The evaluation of relationship between Health Metrics on Cardiovascular Disease

Li Xu, Ph.D.
Assistant Professor
Department of Finance
New Jersey City University

Annie Udhani
Graduate Student
New Jersey City University

Abstract

Cardiovascular diseases such as heart attack or stroke, etc., are increasingly becoming the major factor in human mortality. The early and accurate prediction of such life threats can result in the saving of millions of human lives every year. This capstone project aims to study and understand the relationship between various human health parameters with heart disease. UCI Machine learning heart dataset is used for the purpose of this paper. Various Machine learning algorithms such as decision tree, Linear Kernel SVM, SVM with polynomial kernels, bagging with multiple bagged models are applied to unearth this relationship.

1. Introduction

Heart is a vital organ of the body which keeps all the cells of the body alive by supplying oxygen and essential nutrients and removing carbon-di-oxide and other toxins. It is vitally responsible for everything that gives body life, ranging from the transportation of oxygen to the success of the body's immune system.

Cardiovascular diseases (CVD) are a group of disorders affecting the heart or blood vessels due to build-up of fatty deposits resulting in a heart failure in majority of the cases. There has been a steady rise in the number of cases of people suffering from heart diseases. As per the World Health Organization (WHO) reports, CVD has been a leading cause of death every year with 17.9 million deaths every year. This is a whooping 32% of all the deaths worldwide and around 85% of cardiovascular diseases results in a heart failure.[1] In USA alone, heart diseases result in the death of 659,000 people every year at a rate of one person every 36 seconds and cost around $363B in financial loses.[2] According to CDC, the signs and symptoms of a cardiovascular disease may include chest pain, chest discomfort, dizziness, shortness of breath, fatigue, swollen knees, etc.[3] The key reasons behind cardiovascular diseases include high blood pressure, high cholesterol level, hypertension, obesity, etc.[2]

Expensive diagnosis equipment is required to diagnose and detect cardiovascular diseases in a patient. This group of disorders usually require a patient to go through expensive electrocardiogram tests, Holter monitoring, echocardiogram test, stress test, Cardiac catheterization test, etc. These tests are usually expensive and unaffordable to a greater section of the society. Our project seeks to understand how an inexpensive and non-invasive method of diagnosis namely blood test and chest test can be used to diagnose heart disease without the expertise of a medical professional.

Advancement in computer technologies and data storage techniques, a lot of research data and patient records of hospital are made available. There are many such open public dataset for data analytics and Machine learning enthusiasts to run their model on to improve the physical and financial health of society. We will be working on a 13-feature dataset considered as one of the benchmark datasets for Heart patient collected from four databases: Cleveland, Hungary, Switzerland, and Long Beach and dates to 1998. We will be applying various Machine learning models to this dataset and comparing the results to determine the best model for the purpose of classifying patient into two categories – one who are suffering or may suffer from heart diseases and another category of people who have very less risk of suffering from a heart disease.

2. Literature Review

A lot of work has been already done in the field of detecting cardiovascular disease using various modelling techniques and cleaning approaches. A multitude of all kinds of learning models has been applied in the literature ranging from tree based binary classification to neural network-based classification. Zhang et al. applied Support Vector Machine (SVM) in 2017 to classify clinical data leaving result to the interpretation of others. Guidi et al. in 2014 presented a Clinical Decision Support system evaluating Heart Failure severity and type evaluation and compares the performance of neural network, SVM and random forest.[5] Singh et al. applied Fisher method and generalized Discriminant analysis with binary classifier to detect the coronary disease using the variability of heart rate.[4]

In an interesting paper published by Srinivas et al in IEEE, the detection of cardiovascular disease was performed on a targeted audience that includes the worker of coal mines to study whether reported cardiovascular disease cases are higher in coal mine workers. He used Neural network and Naïve Bayes for the prediction of heart diseases.[6] A similar targeted study was done by Parthiban and Srivatsa which evaluated the relationship between health metrics and diabetes using Naïve Bayes and SVM and achieved good prediction accuracy rate.[7]

Most of the studies here have either used complex models or have used the targeted audience or in some cases targeted disease. In this paper, we will be trying to perform better result in a very simplistic manner by applying simple models with enhanced data cleaning methodologies. We will also be keeping this study very generic to apply it for the betterment of the masses rather a particular section of patients suffering from diabetes. We will also be doing this study on a set of data attribute obtained from UCI Public Dataset to avoid our model from learning the biased data from a section of people who are likely to suffer more from heart diseases due to many factors like their occupation.

3. Data

3.1. Data Description

The dataset used here is a subset of Public Health Dataset. The original dataset has 76 attributes along with the target variable. Most of the research and published articles uses 12 attributes out of them. We are retaining 12 attributes as majorly used in most of the research to predict whether a person is suffering or is likely to suffer with a heart disease or not.

3.2. Data Feature Description

S. No.	Feature Name	Feature Description	Type
1.	**Age**	age of the patient [years]	Independent
2.	**Sex**	Sex of the patient	Independent
3.	**ChestPainType**	chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]	Independent
4.	**RestingBP**	resting blood pressure [mm Hg]	Independent
5.	**Cholestrol**	Serum Cholestrol [mm/dl]	Independent
6.	**FastingBS**	fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]	Independent
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	Independent

		ventricular hypertrophy by Estes' criteria]	
8.	**MaxHR**	maximum heart rate achieved [Numeric value between 60 and 202]	Independent
9.	**ExerciseAngina**	exercise-induced angina [Y: Yes, N: No]	Independent
10.	**Oldpeak**	oldpeak = ST [Numeric value measured in depression]	Independent
11.	**ST_Slope**	the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, down: down sloping]	Independent
12.	**HeartDisease**	output class [1: heart disease, 0: Normal]	Dependent

3.3. Target Variable

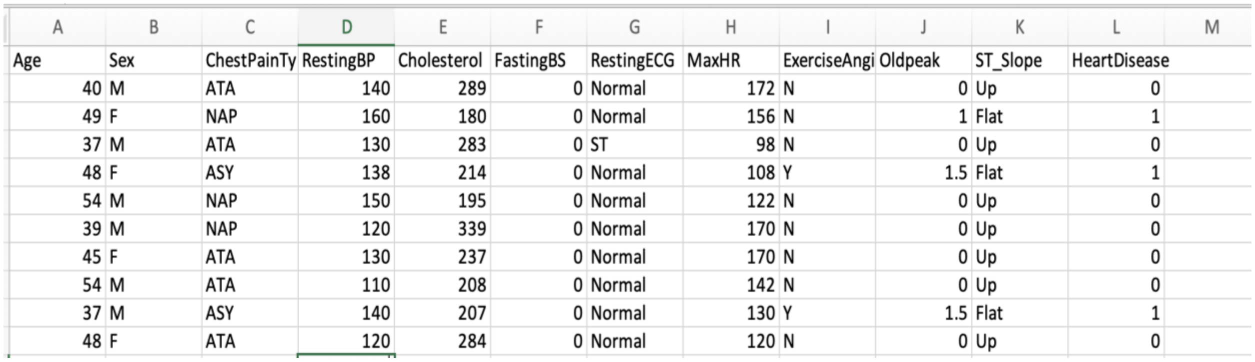
For the context of this dataset and our study, target variable is HeartDisease. We are interested in studying the effect of other body vitals and input parameter like cholesterol, gender, age on the presence of a heart disease in a person.

4. Exploratory Data Analysis (EDA) on dataset

The dataset that we are using here is UCI Machine learning public dataset consisting of **12 features** that we can build models upon. But before building the models, we will perform an exploratory data analysis on it to understand the various aspects of the data features. We will also be counting the **missing variables and outliers** in all the features. We will replace all the missing values for numerical variable with mean and categorical variable with mode, if any. We will also repeat the same operation for outlier if we find any after exploratory analysis.

4.1. Examining Data

The top few rows of the heart failure data are:

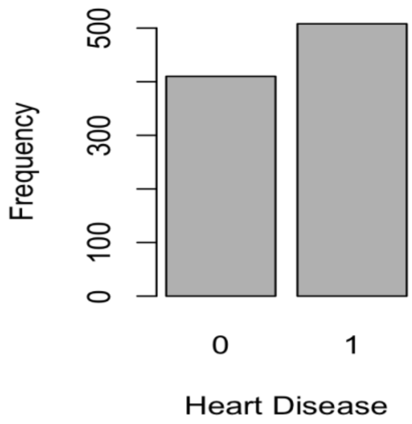


A	B	C	D	E	F	G	H	I	J	K	L	M
Age	Sex	ChestPainTy	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngi	Oldpeak	ST_Slope	HeartDisease	
40	M	ATA	140	289	0	Normal	172	N	0	Up	0	
49	F	NAP	160	180	0	Normal	156	N	1	Flat	1	
37	M	ATA	130	283	0	ST	98	N	0	Up	0	
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1	
54	M	NAP	150	195	0	Normal	122	N	0	Up	0	
39	M	NAP	120	339	0	Normal	170	N	0	Up	0	
45	F	ATA	130	237	0	Normal	170	N	0	Up	0	
54	M	ATA	110	208	0	Normal	142	N	0	Up	0	
37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat	1	
48	F	ATA	120	284	0	Normal	120	N	0	Up	0	

There are 7 numerical columns in this data. The various summary statistics of numeric column are listed in the below table:

S.No.	Columns	Min	Mean	Median	Max
1	Age	28	53.6227876	54	77
2	RestingBP	0	132.259259	130	200
3	Cholesterol	0	198.799564	223	603
4	FastingBS	0	0.23311547	0	1
5	MaxHR	60	136.809368	138	202
6	Oldpeak	-2.6	0.88736383	0.6	6.2
7	HeartDisease	0	0.55337691	1	1

4.2. Examining Target/Dependent Variable



The dependent variable in this case is whether a person has heart disease or not. It’s a binary classification problem where a person can either have the disease (denoted by value 1) or a person doesn’t have the disease (denoted by value 0). There are 508 people having the disease compared to 410 people not having the disease. There is well representation of both the type of records in the data. Additionally, data is almost balanced by output class.

5. Methodology

5.1. Data Cleaning

No values were imputed in the dataset as no missing value or outlier was found in the dataset. Additionally, the target classes were found to be almost balanced, so we didn’t create any synthetic substitution of observation to make it perfectly balanced.

5.2. Data partition

Data was split into training and test dataset. 80% of the observation is used to created train dataset while 20% of the observation is used to create test dataset for validation and accuracy measurement.

5.3. Decision tree

Decision tree is a supervised classification technique where a data is split into multiple sub tree based on the features until the algorithm is left with the pure groups. The feature is selected first which can provide maximum information gain about the classes. All other features are selected in the order of decreasing information gain. The splitting at each feature is controlled by entropy of the dataset based on the classes in the data.

Entropy at each splitting level and information gain is defined as below:

Formula for Entropy:

$$H(s) = -probablity\ of\ \log_2(p+) - -probability\ of\ \log_2(p-)$$

Where

$(p+)$ → % of positive class

$(p-)$→ % of negative class

Formula for Information Gain:

$$Gain(S,A) = H(s)\sum\frac{/Sv/}{/s/}H(Sv)$$

We applied Decision tree classification on the dataset with 10 cross fold validation to minimize the bias in the process. This model was able to classify **87.43%** of test data accurately with a precision of 90.41% and sensitivity of 80.49%.

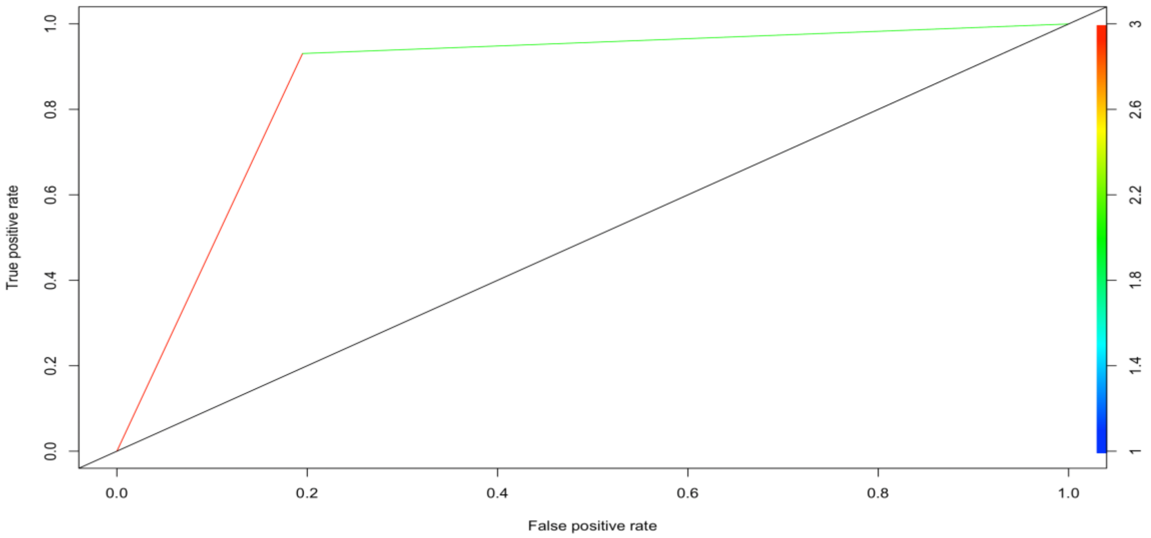


Figure 1: RoC Curve with Decision Tree Classification

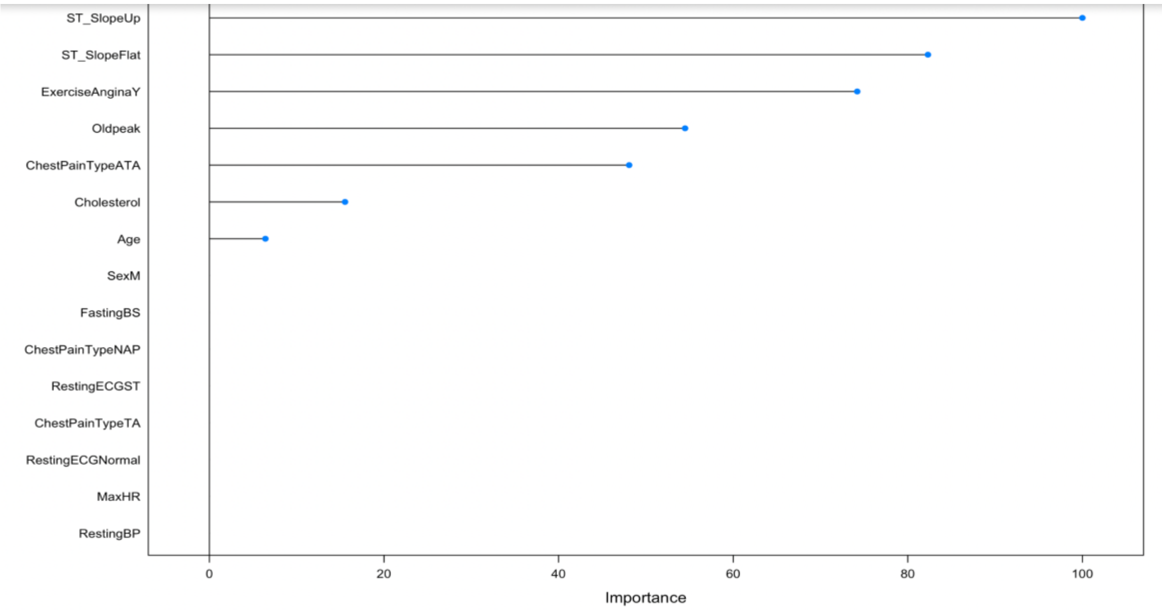


Figure 2: Variable importance with decision tree classification

5.4. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm which tries to create a hyperplane or a set of hyperplanes to separate the data into different categories based on the location of data points respective to the hyperplane. SVM is used with kernels which transforms a data into a high dimensional space. The usage of kernel may modify a non-separable dataset into an easily separable dataset in higher dimension. The common kernels used with SVM are linear, polynomial, and radial kernel.

The choice of the kernel depends on whether the relationship between the class labels and attributes are linear or nonlinear. We will use all the kernels one by one and try to figure out the best performing kernel in classifying our dataset.

5.4.1. SVM with linear kernel

We applied SVM with linear kernel on the dataset with 10 cross fold validation to minimize the bias in the process. This model was able to classify **89.07%** of test data accurately with a precision of 87.80% and sensitivity of 87.80%.

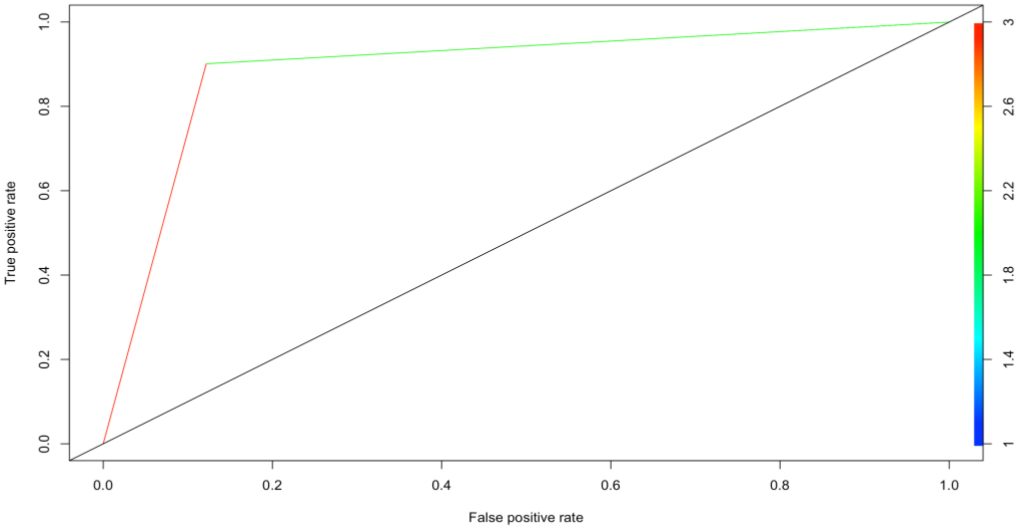


Figure 3: Roc Curve with SVM Linear kernel classification

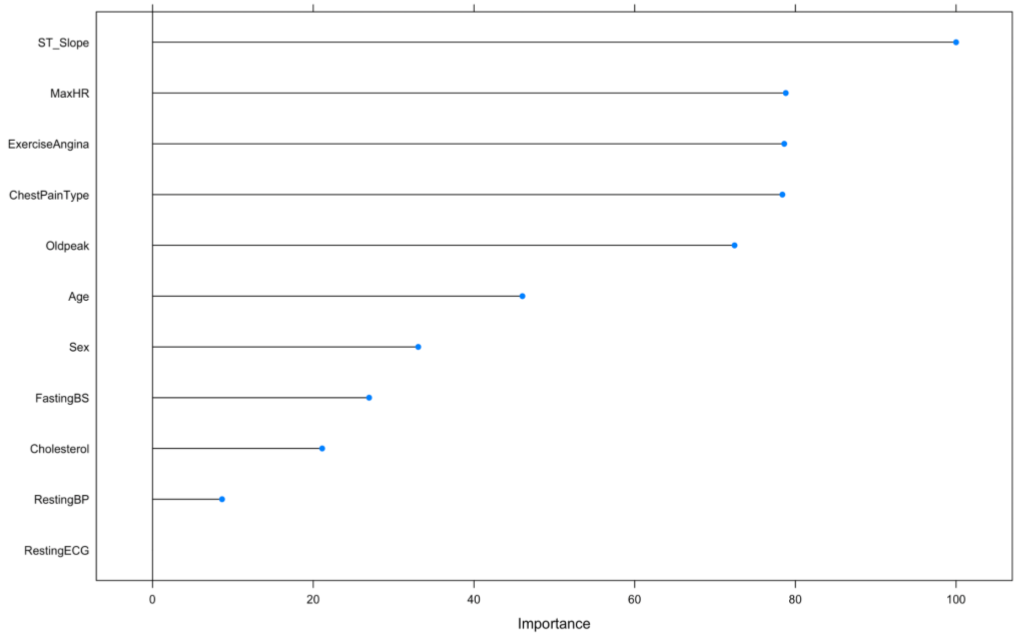


Figure 4: Variable importance with SVM linear kernel classification

5.4.2. SVM with polynomial kernel

We applied SVM with polynomial kernel on the dataset with 10 cross fold validation to minimize the bias in the process. This model was able to classify **89.62%** of test data accurately with a precision of 87.95% and sensitivity of 89.02%.

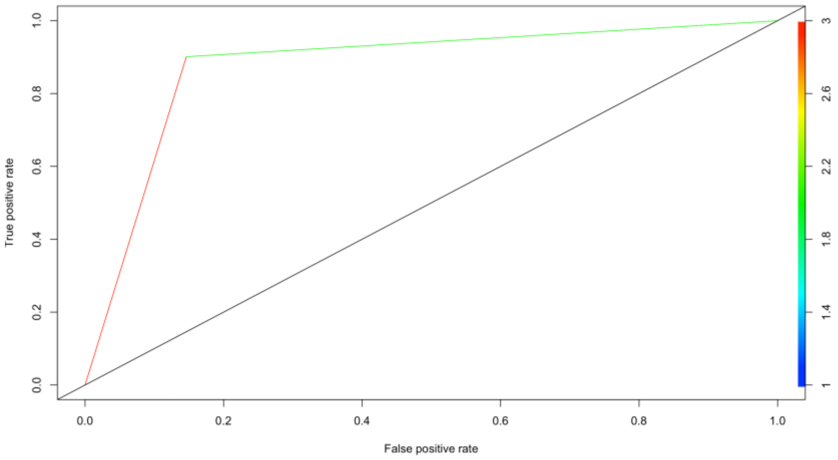


Figure 5: RoC Curve with SVM polynomial kernel classification

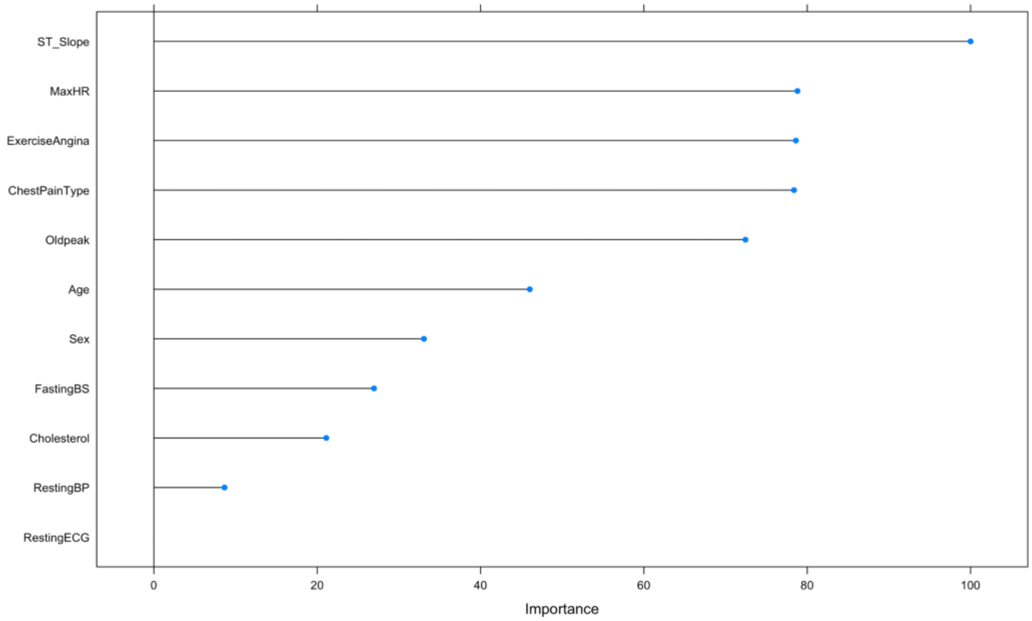


Figure 6: Variable importance with SVM polyomial kernel classification

5.4.3. SVM with radial kernel

We applied SVM with radial kernel on the dataset with 10 cross fold validation to minimize the bias in the process. This model was able to classify **90.16%** of test data accurately with a precision of 90% and sensitivity of 87.80%.

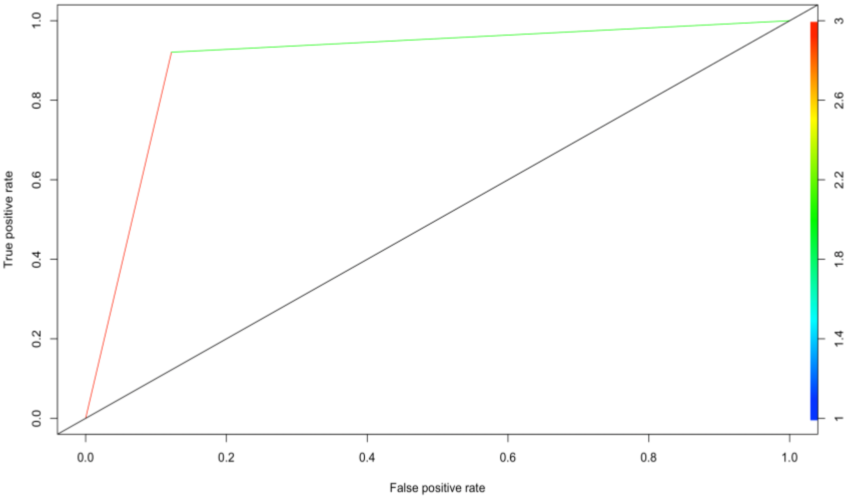


Figure 7: ROC Curve with SVM Radial kernel classification

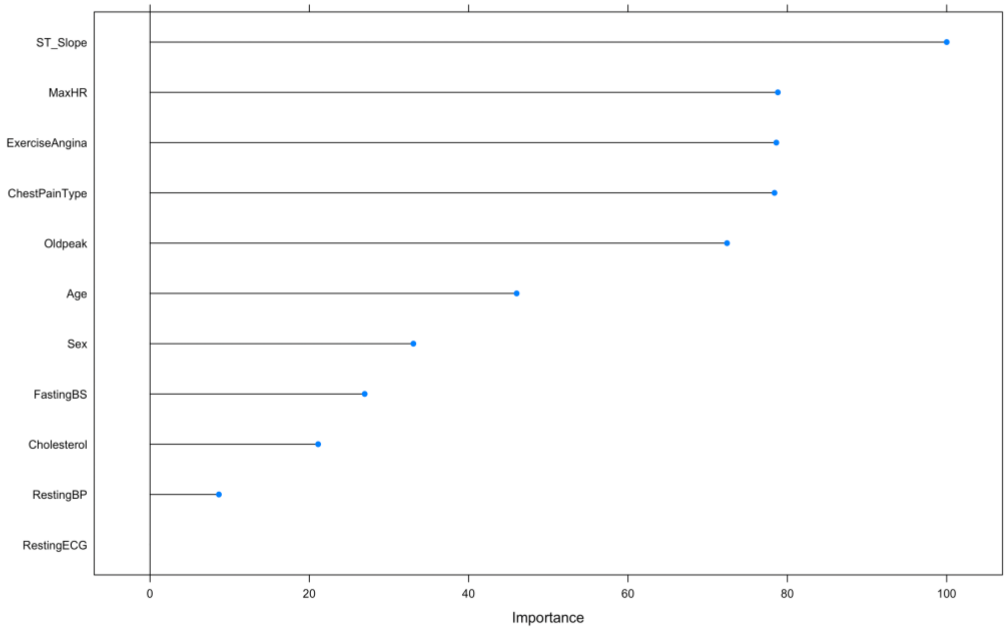


Figure 8: Variable importance with SVM Radial kernel classification

5.5. Ensemble Algorithms

An ensemble method is a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model. Bootstrapping is the general procedure used to reduce the variance of high variance algorithms like decision tree and combining them to output a prediction with low accuracy. We will be focusing on Bagging and Random Forest for the purpose of this dataset.

5.5.1. Bagging Algorithm

Bagging is the application of the Bootstrap procedure to a high-variance machine learning algorithm like decision trees. It allows us to be less concerned with the individual tree overfitting the data. Individual trees are allowed to be grown deep resulting in a high variance and low bias. Ensembling all the fitted model with high variance results in a low variance and low bias overall model.

We applied Bagged Decision tree on the dataset with 10 cross fold validation to minimize the bias in the process. This model was able to classify **88.52%** of test data accurately with a precision of 88.61% and sensitivity of 85.37%.

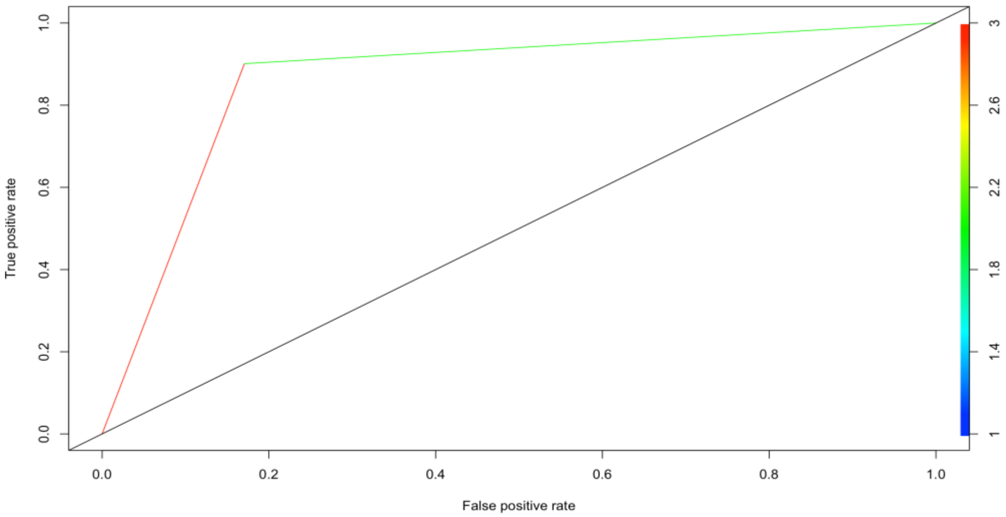


Figure 9: RoC Curve with Bagged Decision tree classification

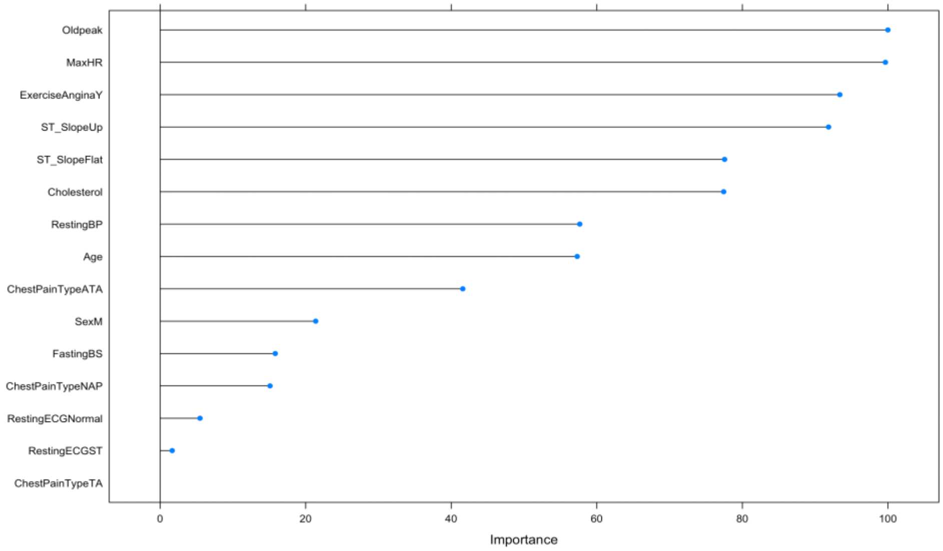


Figure 10: Variable importance with Bagged Decision tree classification

5.5.2. Random Forest

Bagging works best when underlying classification tree are uncorrelated. But most of the bagged trees are correlated as they are allowed to look through all the features for split points. To overcome this, random forest is only allowed to look at a subset of feature while producing the decision tree resulting in uncorrelated trees and shallow trees.

We applied Random Forest on the dataset with 10 cross fold validation to minimize the bias in the process. This model was able to classify **91.26%** of test data accurately with a precision of 91.25% and sensitivity of 89.02%.

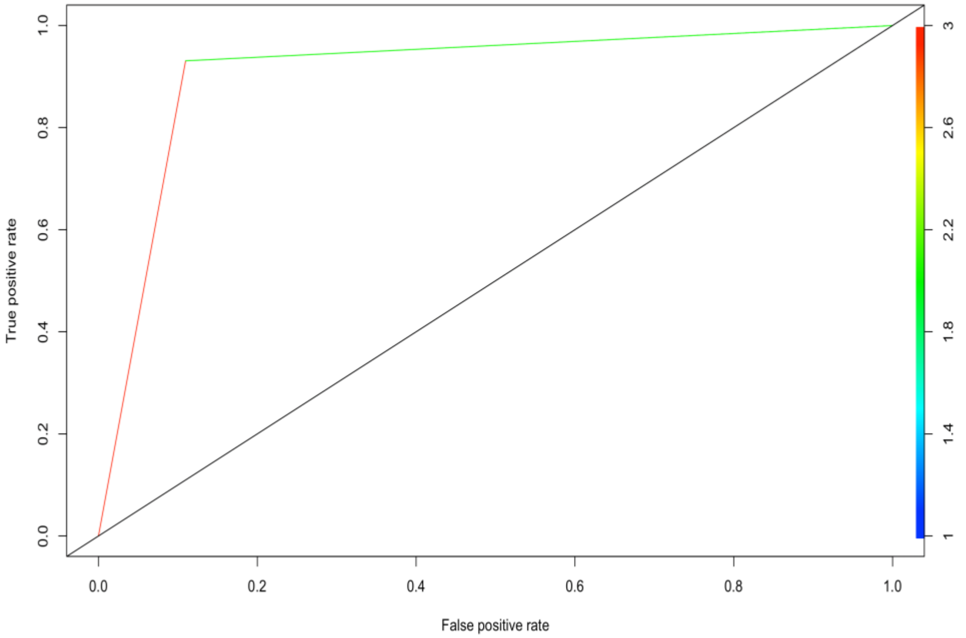


Figure 11: RoC Curve with Random Forest classification

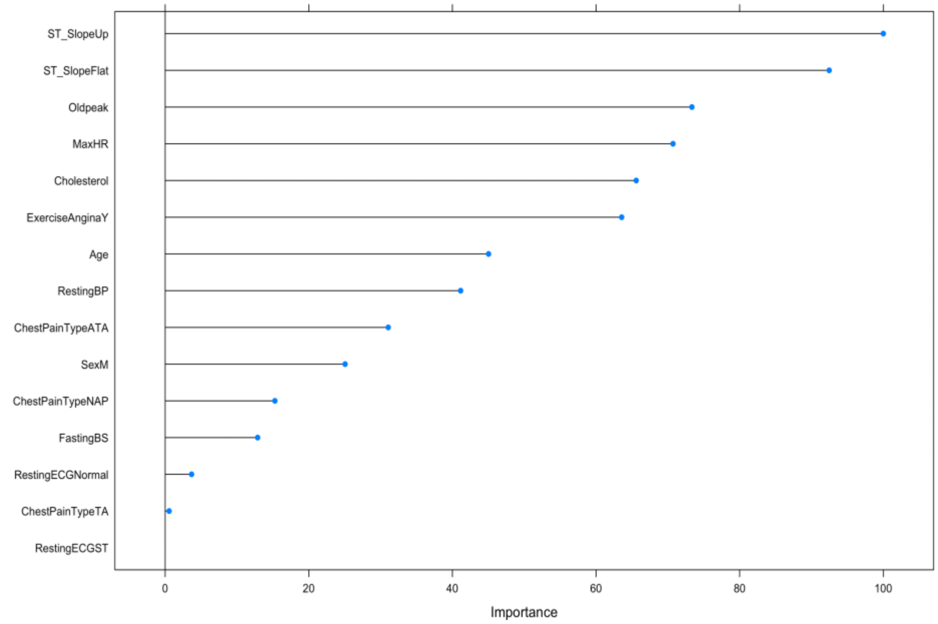


Figure 12: Variable importance with Random Forest classification

6. Result Comparison

Model	Accuracy	Precision	Sensitivity
Decision Tree	87.43%	90.41%	80.49%
SVM with linear kernel	89.07%	87.80%	87.80%
SVM with polynomial kernel	89.62%	87.95%	89.02%
SVM with radial kernel	90.16%	90%	87.80%
Bagged Decision Tree	88.52%	88.61%	85.37%
Random Forest	91.26%	91.25%	89.02%

All model seems to perform well in case of cardiovascular disease prediction, but Random Forest seems to be working exceptionally well here with a 92% accuracy. We can employ random forest model for the purpose of pre-detecting a patient with a heart disease.

7. Conclusion

In our work, we have tried to predict the chance of getting a heart disease using various human health and lifestyle attributes. This work shows that it is possible to diagnose cardiovascular disease early in the cycle. Classifier discussed above can be used in the early detection of heart diseases. The patient can be forewarned to change their lifestyle habit and to focus on health and exercise.

8. ACKNOWLEDMENTS

I am grateful to Dr. Li Xu, Assistant Professor, Department of Finance, New Jersey City University, for guiding me in every aspect of this capstone project from data collection to data analysis and result finding.

9. References

1. *WHO Report: https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1*
2. *CDC: https://www.cdc.gov/heartdisease/facts.htm*
3. *CDC: https://www.cdc.gov/heartdisease/about.htm*
4. *R. S. Singh, B. S. Saini, and R. K. Sunkaria, "Detection of coronary artery disease by reduced features and extreme learning machine," Medicine and Pharmacy Reports, vol. 91, no. 2, pp. 166–175, 2018.*
5. *G. Guidi, M. C. Pettenati, P. Melillo, and E. Iadanza, "A machine learning system to improve heart failure patient assistance," IEEE Journal of Biomedical and Health Informatics, vol. 18, no. 6, pp. 1750–1756, 2014.*
6. *K. Srinivas, 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 Proceedings of 2010 5th International Conference on Computer Science & Education, pp. 1344–1349, IEEE, Hefei, China, August 2010.*
7. *G. Parthiban and S. K. Srivatsa, "Applying machine learning methods in diagnosing heart disease for diabetic patients," International Journal of Applied Information Systems, vol. 3, no. 7, pp. 25–30, 2012.*