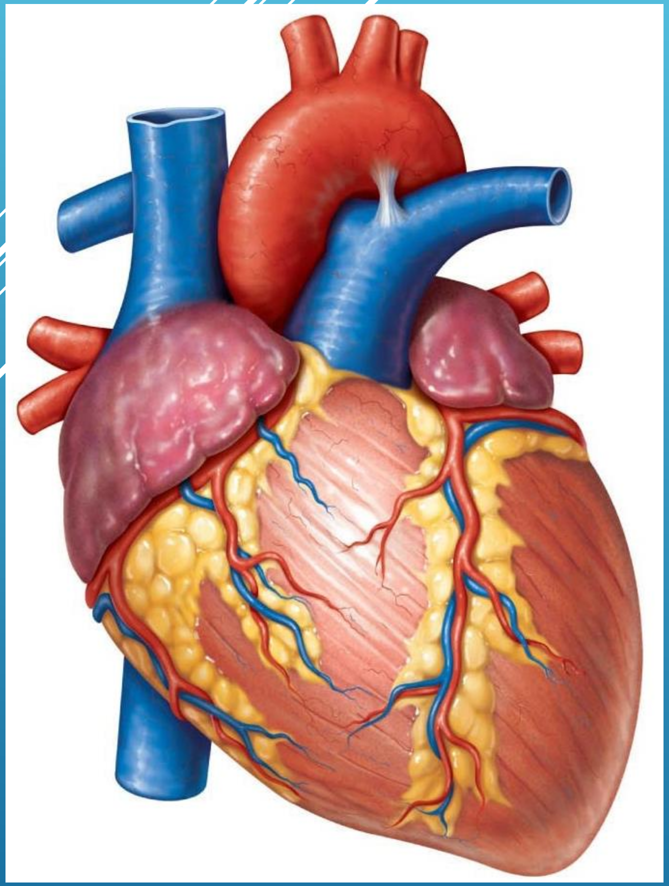
A STUDY ON HEART DISEASE PREDICTION USING DIFFERENT CLASSIFICATION MODELS BASED ON CROSS VALIDATION METHOD

Presented by:

Anirban Ghosh
Undergraduate 3rd Year,
Department of Statistics,
University of Kalyani

Supervised by:

Dr. Sushovon Jana
Department of Applied Statistics,
Maulana Abul Kalam Azad University of Technology



APPSCICON 2022
MAKAUT

INTRODUCTION

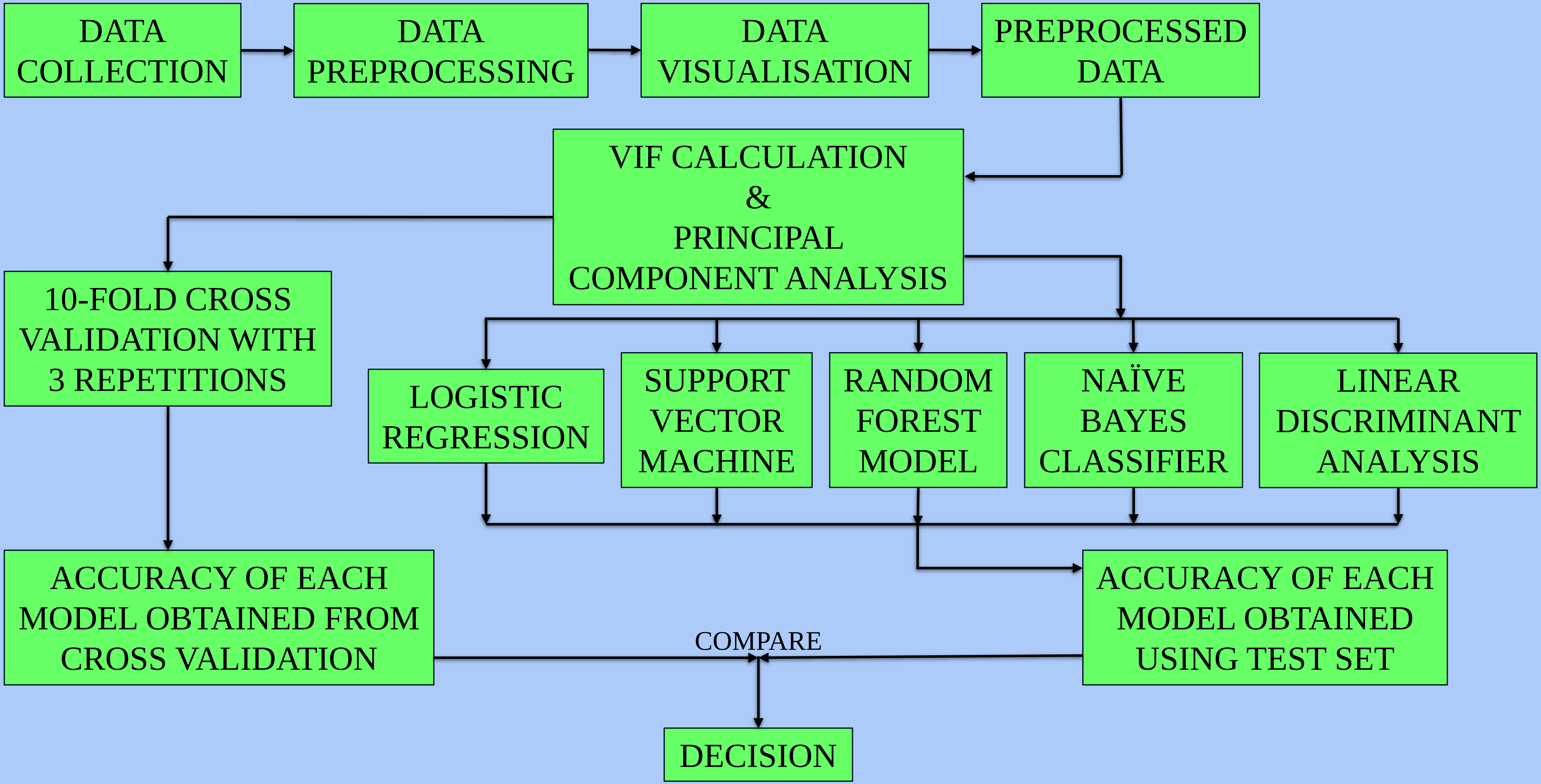
According to World Health Organization (WHO), heart disease is the no. 1 cause of death in world. It is responsible for 16% of total deaths in world [1]. Since 2000, the largest increase in deaths has been for heart disease, rising by more than 2 million to 8.9 million deaths in 2019 [1]. Also in India, 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.

OBJECTIVES

- To compare five classification models and find the best model, out of these, in terms of accuracy.
- To find the importance of variables for classification.

Keywords: - Heart Disease, Logistic Regression model, Support Vector Machine, Random Forest Model, Naïve Bayes classifier, Linear Discriminant Analysis, Cross validation

METHODOLOGY



DATA DESCRIPTION

- The dataset is downloaded from Kaggle website [3]. This is an open-access dataset.
- The dataset is of 918 observations and contains 11 independent variables and a categorical variable, whether there exists heart disease or not, as target variable.
- Out of 11 independent variables, 6 are categorical variables and rest 5 are continuous variables.

Age of the Patient (in years) [Continuous]

Chest Pain Type (TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic) [Categorical]

Serum Cholesterol (mm/dl) [Continuous]

Resting Electrocardiogram results (Normal: Normal, ST: having ST-T wave abnormality, LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria) [Categorical]

Exercise-induced Angina (Y: Yes, N: No) [Categorical]

The slope of the peak exercise ST segment (Up: up sloping, Flat: flat, Down: down sloping) [Categorical]

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
1	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
2	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
3	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
4	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
5	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

Sex of the Patient (M: Male, F: Female) [Categorical]

Resting Blood Pressure (mm Hg) [Continuous]

Fasting Blood Sugar (1: if FastingBS > 120 mg/dl, 0: otherwise) [Categorical]

Maximum heart rate achieved (Numeric Value) [Continuous]

Oldpeak = ST (Numeric value) [Continuous]

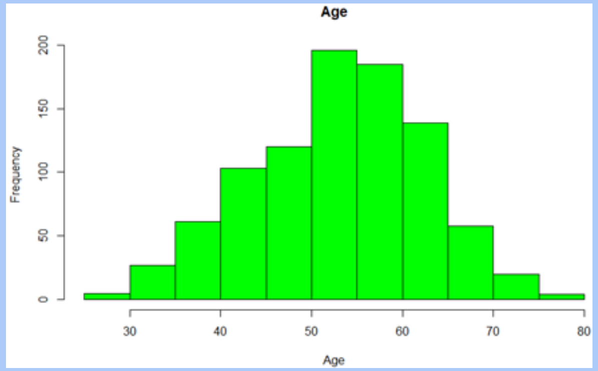
Output class (1: Heart disease, 0: Normal) [Categorical]

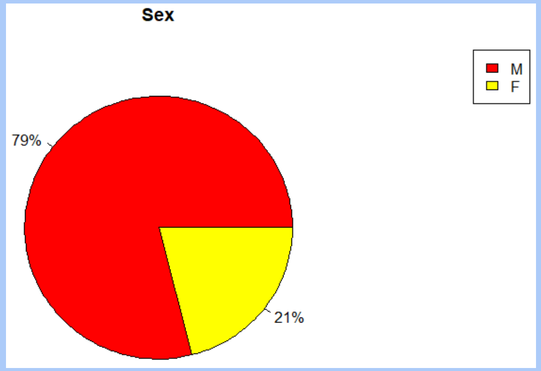
DATA PREPROCESSING

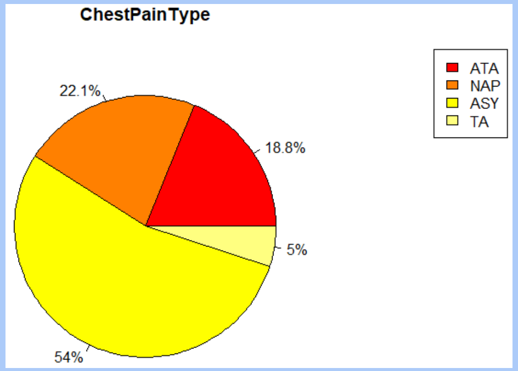
- There is no missing value in our dataset.
- There are some 0 values in the columns RestingBP and Cholesterol. But Resting Blood Pressure and Serum Cholesterol of a person can never be 0. So, these are bad values. These zeros are replaced with median values of the corresponding columns.
- There are some negative values in the column Oldpeak. These negative values are converted to positive.
- It is found that about 77% values of the column FastingBS are 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 the form of a table.

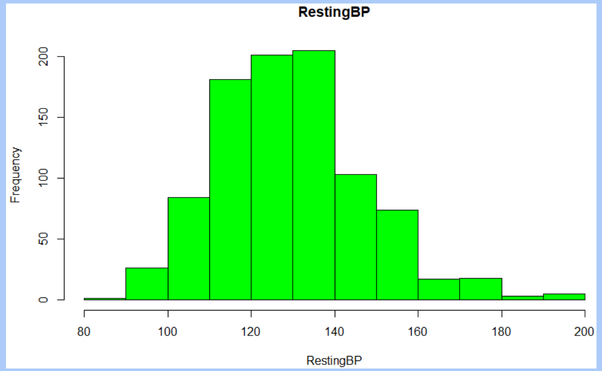
Column Name	Value	Coded Value
Sex	'M'	1
	'F'	2
ChestPainType	'ATA'	1
	'NAP'	2
	'ASY'	3
	'TA'	4
RestingECG	'Normal'	1
	'ST'	2
	'LVH'	3
ExerciseAngina	'Y'	1
	'N'	0
ST_Slope	'Down'	-1
	'Flat'	0
	'Up'	1

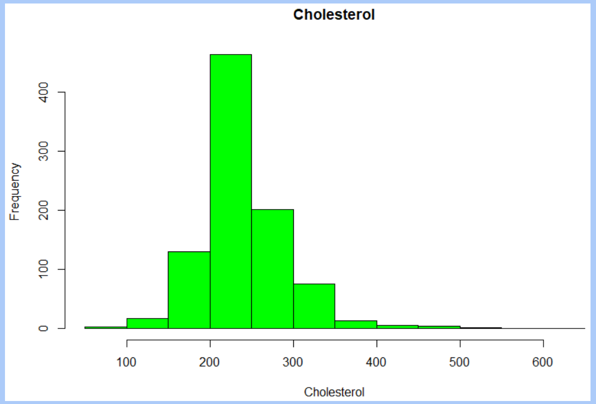
DATA VISUALISATION

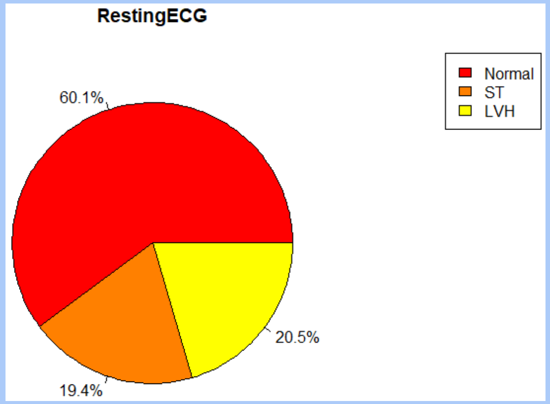


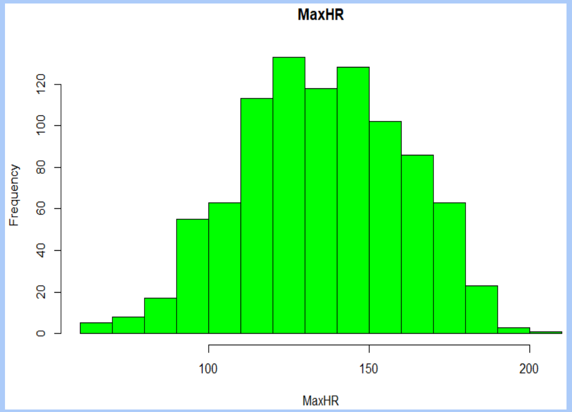


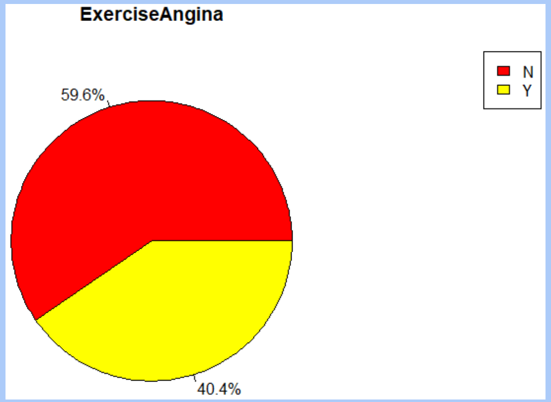


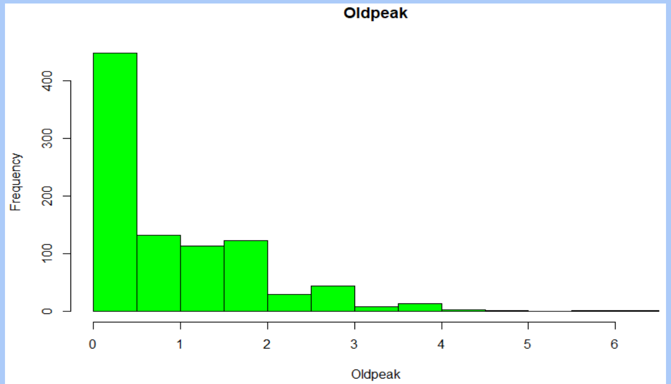


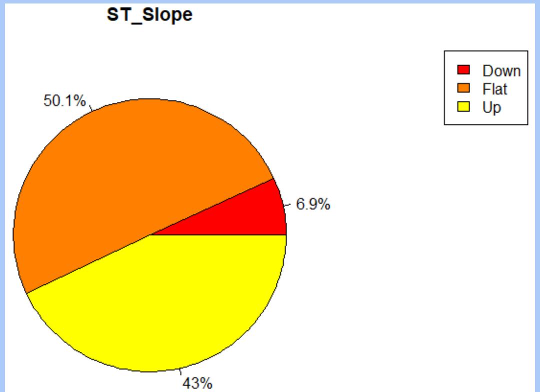


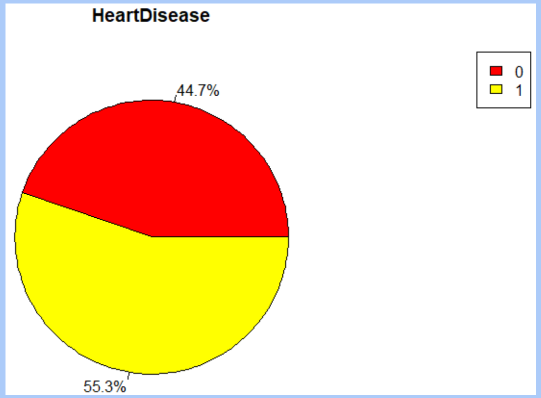












A good interpretation of the dataset can be made using these information and visualisations.

EXPERIMENTAL RESULTS

VIF CALCULATION

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.

PRINCIPAL COMPONENT ANALYSIS

	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

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 are good to fit various classification models.

EXPERIMENTAL RESULTS

We have fitted various model to the train set and predicted the test set. Then 10-fold cross validation is done with 3 repetitions. Then following accuracies are obtained.

ACCURACY OBTAINED USING TEST SET

Model	Accuracy
Logistic Regression	0.832
Support Vector Machine	0.832
Random Forest Model	0.864
Naïve Bayes Classifier	0.832
Linear Discriminant Analysis	0.832

ACCURACY OBTAINED FROM CROSS VALIDATION

	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
LR	0.783	0.818	0.847	0.848	0.877	0.924
SVM	0.793	0.835	0.848	0.852	0.877	0.913
RF	0.772	0.848	0.869	0.869	0.891	0.924
NB	0.761	0.826	0.848	0.842	0.859	0.902
LDA	0.783	0.817	0.852	0.849	0.870	0.935

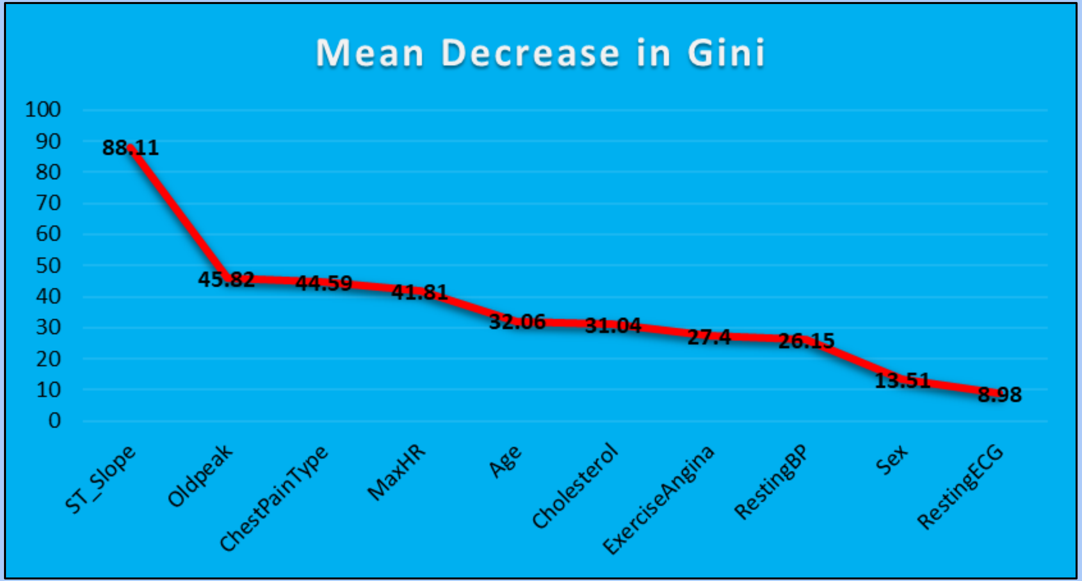
Random Forest model has the highest accuracy of 86.4% when observations of test set is predicted using the fitted models. Resampling also proposes Random Forest as best model with a mean accuracy of 86.9%, which is highest among all models.

EXPERIMENTAL RESULTS

From Random Forest model, variable importance is obtained by 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 below.

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



- ST_Slope has a Mean Decrease in Gini coefficient value of 88.11, which is highest among all predictor variables.
- Oldpeak, ChestPainType and MaxHR also have close values, 45.82, 44.59 and 41.81 respectively.
- RestingECG has the lowest value of 8.98.

SIGNIFICANCE & CONCLUSION

- So, 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.

- There are some advantages of using Random Forest model for classification. Decision tree is highly biased and it has greater variance. But as Random Forest is collection of multiple decision trees, it has less bias and less variance. Also Random Forest solves the problem of overfitting of Decision tree.

- The slope of the peak exercise ST segment is the most important factor for classification of heart disease. Also, old peak, type of chest pain and maximum heart rate should also be seriously considered. Resting electrocardiogram result and sex contributes least for classification.

ACKNOWLEDGEMENT

Primarily I would thank the Almighty for being able to complete this project successfully. I would like to thank School of Applied Science and Technology, MAKAUT for giving me this opportunity to present my project. I would like to express my sincere gratitude to my project supervisor, Dr. Sushovon Jana for staying beside me throughout 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 respectable professors of Department of Statistics, University of Kalyani for their valuable suggestions.

I would like to express my gratitude towards Kaggle website and the creator of the dataset for being the dataset open-access. The dataset is the root of this project.

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

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

THANK YOU!!