Ideation Report

A Method for Improving Prediction of Human Heart Disease

Research Goals and Objectives

The main goal of this research is to develop a heart disease prediction model with improved and enhanced accuracy. The specific objectives are to quickly identify new patients, reduce diagnostic time, reduce heart attacks, and save lives.

Methodology

we describe the proposed method and also explain that the method is defined by the subsequent steps,

(1) The first step is to select the dataset from the machine learning online repositories. There are many online repositories, such as the Cleveland heart disease dataset, Z-Alizadeh Sani dataset, StatLog Heart, Hungarian, Long Beach VA, and Kaggle Framingham dataset.

(2) In the second step, we refined and standardized the collected data sets. These datasets were not gathered in a controlled environment and had erroneous values. Hence, data preprocessing is an essential step for studying data and machine learning. Data normalization means when the risk factors of a dataset have different values. For example, Celsius and Fahrenheit are different measuring units of temperature. The

standardization of data means scaling the risk factors and assigning the values that show the difference between standard deviations from the mean value. It rescales the risk factor value to improve the performance of machine learning classifiers with a standard deviation (σ) of 1 and a mean (μ) of 0. The mathematical form of standardization is given by (1).

Standardization of x

$$\frac{\text{X - Mean of X.}}{\text{Standard deviation of X}} \qquad (1)$$

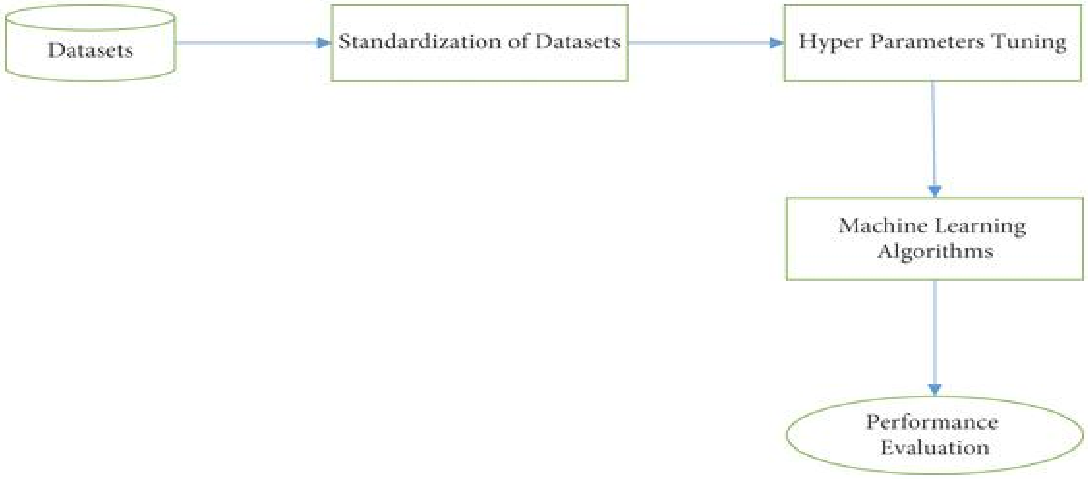
(3) In this step, hyperparameter tuning is performed to select the best value for the hyper parameters and get high accuracy. For this purpose, we used the GridSearchCV method. Before applying machine learning classifiers, we adjust the hyper parameters values of machine learning classifiers to increase their performance. The Scikit-learn GridSearchCV class’s fit approach provides a grid of tuning classification algorithms. It allows each machine learning algorithm to be trained and its corresponding hyper parameters to be adjusted in a single consistent environment. The entire training dataset is then used to achieve a precise model once the adequate values for hyperparameters have been achieved. The 10-fold CV is used to identify the optimum values for the adjustable hyperparameters based on the training dataset. During the CV process, the adjusted hyper parameter values are provided to achieve the overall best classification accuracy.

(4) The fourth step is to apply the machine algorithms (i.e., AdaBoost, logistic regression, extra tree, multinomial Naïve Bayes, support vector machine, linear discriminant analysis, classification and regression tree, random forest, and XGBoost) to the dataset obtained from step 2.

(5) In this step, the prediction model's performance is evaluated using different parameters, such as accuracy, precision, recall, and F-measure. The model that gives the highest prediction accuracy, precision, recall, and F-measures is selected. The accuracy metric assesses the precision or correctness of a machine learning or classifier model's predictions. Mathematically, it is given by equation (2).

Accuracy

$$= \frac{\text{True Positive (TP) + True negative (TN).}}{\text{TP+TN+false negative (FN)+ false positive(FP)}} \qquad (2)$$



Precision measures the predicted positive instances that are true/real positives. Mathematically, it is given by (3).

Precision

$$= \frac{\text{TP (true positives).}}{\text{TP (True Positives) + FP (False positives)}} \qquad (3)$$

Recall evaluates the analysis of the total number of true/real positive instances as affected by the total number of false negative instances. Mathematically, it is given by (4).

Recall

$$= \frac{\text{TP (true Positives).}}{\text{TP (true Positives) + FN (false negatives)}} \qquad (4)$$

An F-Measure is a harmonic mean of precision and recall. It takes the equilibrium between precision and recall, and mathematically, it is given by (5).

F - measure

$$= 2 \times \frac{\text{Precesion} \times \text{Recall.}}{\text{Precession + Recall}} \qquad (5)$$

Experimental Results and Discussion

we discuss our experimental results. We collected the given dataset from repository and refined and standardized it. After standardization, we performed hyperparameter tuning and applied machine learning classifiers. All the classifiers are trained and tested using 10-fold cross-validation. The accuracy of classifiers is also analyzed before and after standardized datasets. For evaluation purposes, the accuracy of the selected classifiers is plotted.

Conclusions

For future work, we plan to use XGBoost for heart disease prediction in children and compare if better accuracy can be achieved. If features are properly managed, then there will be significant performance in the classification of heart disease prediction. In future studies, the outcomes of our proposed methods will serve as the standard performance results on heart disease.