International Journal For Technological Research In Engineering
Volume 9, Issue 10, June-2022 *ISSN (Online): 2347 - 4718*

ANALYSIS & DESIGNING OF HEART DISEASE PREDICTOR

Anushka Jain, Ishika Anand, Prof. Gurpreet Kaur
[1,2]Students, [4]Assistant Professor
Department of Computer Science Engineering
MVSIT, Sonipat, India

1. ABSTRACT

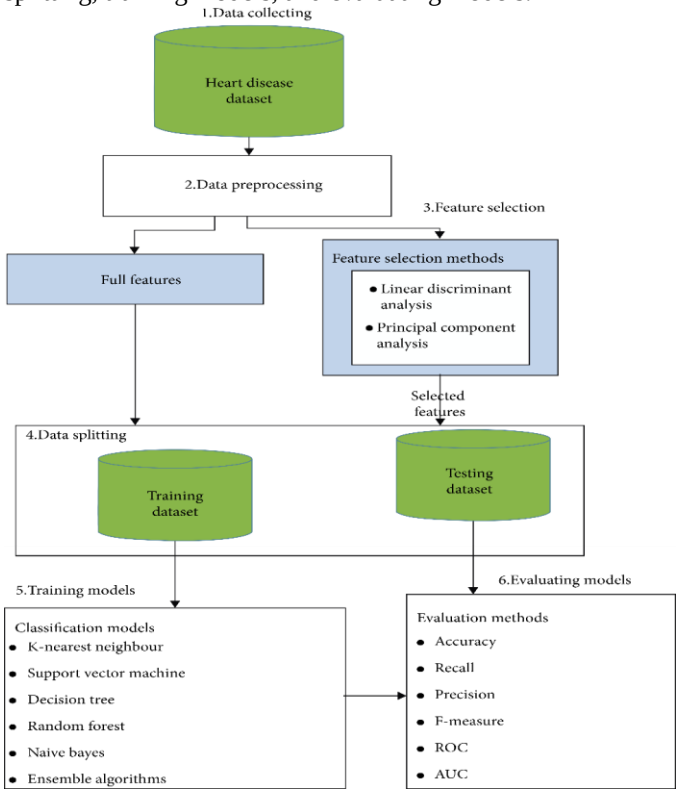
Heart-related diseases or Cardiovascular Diseases (CVDs) are the main reason for a huge number of death in the world over the last few decades and has emerged as the most life-threatening disease. Early and on-time diagnosing of this problem is very essential for preventing patients from more damage and saving their lives. So, there is a need fora reliable, accurate, and feasible system to diagnose such diseases in time for proper treatment. Among the conventional invasive-based techniques, angiography is considered to be the most well-known technique for diagnosing heart problems but it has some limitations. On the other hand, an intelligent computational predictive system is introduced for the identification and diagnosis of cardiac disease. In this study, various machine learning classification algorithms are investigated in order to remove irrelevant and noisy data from extracted feature space, four distinct feature selection algorithms are applied and the results of each feature selection algorithm along with classifiers are analysed
.

2. INTRODUCTION

Nowadays, the cardiac disease is one of the most critical problems relating to human safety. Heart diseases have occurred as one of the most prominent cause of death all around the world. According to World Health Organization, heart associated diseases are responsible for the taking 17.7 million lives every year, 31% of all global deaths. In India too, heart related diseases have become the top cause of death. Heart diseases have killed 1.7 million Indians in 2016, according to the 2016 Global Burden of Disease Report, released on September 15, 2017. Heart related diseases increase the outlay on health care and reduce the efficiency of an individual. Estimates made by the World Health Organization (WHO), suggest that India have lost up to $237 billion, from 2005- 2015, due to heart related or cardiovascular diseases. Thus, reasonable and accurate prediction of heart related diseases is very important. Medical organizations, all around the world, collect data on various health related issues. These data can be oppressed using various machine-learning techniques to gain useful understandings. But the data collected is very massive and, many a times, this data can be very noisy. These datasets, which are too devastating for human minds to comprehend, can be easily explored using various machine-learning techniques. Thus, these algorithms have become very useful, in recent times, to predict the presence or absence of heart related ailments accurately.

3. THE PROPOSED SYSTEM OF PREDICTING HEART DISEASE

The objective of the proposed system technique is to use ensemble techniques to improve the performance of predicting heart disease. Figure 1 describes the architecture of the proposed system. It is structured into six stages, including data collection, data pre processing, feature selection, data splitting, training models, and evaluating models.



3.1 DATA COLLECTION

The heart disease dataset is utilized for training and evaluating models. It consists of 303 records, 16 features, and one target column. The target column includes two classes: 1 indicates heart diseases, and 0 indicates nonheart disease.

3.2 DATA PREPROCESSING

The features are scaled to be in the interval [0, 1]. It is worth noting that missing values are deleted from the dataset.

3.3 DATA SPLITTING

In this step, the heart disease dataset is divided into a 80% training set and a 20% as the testing set. The training set is utilized for training the models, and the testing set is utilized to evaluate the models. Also, ninefold cross-validation is utilized in the training set.

3.4 TRAINING MODELS

Different types of machine learning algorithms: K-Nearest Neighbors, Logistic Regression, Random Forest, and Naïve Bayes are applied to classify heart disease. Also, two types of ensemble techniques: boosting and bagging are applied to classify heart disease:

- Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. In simple words, the dependent variable is binary in nature having data coded as either 1 or 0.
- Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset.
- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. It assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- 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.
- Ensemble techniques are methods that can be utilized to enhance the performance of a classifier. It is an effective classification method that combines a weak classifier with a strong classifier to improve the weak learner's efficiency. There two types of ensemble techniques: boosting and bagging.

1. Boosting means producing a model sequence that aims to correct the errors that have arisen in the models. The dataset is split into different subsets. The classification algorithm is then trained on a sample to create a series of average efficiency models Consequently, based on the previous model's elements not properly classified, new samples are produced. Then, by combining the weak models, the ensemble method increases its efficiency.
2. Bagging refers to taking a replacement training set with multiple subsets and training a model for each subset. The average of the forecast values of the sub models together are as stated by the final performance forecast. A voting procedure for each classification model is then performed as shown in pseudocode of bagging algorithm. Consequently, the classification outcome is determined based on the majority of the average values.

3.5 EVALUATING MODELS

Evaluation of the proposed model is performed focusing on some criteria, namely, accuracy, recall, precision, F-score, ROC, and AUC.

Accuracy is one of the most important performance metrics for classification. It is defined as the proportion between the correct classification and the total sample, as shown in the following equation:

$$\text{Acc} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}.$$

Recall is the small portion of sufficient instances over the overall quantity of applicable instances which have been recovered. The recall equation is shown as follows:

$$\text{recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})}.$$

Precision is identified as follows:

$$\text{precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})}.$$

The F-measure is often referred to as the F1-score as follows, and it measures the mean value of precision and recall:

$$\text{F} - \text{measure} = \frac{(2 * \text{precision} * \text{recall})}{(\text{precision} + \text{recall})}.$$

The receiver operating characteristic curve (ROC) is a graph illustrating the efficiency of a classification algorithm at all classification thresholds. Two parameters are shown in this curve: true positive and false positive. The area under the curve (AUC) is the indicator of a classifier's ability to differentiate among classes and is utilized as a ROC curve description. The greater the AUC is, the greater the model's efficiency is in differentiating between the positive and negative groups.

4. CONCLUSION

After experimenting with four binary classification machine learning algorithms i.e. Random Forest, K-Nearest Neighbors, Logistic Regression, and Naive Bayes, the algorithms that returned the most accurate heart disease predictions was Random Forest Regression algorithm.

Since, Random Forest Regression algorithm is capable of handling large datasets with high dimensionality and works with a goal of reducing the variance and thus enhances the accuracy of the model and prevents the overfitting issue, the model returned fewer false negatives, in other words: fewer false healthy diagnoses when the patients were actually sick. For this case study, I deemed that it was more dangerous to return a false negative, because the consequence could be that a sick patient does not receive the medical treatment they need.

Hence, our Random Forest Regression model worked well with enhanced performance and less errors.

