Cardiovascular Disease Prediction using ML Models

Table of Contents

Introduction

According to the World Health Organization, cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year. Four out of five CVD deaths are due to heart attacks and strokes, and one third of these deaths occur prematurely in people under 70 years of age. Coronary heart disease i.e. coronary artery disease (CAD) is the most common type of heart disease and the main cause of heart attacks.

CAD mortality rate is high; however, if the diagnosis is made early enough, the chance of survival is higher. Emerging machine learning (ML) and data mining (DM) technologies, that are beginning to transform medicine and healthcare, could also improve the diagnosis and treatment of CAD. Predictive models are being developed in order to enable early detection of high-risk patients. These models could save many lives and at the same time contribute to reducing healthcare costs. Scientists estimate that uniformly taken data from a million of patients would be necessary to develop a reliable model.

The dataset under study is the **Heart Disease Dataset**, available on the **IEEE Data Port Platform** [1]. It is curated by combining **5** popular heart disease datasets already available independently, but not combined before. Datasets from Cleveland, Hungaria, Switzerland, Long Beach and Statlog (Heart) are combined over **11** common features thus creating the largest heart disease dataset available so far for research purposes. It has **1190** instances which provide comprehensive and fundamental information for early detection of CAD disease.

The goal of this project is to use machine learning tools to predict whether an individual has coronary heart disease and evaluate the quality of the predictions.

Here are the attributes of the dataset:

Heart Disease Dataset Attribute Description

S.No.	Attribute	Code given	Unit	Data type
1	age	Age	in years	Numeric
2	sex	Sex	1, 0	Binary
3	chest pain type	chest pain type	1,2,3,4	Nominal
4	resting blood pressure	resting bp s	in mm Hg	Numeric
5	serum cholesterol	cholesterol	in mg/dl	Numeric
6	fasting blood sugar	fasting blood sugar	1,0 > 120 mg/dl	Binary
7	resting electrocardiogram results	resting ecg	0,1,2	Nominal
8	maximum heart rate achieved	max heart rate	71–202	Numeric
9	exercise induced angina	exercise angina	0,1	Binary
10	oldpeak =ST	oldpeak	depression	Numeric
11	the slope of the peak exercise ST segment	ST slope	0,1,2	Nominal
12	class	target	0,1	Binary

Description of Nominal Attributes

Attribute	Description
Sex	1 = male, 0= female;
Chest Pain Type	-- Value 1: typical angina -- Value 2: atypical angina -- Value 3: non-anginal pain -- Value 4: asymptomatic
Fasting Blood sugar	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
Resting electrocardiogram results	-- Value 0: normal -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
Exercise induced angina	1 = yes; 0 = no
the slope of the peak exercise ST segment	-- Value 1: upsloping -- Value 2: flat -- Value 3: downsloping
class	1 = heart disease, 0 = Normal

The 11 attributes of heart disease dataset provide data easily measurable by primary health doctors, while 12th attribute, coded as target, classifies the population into two groups: the individuals that have and don't have heart disease. The disease is diagnosed by angiography, a risky and costly procedure.

The project has been developed in 4 stages:

- **Preprocessing** - in order to make the dataset more suitable for analysis, the missing values are checked and treated if necessary; dataset features are analyzed and transformed.
- **Data exploration -** in this stage features from the dataset are analyzed, focusing in particular on correlation between the presence of heart disease and remaining 11 features.
- **Models implementation** – in this stage different supervised ML models are used in order to predict the illness.
- **Evaluation –** all applied models are evaluated and then classified based on the quality of prediction.

Preprocessing

Firstly, I've checked whether there were missing values in the dataset. The results obtained indicated that there weren't missing values and therefore there was no need for data correction.

Secondly, invalid values i.e. values that are out of meaningful domain were checked for 4 attributes (it should have been done for all features, should I had more time): ***age***, ***sex***, ***resting blood pressure*** and ***serum cholesterol level***.

```
## [1] "age"
## [1] 0
## [1] "sex"
## [1] 0
## [1] "resting_bp_s"
## [1] 0
```

All values for attributes ***age***, ***sex*** and ***resting blood pressure*** were within required range, no data improvement was needed.

```
## [1] "with cholesterol level 0"
## [1] 172
```

The ***serum cholesterol*** check for invalid values has shown that 172 individuals had their cholesterol level set to 0. Firstly, I believed that 0 indicated "normal" cholesterol level and a correction was to be made in order to improve the data. I thought of replacing 0 cholesterol level in healthy individuals with a value which is considered "normal", for example something between 125 and 200 mg/dl, according to the US National Library of Medicine (125 – 200) [2].

```
## [1] "healthy, with cholesterol level 0"
## [1] 20
## [1] "ill, with cholesterol level 0"
## [1] 152
```

However, further data examination showed that among individuals with colesterol level set to 0, much more of them had heart disease - **152**, while there were only **20** healthy individuals. Therefore some other type of correction should have to be applied.

Removing **172** rows from the dataset was not an option since it meant reducing the dataset by almost **15%** of its original size.

In this situation I chose to the following: calculate Q_1 and Q_3 of cholesterol level for the subpopulation of patients that had valid cholesterol values in the dataset and subsequently set the cholesterol level for 152 ill individuals to a uniformly dispersed value close to Q_3:

Interval $\left(Q_3 - \left[\frac{Q_3 - Q_1}{8}\right], Q_3 + \left[\frac{Q_3 - Q_1}{8}\right]\right)$,

and for 20 healthy individuals to a uniformly dispersed value close to Q_1:

interval $Q_1 - \left[\frac{Q_3 - Q_1}{8}\right], Q_1 + \left[\frac{Q_3 - Q_1}{8}\right]$

As a result, the cholesterol level for 20 healthy individuals was set to values around 209, and for 152 individuals with heart disease to values around 276.

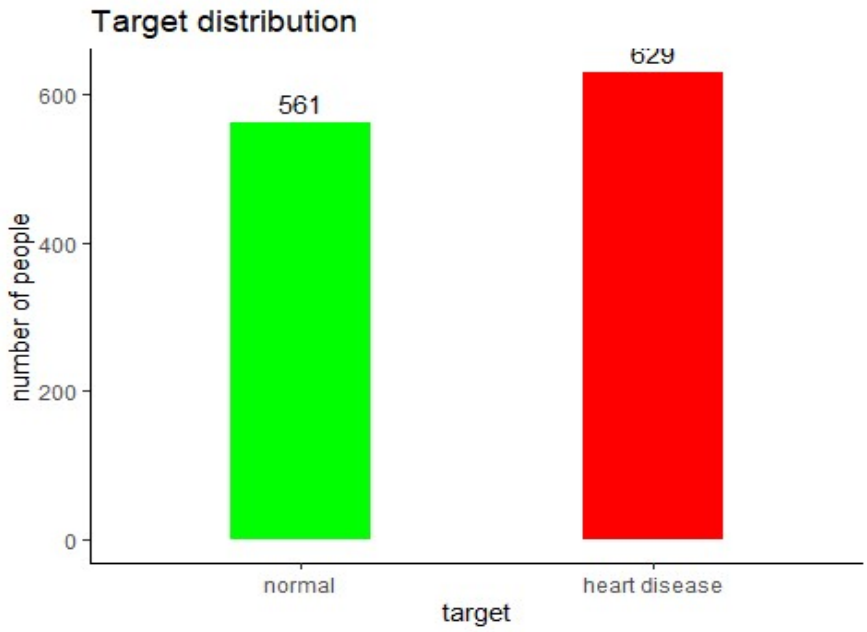
Data exploration

Since the research question is to predict whether an individual has coronary heart disease, the variable **target** is the dependent one in this analysis, and its correlation with the remaining 11 features will be analyzed. This attribute is a discrete variable and prediction of its values will be performed through a classification process.

1. Distribution of outcome

The first analysis of the attribute target has shown that **629** patients out of **1190** have heart disease, while the remaining **561** are healthy.

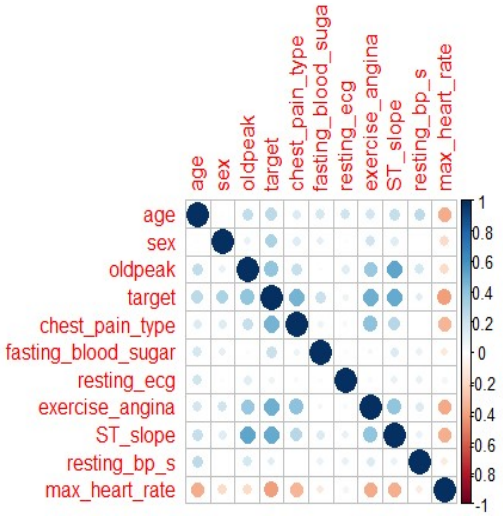
```
## [1] "target"
## .
##        normal heart disease
##           561           629
```



2. Correlation

The correlation coefficients have been calculated and the correlation diagram and matrix displayed. Feature selection was not considered here as the available literature suggests these 11 features to be the indispensable for predicting any coronary disease. In addition, the correlation matrix also suggests that no feature should be excluded as even the ones weakly correlated to the variable ***target*** show high correlation coefficient when related to variables highly correlated to ***target.***

CORRELATION DIAGRAM



In this diagram, positive correlation is marked with different shades of blue, while negative correlation is marked with different shades of brown. Higher the correlation, stronger is the blue or brown shade.

CORRELATION MATRIX

	age	sex	oldpeak	target	chest pane type	fasting blood sugar	resting ECG	exercise angina	ST slope	resting blood pressure	max heart rate
age	1	0.015	0.245	**0.262**	0.149	0.179	0.195	0.188	0238	0.258	-0.369
sex	0.015	1	0.096	**0.311**	0.138	0.111	-0.022	0.194	0.128	-0.006	-0.182
oldpeak	0.245	0.096	1	**0.398**	0.224	0.031	0.126	0.371	0.525	0.176	-0.184
target	**0.262**	**0.311**	**0.398**	1	**0.460**	**0.217**	**0.073**	**0.481**	**0.506**	**0.121**	**-0.413**
chest pane type	0.149	0.138	0.224	**0.460**	1	0.076	0.036	0.403	0.227	0.009	-0.337
fasting blood sugar	0.179	0.111	0.031	**0.217**	0.076	1	0.032	0.053	0.146	0.088	-0.119
resting ECG	0.195	-0.022	0.126	**0.073**	0.036	0.032	1	0.038	0.094	0.096	0.059
exercise angina	0.188	0.194	0.371	**0.481**	0.403	0.053	0.038	1	0.393	0.142	-0.378
ST slope	0.238	0.128	0.525	**0.506**	0.277	0.146	0.094	0.393	1	0.089	-0.351
resting blood pressure	0.258	-0.006	0.176	**0.121**	0.009	0.088	0.096	0.142	0.089	1	-0.101
max heart rate	-0.369	-0.182	-0.184	**-0.413**	-0.337	-0.119	0.059	-0.378	-0.351	-0.101	1

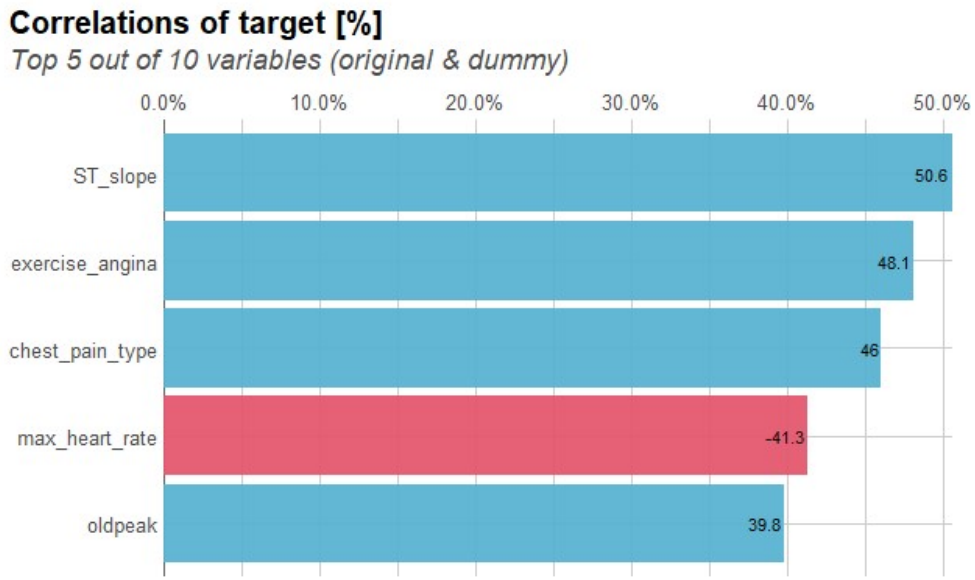
The correlation matrix suggests that the variables are mostly positively correlated, apart from ***maximum heart rate*** which is negatively correlated with all but one variable. This was expected as healthy, younger individuals can reach higher maximal heart rates.

Each variable was analyzed separately: significant statistical values were calculated and their distributions plotted in order to see eventual anomalies. ***Serum cholesterol, resting blood***

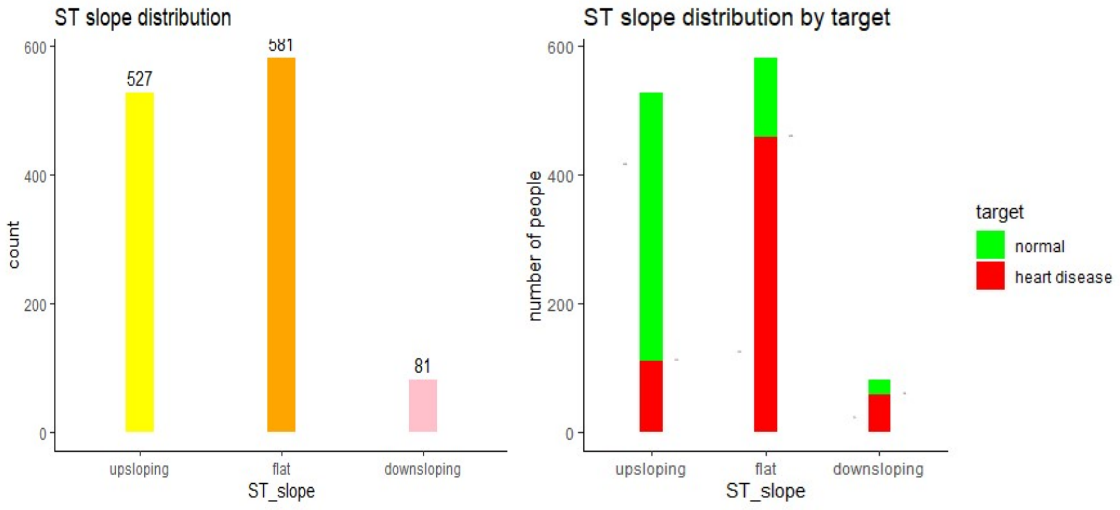
pressure and ***oldpeak*** had outliers, but they were not numerous and could not influence significantly the distribution of these variables.

The correlation of the dependent variable ***target*** to each of the 11 independent ones was studied separately.

Here are the variables with the highest correlation coefficient with ***target***

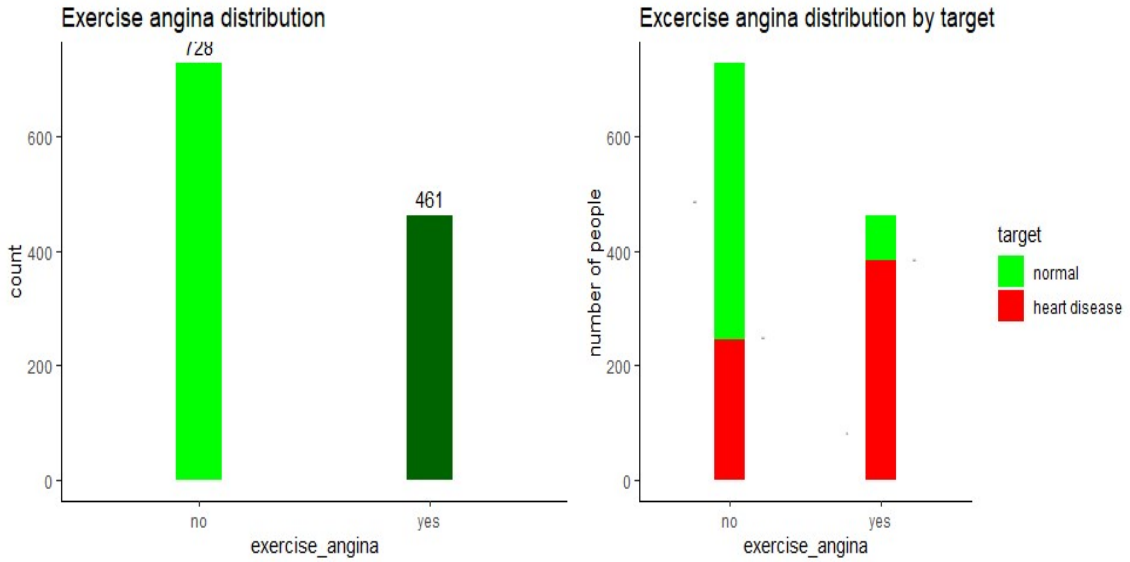


The slope of the peak exercise ST segment (***ST slope***) has the highest correlation with the dependent variable (in ECG a ST Segment represents the interval between ventricular depolarization and repolarization):



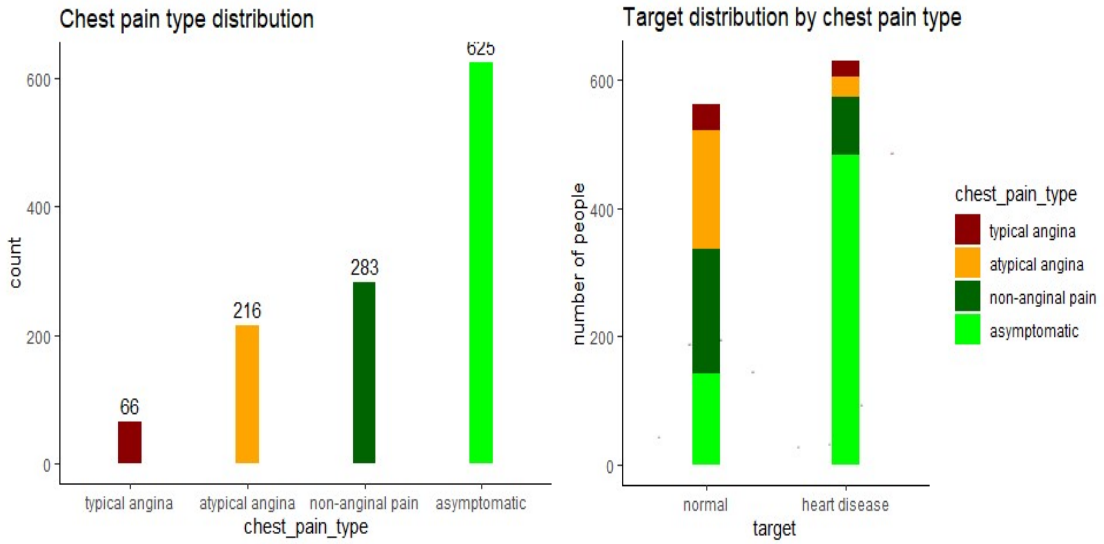
The results show that people with flat ***ST slope*** are more likely to get heart disease, while those with upsloping ***ST slope*** are less likely to be ill.

The exercise induced angina (***exercise angina***) is also highly correlated to ***target*** (angina is a chest pain or discomfort caused when heart muscle doesn't get enough oxygen-rich blood. It is usually triggered by physical activity. When we climb stairs, exercise or walk, our heart demands more blood, but narrowed arteries slow down blood flow).



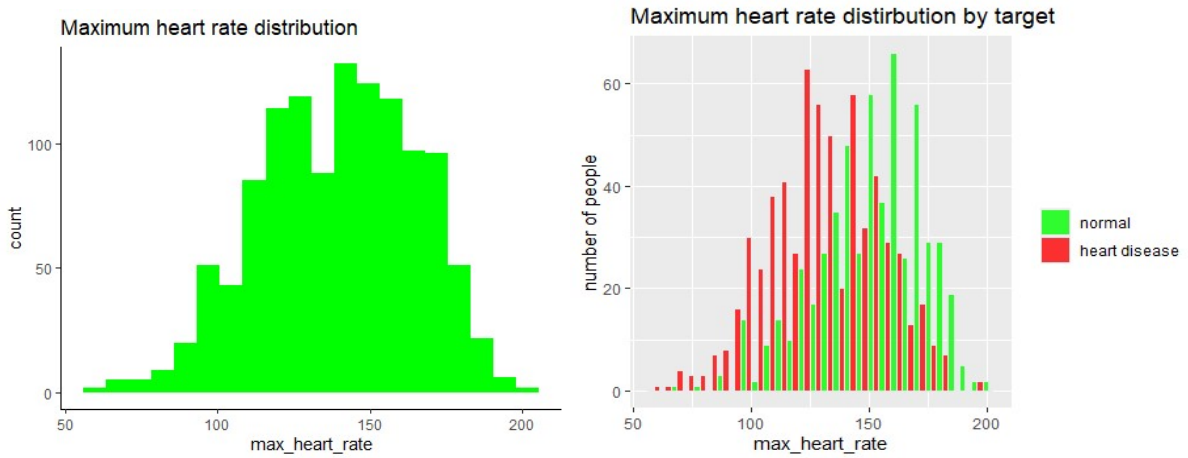
The results indicate that most people with exercise induced angina have coronary heart disease. However, in a small number of healthy individuals exercise can induce angina.

The ***chest pain type*** has also relatively high correlation coefficient with ***target***.



The graphs show that most patients with CAD are asymptomatic, i.e. they do not have chest pain.

The maximum heart rate achieved (***max heart rate***) has high but negative correlation to ***target***, unlike other variables which are positively correlated.



As it was expected, people with heart disease have lower maximal heart rate.

This was only a summary of the most significant correlations betwenne the ***target*** and independent variables. While working on this project, each variable as well as the correlation between them was studied in detail, but due to the lack of time it cannot be presented (as it did not infuence further data analysis).

Models implementation

Five supervised ML models were applied to this dataset in order to predict the values of the target variable: **K Nearest Neighbors** , **Naïve Bayes**, **Support Vector Machine with Linear Kernel**, **Support Vector Machine with Radial Kernel** and **Random Forest**.

R language and R-Studio environment were chosen for implementation of these predictor models. The most important and helpful function in this process was ***train*** function from ***caret*** library] – it was used as umbrella function that covered all R functions and libraries needed for the models' implementation. Thanks to ***train*** function, it was possible to handle all learning models in a uniform manner. At the moment, the ***caret*** library supports "out-of-the box" use of over 230 classification and regression models.

Evaluation

The implemented models are compared using **k-fold cross-validation**. The value of the parameter **k** is set to **10**. The 10-fold cross-validation is executed using ***caret R functions***. In order to achieve exactly the same conditions for comparison in all 10-fold cross-validation scenarios, random generator is set on predefined value 155294099.

The quality of the implemented ML classification models is compared by calculating **sensitivity (or recall), specificity** and **accuracy**, as these are measures mainly in use in medical domain. In order to calculate them, a confusion matrix have to be created for each model. In this case, confusion matrix is for a binary classifier, as we are predicting whether an individual have coronary heart disease or not.

CONFUSION MATRIX

		Actual class	
		P	**N**
Predicted class	**P**	TP	FP
	N	FN	TN

- **Actual class** - is the class from the dataset under study. It is established by angiography, so we have two groups of individuals: the ones that have the coronary heart disease (**P** - positive), and those who are not ill (**N** – negative)
- **Predicted class** is the one obtained by ML algorithms, and it classifies the individuals in two groups: P – they have the disease, N – they don't have the disease
- $\boldsymbol{TP}$ (True Positive) - These are correctly classified cases in which we predicted P (the patients would have the disease), and they actually do have it.
- $\boldsymbol{TN}$ (True Negative) - These are correctly classified cases in which we predicted N (the patients wouldn't have the disease), and they do not have it.
- $\boldsymbol{FP}$ (False Positive) - This is the number of falsely classified cases: we predicted the patients would have the disease, and they actually don't have it.
- $\boldsymbol{FN}$ (False negative) This is also the number of falsely classified cases: we predicted the patients would not be ill, and they actually have the disease.

The calculations of sensitivity, specificity and accuracy are based on the confusion matrix, and this how their values are obtained:

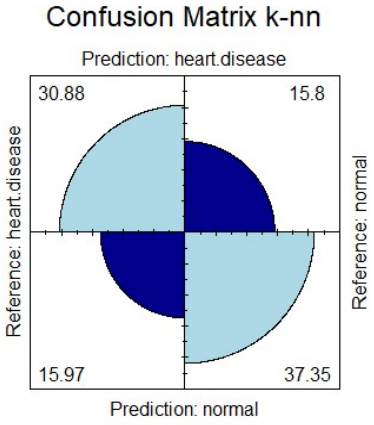
$$Specificity = \frac{TN}{TN + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

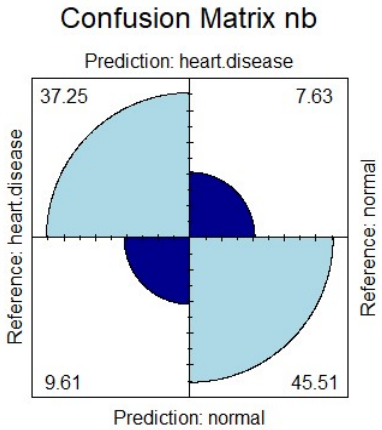
Here are the results of the calculations:

k-NN



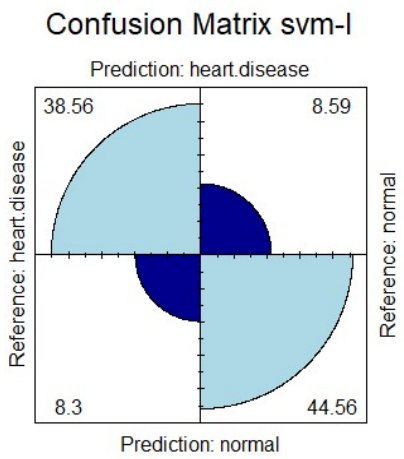
Specificity	0.7027726
Sensitivity	0.6591195
Accuracy	0.6823183

Nb



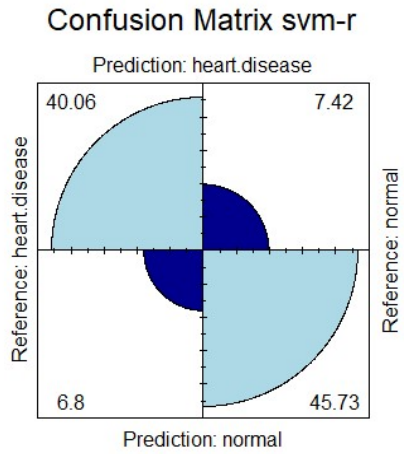
Specificity	0.8563771
Sensitivity	0.7949686
Accuracy	0.8276031

Svm-l



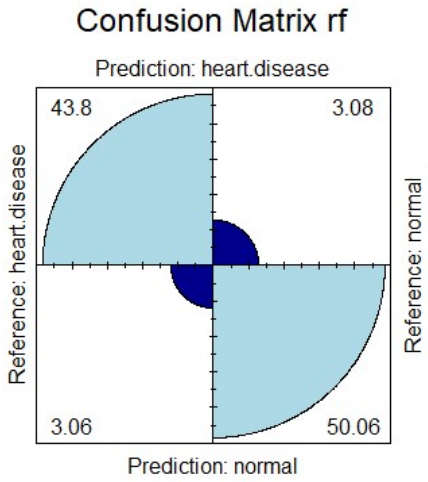
Specificity	0.8384473
Sensitivity	0.8228512
Accuracy	0.8311395

Svm-r



Specificity	0.8604436
Sensitivity	0.8549266
Accuracy	0.8578585

Rf



Specificity	0.9419593
Sensitivity	0.9348008
Accuracy	0.9386051

The graphical representation of confusion matrix of each model suggests that the best results are obtained by Random Forest.

However, the quality of a binary classifier system is often illustrated by ROC (Receiver Operating Characteristic) curve. This type of diagram shows the performance of a classification model at all classification thresholds.

The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

True Positive Rate (TPR) is a synonym for **recall** (or *Sensitivity*) or probability of detection, and is defined as:

$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate (FPR) is also known as probability of false alarm and can be calculated as:

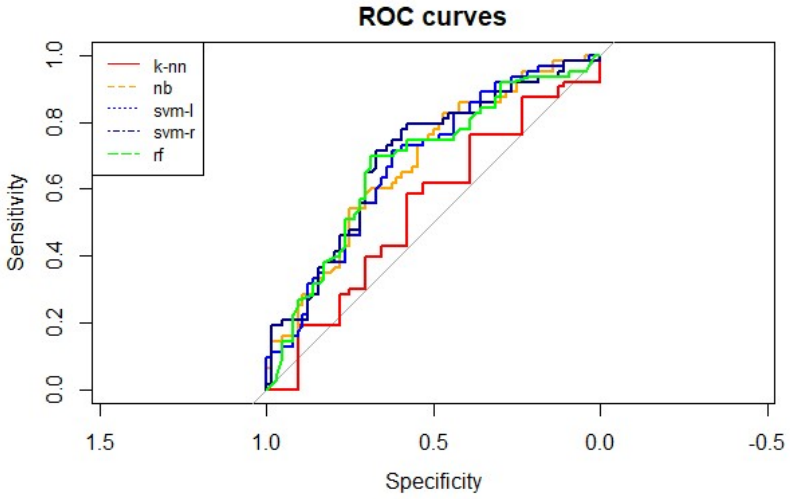
$$FPR = \frac{FP}{TN + FP}$$

Or:

$$FPR = 1 - Specificity$$

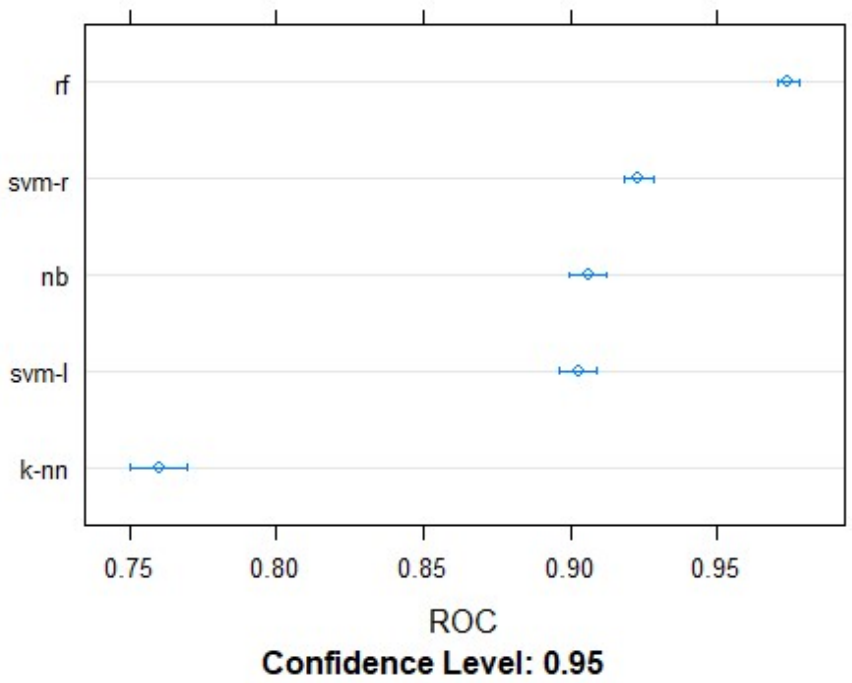
A ROC curve plots TPR vs. FPR at different classification thresholds.

A measure of the prediction quality is the Area Under the ROC Curve (AUC) which measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1). The closer AUC is to 1, better is the prediction.

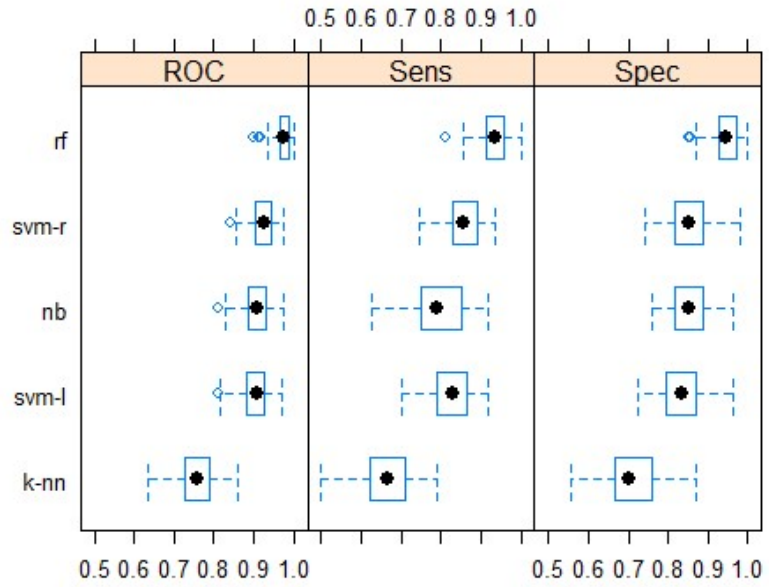


From this graph it is not easy to conclude which model has the largest area under the curve, so we will create other graphic representation to illustrate the quality of all five models.

This graph shows the ROC values with the confidence level of 0.95 (i.e. p=0.05).

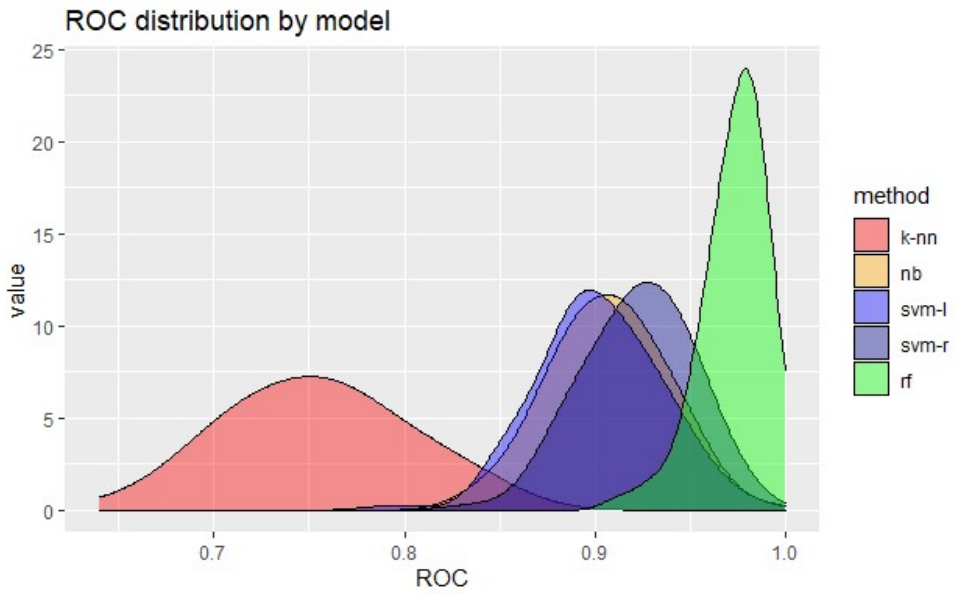


It is obvious that Random Forest is by far the best model as its ROC value is very close to 1. Yet another graph representation confirms these findings.



Besides ROC value at 0.95 confidence level, Random Forest has best results in terms of sensitivity and sensibility.

This result is even more obvious if ROC distributions for all five models are visualized:



Conclusion

The goal of this project wss to predict whether an individual is likely to have coronary heart disease using the 11 features easily measurable by primary health doctor. Should a reliable model be developed, many lives would be saved: earlier diagnosis could be possible, and reduction in healthcare cost as well since angiography, a risky and costly procedure, would be performed only on those who are highly likely to have heart disease.

In this analysis, among 5 classification models the best results by far are obtained by Random Forest model.

However, we should take care not to jump too early into general conclusions and consider the ***No free lunch theorem for supervised learning***. In other words, Random Forest may have good quality for this specific prediction problem, but it is very likely it'll be inadequate for some other problem. In addition, if we change features for the analysis of the same problem, there is no guarantee that Random Forest would be the best prediction model, as in this case.

References

[1] M. Siddhartha, "IEEE data port: Heart disease dataset (comprehensive)." [Online]. Available: https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive

[2] U.S. National Library of Medicine, "Cholesterol levels - what you need to know." [Online]. Available: https://medlineplus.gov/cholesterollevelswhatyouneedtoknow.html

[3] K. P. Murphy, *Machine learning - a probabilistic perspective*. MIT Press, 2012.

[4] R Core Team, *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2019 [Online]. Available: https://www.R-project.org

[5] R Documentation team, "R documentation: Knn." [Online]. Available: https://www.rdocumentation.org/packages/class/versions/7.3-17/topics/knn

[6] W. V. B. Ripley, *Package class*. 2020 [Online]. Available: https://cran.r-project.org/web/packages/class/class.pdf

[7] R Documentation team, "R documentation: NaiveBayes." [Online]. Available: https://www.rdocumentation.org/packages/klaR/versions/0.6-15/topics/NaiveBayes

[8] K. L. C. Roever N. Raabe, *Package klaR*. 2020 [Online]. Available: https://cran.r-project.org/web/packages/klaR/klaR.pdf

[9] R Documentation team, "R documentation: Ksvm." [Online]. Available: https://www.rdocumentation.org/packages/kernlab/versions/0.9-29/topics/ksvm

[10] K. H. A. Karatzoglou A. Smola, *Package kernlab*. 2019 [Online]. Available: https://cran.r-project.org/web/packages/kernlab/kernlab.pdf

[11] R Documentation team, "R documentation: RandomForest." [Online]. Available: https://www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/randomForest

[12] A. L. L. Breiman A. Cutler, *Package randomRorest*. 2018 [Online]. Available: https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

[13] R Documentation team, "R documentation: Train." [Online]. Available: https://www.rdocumentation.org/packages/caret/versions/4.47/topics/train

[14] S. W. M. Kuhn J. Wing, *Package caret*. 2020 [Online]. Available: https://cran.r-project.org/web/packages/caret/caret.pdf

[15] M. Kuhn, "Library caret: Available models." [Online]. Available: https://topepo.github.io/caret/available-models.html

[16] M. R. R. Alizadehsani, "A database for using machine learning and data mining techniques for coronary artery disease diagnosis," *Scientific Data*, vol. 6, no. 227, pp. 1–12, 2019, doi: 10.1002/andp.19053221004. [Online]. Available: https://www.nature.com/articles/s41597-019-0206-3

[17] D. Walpert, "The lack of a priori distinctions between learning algorithms," *Neural Computation*, vol. 8, no. 7, pp. 1341–1390, 1996, doi: 10.1162/neco.1996.8.7.1341.