Heart Disease Prediction Using Machine Learning Algorithm

1st Kibtia Chowdhury
dept. of CSE
United International University (UIU)
Bangladesh
kchowdhury211056@mscse.uiu.ac.bd

2nd Mohammad Nurul Huda
dept. of CSE
United International University (UIU)
Bangladesh
mnh@cse.uiu.ac.bd

I. Background

Heart attack is mentioned to be the main reason for death globally, following to the World Health Organization (WHO) the fatality value due to heart attack is around 17.7 million in 2015 and is estimated to increase by 2021. Each year almost 20 million public die, introducing heart attack as a passing cause of death. Coronary heart disease is the term that recites what happens when your heart's blood execution is blocked or divided by a build-up of fatty elements in the coronary arteries. The major symptoms of coronary heart disease are chest pain (angina),the narrowness of breath, the pain completely the body,feeling weak,feeling ill.Although not everyone has similar symptoms and some people may not have any before coronary heart disease is diagnosed. Investigators and clinicians are reading various heart disease datasets with various machine learning classification algorithms and feature selection procedures to get up practical predictions for heart disease. In this paper, the main motives are to build a model with appear performance and identify which selected features to play a key role in predicting heart disease by using project heart datasets from UCI with Python tool.

II. Objective

Heart attack has been one of the main reasons for ruin global. The heart attack diagnosis has been costly currently, so it is Important to predict the danger of getting a heart attack with chosen features. The feature selection methods could be used as costly techniques to decrease the cost of diagnosis by selecting the important signs. This study aims to predict the classification model and know which selected features play the main role in the prediction of heart disease by using Medical report heart a data set. The accuracy of the random forest algorithm performs the best.

III. System Diagram

In this diagram, I collected a dataset from UCI. After data collected I preprocessing and checked missing values. Then I have divided the dataset 20% for Train and test. After train test split using machine learning algorithm to get a prediction disease.

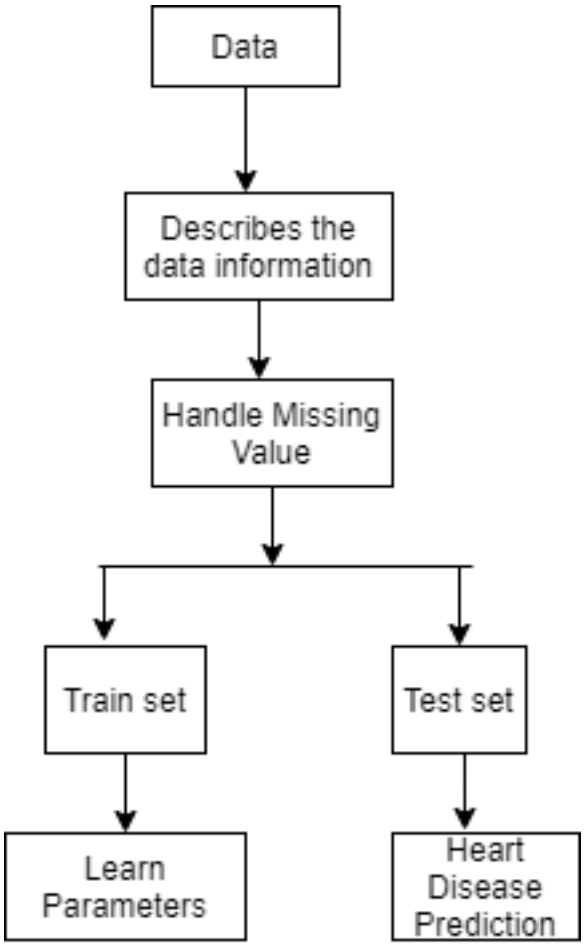

Fig. 1. Heart Disease System Diagram

IV. Dataset

In this study, these heart disease datasets (Cleveland project heart disease) are collected from publicly available source UCI machine learning repository webpage. This dataset consists of 303 instances with 14 attributes.Here sex 0 mean male and 1 means female. And output 0 means do not found heart disease and 1 means found heart disease. Age between 29-77. These characteristics of this heart disease dataset are shown in Table

TABLE I
DESCRIPTION OF ATTRIBUTES IN THE DATASET

Features	Type	Description, Value
Age	Continuous	Patients age , 29 to 77
Sex	Discrete	1 = male; 0 = female
Chest Pain	Discrete	Value 1: typical angina; Value 2: atypical angina; Value 3: non-anginal pain; Value 4: asymptomatic
Trtbps	Continuous	Resting blood pressures of patients measured in mm Hg on admission to the hospital 80-200.
Chol	Continuous	Patient serum cholesterol measured in mg/dl. 85, 100 - 200 - 394, 400-603
Fbs	Discrete	Fasting Blood Sugar, 0 = false (FBS <120 mg/dl) 1 = true (FBS >120 mg/dl)
Restecg	Discrete	Value 0: normal; Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05 mV); Value2: showing probable or definite left ventricular hypertrophy by Estes' criteria
Thalachh	Continuous	Patient maximum heart rate achieved.60-202, Low: below 50, Normal:51-119, High: 120-202
Exng	Discrete	Exercise induced angina. 1 = yes; 0 = no
Oldpeak	Continuous	ST depression made by exercise relative to rest -2.6 to -0.1, 0, 0.1 to 6.2, 120
Slp	Discrete	Peak exercise slope measure, Value 1: upsloping, Value 2: flat, Value 3: down sloping
Caa	Discrete	number of major vessels colored by flourosopy (0-3)
Thall	Discrete	Patient heart rate, 3 = normal; 6=fixed defect; 7 = reversable defect
Output	Discrete	0: No heart disease; 1: Heart disease

1.

V. MACHINE LEARNING ALGORITHM

A. K-Nearest Neighbor's Algorithm (KNN)

It is a normal classifier that cannot fist responses, free to performance and realize, demands short training time, and the whole training set is used for prediction. The K-Nearest Neighbors (K-NN) with weighting parameter has been used for the prediction of heart disease. Among 13 attributes mentioned in the UCI heart disease dataset, the selection of 8 attributes due to simple measurements has been taken into consideration for this study.

B. Naïve Bayes (NB)

The naive Bayes classifier is a generative model for classification. Before the advent of deep learning and its easy-to-use libraries, the Naive Bayes classifier was one of the widely deployed classifiers for machine learning applications. Despite its simplicity, the naive Bayes classifier performs quite well in many applications.

$$P(A|B) = P(B|A) * P(A)/P(B) \tag{1}$$

C. Support Vector Machine (SVM)

SVM works both for classifi

cation and regression challenges. It is mostly used to maximize the margin of a classifier. Support Vector is the data point closer to the hyperplane by influencing the position and orientation of the hyperplane. SVM works by separating the data into distinct classes with a linear hyperplane. That is why it becomes difficult to gather data that is completely separable.

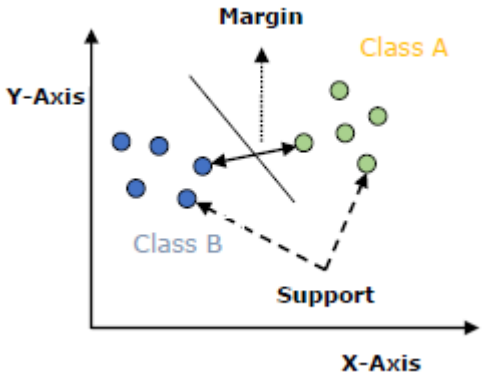

Fig. 2. SVM Hyperplane

D. Ada-Boost Algorithm (ABA)

An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{2}$$

E. Decision Tree Algorithm (DTA)

Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. The decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.

$$Entropy = -\sum_{j=1}^{m} pijlog2pij \tag{3}$$

F. Random Forest Algorithm (RFA)

Random forest extracts a section of decision trees at training time and output the mode prediction of the categories for classification and the mean prediction for regression. The various models in the data are valued by the decision tress. The class prediction is based on the main opinion for classification. The random forest with 5 fold cross validation along with feature selection methods such as chi square and genetic algorithm was used to for the prediction of heart disease.

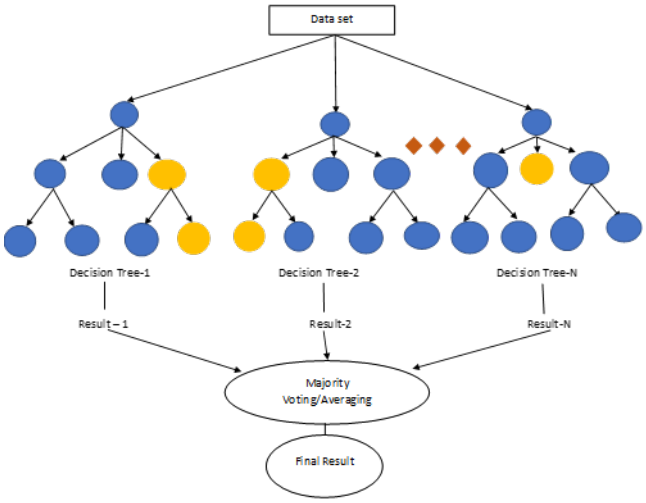

Fig. 3. RFA Classifier

VI. Methodology

A. Data Preprocessing

In the data preprocessing stage, missing values are replaced with mode value based on the particular datasets source. Second, taking into consideration that heart disease patients might have high values of respective attributes (i.e., referred as outliers in the dataset) are not removed.

B. Data Analysis

After data collected to make a classification model, the dataset with 14 attributes is partitioned into training and testing data with a percentage split of 80–20%. The sklearn model selection is used for data splitting, pre-processing steps, classification algorithms such as K-Nearest Neighbor, Support Vector Machine, Random Forest, Naïve Bayes,5 K-fold are evaluated based on the mentioned training and testing data. The different feature selections methods such as correlation matrix.The variable importance estimations such as regression method to calculate variable. The performance of a model on test data is calculated by accuracy, precision,sensitivity/recall,f1-score and specificity in Python tool. Sensitivities and specificity measure the true positives (risk class) and the true negatives (normal class) respectively. Thus the predictive capabilities of the classifiers are measured by sensitivity and specificity values.

VII. Experimental Results

To the best of author knowledge, there is no studies addressed with the combined dataset. Python tool is selected to understand which classification model has a better performance this dataset with 13 features. In this dataset, I got mean, standard deviation, max value, min value, 25%, 50%,75% by using df.describe() function. And I checking the missing value, there is no missing value.The performance measures are in accordance with the accuracy of each classification algorithm. This dataset classification algorithms used, the highest accuracy has been observed in random forest with a 0.80%, and average accuracies (0.83-0.851%) and K-Fold used, the highest accuracy has been observed in SVM with a 0.83% shown in Table 2 and Table 3.

TABLE II
Experimental Results

Algorithms	Train	Test	Precision	Recall	F1 Score
NB	0.77	0.65	0.77	0.57	0.66
DTA	1.0	0.70	0.81	0.63	0.71
SVM	0.86	0.78	0.96	0.66	0.78
RFA	0.94	0.80	0.96	0.69	0.80
ABA	0.97	0.75	0.88	0.66	0.75
KNN	0.79	0.57	0.65	0.57	0.61

TABLE III
5 K-Fold Results

Algorithms	5 K-Fold
NB	0.75
DTA	0.76
RFA	0.82
ADA	0.79
KNN	0.64
SVM	0.83

VIII. Conclusion

This is informed such the accuracy regarding the model builds on the database, prepossessing, analytical tools, and techniques.The modern study displays it is main to select least and great signs to gain the representation when weighed to the use of every feature from the data set.

References

[1] Alexander, Cheryl Ann, and Lidong Wang. "Big data analytics in heart attack prediction." J Nurs Care 6, no. 393 (2017): 2167-1168.

[2] Masethe, Hlaudi Daniel, and Mosima Anna Masethe. "Prediction of heart disease using classification algorithms." In Proceedings of the world Congress on Engineering and computer Science, vol. 2, pp. 22-24. 2014.

[3] Takci, Hidayet. "Improvement of heart attack prediction by the feature selection methods." Turkish Journal of Electrical Engineering Computer Sciences 26, no. 1 (2018): 1-10.

[4] Manikandan, Sushmita. "Heart attack prediction system." In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS), pp. 817-820. IEEE, 2017.

[5] Salman, Issam. "Heart attack mortality prediction: an application of machine learning methods." Turkish Journal of Electrical Engineering Computer Sciences 27, no. 6 (2019): 4378-4389.

[6] Srinivas, K., G. Raghavendra Rao, and A. Govardhan. "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques." In 2010 5th International Conference on Computer Science Education, pp. 1344-1349. IEEE, 2010.

[7] Jabbar, M. A., Priti Chandra, and B. L. Deekshatulu. "Cluster based association rule mining for heart attack prediction." Journal of Theoretical and Applied Information Technology 32, no. 2 (2011): 196-201.

[8] Ahmed, Fizar. "An Internet of Things (IoT) application for predicting the quantity of future heart attack patients." International Journal of Computer Applications 164, no. 6 (2017).