Heart Disease Prediction using Machine Learning

Abhijeet Jagtap[1], Priya Malewadkar[2], Omkar Baswat[3], Harshali Rambade[4]

[1,2,3]*Student, Department of Information Technology, Vidyalankar Institute of Technology, Mumbai, India*

[4]*Professor, Department of Information Technology, Vidyalankar Institute of Technology, Mumbai, India*

***Abstract*: Heart disease is considered as one of the major causes of death throughout the world. It cannot be easily predicted by the medical practitioners as it is a difficult task which demands expertise and higher knowledge for prediction. The healthcare environment is still 'information rich' but 'knowledge poor'. There is a lot of data available within the healthcare systems on the internet. However, there is a lack of effective analysis tools to discover hidden relationships and patterns in data. An automated system in medical diagnosis would enhance medical efficiency and reduce costs. This web application intends to predict the occurrence of a disease based on data gathered from Kaggle and Cleveland foundation medical research particularly in Heart Disease. The goal is to extract the hidden patterns by applying data mining techniques on the dataset, which are noteworthy to heart diseases and to predict the presence of heart disease in patients where the presence is valued on a scale. The prediction of heart disease requires a huge size of data which is too complex and massive to process and analyze by conventional techniques. Our objective is to find out the suitable machine learning technique that is computationally efficient as well as accurate for the prediction of heart disease. Data mining combines Statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases.**

***Keywords*: Prediction, heart disease, machine learning, algorithms, analysis, database, data mining.**

1. Introduction

The highest mortality of both India and abroad is mainly because of heart disease. According to World Health Organization (WHO), heart related diseases are responsible for the taking 17.7 million lives every year, 31% of all global deaths [6]. Hence, this is vital time to check this death rate by identifying the disease correctly in the initial stage. We can use data mining technologies to discover knowledge from the datasets. The discovered knowledge can be used by the healthcare administrators to improve the quality of service. The discovered knowledge can also be used by medical practitioners to reduce the number of adverse drug effect, to suggest less expensive therapeutically equivalent alternatives. Anticipating patient's future behavior on the given history is one of the important applications of data mining techniques that can be used in healthcare management.

A major challenge facing healthcare organizations (hospitals, medical centers) is the provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and/or decision support systems. Healthcare data is massive [8]. It includes patient data, resource management data, and transformed data. Healthcare organizations must have the ability to analyze data. Treatment records of millions of patients can be stored, and computerized and data mining techniques may help in answering several important and critical questions related to health care. Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge-rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Wu, et al proposed that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modelling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

2. Literature survey

Intelligent Heart Disease Prediction System Using Data Mining Techniques: The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining techniques can help remedy this situation. This research has developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques, namely, Decision Trees, Naive Bayes and Neural Network[1].

Smartphone Based Ischemic Heart Disease (Heart Attack) Risk Prediction: An Android based prototype software has been developed by integrating clinical data obtained from patients admitted with IHD (Ischemic Heart Disease). The clinical data from 787 patients has been analyzed and correlated with the risk factors like Hypertension, Diabetes, Dyslipidemia (Abnormal cholesterol), Smoking, Family History, Obesity, Stress and existing clinical symptom which may suggest underlying non-detected IHD. The data is mined with data mining technology and a score is generated. Risks are classified into low, medium and high for IHD[2].

Analysis of Data Mining Techniques for Heart Disease Prediction: Heart disease is considered as one of the major causes of death throughout the world. It is hard to predict for the medical practitioners as it is a difficult task which demands expertise and higher knowledge for prediction. This paper addresses the issue of prediction of heart disease according to input attributes based on data mining techniques. We have investigated the heart disease prediction using KStar, J48, SMO, Bayes Net and Multilayer Perceptron through Weka software. The performance of these data mining techniques is measured by combining the results of predictive accuracy, ROC curve and AUC value using a 6 standard data set as well as a collected data set. Based on performance factor SMO and Bayes Net techniques show optimum performances than the performances of K-Star, Multilayer Perceptron and J48 techniques[3].

Machine Learning Application Predict the Risk of Coronary Artery Atherosclerosis: Coronary artery disease is the leading cause of death in the world. In this research, we propose an algorithm based on the machine learning techniques to predict the risk of coronary artery atherosclerosis. A ridge expectation maximization imputation (REMI) technique is proposed to estimate the missing values in the atherosclerosis databases. A conditional likelihood maximization method is used to remove irrelevant attributes and reduce the size of feature space and thus improve the speed of the learning. The STULONG and UCI databases are used to evaluate the proposed algorithm. The performance of heart disease prediction for two classification models is analyzed and compared to previous work. Experimental results show the improved accuracy percentage of risk prediction of our proposed method compared to other works. The effect of missing value imputation on the prediction performance is also evaluated and the proposed REMI approach performs significantly better than conventional techniques[4].

3. Objectives

- *Easy to use:* The main objective of this project is to develop a platform which will be simple and easy to use, as here one must provide the patient's medical details and based on the features extracted the algorithm will then detect the heart disease and spot its type. As here algorithm does the task hence a well-trained model is less bound to make errors in predicting the heart disease and its type hence, in short accuracy is improved and thereby it also saves time and makes easier for doctors as well as patients to predict whether they are prone to any type of heart disease or not, which is otherwise we difficult to do without doctor's involvement.
- *No human intervention required:* To detect the heart disease one must provide medical details such as age, cholesterol, etc. and here the algorithm will provide the results based on the features extracted and hence here chances of error been made are very minimum since there is no human intervention and it also saves lot of time for the patients or doctors and they can further proceed for treatments or other procedures must faster. This is in case when results are provided faster to them. This can in-turn make the precaution/prevention process of heart treatment a lot faster when it saves doctors and patient the crucial time, so they can go on to further treatments and precautions to be taken to minimize the impact of that heart disease.
- *Not only detect the heart disease type but also suggest precautions:* In this project our aim is not only to find and predict the type of heart disease but pin point towards the precautions to be taken to minimize the impact of the heart disease. Getting suggestions on precautions to be taken will help the doctors and patients to progress easily to further steps in their treatment.
- *Efficient use of available annotated data samples:* There is large consent that successful training of machine learning algorithms requires many thousand annotated training samples. Hence, we use a network and training strategy that relies on the strong use of data pre-processing to use the available annotated samples more efficiently. As medical data is not available in a large bulk (more than or up to thousands of samples, according to machine learning standards) we use data pre-processing to make use of the available data more efficiently. Data pre-processing is an essential to data mining technique that involves transforming raw data into an understandable format. Real-world medical data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data pre-processing is a proven method of resolving such issues. Data pre-processing prepares raw data for further processing.

4. Proposed system

Heart diseases prediction is a web-based machine learning application, trained by a UCI dataset. The user inputs its specific medical details to get the prediction of heart disease for that user. The algorithm will calculate the probability of presence of heart disease. The result will be displayed on the webpage itself. Thus, minimizing the cost and time required to predict the disease. Format of data plays crucial part in this application. At the time of uploading the user data application will check its proper file format and if it not as per need then

ERROR dialog box will be prompted.

Our system will be implementing the following three algorithms:

- Support Vector Machine (SVM)
- Logistic Regression
- Naïve Bayes Algorithm

The algorithms will be trained using the data set obtained from University of California, Irvine. 75% of the entries in the data set will be used for training and the remaining 25% for testing the accuracy of the algorithm. Furthermore, some steps will be taken for optimizing the algorithms thereby improving the accuracy. These steps include cleaning the dataset and data pre-processing. The algorithms were judged based on their accuracy and it is observed that the SVM is the most accurate out of the three with 64.4% efficiency. Hence, it is selected for the main application.The main application is a web application which accepts the various parameters from the user as input and computes the result. The result is displayed along with the accuracy of prediction.

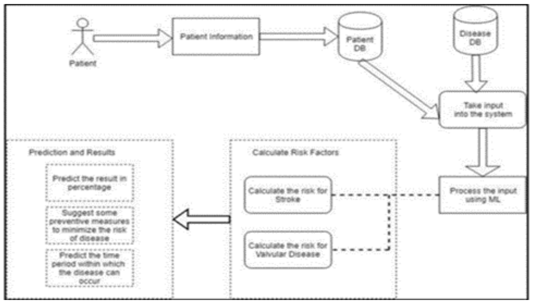
Patient
Patient Information
Patient DB
Disease DB
Take input into the system
Process the input using ML
Calculate Risk Factors
Calculate the risk for Stroke
Calculate the risk for Valvular Disease
Prediction and Results
Predict the result in percentage
Suggest some preventive measures to minimize the risk of disease
Predict the time period within which the disease can occur

Fig. 1. Block diagram for heart disease prediction

A. Website

The system will consist of a website, where users will register themselves for getting the report of health of their heart in terms of predictive analysis about their heart disease. User will have to fill a form initially for registration. Then user will get redirected to the profile page where they will have to complete their profile by filling all the information related to their heart. After submitting the health information the patient will be able to have look at the report where they will be knowing the status or risk of their heart in terms of percentage. If the user will have risk greater than 60% then user will be redirected to another form where he will have to enter additional symptoms so that system will give prediction about the category of heart disease from two most common categories i.e. CAD (Coronary Artery Disease) and Valvular disease

B. Database

The server will be using MySQL database. The system‘s database consists of following tables.

- *Users table* – This table will consist of all the user information which includes user’s name, e-mail id, phone number, address, etc.
- *Medical history table* – This table will consist of all the health related information of users which is related to heart that includes attributes such as age, gender, resting blood pressure, cholesterol, fasting blood sugar, old peak, etc.

C. Machine learning algorithm

The machine learning algorithm will be used to predict the risk of heart disease in terms of percentage.

5. Methodology

As per the data and information we have gathered, we found that these following tasks must be carried out in order to get much accurate predictions. The tasks that we are going to carry out are as follows.

- Data Preprocessing: The dataset we obtained is not completely accurate and error free. Hence, we will first carry out the following operations on it.
- Data Cleaning: NA values in the dataset is the major setback for us as it will reduce the accuracy of the prediction profoundly so, we will remove the fields which does not have values. We will substitute it with the mean value of the column. This way, we will remove all the values in the data set.
- Feature Scaling: Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without feature scaling. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be scaled so that each feature contributes approximately proportionately to the final distance. So we will scale the various fields in order to get them closer in terms of values. e.g. Age has just two values i.e. 0,1 and cholesterol has high values like 100. So, in order to get them closer to each other we will need to scale them.
- Factorization: In this section, we assigned a meaning to the values so that the algorithm doesn’t confuse between them. For example, assigning meaning to 0 and 1 in the age section so that the algorithm doesn’t consider 1 as greater than 0 in that section.
- Support Vector Machine: Support vector machine (SVM) are supervised learning method that analyze data used for classification and regression analysis. It is given a set of training data, marked as belonging to either one of two categories, an SVM training algorithm then builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space

and predicted to belong to a category based on which side of the gap they fall. The points are separated based on hyper plane that separate them. When data are not labeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. In the project, we have used this algorithm to classify the patients into groups according to the risk posed to them based on the parameters provided. It was observed that: Naïve Bayes had 60% accuracy, logistic regression had 61.45% and SVM had 64.4%. Hence SVM was selected as the most efficient algorithm for the web application

6. Conclusion

This paper presents an overview of Heart Disease Prediction using Machine Learning.

References

[1] Sellappan Palaniappan, Rafiah Awang "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IEEE, July 2015

[2] M. Raihan, Saikat Mondal, Arun More, Md. Omar Faruqe Sagor, Gopal Sikder, Mahbub Arab Majumder, Mohammad Abdullah Al Manjur and Kushal Ghosh "Smartphone Based Ischemic Heart Disease (Heart Attack) Risk Prediction using Clinical Data and Data Mining Approaches, a Prototype Design", September 2014.

[3] Marjia Sultana, Afrin Haider and Mohammad Shorif Uddin "Analysis of Data Mining Techniques for Heart Disease Prediction", May 2015.

[4] Soodeh Nikan, Femida Gwadry-Sridhar, and Michael Bauer "Machine Learning Application to Predict the Risk of Coronary Artery Atherosclerosis", IEEE, August 2016

[5] Sanjay Kumar Sen Asst. Professor, Computer Science & Engg. Orissa Engineering College, Bhubaneswar, Odisha – India." Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms" International Journal of Engineering and Computer Science. Volume 6 Issue 6, June 2017

[6] V.V. Ramalingam, Ayantan Dandapath, M Karthik Raja "Heart disease prediction using machine learning tech : A survey" International Journal of Engineering & Technology, 7 (2.8), April 2018.

[7] Heart Disease Dataset - https://www.kaggle.com/c/heart-disease dated: Sept 2018

[8] K. Srinivas, B. Kavihta Rani, A. Govrdhan "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attack" IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 02, 2010.

[9] Heart Disease Data Set https://archive.ics.uci.edu/ml/datasets/heart+Disease