time as compared to other algorithms.
- 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.

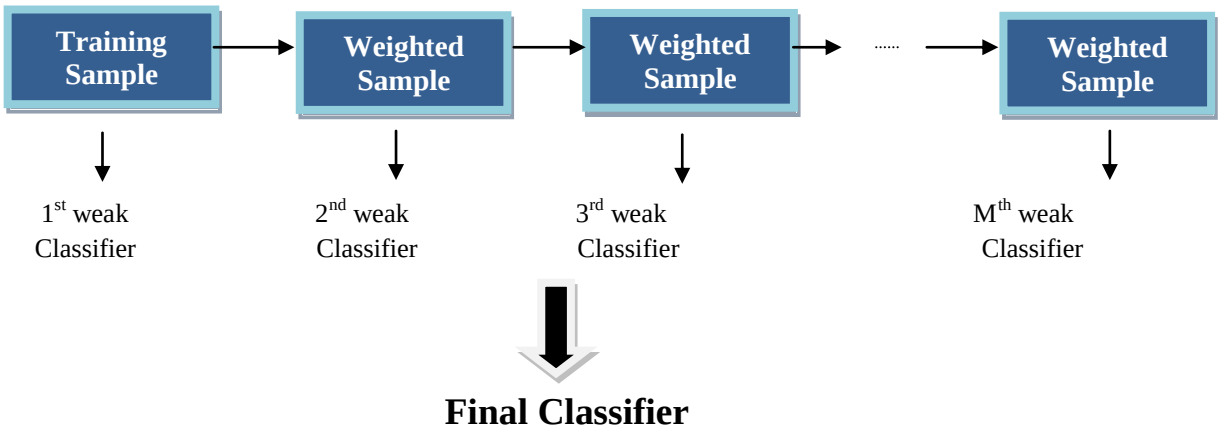
Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

- **Step-1:** Select random K data points from the training set.
- **Step-2:** Build the decision trees associated with the selected data points (Subsets).
- **Step-3:** Choose the number N for decision trees that you want to build.
- **Step-4:** Repeat Step 1 & 2.
- **Step-5:** For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

1.3 XGBoost (Extreme Gradient Boosting)

XGBoost is an implementation of Gradient Boosted decision trees. This library was written in C++. It is a type of Software library that was designed basically to improve speed and model performance. It has recently been dominating in applied machine learning. XGBoost models majorly dominate in many Kaggle Competitions. In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. Weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.



1.4 XGBoost Features

The library is laser-focused on computational speed and model performance, as such, there are few frills.

1.4.1 Model Features

Three main forms of gradient boosting are supported:

- Gradient Boosting
- Stochastic Gradient Boosting
- Regularized Gradient Boosting

1.4.2 System Features

For use of a range of computing environments this library provides.

- Parallelization of tree construction.
- Distributed Computing for training very large models.
- Cache Optimization of data structures and algorithm.

II. LITERATURE REVIEW

Prediction of Heart Disease by Using Machine Learning

Authors: Rohit Murty, Satish Patle, Saurabh Bute, Sneha Bhilkar, Durga Wanjari- 2020

With the rampant increase in the heart stroke rates at juvenile ages, we need to put a system in place to be able to detect the symptoms of a heart stroke at an early stage and thus prevent it. It is impractical for a common man to frequently undergo costly tests like the ECG and thus there needs to be a system in place which is handy and at the same time reliable, in predicting the chances of a heart disease. Thus, we propose to develop an application which can predict the vulnerability of a heart disease given basic symptoms like age, sex, pulse rate etc. The machine learning algorithm neural networks has proven to be the most accurate and reliable algorithm and hence used in the proposed system.

Analysis Of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset

Authors: Tapas Ranjan Baitharua, Subhendu Kumar Panib- 2016

Accuracy in data classification depends on the dataset used for learning. Now-a-days the most important cause of death for both men and women is due to the Liver Problem. The healthcare industry collects a huge amount of data which is not properly mined and not put to the optimum use. Discovery of these hidden patterns and relationships often goes unexploited. Our research focuses on this aspect of medical diagnosis by learning pattern through the collected data of Liver disorder to develop intelligent medical decision support systems to help the physicians. In this paper, we propose the use decision trees J48, Naive Bayes, ANN, ZeroR, 1BK and VFI algorithm to classify these diseases and compare the effectiveness, correction rate among them. Detection of Liver disease in its early stage is the key of its cure. It leads to better performance of the classification models in terms of their predictive or descriptive accuracy, diminishing of computing time needed to build models as they learn faster, and better understanding of the models. In this paper, a comparative analysis of data classification accuracy using Liver disorder data in different scenarios is presented. The predictive performances of popular classifiers are compared quantitatively.

Improving Disease Prediction by Machine Learning

Authors: Smriti Mukesh Singh1, Dr. Dinesh B. Hanchate2 – 2018

These days utilization of Big Data is expanding in biomedical and human services groups, exact investigation of medicinal information benefits early malady discovery, quiet care and group administrations. Fragmented therapeutic information lessens examination precision. The machine learning calculations are proposed for successful expectation of ceaseless infection. To beat the trouble of deficient information, Genetic algorithm will be utilized to remake the missing information. The dataset comprises of structured data and unstructured data. To extract features from unstructured data RNN algorithm will be utilized. Framework proposes SVM calculation and Naive Bayesian calculation for sickness expectation utilizing unstructured and structured information individually from hospital information. Community Question Answering (CQA) system is additionally proposed which will foresee the inquiry and answers and will give proper responses to the clients. For that, two calculations are proposed KNN and SVM. KNN algorithm will perform classification on answers and SVM calculation will perform classification on answers. It will help client to discover best inquiries and answers identified with infections.

Heart Disease Prediction System Using Data Mining Techniques

Authors: Abhishiek Taneja – 2015

In today's modern world cardiovascular disease is the most lethal one. This disease attacks a person so instantly that it hardly gets any time to get treated with. So, diagnosing patients correctly on timely basis is the most challenging task for the medical fraternity. A wrong diagnosis by the hospital leads to earn a bad name and loosing reputation. At the same time treatment of the said disease is quite high and not affordable by most of the patients particularly in India. The purpose of this paper is to develop a cost-effective treatment using data mining technologies for facilitating data base decision support system. Almost all the hospitals use some hospital management system to manage healthcare in patients. Unfortunately, most of the systems rarely use the huge clinical data where vital information is hidden. As these systems create huge amount of data in varied forms but this data is seldom visited and remain untapped. So, in this direction lots of efforts are required to make intelligent decisions. The diagnosis of this disease using different features or symptoms is a complex activity. In this paper using varied data mining technologies an attempt is made to assist in the diagnosis of the disease in question.

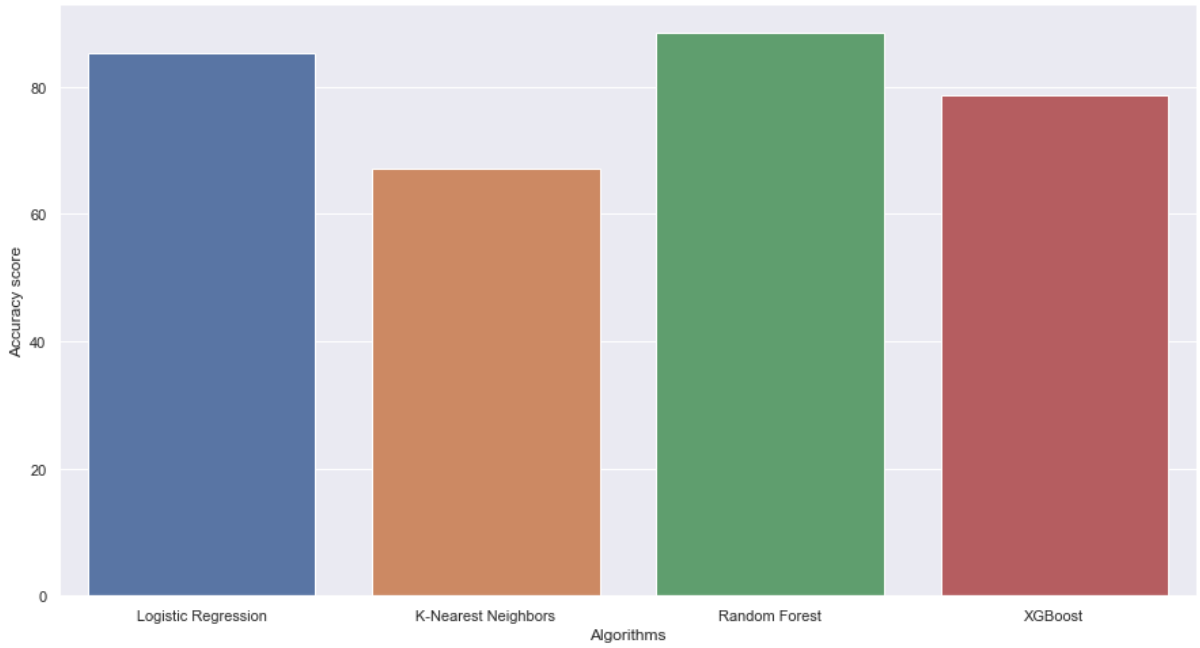
Improved Study of Heart Disease Prediction System Using Data Mining Classification Techniques

Authors: Chaitrali S. Dangare, Sulabha S. Apte- 2016

The Healthcare industry is generally "information rich", but unfortunately not all the data are mined which is required for discovering hidden patterns & effective decision making. Advanced data mining techniques are used to discover knowledge in database and for medical research, particularly in heart disease prediction. This paper has analysed prediction systems for Heart disease using a greater number of input attributes. The system uses medical terms such as sex, blood pressure, cholesterol like 13 attributes to predict the likelihood of patient getting a heart disease. Until now, 13 attributes are used for prediction. This research paper added two more attributes i.e., obesity and smoking. The data mining classification techniques, namely Decision Trees, Naive Bayes, and Neural Networks are analysed on heart disease database. The performance of these techniques is compared, based on accuracy. As per our results accuracy of Neural Networks, Decision Trees, and Naive Bayes are 100%, 99.62%, and 90.74% respectively. Our analysis shows that out of these three classification models Neural Networks predicts heart disease with highest accuracy.

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target	
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

303 rows × 14 columns



The accuracy score Logistic Regression is: 85.25 %.

The accuracy score K-Nearest Neighbors is: 67.21 %.

The accuracy score Random Forest is: 88.52 %.

The accuracy score XGBoost is: 78.69 %.

RANDOM FOREST GIVES GOOD ACCURACY.

III. CONCLUSION AND FUTURE WORK

In this paper a reliable multi process method, Machine Learning concept to build a Heart disease risk prediction system is proposed and Evaluate High accuracy had done comparatively Existing system. Heart disease has become the leading cause of death worldwide. The most effective way to reduce such deaths is to detect its symptoms earlier. The overall aim is to define various data mining techniques useful in effective heart disease prediction. Efficient and accurate prediction with a lesser number of attributes and tests is our goal. In this study, I consider only 14 essential attributes. I applied four data mining classification techniques, K-nearest neighbor and random forest. The data were pre-processed and then used in the model. K-nearest neighbor and random forest are the algorithms showing the best results in this model. I found the accuracy after implementing four algorithms to be highest in K-nearest neighbors ($k = 7$). We can further expand this research incorporating other data mining techniques such as time series, clustering and association rules. Considering the limitations of this study, there is a need to implement more complex and combination of models to get higher accuracy for early prediction of heart disease.

REFERENCES

[1] Baitharu, Tapas Ranjan, and Subhendu Kumar Pani. "Analysis of data mining techniques for healthcare decision support system using liver disorder dataset." *Procedia Computer Science* 85 (2016): 862-870.

[2] Gavhane, Aditi, GouthamiKokkula, Isha Pandya, and Kailas Devadkar. "Prediction of heart disease using machine learning." In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 1275-1278. IEEE, 2018.

[3] Singh, Smriti Mukesh, and Dinesh B. Hanchate. "Improving disease prediction by machine learning." *Int J Res EngTechnol* 5, no. 6 (2018): 1542-1548.

[4] Taneja, Abhishek. "Heart disease prediction system using data mining techniques." *Oriental Journal of Computer science and technology* 6, no. 4 (2013): 457-466.

[5] Dangare, Chaitrali S., and Sulabha S. Apte. "Improved study of heart disease prediction system using data mining classification techniques." *International Journal of Computer Applications* 47, no. 10 (2012): 44-48.

[6] Thomas, J., and R. Theresa Princy. "Human heart disease prediction system using data mining techniques." In *2016 international conference on circuit, power and computing technologies (ICCPCT)*, pp. 1-5. IEEE, 2016.

[7] Kaur, Beant, and Williamjeet Singh. "Review on heart disease prediction system using data mining techniques." *International journal on recent and innovation trends in computing and communication* 2, no. 10 (2014): 3003-3008.

[8] Meyer, Alexander, Dina Zverinski, Boris Pfahringer, JörgKempfert, Titus Kuehne, Simon H. Sündermann, Christof Stamm, Thomas Hofmann, Volkmar Falk, and Carsten Eickhoff. "Machine learning for real-time prediction of complications in critical care: a retrospective study." *The Lancet Respiratory Medicine* 6, no. 12 (2018): 905-914.

[9] Rajkomar, Alvin, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. "Ensuring fairness in machine learning to advance health equity." *Annals of internal medicine* 169, no. 12 (2018): 866-872.

[10] Rajamhoana, S. P., C. Akalya Devi, K. Umamaheswari, R. Kiruba, K. Karunya, and R. Deepika. "Analysis of neural networks-based heart disease prediction system." In *2018 11th International Conference on Human System Interaction (HSI)*, pp. 233-239. IEEE, 2018.

[11] Ramalingam, V. V., AyantanDandapath, and M. Karthik Raja. "Heart disease prediction using machine learning techniques: a survey." *International Journal of Engineering & Technology* 7, no. 2.8 (2018): 684-687.

[12] Kohli, Pahulpreet Singh, and Shriya Arora. "Application of machine learning in disease prediction." In *2018 4th International conference on computing communication and automation (ICCCA)*, pp. 1-4. IEEE, 2018.

[13] Marimuthu, M., M. Abinaya, K. S. Hariesh, K. Madhankumar, and V. Pavithra. "A review on heart disease prediction using machine learning and data analytics approach." *International Journal of Computer Applications* 181, no. 18 (2018): 20-25.

[14] Beyene, Chala, and Pooja Kamat. "Survey on prediction and analysis the occurrence of heart disease using data mining techniques." *International Journal of Pure and Applied Mathematics* 118, no. 8 (2018): 165-174.

[15] Khourdifi, Youness, and Mohamed Bahaj. "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization." *International Journal of Intelligent Engineering & Systems* 12, no. 1 (2019): 242-252.