Heart Disease Prediction Using Random Forest and Logistic Regression

Aamir Khan*, Dr.Sanjay Jain
ITM University, Gwalior, Madhya Pradesh, India
CSA Department
helloaamirkhan123@gmail.com

Abstract

The data mining (DM) is a process that deals with mining of valuable information from the rough data. The method of prediction analysis (PA) is implemented for predicting the future possibilities on the basis of current information. This research work is planned on the basis of predicting the heart disease. The coronary disorder can be forecasted in different phases in which pre-processing is done, attributes are extracted and classification is performed. The hybrid method is introduced on the basis of RF and LR. The Random Forest classification is adopted to extract the attributes and the classification process is carried out using logistic regression. The analysis of performance of introduced system is done with regard to accuracy, precision and recall. It is indicated that the introduced system had provided accuracy around 95% while predicting the heart disease.

Keywords:- Heart Disease prediction, MLP, Decision Tree, Naïve Bayes , Random Forest, Logistic Regression

I. INTRODUCTION

A technology utilized to analyze the data is recognized as DM that assists in recognizing the patterns from data set with the help of diverse data mining tools and methods. The least user input and efforts are ensured to recognize the patterns in automatic manner using data mining. DM is proved efficient tool for handling decision making and predicting the future market trends. Various applications make the implementation of DM tools and methods in effective manner. The data mining is extensively utilized in various organizations to analyze the data so that the complex environment can be tackled [1]. The mining tools and methods are exploited in several trading applications with the objective of computing diverse trends and patterns of market and analyzing the fast and effectual market trend.

Diverse kinds of methods adopted in DM are defined as.

a. Association: This method is focused on recognizing a particular pattern using the association among particular items of similar transaction. To illustrate, the association method is employed to find the association of several features utilized to perform the analysis while predicting the heart disease. All the risk factors required to predict the disease are exploited for classifying the patients affected with such coronary disorder [2].

b. Classification: This is another standard DM method planned depending upon the ML (machine learning). This technique classifies each object in the data set into one predefined set of class. This technique makes the utilization of distinct mathematical method.

c. Clustering: The major intend of this approach is to cluster the objects with similar property for producing a valuable cluster with the help of an automatic method. The clustering methods assist in defining the classes and the objects available in them. Additionally, the predefined classes assign the classification objects. For example, the clustering is employed for clustering the list of patients having similar risk factors after predicting the condition of heart. Consequently, the patients with high glucose and pertinent danger factors are separated [3].

d. Prediction: This DM method is implemented to investigate the association not only among independent variables but also the dependent and independent variables. To illustrate, the predictive analysis methods are applied to forecast the profit for future in case the deal is viewed as autonomous variable and benefit is treated as a reliant variable in deals. Moreover, an appropriate regression curve can be drawn in order to predict the benefit dependent on the given authentic deal and benefit information.

1.1. Predictive Analysis in Data Mining

The past data and knowledge can be analyzed effectively using a number of statistical trends and methods ranging from ML and predictive modeling to DM in PA (prediction analysis) method. The PA method is utilized to predict any unknown future events [4]. Any kinds of risks and opportunities are recognized by PA as using it the patterns of historical business data can be employed on the basis of business aspect. The risk assessment can be obtained or any type of potential threat can be recognized by capturing the association among several factors. The effectual decision-making stages can be executed to guide the business. The predictive analysis is described on the basis of prediction modeling and forecasting .In last decades, several PA models are developed for achieving the prediction. Three classes of these models are defined as:

a. Predictive Models: These systems are executed to investigate the association among several attributes available in the gathered data. This model is useful for assessing the similarities among a group of units. It ensures that the same features, that a group represented, are available [5].

b. Descriptive Models: These models assist in recognizing and evaluating the associations amid diverse features of unit. Afterward, these features are classified into certain groups using descriptive models [6]. Unlike the other systems, this model has potential to compare and forecast the data depending upon their association among multiple behaviors of units.

c. Decision Models: This model is deployed to define and discover the association among different kinds of information components accessible in the given dataset. The model is clarified in this dataset. The choice composition is defined in order to classify the known and

predicted outcome and its classification is carried out this model. Numerous attributes of dataset are considered to recognize and forecast the results of decisions [7].

II. Literature Review

Anjan Nikhil Repaka, et.al (2019) discussed various sources led to cause any kind of coronary disease and the gathering of data was done from these types of sources [8]. Due to this, the structure of database was generated. The issues related to predict the heart disease were resolved using NB (Naive Bayesian) technique and AES (Advanced Encryption Standard) algorithm that assisted in designing the SHDP (Smart Heart Disease Prediction). The accuracy achieved from the suggested NB approach was calculated 89% and superior to traditional model. Additionally, the AES performed better in contrast to other models with regard to security.

Aditi Gavhane, et.al (2018) suggested an application that employed the essential indications for foreseeing the weakness of a coronary illness [9]. The exactness and unwavering quality NNs was found greater. Thus, this technique was utilized in the suggested approach. The MLP (Multilayer Perceptron) model was employed as the CAD (computer aided design) to provide the predictive results so that the condition of user was computed. The ML techniques were assisted in predicting the disease as these techniques were developed rapidly. Thus, due to the superior efficacy and accuracy, the MLP model was implemented in the suggested approach. The suggested approach provided optimal result on the basis of input that the user inserted. The exploitation of this kind of algorithms by numerous people resulted in maximizing the awareness about the current heart status. Therefore, the number of patients infected with coronary disease was mitigated.

Aakash Chauhan, et.al (2018) analyzed that the number of people suffered due to the cardiovascular diseases were maximized at rapid level in India [10]. The reason behind death in India was predicted heart disorder in the upcoming years. Thus, it was essential to diminish its impact. Thus, a heart disease prediction system was put forward for investigating the risk of heart disease accurately. The DM (data mining) methods were deployed to construct a novel system for predicting the heart disease. The example development approach was applied on the patients' record for producing substantial association rules. The presented approach assisted the physicians in exploring the data and forecasting the heart disease in accurately manner.

C. Sowmiya, et.al (2017) introduced the evaluation of forecasting the heart disease with the analysis of potential of 9 diverse classifiers. Different research studies made the utilization of these methods [11].This approach had adopted the SVM (support vector machine) and apriori algorithms for predicting the heart disease. This approach focused on gathering and deploying medical profiles on the basis of several factors. The probability of occurrence of cardiovascular disease in patients was predicted in this approach. The heart disease was to be detected and prevented that was the major concern of medical sector. It was observed that the introduced approach was more accurate and efficient as compared to earlier methods.

Rashmi G Saboji, (2017) developed a novel system in order to forecast the heart disease on the basis of certain features with the help of the healthcare data [12]. The fundamental intend of this approach was to foresee the analysis of coronary illness dependent on modest number of qualities. Apache Spark was applied to implement the RF (random forest) method so as a solution was offered for the forecasting. This solution was utilized on highly scalable scenario to make the decision using the developed system that generated an opportunity to the health care analysts. The developed system yielded accuracy up to 98%. Moreover, the developed system had generated optimal outcomes as compared to the NB (Naïve-Bayes) algorithm.

III. Research Methodology

The heart is the essential muscular organ of human being. This part pumps the blood from the blood vessels of circulatory system. Thus, the life of human beings is highly dependent on the heart. The occurrence of any kind of disease in heart causes impact on other parts of the human body. The DM (data mining) is implemented in order to extract the information based on computer from enormous datasets. Various communities make the deployment of DM tools and methods. The medical sector utilizes the DM tools to forecast various diseases. According to the reports of World Health organization, a millions of people are affected with the coronary diseases. The medical communities have recorded the information in detail regarding heart patients manually. The physicians require only electronic records. The data mining methods can easily convert the DM methods into manual records. Various risk factors cause the heart diseases in patients.

Different stages to predict the heart disease are defined as:-

A. Data Acquisition: This stage collects the data from a variety of clinical organizations for conducting the experiments.

B. Data preprocessing: The data is pre-processed to implement the ML methods with the objective of completeness and performing a useful analysis on the data. First of all, the missing values are marked in the data using a numerical cleaner filter. These values are set to a defined default value in order to clean the huge or small sized numeric data. Thereafter, a filter is deployed for marking and detecting the missing values and replacing them with mean value of data distribution. A clean data free of noise is offered to enhance the efficacy of training model in the process of selecting attributes and to eliminate the irrelevant attributes from the dataset.

C. Feature selection: This process employs a subset of highly distinguished attributes for diagnosing the disease. This stage emphasized on choosing the discriminating attributes

which come under the available classes. The attributes are selected in two stages. Firstly, the attribute evaluator method is applied to compute the attributes of dataset on the basis of the output class. Subsequently, search technique that employs several groups of features to choose an optimal set for dealing with the issue of classification. The RF (random forest) algorithm is implemented so as the attributes are chosen. This algorithm has taken 100 as the estimator value and its major objective is to produce a tree structure of the most relevant features. The most relevant or significant attributes used to predict the coronary disease are selected using this algorithm.

D. Classification: The given attributes are classified to predict the disease by mapping the chosen attributes to the training model. As a multi-class issue, this process is carried out and the medical data is classified among 4 diverse classes. Every separate class expresses the category of coronary disease. The classified is carried out using LR (logistic regression) algorithm. This algorithm employs the input of the attributes whose extraction is done. The LR classifier is based on the probability assists in calculating the probability and this probability is applied to categorize the data into certain classes. This research work describes two classes: having coronary disease and normal. This implies that the person has a likelihood of occurrence of cardiovascular disease or not. The extracted attributes are fed as input in LR algorithm. This type of regression has potential to predict the probability of occurrence of an event for which data must have robustness for a logistic function. Similar to the several kinds of regression analysis, distinct predictive variables such as numerical or categorical are employed in logistic regression algorithm.

The hypothesis of LR is defined as:

$$h_\theta(x) = g(\theta^T x)$$

Here, the function g is the sigmoid function that is defined as:

$$g(z) = \frac{1}{1 + e^{-z}}$$

The special properties are comprised in the sigmoid function to offer the values in range [0,1]. The cost function for logistic regression is described as:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}[-y^{(i)}\log\left(h_e\left(x^{(i)}\right)\right) - \left(1 - y^{(i)}\right)\log\left(1 - h_e\left(x^{(i)}\right)\right)]$$

A built-in function called $fmin\_bfgs^2$ is utilized to investigate the minimum of this cost function in Machine Learning. This function assists in discovering the finest parameters θ for the cost function of logistic regression using which a fixed dataset having x and y as values is generated. This classifies the person having a likelihood of occurrence of disease or not.

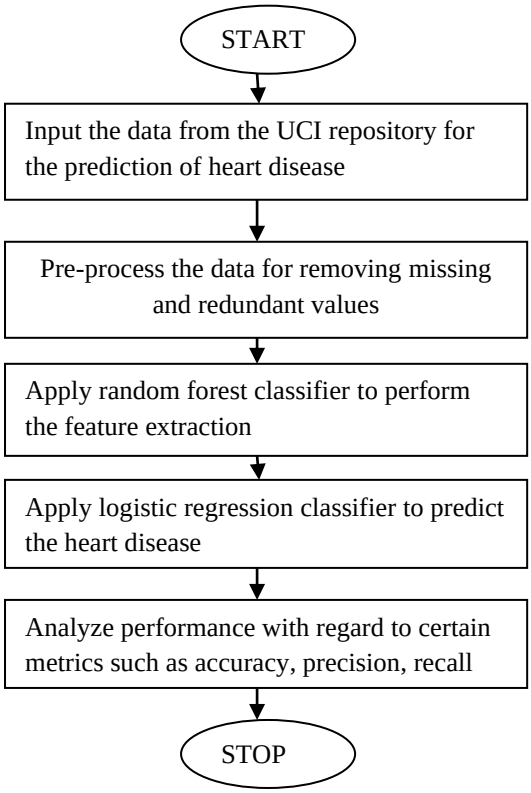

Figure 1: Proposed Methodology

Ⅳ. Result And Discussion

The clevaland is an extensively utilized dataset to predict the heart disease. There were 76 features utilized in this dataset, but only 14 are employed in the experimentation. These features are age, sex, cp, trestbps, chol. etc. and other predicted attributes.

This research work makes the implementation of several techniques and compares them to predict the coronary disease. A comparative analysis is conducted on DT (decision tree), NB (naïve bayes), MLP (Multilayer perception), Ensemble classification that integrates RF, NB and baysian belief models and suggested models with regard to accuracy, precision and recall.

Table 1: Analysis based on Accuracy

Models	Accuracy
Decision Tree	75.41 percent
Naïve Bayes	83.61 percent
Multilayer perceptron	83.61 percent
Ensemble Method	85.25 percent
Proposed Method	95.08 percent

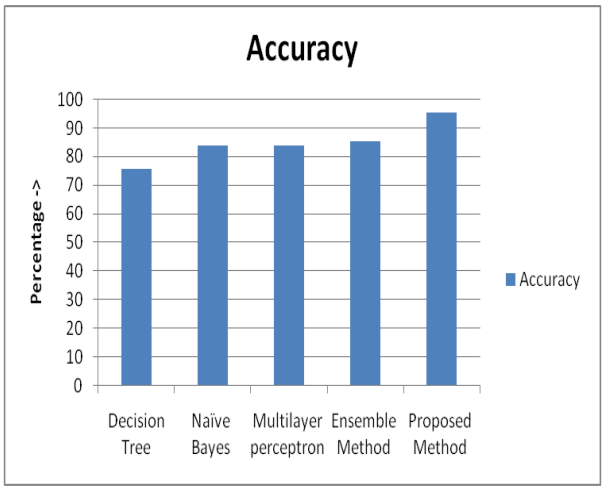

Figure 2: Analysis based on Accuracy

The figure 2 represents the comparison of several algorithms including DT, NB, MLP and EL with suggested models on the basis of accuracy. This reveals that the suggested model provides higher accuracy around 95% in comparison with other algorithms while predicting the heart disease.

Table 2: Analysis based on Precision

Models	Precision
Decision Tree	75 percent
Naïve Bayes	84 percent
Multilayer perceptron	85 percent
Ensemble Method	86 percent
Proposed Method	95 percent

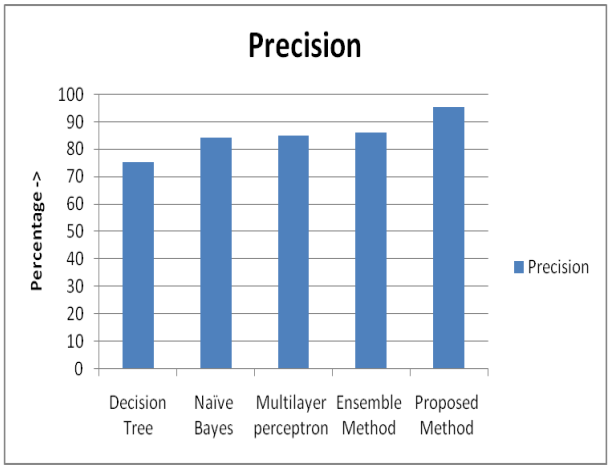

Figure 3: Analysis based on Precision

The figure 3 illustrates the comparison of several algorithms namely DT, NB, MLP, EL with the suggested models with regard to precision. This demonstrates the suggested system yields superior precision of 95% in comparison with other techniques for predicting the coronary disease.

Table 3: Analysis based on Recall

Models	Precision
Decision Tree	75 percent
Naïve Bayes	84 percent
Multilayer perceptron	84 percent
Ensemble Method	85 percent
Proposed Method	95 percent

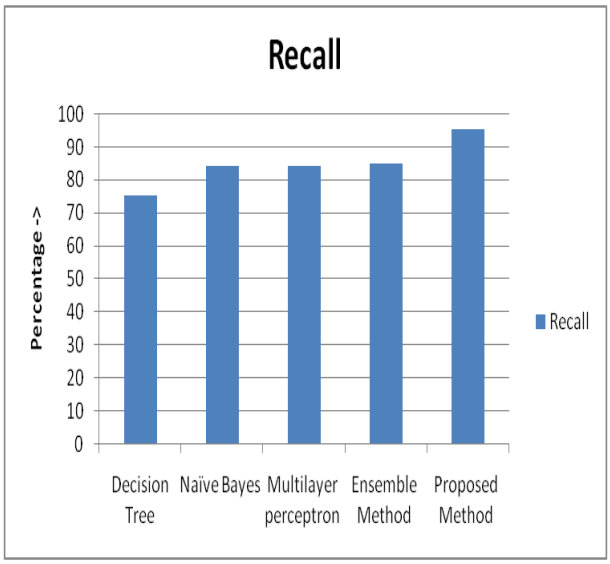

Figure 4: Analysis based on Recall

The figure 4 depicts that various models such as DT, NB, MLP and ensemble are compared with the suggested systems with regard to recall. This indicates that the suggested system provides greater recall up to 95% over other models to predict the heart disease.

Conclusion

This work summarized that it is challenging task to predict the heart disease as an enormous amount of attributes are present. The testing of several techniques such as DT, NB, MLP and ensemble classifier is executed in forecasting the coronary disease. A new framework is developed that integrates the RF (random forest) with LR (logistic regression) to predict the heart disease. The random forest algorithm is implemented to extract the attributes and the classification is carried out using LR. The precision, recall and accuracy obtained from the suggested system are calculated 95%.

References

[1] Sellappan Palaniappan and Rafiah Awang, “Intelligent Heart Disease Prediction System using Data Mining Techniques”, International Journal of Computer Science and Network Security, Vol. 8, No. 8, pp. 1-6, 2008.

[2] Franck Le Duff, Cristian Munteanb, Marc Cuggiaa and Philippe Mabob, “Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method”, Studies in Health Technology and Informatics, Vol. 107, No. 2, pp. 1256-1259, 2004.

[3] W.J. Frawley and G. Piatetsky-Shapiro, “Knowledge Discovery in Databases: An Overview”, AI Magazine, Vol. 13, No. 3, pp. 57-70, 1996.

[4] HeonGyu Lee, Ki Yong Noh and Keun Ho Ryu, “Mining Bio Signal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV”, Proceedings of International Conference on Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66, 2007.

[5] Kiyong Noh, HeonGyu Lee, Ho-Sun Shon, Bum Ju Lee and Keun Ho Ryu, “Associative Classification Approach for Diagnosing Cardiovascular Disease”, Intelligent Computing in Signal Processing and Pattern Recognition, Vol. 345, pp. 721-727, 2006.

[6] Latha Parthiban and R. Subramanian, “Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm”, International Journal of Biological, Biomedical and Medical Sciences, Vol. 3, No. 3, pp. 1-8, 2008.

[7] Niti Guru, Anil Dahiya and Navin Rajpal, “Decision Support System for Heart Disease Diagnosis using Neural Network”, Delhi Business Review, Vol. 8, No. 1, pp. 1-6, 2007.

[8] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, “Design And Implementing Heart Disease Prediction Using Naives Bayesian”, 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)

[9] Aditi Gavhane, Gouthami Kokkula, Isha Pandya, Prof. Kailas Devadkar, “Prediction of Heart Disease Using Machine Learning”, 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)

[10] Aakash Chauhan, Aditya Jain, Purushottam Sharma, Vikas Deep, “Heart Disease Prediction using Evolutionary Rule Learning”, 2018, 4th International Conference on Computational Intelligence & Communication Technology (CICT)

[11] C. Sowmiya, P. Sumitra, “Analytical study of heart disease diagnosis using classification techniques”, 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)

[12] Rashmi G Saboji, “A scalable solution for heart disease prediction using classification mining technique”, 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)

[13] Ankita Dewan, Meghna Sharma, “Prediction of heart disease using a hybrid technique in data mining classification”, 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)

[14] Aditi Gavhane, GouthamiKokkula, Isha Pandya, Prof. Kailas Devadkar, “Prediction of Heart Disease Using Machine Learning”, 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)

[15] M. A. Jabbar, Shirina Samreen, “Heart disease prediction system based on hidden naïve bayes classifier”, 2016 International Conference on Circuits, Controls, Communications and Computing (I4C)

[16] Purushottam, Kanak Saxena, Richa Sharma, “Efficient heart disease prediction system using decision tree”, 2015, International Conference on Computing, Communication & Automation

[17] Aakash Chauhan, Aditya Jain, Purushottam Sharma, Vikas Deep, “Heart Disease Prediction using Evolutionary Rule Learning”, 2018, 4th International Conference on Computational Intelligence & Communication Technology (CICT)

[18] C. Sowmiya, P. Sumitra, “Analytical study of heart disease diagnosis using classification techniques”, 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)

[19] Rashmi G Saboji, “A scalable solution for heart disease prediction using classification mining technique”, 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)