Heart Disease Identification Method using Machine Learning Classification in E Healthcare

Ms. Chintalappagari Mani

Dept of Computer Science, Sri Venkateswara 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 Neighbors, 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 occurs 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.

II. LITERATURE REVIEW

Prediction of heart disease by using machine learning

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

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 are 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

Smriti Mukesh Singh, Dr. Dinesh B. Hanchate – 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

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

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.

Problem Statement

- 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.
- The heart damages 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.

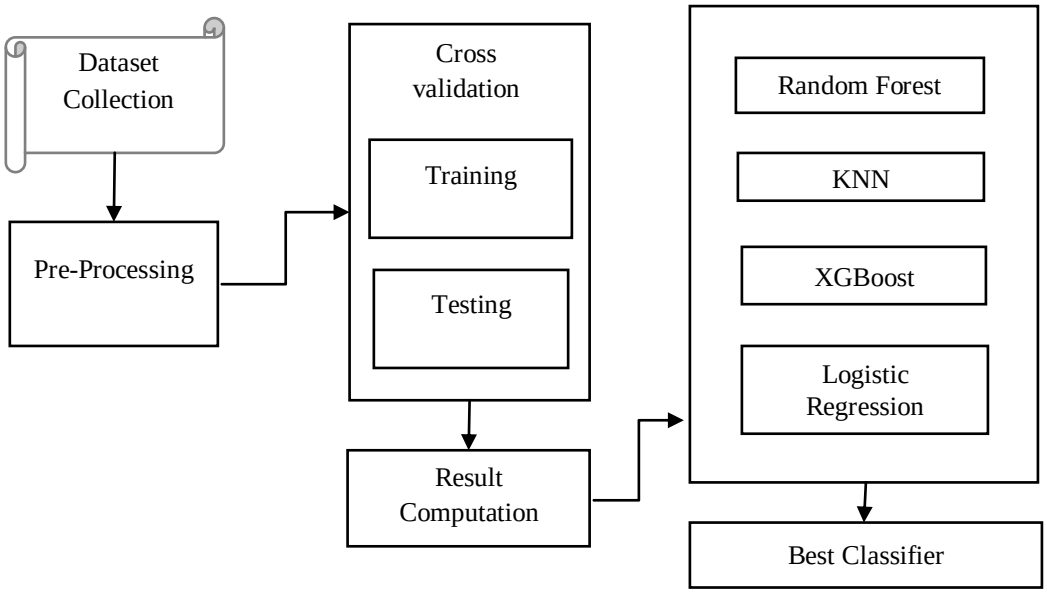
III. PROPOSED SYSTEM

- Our proposed system involves Logistic regression, XG Boost, Random Forest classifier, KNN Algorithm in Machine Learning concept used to train the dataset.
- Thus, preventing heart diseases has become more than necessary. Good data-driven systems for predicting heart diseases can improve the entire research and prevention process, making sure that more people can live healthy lives.
- This is where Machine Learning comes into play. Machine Learning helps in predicting the heart diseases, and the predictions made are quite accurate.
- The project involved analysis of the heart disease patient dataset with proper data processing. Then, different models were trained and predictions are made with different algorithms KNN, Decision Tree, Random Forest,Logistic Regression etc.
- This is the jupyter notebook code and dataset I've used for my Kaggle kernel 'Binary Classification with Sklearn and Keras'

Advantages

- Easy detection of the Cardiac disease with the concluded technique.
- Time consuming.
- Best accuracy Model helps in better treatment as early.

Detection of best Model will quick the treatment which is life saving



IV. IMPLEMENTATION

- Dataset Collection and pre-processing
- Train the model
- Evaluation
- Comparison of existing model
- Performance analysis

4.1 Dataset Collection and Pre-processing

A dataset (or data set) is a collection of data, usually presented in tabular form. Each column represents a particular variable. Each row corresponds to a given member of the dataset in question. It lists values for each of the variables, such as height and weight of an object. Each value is known as a datum.

We have chosen to use a publicly-available Healthcare dataset which contains a relatively small number of inputs and cases. The data is arranged in such a way that will allow those trained in medical disciplines to easily draw parallels between familiar statistical and novel ML techniques. Additionally, the compact dataset enables short computational times on almost all modern computers.

The sklearn preprocessing package provides several common utility functions and transformer classes to change raw feature vectors into a representation that is more suitable for the downstream estimators.

4.2 Train the Model

This stage is to form evaluation the models based on the input data. For our purpose of study, we are going to implement to train the model using four types of algorithms in machine learning to predict heart disease.

The K-nearest neighbours algorithm is a supervised classification algorithm method. It classifies objects dependant on nearest neighbour. It is a type of instance-based learning. The calculation of distance of an attribute from its neighbours is measured using Euclidean distance. It uses a group of named points and uses them on how to mark another point. The data are clustered based on similarity amongst them, and is possible to fill the missing values of data using K-NN. Once the missing values are filled, various prediction techniques apply to the data set. It is possible to gain better accuracy by utilizing various combinations of these algorithms.

Random forest algorithm is a supervised classification algorithmic technique. In this algorithm, several trees create a forest. Each individual tree in random forest lets out a class expectation and the class with most votes turns into a model's forecast. In the random forest classifier, the greater number of trees gives higher accuracy. The three common methodologies are:

- Forest RI (random input choice);
- Forest RC (random blend);
- Combination of forest RI and forest RC.

It is used for classification as well as regression task, but can do well with classification task, and can overcome missing values. Besides, being slow to obtain predictions as it requires large data sets and more trees, results are unaccountable.

Random forest algorithm has obtained an accuracy of 91.6% with Cleveland dataset in Using People's dataset, an accuracy of 97% was achieved.

4.3 Evaluation

This stage is to form evaluation the models based on the input data. For our purpose of study, we are going to implement the model using XGboost classifier algorithm in Machine Learning.

We will split the data set into test and train set. After splitting the data first have to train the data and test the data using XGboost classifier, Logistic regression, KNN, Random Forest in Machine learning techniques.

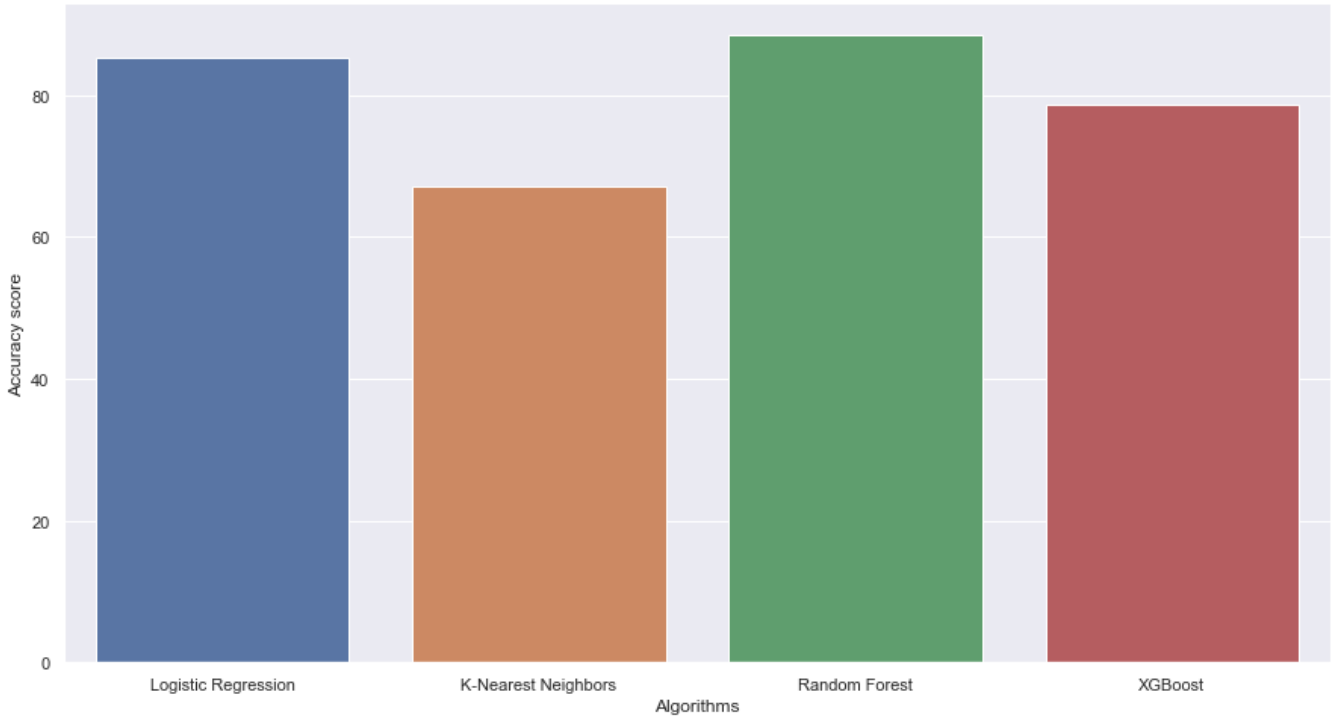
4.4 Comparison of Existing model

This module includes comparison of existing system algorithm accuracy and our proposed model accuracy. Our aim is to improve the accuracy score.

4.5 Performance Analysis

The next stage is to predict the results using Classifier. The best method for the training and test data set is definitely given has the best results for Classification Accuracy and Recall for both validation cases. Now we forward this Random Forest classifier to next stage to predict the disease that may further lead to a cardiac arrest.

The results are compared using a confusion matrix. The consistency of a classification model can be well visualized with a tabular form also called Confusion Matrix (or "classifier") which shows its results over a set of known test data.




```
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.

V. 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.