A STUDY ON HEART DISEASE PREDICTION USING DIFFERENT CLASSIFICATION MODELS BASED ON CROSS VALIDATION METHOD

By

ANIRBAN GHOSH

Department of Statistics

University of Kalyani

Under The Supervision of

Dr. Sushovon Jana

Assistant Professor

Maulana Abul Kalam Azad University of Technology

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. Through VIF calculation and Principal Component Analysis, we show that there is no multicollinearity in the data. Then 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. 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.

ACKNOWLEDGEMENT

Primarily I would thank the Almighty for being able to complete this project successfully. I would like to express my sincere gratitude to my project supervisor, **Dr. Sushovon Jana** for staying beside me through the whole course of the project. Without his help, knowledge, patience, practical advice and continuous insightful feedbacks, this project would not have been possible. I would extend my thanks to the respected professors of Department of Statistics, University of Kalyani for their valuable suggestions.

I would like to thank my parents and my friends for support in various fields of this project.

CONTENT

1. 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. 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. 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.

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.

2. DATA DICTIONARY

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.

2.1 Source:

The dataset is downloaded from https://www.kaggle.com/fedesoriano/heart-failure-prediction. This dataset was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes. Every dataset used can be found under the Index of heart disease datasets from UCI Machine Learning Repository on the following link: https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/.

2.2 Description of Variables:

The variables under study are as follows: -

1. **Age:** Age of the patient [Years]
2. **Sex:** Sex of the patient [M: Male, F: Female]
3. **ChestPainType:** Chest Pain Type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. **RestingBP:** Resting Blood Pressure [mm Hg]
5. **Cholesterol:** Serum Cholesterol [mm/dl]
6. **FastingBS:** Fasting Blood Sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. **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]
8. **MaxHR:** Maximum heart rate achieved [Numeric value between 60 and 202]
9. **ExerciseAngina:** Exercise-induced Angina [Y: Yes, N: No]
10. **Oldpeak:** Oldpeak = ST [Numeric value measured in depression]
11. **ST_Slope:** The slope of the peak exercise ST segment [Up: up sloping, Flat: flat, Down: down sloping]

12.**HeartDisease:** Output class [1: Heart disease, 0: Normal]

2.3 Head of Raw Data: The first six rows of our raw data are

Table 1: Head of Raw Data

Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0
39	M	NAP	120	339	0	Normal		N	0.0	Up	0

2.4 Summary of Raw Data: The summary of our raw data is

Table 2: Summary of The Variables

	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.	77.00	200.0	603.0	202.0	6.2000

Table 3: Summary of The Attributes

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

3. DATA PRE-PROCESSING

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.

3.1 **Cleaning Missing Values:** We can see that there are no missing values in our data.

3.2 **Manipulating Bad Values:** 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.

3.3 **Dropping Unnecessary Columns:** 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.

3.4 **Coding Attributes to Numbers:** Values of some columns are categorical variables. So, we code them into numbers. The changes are as follows: -

Table 4: List of Coding into Numeric Values

Column Name	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

4. DATA INFORMATION

4.1 Datatypes: Datatypes of our processed data is as follows: -

Table 5: Datatypes of The Dataset

Column Name	Datatype
Age	int 40 49 37 48 54 39 45 54 37 48 ...
Sex	Factor w/ 2 levels "1","2": 1 2 1 2 1 1 2 1 1 2 ...
ChestPainType	Factor w/ 4 levels "1","2","3","4": 1 2 1 3 2 2 1 1 3 1 ...
RestingBP	int 140 160 130 138 150 120 130 110 140 120 ...
Cholesterol	int 289 180 283 214 195 339 237 208 207 284 ...
RestingECG	Factor w/ 3 levels "1","2","3": 1 1 2 1 1 1 1 1 1 1 ...
MaxHR	int 172 156 98 108 122 170 170 142 130 120 ...
ExerciseAngina	Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 2 1 ...
Oldpeak	num 0 1 0 1.5 0 0 0 0 1.5 0 ...
ST_Slope	Factor w/ 3 levels "-1","0","1": 3 2 3 2 3 3 3 3 2 3 ...
HeartDisease	Factor w/ 2 levels "0","1": 1 2 1 2 1 1 1 1 2 1 ...

4.2 Correlation Plot: Correlation plot is a very useful visualisation tool for expressing correlation between variables as a coloured diagram. Correlation plot of the variables of our dataset is

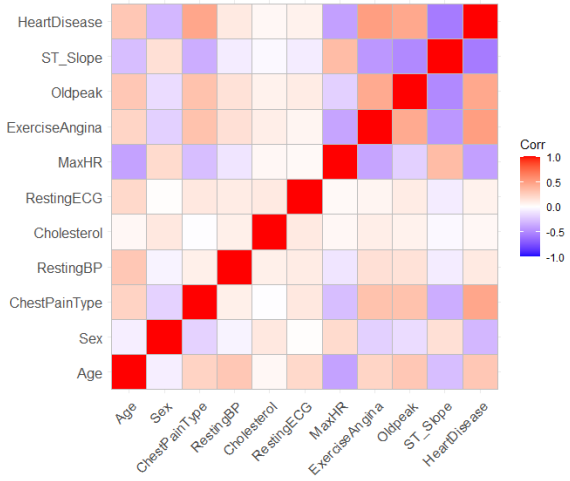
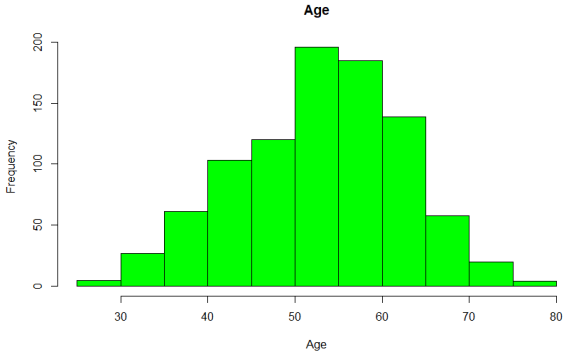
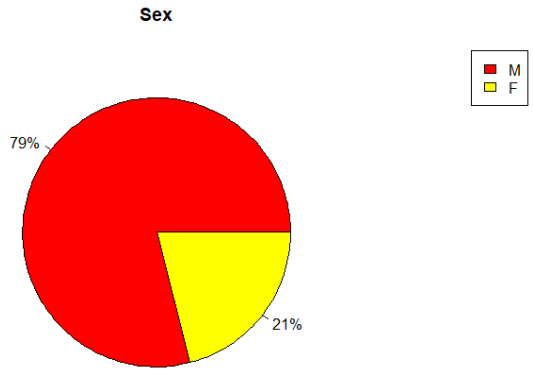
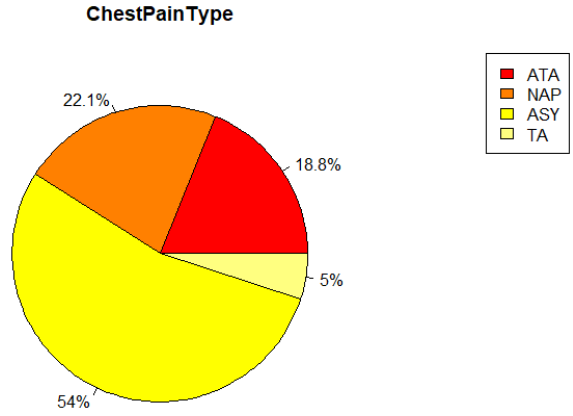


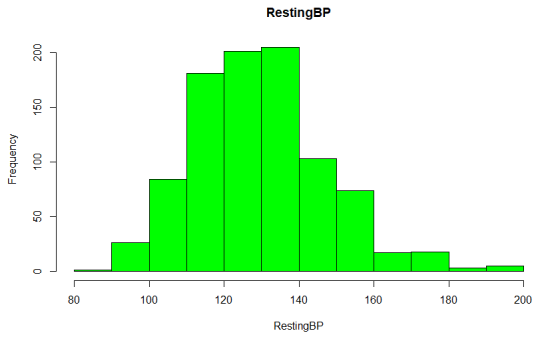
Figure 1: Correlation Plot of The Dataset

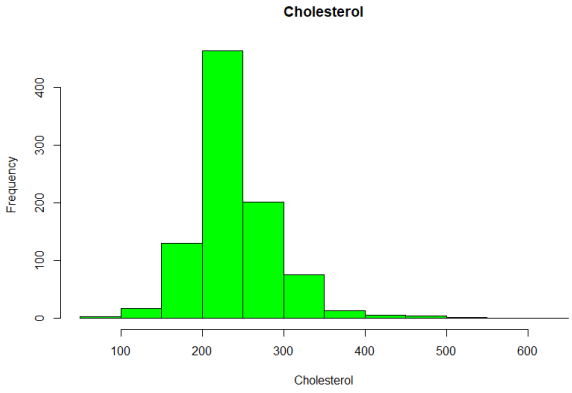
4.3 **Visualisations:** Bar diagrams and Pie charts of continuous and categorical variables respectively are

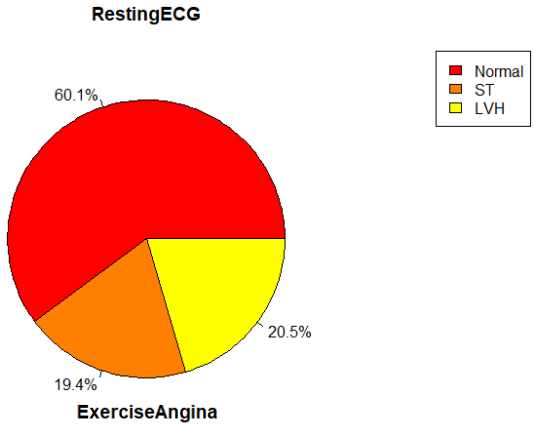


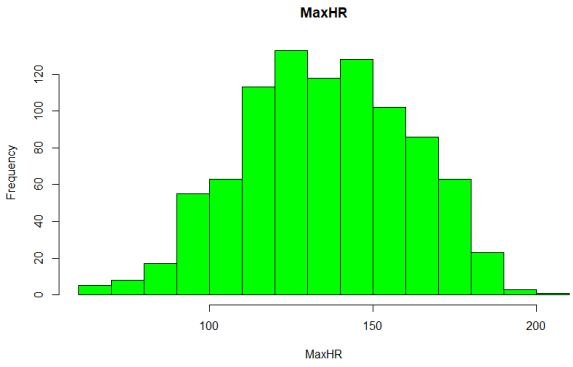


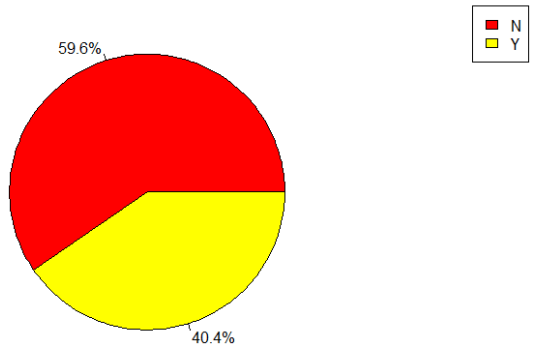












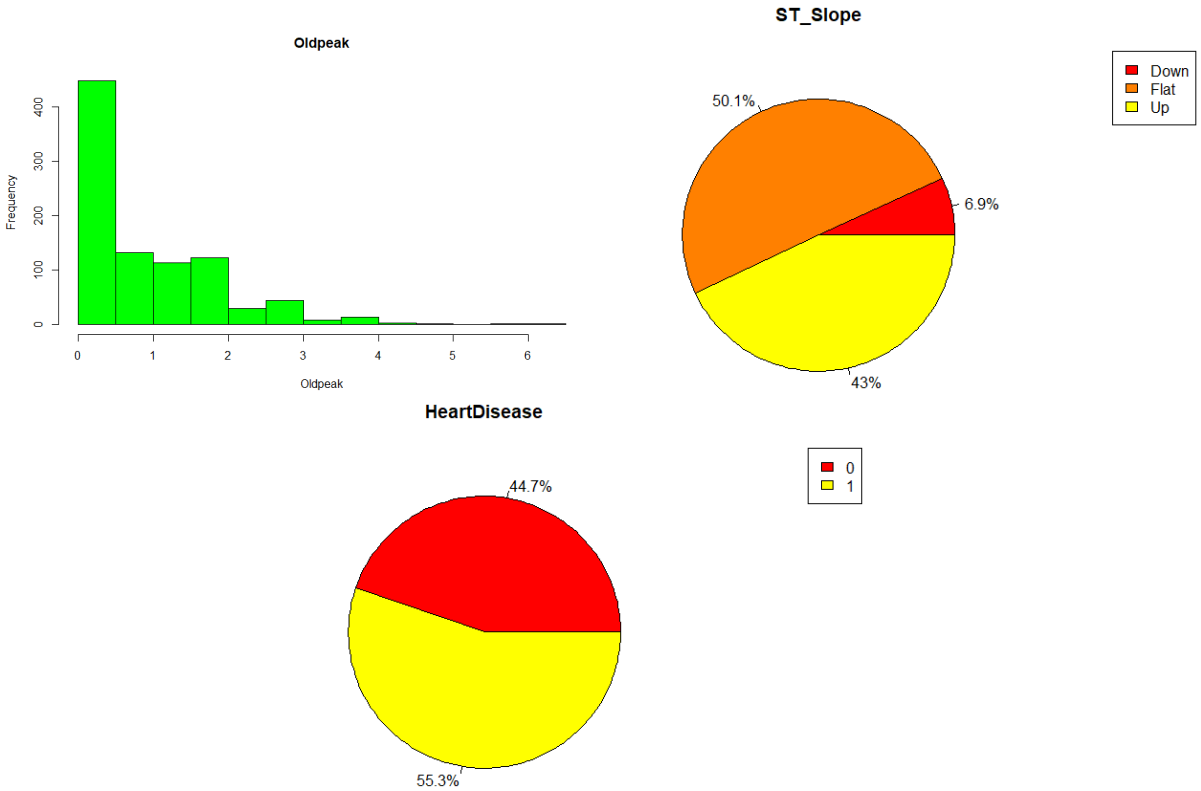


Figure 2: Visualisations of Different Variables of The Dataset

A good interpretation of the dataset can be made using these informations and visualisations.

5. METHODOLOGIES

5.1 **Correlation:** Correlation is a statistical measure which describes how much two variables are linearly related to each other i.e., how much they change together at a constant rate. Though nothing can be said about the realtionship is causal or not. Correlation is measured by correlation coefficient,

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y))}}$$

5.2 **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 general, multicollinearity can lead to wider confidence intervals that produce less reliable probabilities in terms of the effect of independent variables in a model. One might not able to trust the p-values to identify independent variables that are statistically significant.

5.3 **VIF:** Variance Inflation Factor or VIF is a measure of amount of multicollinearity in a dataset. Mathematically, the VIF for a regression model variable is the ratio of the overall model variance to the variance of a model includes only that single independent variable. VIF is measured by

$$VIF_j = \frac{1}{1-R_j^2},$$

where R_j^2 is the multiple correlation coefficient between j^{th} 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.

5.4 **Principal Component Analysis:** The principal components of a collection of points in a real coordinate space are a sequence of p unit vectors, where the i^{th} vector is the direction of a line that best fits the data while being orthogonal to the first (i-1) vectors. Mathematically, it can be shown that the principal components are eigenvectors of the covariance matrix of the data.

PCA is used in exploratory data analysis and for making predictive models. 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.

5.5 **Logistic Regression:** The application of Logistic regression model has been increased significantly in various domain. 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.

$$P(Y=1|\mathbf{X}=\mathbf{x}) = \frac{1}{1+e^{-\boldsymbol{\theta}^T x}} = \sigma\,(\boldsymbol{\theta}^T \boldsymbol{x})$$

$$P(Y=0|\mathbf{X}=\mathbf{x}) = 1- P(Y=1|\mathbf{X}=\mathbf{x})$$

Sigmoid function always gives value in the [0, 1] range. Now we need to estimate the parameter $\boldsymbol{\theta}$ such that $P(Y=1|\mathbf{X}=\mathbf{x})$ is large when **x** belongs to the positive class and small when **x** belongs to the negative class. The parameter $\boldsymbol{\theta}$ can be estimated by Maximum Likelihood estimation.

5.6 **Support Vector Machine:** 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 categorised 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. Then a separator is estimated from the data. The data should be transformed in such a way that the separator could be expressed as a hyperplane. In our case, we have used Radial SVM. Radial Kernel is of form

$$\mathrm{K(x, y)} = e^{-\gamma \sum_{j=1}^{P} (x_{ij} - y_{ij})^2},$$

where γ is the tuning parameter which accounts for the smoothness of the decision boundary and controls the variance of the model. The equation of SVM becomes

$$\mathrm{f(x)} = \beta_0 + \sum_{i \in S} \alpha_i K(x_i, y_i)$$

5.7 **Random Forest Model:** Random forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision tress at training time. Random forests correct for decision trees' habit of overfitting to their training set.

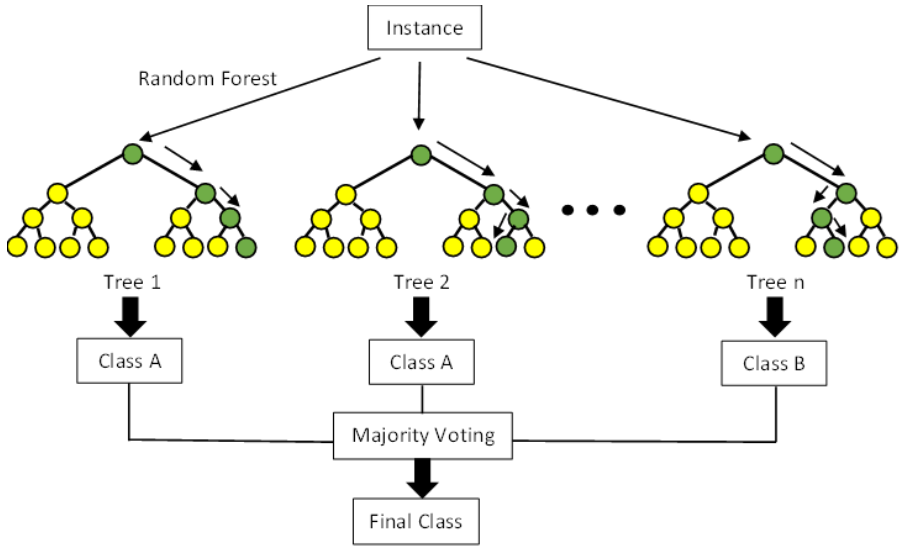


Figure 3: Simplification of Random Forest Model

One 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.

5.8 **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 text classification, spam filtration, sentiment analysis, recommendation system etc. From Bayes' theorem

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

As features of **X** are independent, $P(\mathbf{X}|y) = P(x_1|y) * P(x_2|y) * P(x_3|y) * \dots * P(x_n|y) = \prod_{i=1}^{n} P(x_i|y)$. So,

$$P(y|\mathbf{X}=\mathbf{x}) \propto P(y) * \prod_{i=1}^{n} P(x_i|y).$$

Now the goal of Naïve Bayes is to choose the class y with the maximum probability.

$$y = \text{argmax}_y[P(y) * \prod_{i=1}^{n} P(x_i|y)]$$

5.9 **Linear Discriminant Analysis:** Linear discriminant analysis is used as a tool for classification, dimension reduction and data visualisation. Despite its simplicity, LDA often produces robust, decent and interpretable classification results. To incorporate classification by LDA, we consider a random variable **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 **x** to class j is allocated if **x** is in region j following Bayesian rule or Maximum Likelihood rule according to the class prior probabilities are assumed or not respectively.

5.10 **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.

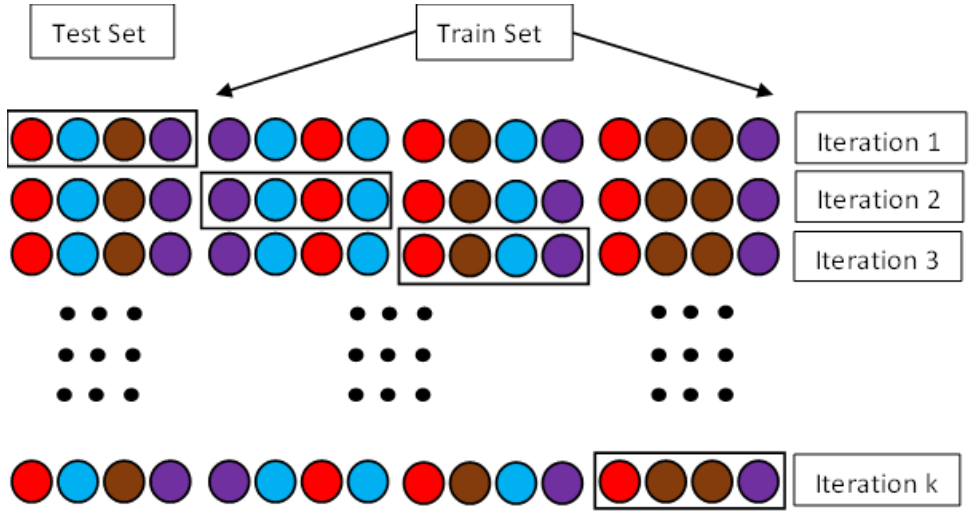


Figure 4: k-Fold Cross-validation

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.

6. COMPUTATIONS

6.1 **Calculation of VIF:** The Variance Inflation Factors of the independent variables of our dataset are

Table 6: VIFs of Independent Variables of The Dataset

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

We can see that the VIFs are very close to 1. So, there is no significant multicollinearity in the data.

6.2 **Analyzing Principal Components:** We see if we can reduce the dimension of the dataset by PCA.

Table 7: Information of Principal Components of The Data

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

Cumulative Proportion	0.8229500	0.8788400	0.9255300	0.9660000	1.0000000

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 can fit various classification models to predict heart disease to the dataset.

6.3 **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.

6.4 **Fitting Logistic Regression:** Logistic regression is fitted to the train set. Obtained Logistic regression model is

$$\text{HeartDisease} = \sigma\,(0.1373445 + 0.0194577\text{Age} - 1.6193013\text{Sex} + 0.7984934\text{ChestPainType} - 0.0008461\text{RestingBP} + 0.0041048\text{Cholesterol} - 0.0172551\text{RestingECG} - 0.0138658\text{MaxHR} + 1.0283600\text{ExerciseAngina} + 0.4266932\text{Oldpeak} - 1.6150676\text{ST\_Slope}).$$

Here p-values of some variables are not significant. So, we modify the model using StepAIC function from the library MASS. The modified model is

$$\text{HeartDisease} = \sigma\,(0.051426 + 0.018804\text{Age} - 1.620636\text{Sex} + 0.797416\text{ChestPainType} + 0.004063\text{Cholesterol} - 0.013890\text{MaxHR} + 1.024271\text{ExerciseAngina} + 0.426590\text{Oldpeak} - 1.614436\text{ST\_Slope}).$$

Predicting the classes for test sets with this model, we have got the following confusion matrix

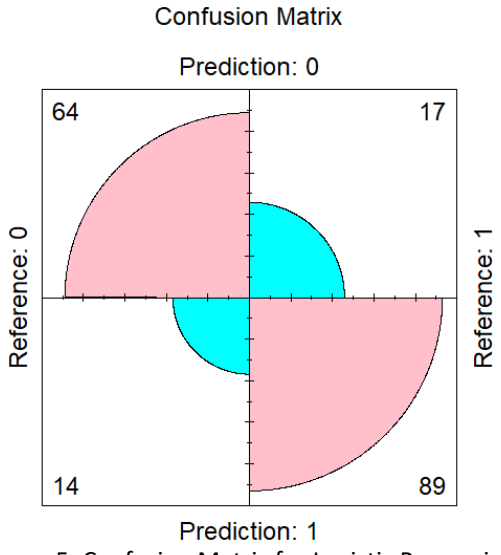


Figure 5: Confusion Matrix for Logistic Regression

Residual sum of square of predicted and actual classes of test set is 31.

Accuracy, precision, F-score and recall of our model are as following: -

Table 8: Performance Metric for Logistic Regression Model

Accuracy	Precision	F-Score	Recall
0.832	0.821	0.805	0.790

6.5 **Fitting Support Vector Machine:** Support Vector Machine is fitted to the train set. We use Radial kernel here. We predict the classes for test set and the confusion matrix is

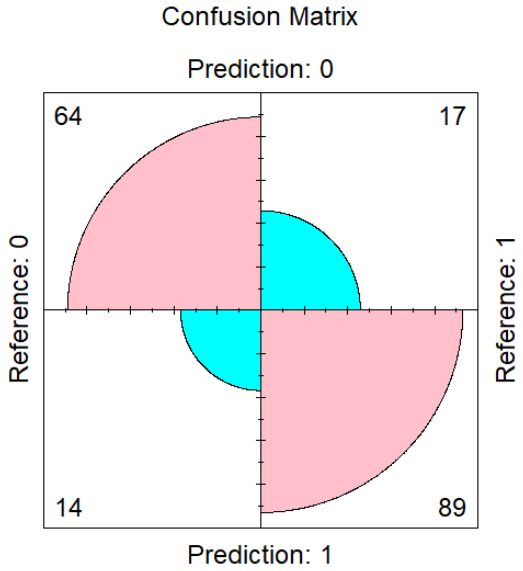


Figure 6: Confusion Matrix for Support Vector Machine

Residual sum of square of predicted and actual classes of test set is 31.

Accuracy, precision, F-score and recall of our model are as following: -

Table 9: Performance Metric for SVM Model

Accuracy	Precision	F-Score	Recall
0.832	0.821	0.805	0.790

6.6 **Fitting Random Forest Model:** We fit Random Forest model to the train set and obtain the importance of the variables

Table 10: 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

The confusion matrix after predicting the classes using test set is

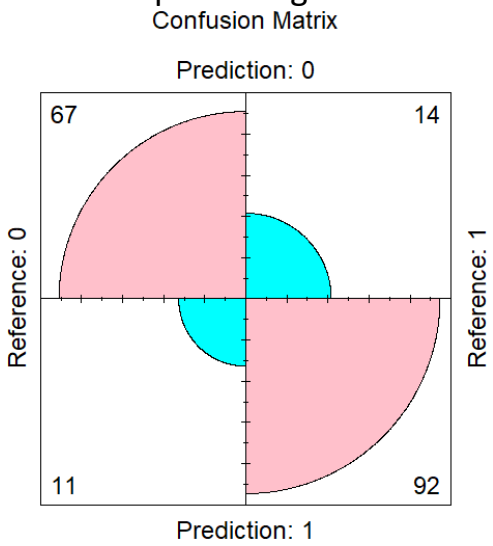


Figure 7: Confusion Matrix for Random Forest Model

Residual sum of square of predicted and actual classes of test set is 25.

Accuracy, precision, F-score and recall of our model are as following: -

Table 11: Performance Metric for Random Forest Model

Accuracy	Precision	F-Score	Recall
0.864	0.859	0.843	0.827

6.7 **Fitting Naïve Bayes Classifier:** Naïve Bayes classifier is fitted to the train set.

We predict the classes of the test data using the fitted model. The confusion matrix is

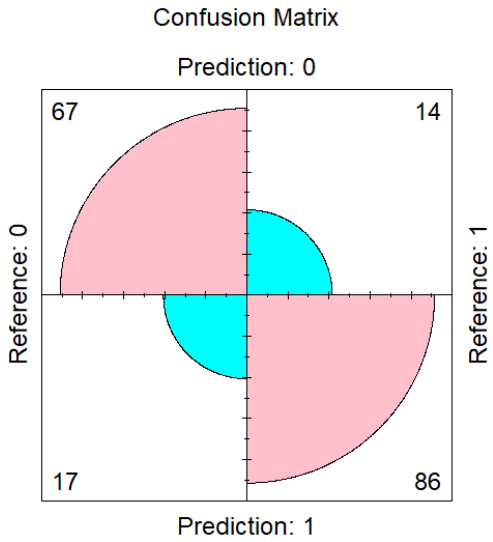


Figure 8: Confusion Matrix for Naïve Bayes Classifier

Residual sum of square of predicted and actual classes of test set is 31.

Accuracy, precision, F-score and recall of our model are

Table 12: Performance Metric for Naïve Bayes Classifier

Accuracy	Precision	F-Score	Recall
0.832	0.798	0.812	0.827

6.8 **Classification Using Linear Discriminant Analysis:** We have tried to do classification using Linear Discriminant Analysis. We have fitted the model to train set. Coefficients of linear discriminants are

Table 13: Coefficients of Linear Discriminants

Variables	LD1
Age	0.0109491646
Sex	-0.8175233015
ChestPainType	0.4746370238
RestingBP	-0.0006371966
Cholesterol	0.0019048236
RestingECG	0.0081014613
MaxHR	-0.0082933683
ExerciseAngina	0.7040684446
Oldpeak	0.1951063916
ST_Slope	-0.9941042701

We predict the classes of the test data using the obtained model. The confusion matrix is

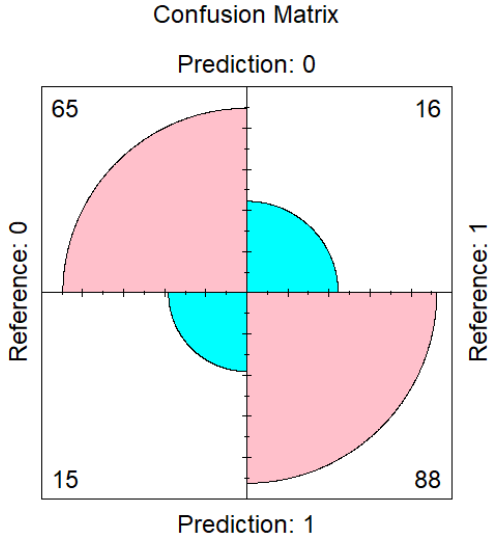


Figure 9: Confusion Matrix for Classification using Linear Discriminant Analysis

Residual sum of square of predicted and actual classes of test set is 31. Accuracy, precision, F-score and recall of our fitted model are

Table 14: Performance Metric for Classification Using LDA

Accuracy	Precision	F-Score	Recall
0.832	0.812	0.807	0.802

6.9 **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. We obtained the following results: -

Table 15: Accuracy Table from Cross-validation

	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
LR	0.7826087	0.8179348	0.8469900	0.8479097	0.8767320	0.9239130
SVM	0.7934783	0.8351648	0.8478261	0.8522137	0.8767320	0.9130435
RF	0.7717391	0.8478261	0.8688485	0.8692985	0.8913043	0.9239130
NB	0.7608696	0.8260870	0.8478261	0.8416746	0.8586957	0.9021739
LDA	0.7826087	0.8174570	0.8524845	0.8486025	0.8695652	0.9347826

Table 15: Kappa Table from Cross-validation

	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
LR	0.5557702	0.6309042	0.6899008	0.6909682	0.7477184	0.8456376
SVM	0.5789981	0.6652348	0.6934724	0.6989110	0.7480356	0.8456376
RF	0.5278592	0.6909034	0.7316735	0.7337102	0.7789524	0.8463740
NB	0.5065822	0.6446161	0.6859994	0.6775783	0.7133269	0.8024809
LDA	0.5579049	0.6300594	0.7015150	0.6924890	0.7366367	0.8667310

The results are visualized by box plot and dot plot.

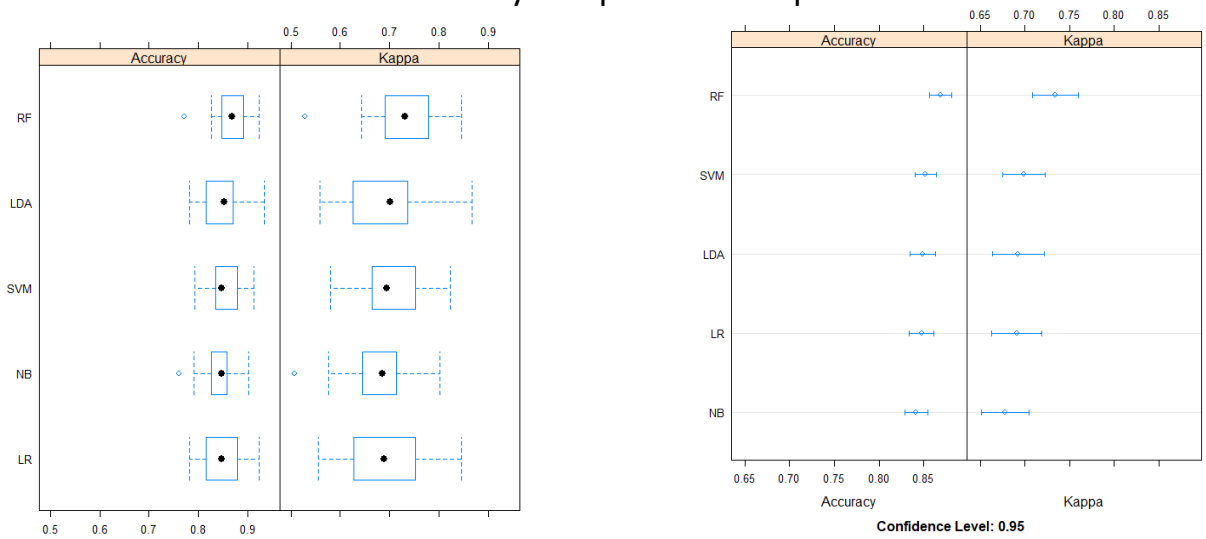


Figure 10: Box plot and Dot plot for Results of Cross-validation

Here it is clear that Random Forest model gives best performance.

7. CONCLUSIONS

We compare the accuracies of various model using the table

Table 16: Comparison of Performances of Various Models

Model	Accuracy	Residual Sum of Squares
Logistic Regression	0.832	31
Support Vector Machine	0.832	31
Random Forest Model	0.864	25
Naïve Bayes Classifier	0.832	31
Linear Discriminant Analysis	0.832	31

- From this table, it can be concluded that the Random Forest model gives the best predictions of existence of heart disease. From the results of 10- fold cross-validation, it can be said 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 in terms of correctness of prediction.

- From the variable importance (Table 9) 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.

8. MEDICAL AND STATISTICAL SIGNIFICANCE

- Medically, the slope of the peak exercise ST segment is the most important factor for taking preventive action against heart disease. Also, old peak, type of chest pain and maximum heart rate should also be seriously considered. Various medical research papers and journals (A. Yazdani, K.D. Varathan, Y.K. Chiam, A.W. Malik, W.A.W. Ahmad: 2021, M. Hori, H. Okamoto: 2012) also support this fact.

- Statistically, it can be concluded that Random Forest is the best model for prediction of heart disease for a person. This can be explained with some theory. Random Forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. Rigorous study of statistical research papers (M. Rua: 2021, M. Pal, S. Parija, 2021) supports the fact of good performance of Random Forest model in medical domain.

So, Random Forest model can be used with good confidence for various medical purposes.

REFERENCES

1. C. Zhu, C.U. Idemudia, W. Feng: Improved Logistic Regression Model for Diabetes Prediction by Integrating PCA and K-means Techniques, Informatics in Medicine Unlocked, Volume 17, 2019, 100179, ScienceDirect

2. A. Yazdani, K.D. Varathan, Y.K. Chiam, A.W. Malik, W.A.W. Ahmad: A Novel Approach for Heart Disease Prediction Using Strength Scores with Significant Predictors, BMC Medical Informatics and Decision Making, Volume 21, Article Number: 194 (2021)

3. M. Hori, H. Okamoto: Heart Rate as A Target of Treatment of Chronic Heart Failure, Journal of Cardiology, ScienceDirect

4. M. Pal, S. Parija: Prediction of Heart Disease using Random Forest, Journal of Physics: Conference Series 1817 012009

5. M. Rua: Heart Disease Prediction Random Forest Classifier, Kaggle

6. C.A. Brawner, J.K. Ehrman, J. R. Schairer, J.J. Kao, S.J. Keteyian: Predicting Maximum Heart Rate among Patients with Coronary Heart Disease Receiving Beta-adrenergic Blockade Therapy, PMID: 15523326, National Library of Medicine

7. T. Evgeniou, M. Pontil: Support Vector Machines: Theory and Applications, ResearchGate

8. P. Kaviani, S. Dhotre: Short Survey on Naïve bayes Algorithm, ResearchGate

9. Kshirsagan, A.M. Multivariate Analysis

10. R.A. Johnson and D.W. Wichern: Applied Multivariate Statistical Analysis

11. https://www.wikipedia.org

APPENDIX (R CODE)

```
sum(is.na(heart_df))

heart_df_coded <- heart_df %>%

  mutate(Sex=replace(Sex, Sex=='M', 1))

heart_df_coded <- heart_df_coded %>%

  mutate(Sex=replace(Sex, Sex=='F', 2))

heart_df_coded$Sex <- as.integer(heart_df_coded$Sex)

heart_df_coded <- heart_df_coded %>%

  mutate(ChestPainType=replace(ChestPainType, ChestPainType=='ATA', 1))

heart_df_coded <- heart_df_coded %>%

  mutate(ChestPainType=replace(ChestPainType, ChestPainType=='NAP', 2))

heart_df_coded <- heart_df_coded %>%

  mutate(ChestPainType=replace(ChestPainType, ChestPainType=='ASY', 3))

heart_df_coded <- heart_df_coded %>%

  mutate(ChestPainType=replace(ChestPainType, ChestPainType=='TA', 4))

heart_df_coded <- heart_df_coded %>%

  mutate(RestingECG=replace(RestingECG, RestingECG=='Normal', 1))

heart_df_coded <- heart_df_coded %>%

  mutate(RestingECG=replace(RestingECG, RestingECG=='ST', 2))

heart_df_coded <- heart_df_coded %>%

  mutate(RestingECG=replace(RestingECG, RestingECG=='LVH', 3))

heart_df_coded$RestingECG <- as.integer(heart_df_coded$RestingECG)

heart_df_coded <- heart_df_coded %>%

  mutate(ExerciseAngina=replace(ExerciseAngina, ExerciseAngina=='Y',
1))

heart_df_coded <- heart_df_coded %>%

  mutate(ExerciseAngina=replace(ExerciseAngina, ExerciseAngina=='N',
0))

heart_df_coded <- heart_df_coded %>%

  mutate(ST_Slope=replace(ST_Slope, ST_Slope=='Down', -1))

heart_df_coded <- heart_df_coded %>%

  mutate(ST_Slope=replace(ST_Slope, ST_Slope=='Flat', 0))

heart_df_coded <- heart_df_coded %>%
```

```
  mutate(ST_Slope=replace(ST_Slope, ST_Slope=='Up', 1))
heart_df_coded$ST_Slope <- as.integer(heart_df_coded$ST_Slope)
######## VISUALIZATION ########
# CORRELATION MATRIX
cor_mat <- round(cor(heart_df_coded), 2)
head(cor_mat)
library(ggcorrplot)
ggcorrplot(cor_mat)
LiR_model <- lm(HeartDisease~., data = heart_df_coded)
vif(LiR_model)
model_LR <- glm(HeartDisease ~., data = train_set, family = 'binomial')
modified_model_LR <- stepAIC(model_LR)
model_SVM <- svm(HeartDisease~., data = train_set, type = 'C')
set.seed(1111)
model_RF <- randomForest(HeartDisease~., data = train_set)  #we need to
change the target variable
model_NB <- naive_bayes(HeartDisease~., data = train_set)  #we need to
change the target variable
model_LDA <- lda(HeartDisease~., data = train_set)
control <- trainControl(method = 'repeatedcv', number = 10, repeats =
3)
set.seed(345)
model_LR_CV <- train(as.factor(HeartDisease) ~., data = heart_df_coded,
trControl = control, method = 'glm', family = 'binomial')
set.seed(345)
model_SVM_CV <- train(as.factor(HeartDisease) ~., data =
heart_df_coded, trControl = control, method = 'svmRadial')
set.seed(345)
model_RF_CV <- train(as.factor(HeartDisease) ~., data = heart_df_coded,
trControl = control, method = 'rf')
model_NB_CV <- train(as.factor(HeartDisease) ~., data = heart_df_coded,
trControl = control, method = 'nb')
model_LDA_CV <- train(as.factor(HeartDisease) ~., data =
heart_df_coded, trControl = control, method = 'lda')
results <- resamples(list(LR=model_LR_CV, SVM=model_SVM_CV,
RF=model_RF_CV, LDA=model_LDA_CV))
summary(results)
```