A Study on Heart Disease Prediction using Different Classification Models based on Cross Validation Method

Anirban Ghosh
Department of Statistics
University of Kalyani
West Bengal, India

Sushovon Jana
Department of Applied Statistics
Maulana Abul Kalam Azad University of Technology
West Bengal, India

***Abstract*— Heart disease causes the greatest number of deaths in world. A large number of people cannot recognize it in early stage. In this study, our goal is to find a good model for prediction of heart disease. The dataset consists of 918 observations, out of which, 508 have heart disease and 410 are normal. To find the best model, we compare five classification models i.e., Logistic Regression model, Support Vector Machine, Random Forest model, Naïve Bayes classifier and Linear Discriminant Analysis to predict if a person has heart disease or not. We compare the models using 10-fold cross-validation method with three repetitions. The study proposes Random Forest model as the most appropriate predictor of heart disease mean accuracy of 86.93%, which is the highest among all models. The slope of the peak exercise ST segment is the most important subject to predict heart disease. Old peak, type of chest pain and maximum heart rate achieved are also important for predicting heart disease.**

Keywords*— Heart disease; Logistic Regression; Support Vector Machine; Random Forest Model; Naïve Bayes Classifier; Linear Discriminant Analysis; Cross Validation

I. INTRODUCTION

According to World Health Organization (WHO), heart disease is the no. 1 cause of death in world. It is responsible for 16% of total deaths in world [1]. Since 2000, the largest increase in deaths has been for heart disease, rising by more than 2 million to 8.9 million deaths in 2019 [1]. Also in India, heart disease is the leading cause of death. According to Global Burden of Disease, 24.8% of all deaths in India is due to heart disease [2]. Heart disease may happen for various reasons. Most common heart disease is coronary artery disease, which happens due to building up of fatty plaques in arteries (atherosclerosis). Heart disease can show various symptoms like chest pain, suffocation, weakness and many more according to the type of heart disease. It can be prevented by maintaining proper diet, following healthy lifestyle, doing regular exercise etc. Though a great amount of statistical and scientific researches is being done, heart disease continues to be the largest killer of world. By early detection of heart disease and proper treatment, chance of survival of a heart disease patient can be increased.

We have analyzed a dataset of 918 observations containing 11 independent variable and whether there is heart disease or not. Through VIF calculation and Principal Component Analysis, we have found that no significant multicollinearity exists among the variables. So, we have fitted some classification models to predict heart disease of a person and compared the accuracy of different models. We have used R programming language as a tool for these purposes.

II. RELATED WORKS

In this study, a comparative analysis has been done among various Machine Learning classification algorithms. Random Forest becomes the best model among these models.

Reference [4] used Random Forest model for prediction of heart disease. They obtained accuracy of 86.9% with sensitivity value 90.6% and specificity value 82.7%.

Reference [5] authors have proposed a Logistic Regression model for Diabetes prediction by integrating PCA and K-means techniques. This model shows high accuracy of 97.40%.

Reference [6] obtained the slope of peak exercise ST segment, old peak, chest pain type etc. as significant subject for predicting heart disease. This study provided a significant contribution in computing strength scores with significant predictors in heart disease prediction.

Reference [7] compared four classification algorithms i.e., Naïve bayes, random forest, Linear regression, Decision tree to predict the heart disease. Among these algorithms Random Forest gives best accuracy of 90.16% compared to other algorithms.

Reference [8] authors proposed hybrid Random Forest model for prediction of cardiovascular disease. The study showed accuracy of 88.7% for prediction of CVD using the proposed model.

III. METHODOLOGY

This section includes the dataset description, data pre-processing techniques and classification algorithms. R programming language is used for analyzing the data.

A. Dataset Description

The heart disease dataset collected from [3] is used for this study. The dataset is of 918 observations and contains 11 independent variable and a categorical variable, whether there exists heart disease or not, as target variable. The variables of the dataset are

- Age: Age of the patient [Years]
- Sex: Sex of the patient [M: Male, F: Female]

- ChestPainType: Chest Pain Type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
- RestingBP: Resting Blood Pressure [mm Hg]
- Cholesterol: Serum Cholesterol [mm/dl]
- FastingBS: Fasting Blood Sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
- RestingECG: Resting Electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
- MaxHR: Maximum heart rate achieved [Numeric value between 60 and 202]
- ExerciseAngina: Exercise-induced Angina [Y: Yes, N: No]
- Oldpeak: Oldpeak = ST [Numeric value measured in depression]
- ST_Slope: The slope of the peak exercise ST segment [Up: up sloping, Flat: flat, Down: down sloping]
- HeartDisease: Output class [1: Heart disease, 0: Normal]

B. *Data Preprocessing*

Some pre-processing is required to make the data usable. We need to clean the data and code the attributes to numbers to fit classification models.

Firstly, we can see that there are no missing values in our data.

Secondly, there are some 0 values in the columns RestingBP and Cholesterol. But Resting Blood Pressure and Serum Cholesterol of a person can never be 0. So, these are bad values. These zeros are replaced with median values of the corresponding columns. Also, there are some negative values in the column Oldpeak. These negative values are converted to positive.

It is found from the summary of the raw data that about 77% values of the column FastingBS is 0. So, this column will not impact greatly on classification. So FastingBS column is dropped.

Values of some columns are categorical variables. So, we code them into numbers. The changes are shown in Table III.

Now the dataset is ready for analysis.

C. *Multicollinearity*

Multicollinearity is a statistical measure which measures the inter-correlations between the independent variables of the data. For classification, non-existence of multicollinearity is required. If multicollinearity exists, skewed or misleading results can be obtained, when we study the power of each variable independently to predict or interpret the dependent variable using a statistical model. In presence of multicollinearity, to find out the effect of independent variables in a statistical model, wider confidence interval with less accurate probabilities can be produced. One might not able to trust the p-values to identify independent variables that are statistically significant.

TABLE I. SUMMARY OF CONTINUOUS VARIABLES OF RAW DATA

	Age	RestingBP	Cholesterol	MaxHR	Oldpeak
Count	918	918	918	918	918
Min.	28.00	0.0	0.0	60.00	2.6000
1st Qu.	47.00	120.0	173.2	120.0	0.0000
Median	54.00	130.0	223.0	138.0	0.6000
Mean	53.51	132.4	198.8	136.8	0.8874
3rd Qu.	60.00	140.0	267.0	156.0	1.5000
Max.	77.0	200.0	603.0	202.0	6.2000

TABLE II. SUMMARY OF CATEGORICAL VARIABLES OF RAW DATA

Sex	F: 193 M: 725
ChestPainType	ASY: 496 ATA: 173 NAP: 203 TA: 46
FastingBS	0: 704 1:214
RestingECG	LVH: 188 Normal: 552 ST: 178
ExerciseAngina	N: 547 Y: 371
ST_Slope	Down: 63 Flat: 460 Up: 395
HeartDisease	0: 410 1: 508

TABLE III. LIST OF CODING INTO NUMERIC VALUES

Column Name	Actual Value	Coded Value
Sex	‘M’	1
	‘F’	2
ChestPainType	‘ATA’	1
	‘NAP’	2
	‘ASY’	3
	‘TA’	4
RestingECG	‘Normal’	1
	‘ST’	2
	‘LVH’	3
ExerciseAngina	‘Y’	1
	‘N’	0
ST_Slope	‘Down’	-1
	‘Flat’	0
	‘Up’	1

D. *Variance Inflation Factor (VIF)*

Variance Inflation Factor or VIF is a measure of amount of multicollinearity in a dataset. Mathematically, VIF is measured by

$$\mathrm{VIF}_j = \frac{1}{1-\mathrm{R}_j^2}, \qquad (1)$$

where R_j^2 is the multiple correlation coefficient between jth and other independent variables. A large value of VIF indicates high existence of multicollinearity in the variables. Generally, we consider existence of multicollinearity if VIF is greater than 5 or 10, according to the situation. VIF value of less than 5 will generally be considered as non-existence of multicollinearity.

E. Principal Component Analysis (PCA)

PCA is a dimension reduction technique for the data, which has a large numbers of predictor variables. PCA is mainly used for two purposes- dimensionality reduction and checking existence of multicollinearity. Using principal component analysis if number of variables of the datasets can be reduced i.e., if the variance of many variables can be explained by some few principal components, then it can be concluded that multicollinearity should exist there.

F. Logistic Regression

Logistic regression is used when the objective is to classify the target variable into two or more categories. In binary Logistic regression model, the target variable is classified into two classes i.e., 0 and 1, which in our case refers to negative or positive respectively for heart disease. For fitting Logistic regression, the Sigmoid function is used to estimate the probability of the data point belonging to the positive class.

G. Support Vector Machine (SVM)

SVM is a supervised machine learning algorithm that classifies cases by finding a separator. SVM first maps data to a high-dimensional feature space so that data points can be categorized even when the data points are not linearly or otherwise separable. Mapping data in higher dimensional space is called Kernelling. The mathematical function used for mapping is called a Kernel function. There are various types of Kernel functions. In this study, Radial Kernel is used to classify heart disease.

H. Random Forest Model

Random Forest method is a supervised machine learning algorithm for classification. It constructs multiple decision trees at a time. The decision is made on the majority of the decisions in the decision trees. The advantage of random forest over decision tree is random forest is free from the problem of high bias and low variance of decision tree. It also solves the overfitting problem of decision tree. Another advantage of using Random Forest model as a classifier is we get variable importance, which helps to understand the impact of various variables on classification, as an output.

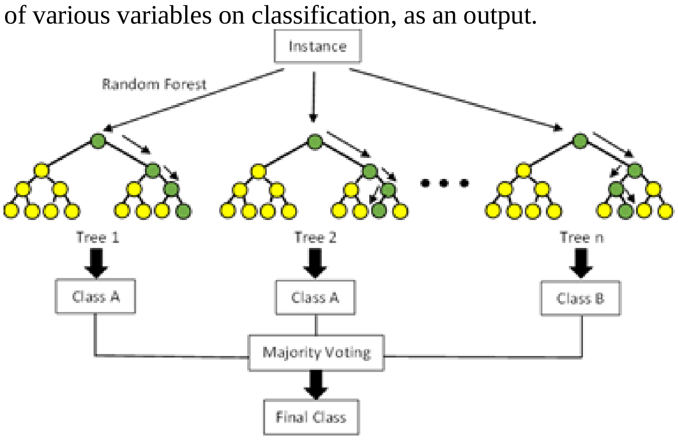

Fig. 1. Simplification of Random Forest Model

I. Naïve Bayes Classifier

Naïve Bayes classifiers are a family of simple 'probabilistic classifiers' based on applying Bayes' theorem with strong independence assumptions between the features. This model is used for binary classification, text classification, spam filtration, sentiment analysis, recommendation system etc. The Bayes' theorem is

$$P(y|\mathbf{X}=\mathbf{x}) = \frac{P(\boldsymbol{X}|y) * P(y)}{P(\boldsymbol{X})}. \tag{2}$$

J. Linear Discriminant Analysis (LDA)

Linear discriminant analysis is used as a tool for classification, dimension reduction and data visualisation. Despite its simplicity, LDA often produces powerful, reasonable and interpretable classification results. To incorporate classification by LDA, we consider a random variable $\mathbf{X}$ comes from one of the K classes with density $f_k(\mathbf{x})$ on $\mathbb{R}^p$. A discriminant rule tries to divide the data space into K disjoint regions $\mathbb{R}_1$, $\mathbb{R}_2$, …, $\mathbb{R}_K$ that represent all classes. Now $\mathbf{x}$ to class j is allocated if $\mathbf{x}$ is in region j following Bayesian rule or Maximum Likelihood rule according to the class prior probabilities are assumed or not respectively.

K. Cross Validation

Cross-validation is a resampling method that uses different partitions of the data to test and train a model on different iterations. It is mainly used in setting where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. At first, the shuffled dataset should be split into k groups. Then for each iteration, each of the k groups is to be considered as test set and the model should be trained over the remaining k-1 groups. Then we should summaries the outputs.

L. Performance Metrics

TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative.

$$\text{Precision} = TP / (TP + FP) \tag{3}$$

$$\text{Recall} = TP / (TP + FN) \tag{4}$$

$$\text{F-Score} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \tag{5}$$

IV. EXPERIMENTAL RESULTS

A. Calculation of VIF

The Variance Inflation Factors of the independent variables of our dataset are given in Table IV.

B. Analyzing Principal Components

Analyzing PCA, obtained results are shown in Table V.

C. Train-Test Splitting

We split our dataset into train set and test set. 80% of total data is used to train the model and remaining rows are used to test the performance of the model. There are 734 and 184 observations respectively in the train and test set.

D. Performance of models using Test Data

Five proposed models i.e., Logistic Regression, SVM, Random Forest model, Naïve Bayes Classifier and LDA are fitted to the train set. Then the observations of test set are predicted using these models. Thus, obtained performance metrices for each model are shown in Table VI.

E. Accuracy Obtained from Cross Validation

We compared the models using resampling method. 10-Fold Cross-validation is used here. We repeated 10-fold cross-validation 3 times. So, total number of resamples is 30.

Accuracy results obtained from cross-validation is shown in Table VII.

F. *Variable Importance*

From Random Forest model, variable importance is obtained in terms of Mean Decrease in Gini coefficient. It is produced simultaneously during training of the model. Mean Decrease in Gini coefficients for each predictor variable obtained from the Random Forest model fitted to the training data is shown in Table VIII.

TABLE IV. VARIANCE INFLATION FACTORS OF INDEPENDENT VARIABLES

Variables	VIF
Age	1.361663
Sex	1.092017
ChestPainType	1.258605
RestingBP	1.100360
Cholesterol	1.038561
RestingECG	1.090604
MaxHR	1.428407
ExerciseAngina	1.455541
Oldpeak	1.539348
ST_Slope	1.622914

TABLE V. INFORMATION OF PRINCIPAL COMPONENTS

	Standard Deviation	Proportion of Variance	Cumulative Proportion
PC1	1.829433	0.304260	0.304260
PC2	1.14266	0.11870	0.42295
PC3	1.023245	0.095180	0.518140
PC4	0.9901791	0.0891300	0.6072700
PC5	0.9213231	0.0771700	0.6844400
PC6	0.908279	0.075000	0.759440
PC7	0.8358615	0.0635100	0.8229500
PC8	0.7840512	0.0558900	0.8788400
PC9	0.7166974	0.0467000	0.9255300
PC10	0.6671866	0.0404700	0.9660000
PC11	0.6115684	0.0340000	1.0000000

TABLE VI. PERFORMANCE METRICS PREDICTING OBSERVATIONS OF TEST SET

Model	Accuracy	Precision	Recall	F-Score
Logistic Regression	0.832	0.821	0.790	0.805
Support Vector Machine	0.832	0.821	0.790	0.805
Random Forest Model	0.864	0.859	0.827	0.843
Naïve Bayes Classifier	0.832	0.798	0.827	0.812
Linear Discriminant Analysis	0.832	0.812	0.802	0.807

TABLE VII. ACCURACY FROM CROSS VALIDATION

Model	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
LR	0.7826	0.8179	0.8470	0.8479	0.8767	0.9239
SVM	0.7935	0.8352	0.8478	0.8522	0.8767	0.9130
RF	0.7717	0.8478	0.8688	0.8693	0.8913	0.9239
NB	0.7609	0.8261	0.8478	0.8417	0.8587	0.9022
LDA	0.7826	0.8175	0.8525	0.8486	0.8696	0.9348

TABLE VIII. VARIABLE IMPORTANCE

Variables	Mean Decrease in Gini
ST_Slope	88.109127
Oldpeak	45.819191
ChestPainType	44.594254
MaxHR	41.805309
Age	32.061294
Cholesterol	31.044686
ExerciseAngina	27.399798
RestingBP	26.154423
Sex	13.510075
RestingECG	8.977059

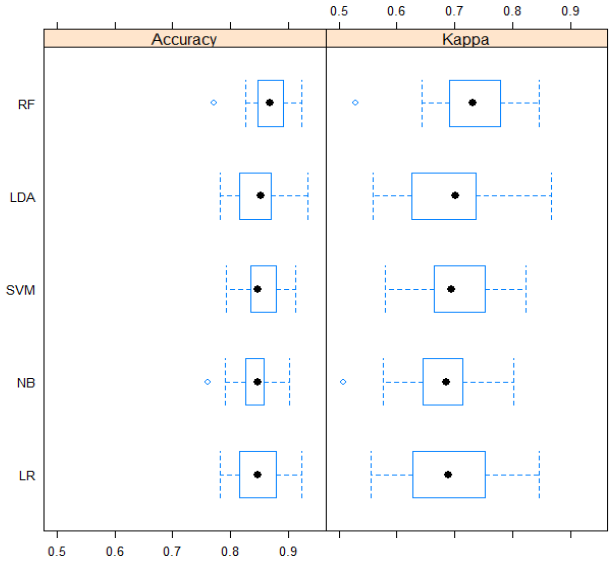

Fig. 2. Boxplot for Results of Cross Validation

V. DISCUSSION

From Table IV, it is seen that the VIFs are very close to 1. So, there is no significant multicollinearity in the data. Also, Table V shows that to explain 95% variance, 10 out of 11 principal components is required. So, no significant dimension reduction is possible. This also indicates non-existence of multicollinearity, which supports the information obtained from the VIF values. So, we were good to fit various classification models to predict heart disease to the dataset.

We got an accuracy of 86.4% when Random Forest model is used to predict test data. Resampling gives mean and median accuracy of 86.93% and 86.88% respectively when 10-fold cross-validation is used with 3 repetitions.

From Random Forest model we get the importance of various variables to make predictions. We get the value of mean decrease in Gini coefficient of ST_Slope as 88.109, which is the highest among all independent variables. This value is lowest for RestingECG which is 8.977.

VI. CONCLUSION

From Table VI, it can be concluded that Random Forest model gives the best prediction of existence of heart disease when predictions are made using test data. From Table VII and Fig.2, it is seen that resampling also supports the fact of the Random Forest being the best model out of our experimented models. Rest of the models give more or less similar performance.

From the variable importance (Table 6) obtained from the Random Forest model, it can be interestingly noted that ST_Slope is the most important factor for prediction of heart disease. Oldpeak, ChestPainType, MaxHR are the next important variables respectively with close importance value. RestingECG and Sex are the least important variables.

ACKNOWLEDGMENT

Primarily the authors would like to thank the Almighty for being able to complete this work successfully. The authors would like to express their sincere gratitude to the respected professors of Department of Statistics, University of Kalyani and Department of Applied Statistics, MAKAUT WB for their cooperation to this project.

The authors would like to extend gratitude towards Kaggle website and creator of the dataset for keeping the dataset open-accessed. The data set is the spine of the study.

The authors would like to thank their parents, family members and friends for continuous support in various field of the study.

REFERENCES

[1] https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] https://www.downtoearth.org.in/blog/health/india-s-burden-of-heart-diseases-study-says-elderly-women-more-at-risk-74993

[3] https://www.kaggle.com/fedesoriano/heart-failure-prediction

[4] M. Pal and S. Parija, “Prediction of Heart Diseases using Random Forest”, Journal of Physics: Conference Series 1817 012009, 2021

[5] C. Sh. Zhu, C. U. Idemudia and W.F. Feng, “Improved Logistic Regression model for Diabetes prediction by integrating PCA and K-means techniques”, Informatics in Medicine Unlocked 17 (2019) 100179

[6] A. Yazdani, K.D. Varathan, Y.K. Chiam, A.W. Malik and W.A.W. Ahmad, “A novel approach for Heart Disease prediction using Strength Scores with significant predictors”, BMC Medical Informatics and Decision Making 21 (2021)

[7] A. Rajdhan , A. Agarwal , M. Sai , D. Ravi and P. Ghuli, “Heart Disease prediction using Machine Learning”, International Journal of Engineering Research and Technology 09 (2020)

[8] A. Rairikar, V. Kulkarni, V. Sabale, H. Kale and A. Lamgunde, “Heart Disease prediction using Data Mining techniques”, International Conference on Intelligent Computing and Control (I2C2), 2017