Heart Disease Prediction System

Chandra Sekhar Polisetti

29 January, 2022

Contents

1 Introduction **4**

2 Overview **4**

3 Data Preparation **5**

3.1 Data Download 5

3.2 Data Cleanup 5

4 Analysis **7**

4.1 Heart Disease dataset fetures 7

4.2 Heart Disease Distribution 8

4.3 Heart Disease by Sex 9

4.4 Heart Disease by Chest Pain 10

4.5 Heart Disease by Fasting Blood Sugar 10

4.6 Heart Disease by Resting Electro Cardiographic Results 11

4.7 Heart Disease by Exercise Induced Angina 12

4.8 Heart Disease by Slope 13

4.9 Heart Disease by Number of major vessels (0-3) colored by fluoroscopy 14

4.10 Heart Disease by Thalium heart scan 15

4.11 Heart Disease vs Resting Blood Pressure 15

4.12 Heart Disease vs Serum Cholestoral 16

4.13 Heart Disease vs Maximum Heart Rate Achieved 17

4.14 Heart Disease vs ST depression induced by exercise relative to rest 18

4.15 Heart Disease vs Age 19

5 Methods **21**

5.1 Data Partition for Training & Testing 21

5.2 Algorithm Evaluation 22

5.3 Model 1 - Novice Heart Disease Model 22

5.4 Model 2 - Logistic Regression Heart Disease Model 23

5.5 Model 3 - KNN Heart Disease Model 26

5.6 Model 4 - Classification Tree Heart Disease Model 28

5.7 Model 5 - Random Forest Heart Disease Model 31

5.8 Model 6 - Ensemble Model 34

6 Results **36**

7 Conclusion **37**

1 Introduction

As per Centers for Disease Control and Prevention (CDC), in 2018 30.3 million U.S. adults were diagnosed with heart disease. Every year, about 647,000 Americans die from heart disease, making it the leading cause of death in the United States.

Heart disease causes 1 out of every 4 deaths.

Due to the limitations of the medical resources and high cost of medical services quality diagnosis is un anviable to many and an effective heart disease prediction system is much needed.

An effective Heart Disease Prediction System to predict the presence of Heart Disease helps in diagnosing the heart disease and there by providing the right treatment to the patients on time.

This project presents you the steps involved in building an effective Heart Disease Prediction Model.

2 Overview

Predicting whether the patient is Healthy (No heart disease) or Unhealthy (has heart disease) with an effective machine learning Model is the goal of this project.

We will build few machine learning algorithms and pick the algorithm that performs better and meets the sensitivity and specificity expectations.

We start with downloading and cleaning the data, followed by analysis of the data given taken from the UCI machine learning repository.

Predicting the presence of heart disease by classifying the output as Unhealthy or Healthy is a Classification problem. Based on the characteristics of the data few machine learning algorithms which classifies the output are used to predict the heart disease.

All the algorithm performances would be analyzed and presented in the results section, moreover we also recommended algorithms based on the sensitivity and specificity expectations.

Here is the High level process followed to build the machine learning model, and is also the outline of this report, each step will be dealt with great detail in the subsequent sections

1. Data Preparation - Data Preparation, Cleanup and Wrangling.
2. Analysis - Analyze the data to gain insights into the data.
3. Methods - Various machine learning algorithms would used to predict heart disease
4. Results - Results of various algorithms used would be discussed
5. Conclusion - Conclude the report with limitations and future work

3 Data Preparation

Data Preparation downloads and wrangles the data to facilitate the data analysis. Primary focus would be to download, clean the data.

3.1 Data Download

Heart disease dataset is downloaded form the UCI machine learning repository and here is the path to the repository, http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data.

Note that there are 3 other unprocessed datasets in the UCI machine learning repository which were not considered. We have taken the processed cleaveland dataset for this project.

3.2 Data Cleanup

Lets look at the dimensions of the dataset

```
## [1] 303  14
```

As we can see there are 14 dimensions and the total number of rows are 304. Total number of rows , 303, are quite low for the 14 dimensions we have to get a good predicting power. Moreover as the number of dimensions are high certain machine learning algorithms like knn may not be suitable.

Let's look at the first 6 rows of the data.

Table 1: First 6 rows of data from heart disease dataset

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14
63	1	1	145	233	1	2	150	0	2.3	3	0.0	6.0	0
67	1	4	160	286	0	2	108	1	1.5	2	3.0	3.0	2
67	1	4	120	229	0	2	129	1	2.6	2	2.0	7.0	1
37	1	3	130	250	0	0	187	0	3.5	3	0.0	3.0	0
41	0	2	130	204	0	2	172	0	1.4	1	0.0	3.0	0
56	1	2	120	236	0	0	178	0	0.8	1	0.0	3.0	0

As we can see that there are no column names so let's fix the column names and re look at the data.

Table 2: First 6 rows of heart disease dataset

age	sex	cp	restbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	hd
63	1	1	145	233	1	2	150	0	2.3	3	0.0	6.0	0
67	1	4	160	286	0	2	108	1	1.5	2	3.0	3.0	2
67	1	4	120	229	0	2	129	1	2.6	2	2.0	7.0	1
37	1	3	130	250	0	0	187	0	3.5	3	0.0	3.0	0
41	0	2	130	204	0	2	172	0	1.4	1	0.0	3.0	0
56	1	2	120	236	0	0	178	0	0.8	1	0.0	3.0	0

Lets look at the structure of the dataset

```
## 'data.frame':    303 obs. of  14 variables:
##  $ age    : num  63 67 67 37 41 56 62 57 63 53 ...
##  $ sex    : num  1 1 1 1 0 1 0 0 1 1 ...
##  $ cp     : num  1 4 4 3 2 2 4 4 4 4 ...
##  $ restbps: num  145 160 120 130 130 120 140 120 130 140 ...
##  $ chol   : num  233 286 229 250 204 236 268 354 254 203 ...
##  $ fbs    : num  1 0 0 0 0 0 0 0 0 1 ...
##  $ restecg: num  2 2 2 0 2 0 2 0 2 2 ...
##  $ thalach: num  150 108 129 187 172 178 160 163 147 155 ...
##  $ exang  : num  0 1 1 0 0 0 0 1 0 1 ...
##  $ oldpeak: num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
##  $ slope  : num  3 2 2 3 1 1 3 1 2 3 ...
##  $ ca     : chr  "0.0" "3.0" "2.0" "0.0" ...
##  $ thal   : chr  "6.0" "3.0" "7.0" "3.0" ...
##  $ hd     : int  0 2 1 0 0 0 3 0 2 1 ...
```

Based on the above data we need to make the below changes to the data structure

- Sex column has values 0 and 1 to represent female and male patients, for better readability lets mark them as F for female and M for male values.
- There are missing values in ca and thal, we need to convert them into NA's first and later we will discuss on what to do with the values.
- cp, fbs, restecg , exang , slop needs to be converted into factors
- hd values has 0 represent healthy heart with no heart disease , and 1 ,2 and 3 values represent unhealthy heart. As our main goal is to predict the presence of heart disease, hd 0 values would be converted into as $Healthy$ and hd values $1, 2 or 3$ as $Unhealthy$.

Now after all the above changes here is the modified data structure of the dataset

```
## 'data.frame':    303 obs. of  14 variables:
##  $ age    : num  63 67 67 37 41 56 62 57 63 53 ...
##  $ sex    : Factor w/ 2 levels "F","M": 2 2 2 2 1 2 1 1 2 2 ...
##  $ cp     : Factor w/ 4 levels "1","2","3","4": 1 4 4 3 2 2 4 4 4 4 ...
```

```
##  $ restbps: num  145 160 120 130 130 120 140 120 130 140 ...
##  $ chol   : num  233 286 229 250 204 236 268 354 254 203 ...
##  $ fbs    : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 1 2 ...
##  $ restecg: Factor w/ 3 levels "0","1","2": 3 3 3 1 3 1 3 1 3 3 ...
##  $ thalach: num  150 108 129 187 172 178 160 163 147 155 ...
##  $ exang  : Factor w/ 2 levels "0","1": 1 2 2 1 1 1 1 2 1 2 ...
##  $ oldpeak: num  2.3 1.5 2.6 3.5 1.4 0.8 3.6 0.6 1.4 3.1 ...
##  $ slope  : Factor w/ 3 levels "1","2","3": 3 2 2 3 1 1 3 1 2 3 ...
##  $ ca     : Factor w/ 4 levels "0","1","2","3": 1 4 3 1 1 1 3 1 2 1 ...
##  $ thal   : Factor w/ 3 levels "3","6","7": 2 1 3 1 1 1 1 1 3 3 ...
##  $ hd     : Factor w/ 2 levels "Healthy","Unhealthy": 1 2 2 1 1 1 2 1 2 2 ...
```

Now that the data structure has been cleaned up to facilitate the analysis, lets work on the missing values.

There are 6 rows with missing values, and there are 303 number of rows in the dataset so the percentage of missing values in the dataset is 1.980198. As this percent is very low we are ignoring the missing values.

4 Analysis

Now that the data is cleaned up and the missing values have been deleted we are ready for data analysis.

This section would mainly analyze data by looking at each of the independent variables.

4.1 Heart Disease dataset fetures

Here is a quick look at dataset and its sample data

Dimensions of the dataset

```
## [1] 297  14
```

Here is some sample data

Table 3: First 6 rows of data from heart disease dataset

age	sex	cp	restbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	hd
63	M	1	145	233	1	2	150	0	2.3	3	0	6	Healthy
67	M	4	160	286	0	2	108	1	1.5	2	3	3	Unhealthy
67	M	4	120	229	0	2	129	1	2.6	2	2	7	Unhealthy
37	M	3	130	250	0	0	187	0	3.5	3	0	3	Healthy
41	F	2	130	204	0	2	172	0	1.4	1	0	3	Healthy
56	M	2	120	236	0	0	178	0	0.8	1	0	3	Healthy

As we can see from the above , each row is a particular patients metrics, namely age,sex,cp,restbps,chol,fbs,restecg,ex
and their associated heart disease condition such as Healthy or Unhealthy.

Lets look at individual attributes and how the the values are distributed to see the variability

4.2 Heart Disease Distribution

Lets look at how the presence of heart disease in the dataset

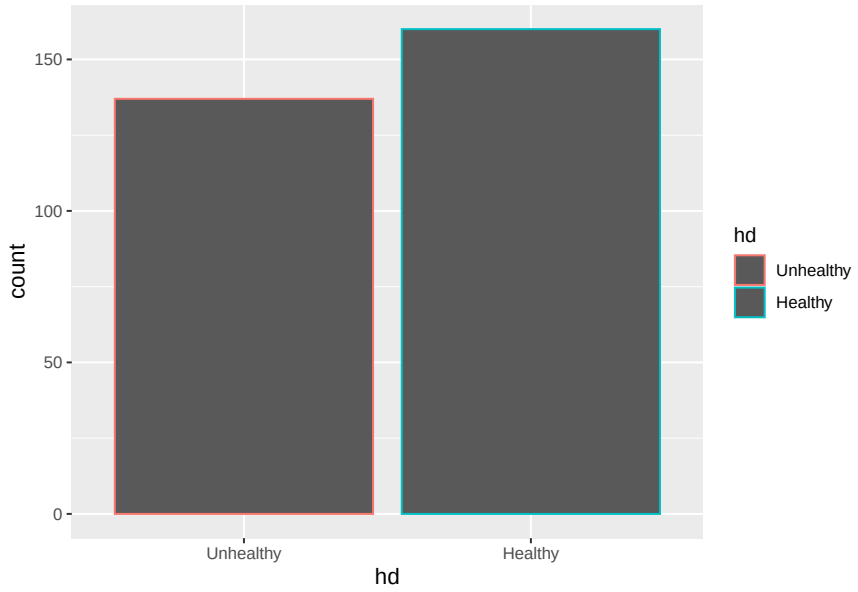


Figure 1: Heart Disease Distribution

Prevalence of heart disease is NA.

4.3 Heart Disease by Sex

Lets look at how Sex values are distributed among Healthy and Unhealthy categories

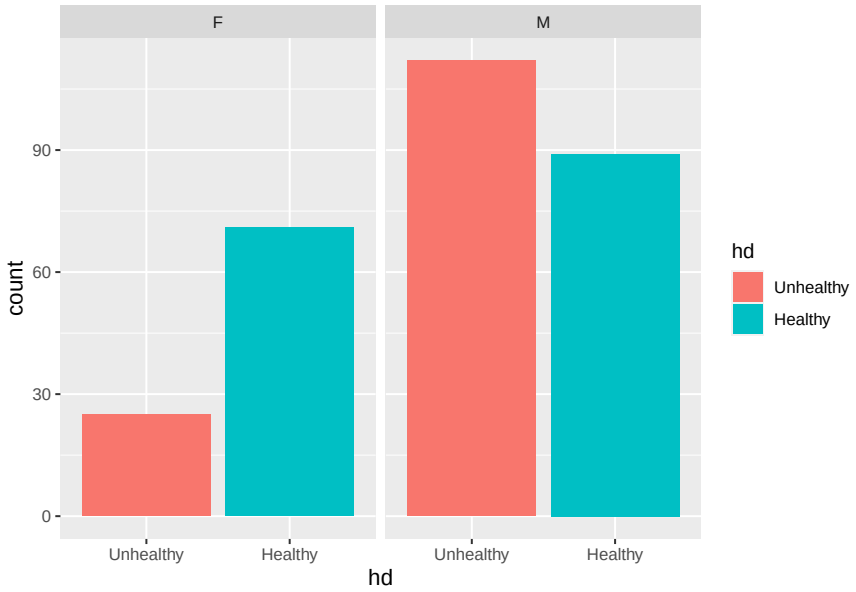


Figure 2: Heart Disease by Sex

Table 4: Heart Disease by Sex

	F	M
Unhealthy	25	112
Healthy	71	89

Lets look at the prevalence of Heart disease by gender.

Table 5: Heart Disease Prevelence By Gender

sex	prevelance
F	0.2604167
M	0.5572139

This dataset was taken from patients in Cleveland, USA .

As we can see from the above the prevalence of heart disease in men is 0.5572139 and women is 0.2604167.
The prevalence of heart disease given in the CDC website for USA is 0.223 in Women and 0.244 for Men, by these numbers Women prevalence seems to be close enough with the dataset, but for Men,

prevalence is way off with CDC prevalence, one reason for this could be that the CDC data was based on the general public and is not from the patients who have had heart symptoms and might not be from the heart hospitals, as the data from Cleveland dataset was taken from patients from cardiologists, meaning these patients had symptions to start with and various tests were performed, this high prevalence for Men in the data set might not be an issue, this could be verified by looking at the data from cdc website but it is out of scope for this project.

4.4 Heart Disease by Chest Pain

Lets look at cp, chest pain, and see how the values are distributed among Healthy and Unhealthy categories.

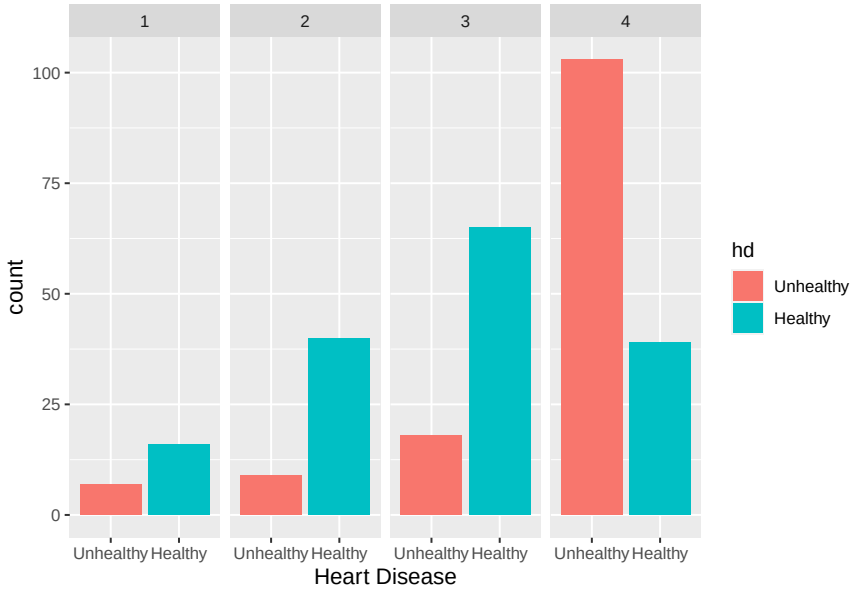


Figure 3: Heart Disease by Chest Pain

Table 6: Heart Disease by Chest Pain Categories

	1	2	3	4
Unhealthy	7	9	18	103
Healthy	16	40	65	39

As we can see from the above the proportion of Unhealthy patients varies across cp categories, and we could make use of these variation while predicting the Heart Disease.

4.5 Heart Disease by Fasting Blood Sugar

Lets look at fbs, fasting blood sugar, and see how the values are distributed among Healthy and Unhealthy categories.

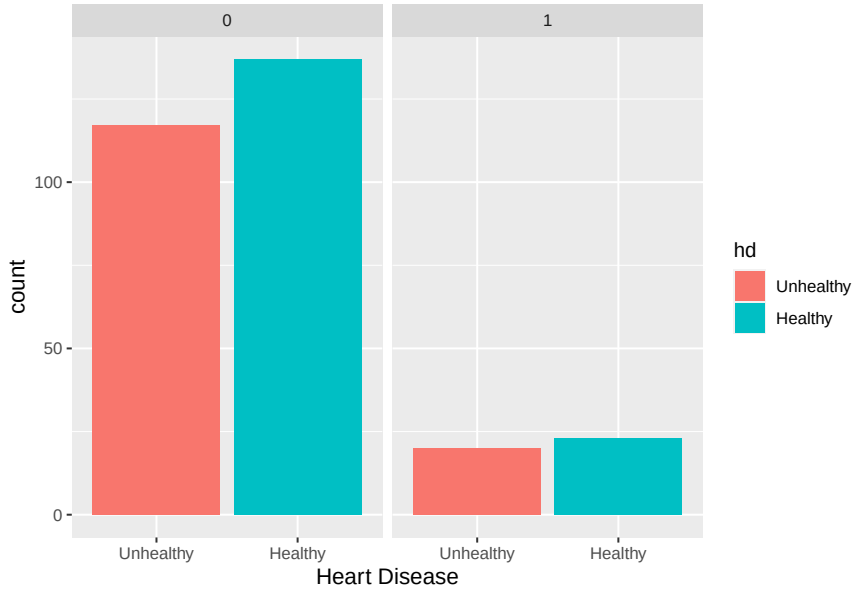


Figure 4: Heart Disease by Fasting Blood Sugar

Table 7: Heart Disease by Fasting Blood Sugar

	0	1
Unhealthy	117	20
Healthy	137	23

As we can see from the above the proportion of Unhealthy patients varies across fbs categories, and we could make use of these variation while predicting the Heart Disease.

4.6 Heart Disease by Resting Electro Cardiographic Results

Lets move on to the next column, restecg, resting electro cardiographic results, and see how the values are distributed among Healthy and Unhealthy categories.

Table 8: Heart Disease by Resting Electro Cardiographic Results

	0	1	2
Unhealthy	55	3	79
Healthy	92	1	67

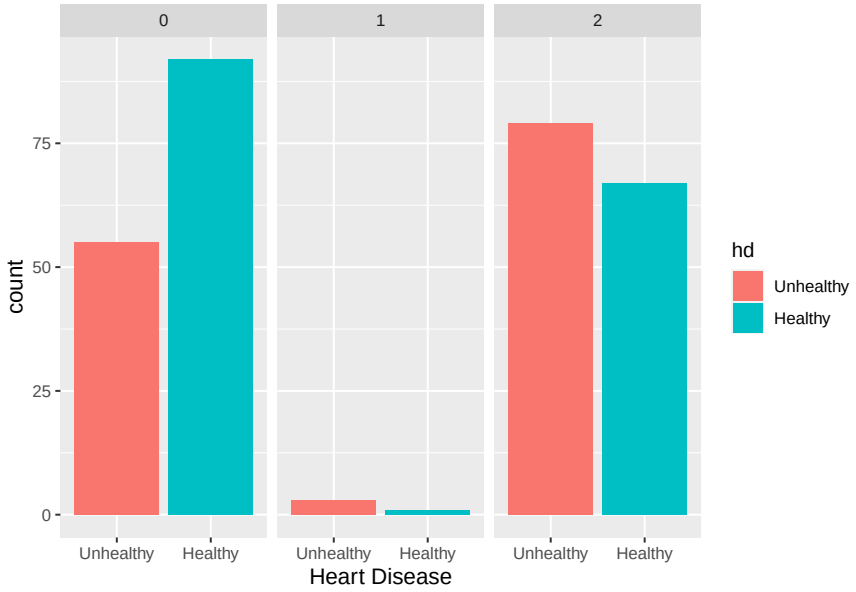


Figure 5: Heart Disease by Resting Electro Cardiographic Results

As we can see from the above the proportion of Unhealthy patients varies across restecg categories, and we could make use of these variation while predicting the Heart Disease. Note that restecg category 1 values have very few values (1 and 3) for Healthy and Unhealthy categories and these values could adversely impact the result due to the low values so we need to handle this in the methods section.

4.7 Heart Disease by Exercise Induced Angina

Lets look at exang, exercise induced angina, and see how the values are distributed among Healthy and Unhealthy categories.

Table 9: Heart Disease by Exercise Induced Angina

	0	1
Unhealthy	63	74
Healthy	137	23

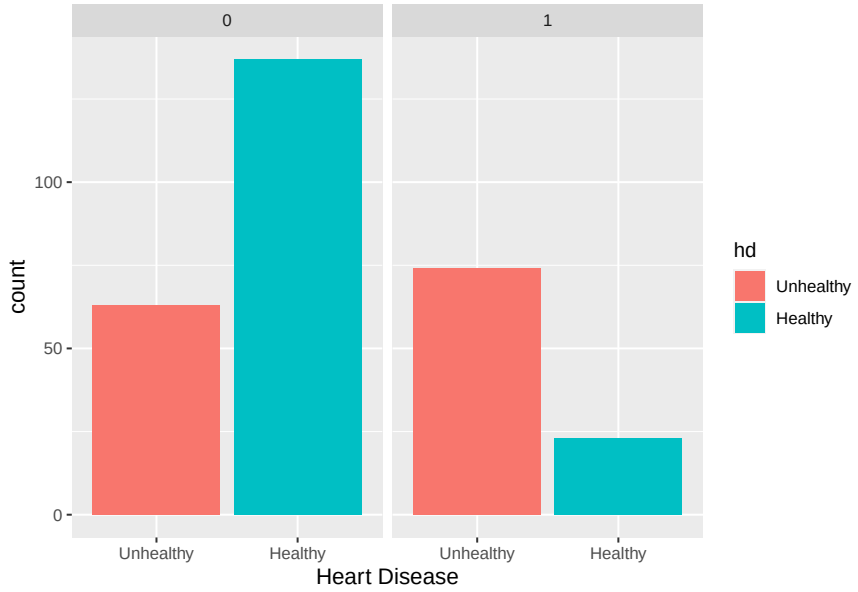


Figure 6: Heart Disease by Exercise Induced Angina

As we can see from the above the proportion of Unhealthy patients varies across exang categories, and we could make use of these variation while predicting the Heart Disease.

4.8 Heart Disease by Slope

Lets look at slope, the slope of the peak exercise ST segment, and see how the values are distributed among Healthy and Unhealthy categories.

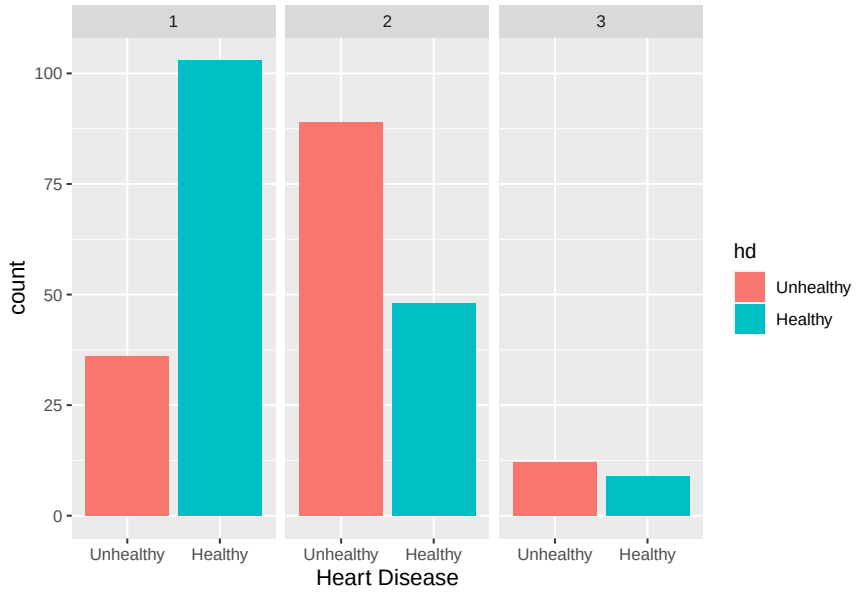


Figure 7: Heart Disease by Slope

Table 10: Heart Disease by Slope

	1	2	3
Unhealthy	36	89	12
Healthy	103	48	9

As we can see from the above the proportion of Unhealthy patients varies across slope categories, and we could make use of these variation while predicting the Heart Disease.

4.9 Heart Disease by Number of major vessels (0-3) colored by fluoroscopy

Lets look at ca, number of major vessels (0-3) colored by fluoroscopy, and see how the values are distributed among Healthy and Unhealthy categories.

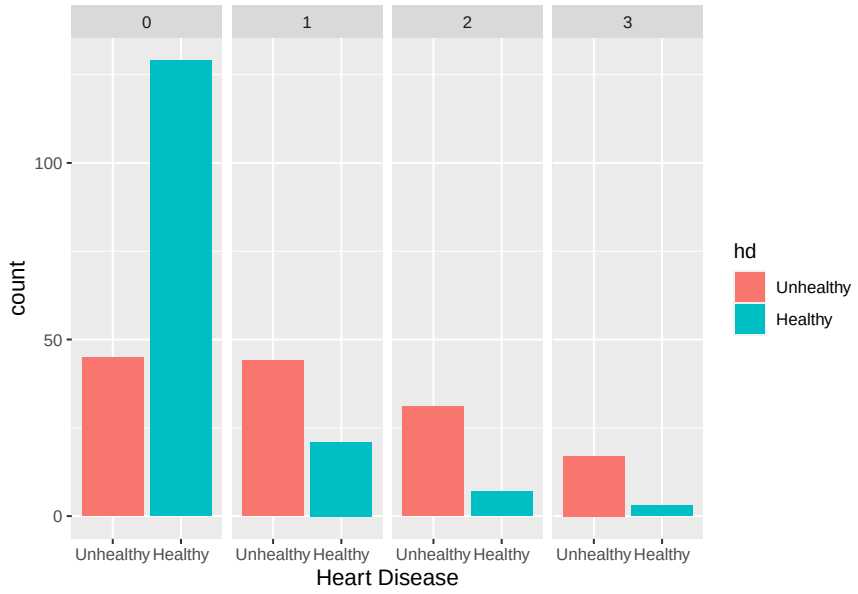


Figure 8: Heart Disease by Number of major vessels (0-3) colored by fluoroscopy

Table 11: Heart Disease by Number of major vessels (0-3) colored by fluoroscopy

	0	1	2	3
Unhealthy	45	44	31	17
Healthy	129	21	7	3

As we can see from the above the proportion of Unhealthy patients varies across *ca* categories, and we could make use of these variation while predicting the Heart Disease. Note that there are few values for ca category 2 and 3 for Healthy heart category.

4.10 Heart Disease by Thalium heart scan

Lets move on to the next column, thal, thalium heart scan, and see how the values are distributed among Healthy and Unhealthy categories.

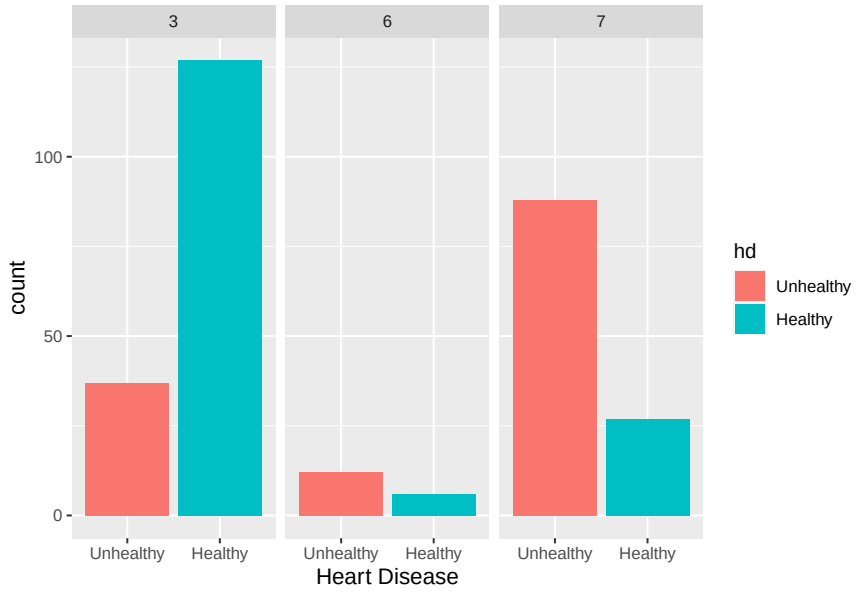


Figure 9: Heart Disease by Thalium heart scan

Table 12: Heart Disease by Thalium heart scan

	3	6	7
Unhealthy	37	12	88
Healthy	127	6	27

As we can see from the above the proportion of Unhealthy patients varies across *thal* categories, and we could make use of these variation while predicting the Heart Disease. Note that there are few values for thal category 6 under the Healthy heart category.

As we are done with variables of type factor, lets move on to the variables with numerical data and analyze the data.

4.11 Heart Disease vs Resting Blood Pressure

Lets look at how resting blood pressure effects Heart Disease

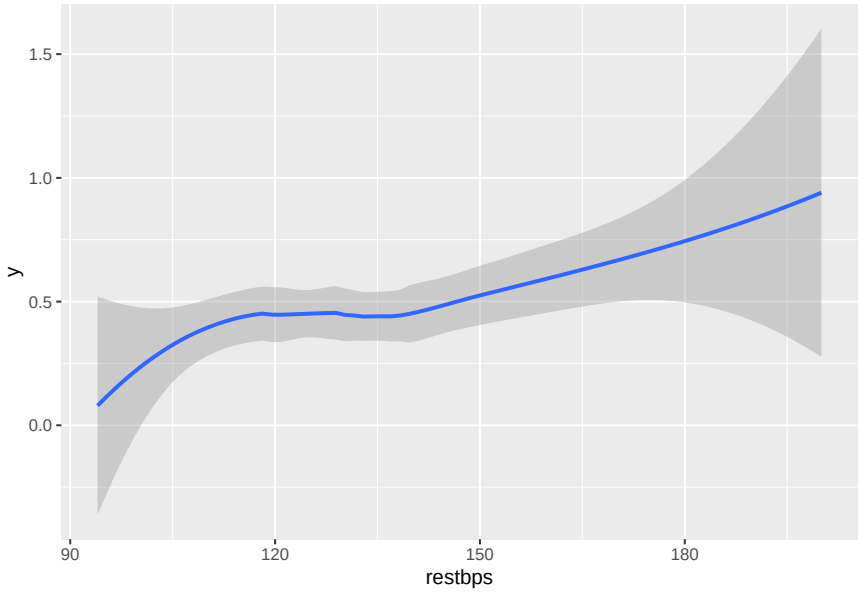


Figure 10: Heart Disease vs Blood Pressure

As we can see from the above, lower resting blood pressure is associated with *Healthy* and the higher the blood pressure readings are associated with the *Unhealthy*. We can clearly see a trend between resting blood pressure and the heart disease.We could make use of this variable in predicting the heart disease.

4.12 Heart Disease vs Serum Cholestoral

Lets look at how serum cholesterol effects Heart Disease

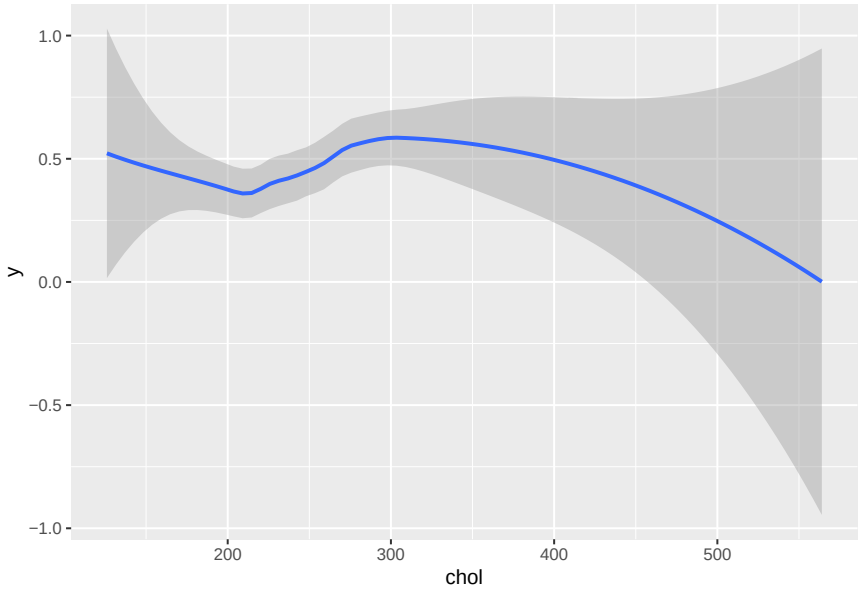


Figure 11: Heart Disease vs Serum Cholestoral

As we can see from the above figure, there is a moderate effect of serum cholesterol on the heart disease.We can make use of this variable in predicting the heart disease.

4.13 Heart Disease vs Maximum Heart Rate Achieved

Lets look at how maximum heart rate achieved effects Heart Disease

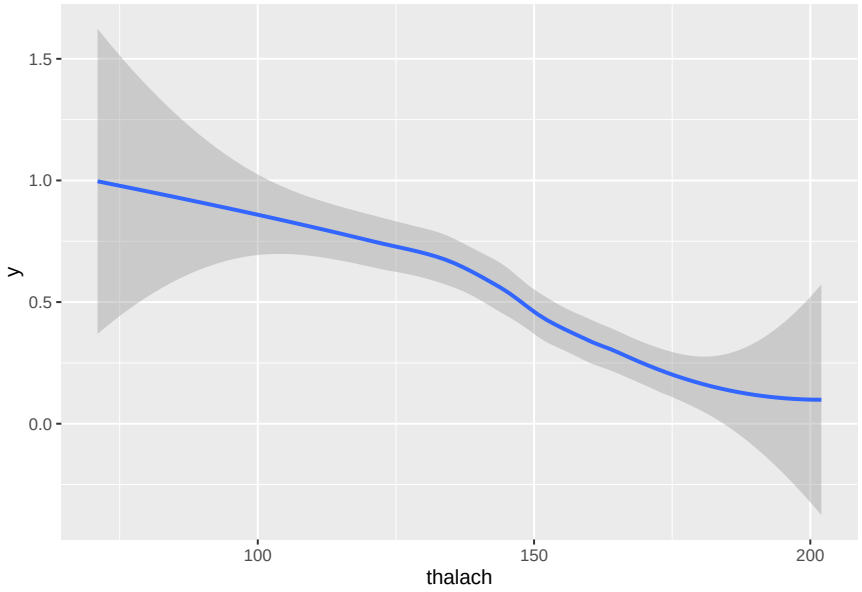


Figure 12: Heart Disease vs Maximum Heart Rate Achieved

As we can see from the above figure, lower values of maximum heart rate achieved (< 150) are associated with heart disease than the higher values.We can make use of this variable in predicting the heart disease.

4.14 Heart Disease vs ST depression induced by exercise relative to rest

Lets look at how ST depression induced by exercise relative to rest effects Heart Disease

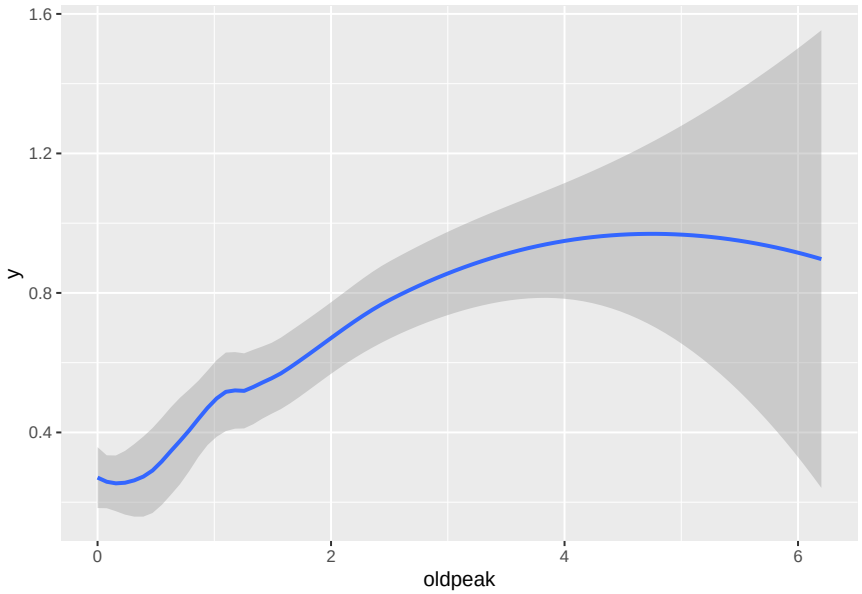


Figure 13: ST depression induced by exercise relative to rest

As we can see from the above figure, higher values of ST depression induced by exercise relative to rest are associated with heart disease than the lower values.We can make use of this variable in predicting the heart disease.

4.15 Heart Disease vs Age

Lets look at how age effects Heart Disease

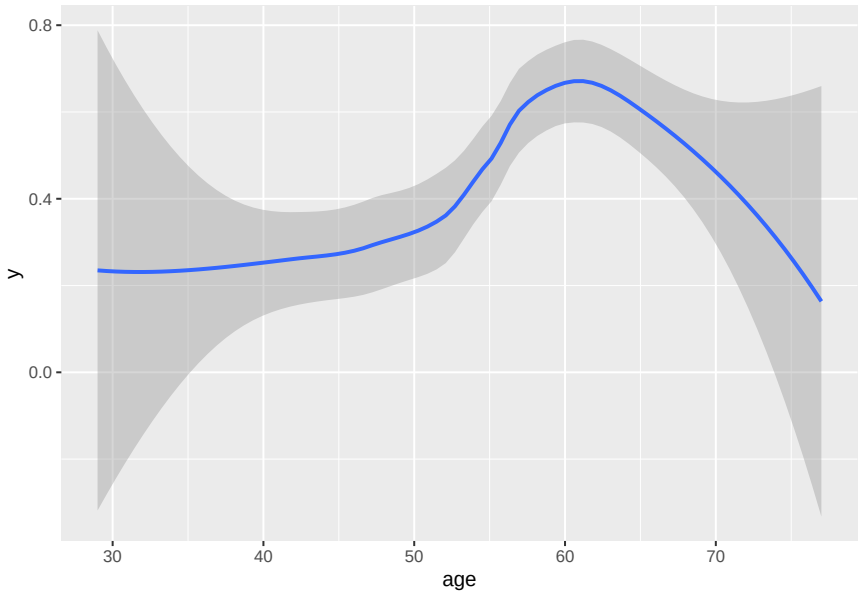


Figure 14: Heart Disease vs Age

As we can see from the above figure, there is a clear trend between age and heart disease, presence of heart disease is apparent in patients between 50 and 70 years old. We can make use of this variable in predicting the heart disease.

Summary of the analysis is that we have seen changes in proportion of Unhealthy values across categories for all of the factor variables, and for the numerical variables have shown the trend associated with the heart disease.

Variable importance and its contribution to the prediction power would become clear in the Methods section.

As we are now done with the analysis of the independent variables, let jump on to the methods section to build various algorithms and see which algorithm fits the best for predicting the heart disease presence.

5 Methods

In this section the primary focus would be on predicting the heart disease using various machine learning algorithms and choose the one which performs better.

The following is the high level outline of this section, each step would be detailed out in the subsequent sections.

1) Package Installation & Loading
2) Data Partition for Training & Testing
3) Algorithm evaluation criteria
4) Model 1 - Novice Heart Disease Model
5) Model 2 - Logistic Regression Heart Disease Model
6) Model 3 - KNN Heart Disease Model
7) Model 4 - Classification Tree Heart Disease Model
8) Model 5 - Random Forest Heart Disease Model
9) Model 6 - Ensemble Model

5.1 Data Partition for Training & Testing

Before we start building the algorithms we need to split the data for training and testing the algorithm.

Training data will only be used for training and optimizing the algorithm and the testing data will be exclusively used for testing the optimized algorithm.

Following are the steps that are performed for data partitioning

1) 20% of the data rows from the heart disease dataset (heart_disease_ds) are randomly selected and placed in test_set dataset, and this will be kept aside for performing the validation of each Heart Disease Prediction Model we build in the subsequent sections. This dataset will not be used for training and optimizing the model. This dataset is not used for training mainly to avoid over fitting the data.
2) The remaining 80% of the data rows from the heart disease dataset are brought into the train_set dataset.This dataset is mainly used for training and optimizing each of the algorithms we build.

After all the above steps the following datasets would be obtained

1) train_set - This dataset has 237 records and this will be primarily used for training and optimizing the machine learning algorithms.
2) test_set - This dataset has 60 and this will only be used to get the final predictions for each algorithm.

5.2 Algorithm Evaluation

As we are dealing with a classification problem we will be using sensitivity , specificity and accuracy to evaluate the performance of the algorithms.

The high level process we follow for Training & Optimization, and Validation is listed below,

- We use cross validation with 10 folds on train_set,to train and optimize the algorithm.
- We take the parameters that optimized the algorithm from the above step and use them to re-train the algorithm using the entire train_set. Retraining the algorithm with optimized parameters will make sure that the algorithm is exposed to the entire train_set.
- The retrained model in the above step is used to validate the data in test_test.
- Algorithm accuracy,sensitivity and specificity values are published in the respective sections but a detailed analysis will be performed in the Results section.

5.3 Model 1 - Novice Heart Disease Model

Before getting into various machine learning models, lets work on a rudimentary model which predicts the presence of heart disease using the prevalence of the heart disease in the dataset.

Here is the model we are using.

$$p(\mathbf{x}) = \Pr(Y = "Unhealthy" \mid \mathbf{X} = \mathbf{x}) = p("Unhealthy") = 0.4599156$$

In the above formula we are using the prevalence of heart disease in the train_set as the probability of predicting the patient as Unhealthy, we are not making use of any of the predictors for this model. We have taken the prevalence from train_set to avoid over-fitting.

Lets see how this has performed by looking at the confusion matrix.

```
##            Reference
## Prediction  Unhealthy Healthy
##   Unhealthy        17      16
##   Healthy          11      16
```

```
##          Sensitivity          Specificity       Pos Pred Value
##            0.6071429            0.5000000            0.5151515
##       Neg Pred Value            Precision               Recall
##            0.5925926            0.5151515            0.6071429
##                   F1           Prevalence       Detection Rate
##            0.5573770            0.4666667            0.2833333
## Detection Prevalence    Balanced Accuracy
##            0.5500000            0.5535714
```

As seen from the above results, overall accuracy of the algorithm is 0.55, where as the sensitivity is 0.6071429, that is 60.7142857 percent of patients who are Unhealthy will be predicted as Unhealthy, and its specificity is 0.5, that is 50 percent of patients who are Healthy will be predicted as Healthy.

The above metrics are same as flipping a coin, and we can take this as the baseline and try to beat this algorithm.

5.4 Model 2 - Logistic Regression Heart Disease Model

The next simple model to the Novice Model is linear model, and moreover linear models are easy to interpret and often taken as a baseline model before getting into the complex models.

Lets build a linear model to predict the heart disease.

Logistic Regression is the linear model which is used for Classification problems and lets build this model for Heart Disease Prediction.

Our goal in building a machine learning model is to estimate the below conditional probability for any given value of x. That is the probability of a patient being Unhealthy given X = x, in our case X is a multi-dimensional vector with all the independent variables in the heart disease dataset, and X = x is one row of this vector.

$$\Pr(Y = "Unhealthy" \mid X = x)$$

Logistic regression fits the the below linear model to find out the expected value of the above probability assuming a linear relationship between the Y and the independent variables.

$$g\left\{\Pr(Y = "Unhealthy" \mid X = x)\right\} = \beta_0 + \beta_1 x$$

Linear coefficients are fitted with the train_set data.

Here the logistic transformation g(p) is given below

$$g(p) = \log \frac{p}{1-p}$$

5.4.1 Train Model

Lets fit the logistic regression model using cross validation with 10 folds and look at the fitted model.

Maximum likelihood is used to fit the linear model. The coefficients of the fitted model are given below. Note that the coefficients are in log(odds of Unhealthy heart).

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6896  -0.3808   0.1454   0.5030   3.0982
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.270703   3.188564   2.280  0.02259 *
## age          0.018455   0.028211   0.654  0.51300
## sexM        -1.883959   0.613389  -3.071  0.00213 **
```

```
## cp2         -1.454022   0.917422  -1.585  0.11299
## cp3         -0.625054   0.768400  -0.813  0.41596
## cp4         -2.467096   0.781699  -3.156  0.00160 **
## restbps     -0.028672   0.012902  -2.222  0.02626 *
## chol        -0.008014   0.004488  -1.786  0.07411 .
## fbs1         0.533154   0.643610   0.828  0.40746
## restecg1    -0.568488   2.885160  -0.197  0.84380
## restecg2    -0.201334   0.424276  -0.475  0.63512
## thalach      0.019768   0.012402   1.594  0.11097
## exang1      -0.617129   0.498480  -1.238  0.21571
## oldpeak     -0.233478   0.266669  -0.876  0.38128
## slope2      -1.295017   0.536218  -2.415  0.01573 *
## slope3      -1.329498   1.095662  -1.213  0.22497
## ca1         -2.365343   0.565284  -4.184 2.86e-05 ***
## ca2         -3.145813   0.788242  -3.991 6.58e-05 ***
## ca3         -2.560277   1.133837  -2.258  0.02394 *
## thal6        0.452021   0.839462   0.538  0.59026
## thal7       -0.865555   0.489074  -1.770  0.07676 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 327.03  on 236  degrees of freedom
## Residual deviance: 157.86  on 216  degrees of freedom
## AIC: 199.86
##
## Number of Fisher Scoring iterations: 6
```

Here is the final fitted model using the logistic regression

$$g\left\{\Pr(Y = \text{"}Unhealthy\text{"} \mid X = x)\right\} = 7.270703 + 0.018455age - 1.883959sexM - 1.454022cp2 - 0.625054cp3$$
$$- 2.467096cp4 - 0.028672restbps - 0.008014chol + 0.533154fbs1 - 0.568488restecg1 - 0.201334restecg2$$
$$+ 0.019768thalach - 0.617129exang1 - 0.233478oldpeak - 1.295017slope2 - 1.329498slope3 - 2.365343ca1$$
$$- 3.145813ca2 - 2.560277ca3 + 0.452021thal6 + -0.865555thal7$$

As the results of the fit is a random variable we used cross validation to get the expected values of the accuracy. In logistic regression there are no parameters to the model and hence we used cross validation to get the expected value of the accuracy. Here are the accuracy values we got in various folds of the cross validation.

Table 13: Accuracy of the fit in various folds during cross validation

Accuracy	Kappa	Resample
0.7500000	0.4965035	Fold01
0.8750000	0.7500000	Fold02
0.7916667	0.5714286	Fold03
0.9130435	0.8257576	Fold04
0.8695652	0.7376426	Fold05
0.8333333	0.6643357	Fold06
0.7916667	0.5774648	Fold07
0.6521739	0.2977099	Fold08
0.8750000	0.7500000	Fold09
0.9583333	0.9154930	Fold10

Estimated accuracy of the model is 0.8309783. This is same as the mean accuracy of all the folds in the cross validation table as shown above.

As we have seen the estimated performance of the model, we can refit the model with the entire train_set so that the model will get more data to train on, as we have only used partial data to fit the model during the cross validation.

Lets refit the model with the train_set and look at McFadden's Pseudo R Squared, see below for the formula , this would give us an estimate of the R Squared, by which we can check what percent of the variance in the output is explained by the Logistic Regression Model.

$$McFadden'sPseudoR^2 = [LL(Null) - LL(Proposed)]/LL(Null)$$

Here LL(Null) is the log likelihood of the Null model and LL(Proposed) is the log likelihood of the logistic regression model.

$McFadden'sPseudoR^2$ value by using the above formula is 0.5172917 and p_value is 0. The fitted model explains the 51.7291665 percent of the variance in the output and the p_value 0 tell us that the R Squared value is statistically significant.

5.4.2 Validate Model

Lets validate our model by predicting the presence of heart disease in the test_test and look at the results

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction  Unhealthy Healthy
##   Unhealthy        25       4
##   Healthy           3      28
##
##                Accuracy : 0.8833
```

```
##                  95% CI : (0.7743, 0.9518)
##     No Information Rate : 0.5333
##     P-Value [Acc > NIR] : 7.387e-09
##
##                   Kappa : 0.7661
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.8929
##             Specificity : 0.8750
##          Pos Pred Value : 0.8621
##          Neg Pred Value : 0.9032
##              Prevalence : 0.4667
##          Detection Rate : 0.4167
##    Detection Prevalence : 0.4833
##       Balanced Accuracy : 0.8839
##
##        'Positive' Class : Unhealthy
##
```

As seen from the above results, overall accuracy of the algorithm is 0.8833333, where as the sensitivity is 0.8928571, that is 89.2857143 percent of patients who are Unhealthy will be predicted as Unhealthy, and its specificity is 0.875, that is 87.5 percent of patients who are Healthy will be predicted as Healthy.

Here is the compasion of Logistic Regression model with the Novice model

Table 14: Model Comparison Table

method	accuracy	Sensitivity	Specificity	F1
Model 1 - Guessing	0.5500000	0.6071429	0.500	0.557377
Model 2 - Logistic Regression	0.8833333	0.8928571	0.875	0.877193

There metrics are far better than the Novice model.

Lets build few more models and see whether the accuracy improves.

5.5 Model 3 - KNN Heart Disease Model

Lets build KNN (k nearest neighbors) model and see whether the accuracy improves. In knn we estimate the below conditional probabilities

$$p(x_1, x_2) = \Pr(Y = 1 \mid X_1 = x_1, X_2 = x_2).$$

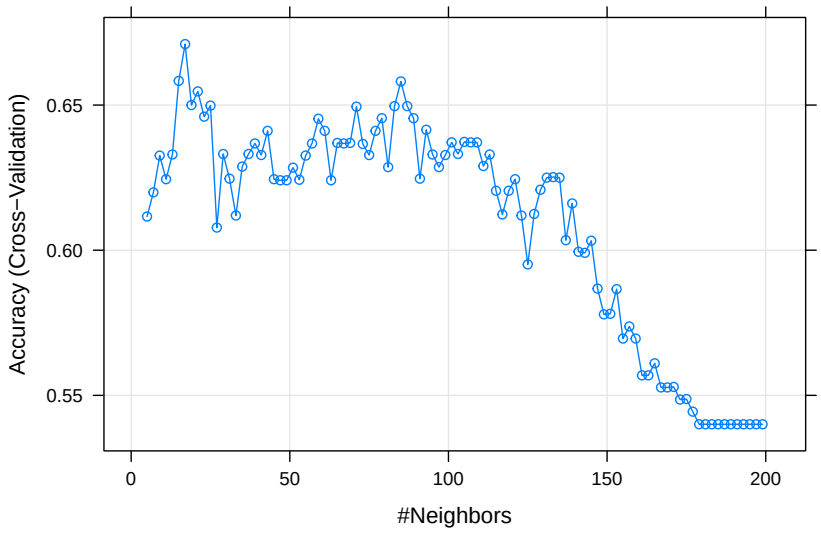
First we define the distance between all observations based on the features. Then, for any point (x_1, x_2) for which we want an estimate of $p(x_1, x_2)$, we look for the k nearest points to (x_1, x_2) and then take an average of the Unhealthy and Healthy outputs associated with these points.

To implement the algorithm, we can use the `knn3` function from the **caret** package.

5.5.1 Train & Optimize

Lets train the knn model on the train_set using 10 fold cross validation and k values 5, 8, 11, 14, 17, 20, 23, 26, 29, 32, 35, 38, 41, 44, 47, 50, 53, 56, 59, 62, 65, 68, 71, 74, 77, 80, 83, 86, 89, 92, 95, 98, 101, 104, 107, 110, 113, 116, 119, 122, 125, 128, 131, 134, 137, 140, 143, 146, 149, 152, 155, 158, 161, 164, 167, 170, 173, 176, 179, 182, 185, 188, 191, 194, 197, 200 for tuning.

Here is the accuracy of the knn model vs k-values graph from cross validation



From the above figure we can see that the k-value that performed better in the cross validation is 17 and its accuracy is 0.6710145.

5.5.2 Validate Model

Lets refit the knn model on the train_set with the optimal k-value obtained in the above step, k = 17, as it will give the model an opportunity to train on the entire train_set, and use the model to validate the model on the validation dataset (test_set).

Here are the confusion matrix results after validation.

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction  Unhealthy Healthy
##   Unhealthy        16       7
##   Healthy          12      25
##
##                Accuracy : 0.6833
##                  95% CI : (0.5504, 0.7974)
##     No Information Rate : 0.5333
```

```
##     P-Value [Acc > NIR] : 0.01309
##
##                   Kappa : 0.3567
##
##  Mcnemar's Test P-Value : 0.35880
##
##             Sensitivity : 0.5714
##             Specificity : 0.7812
##          Pos Pred Value : 0.6957
##          Neg Pred Value : 0.6757
##              Prevalence : 0.4667
##          Detection Rate : 0.2667
##    Detection Prevalence : 0.3833
##       Balanced Accuracy : 0.6763
##
##        'Positive' Class : Unhealthy
##
```

As seen from the above results, overall accuracy of the algorithm is 0.6833333, where as the sensitivity is 0.5714286, that is 57.1428571 percent of patients who are Unhealthy will be predicted as Unhealthy, and its specificity is 0.78125, that is 78.125 percent of patients who are Healthy will be predicted as Healthy.

Here is the KNN Model comparison with other models

Table 15: Model Comparison Table

method	accuracy	Sensitivity	Specificity	F1
Model 1 - Guessing	0.5500000	0.6071429	0.50000	0.557377
Model 2 - Logistic Regression	0.8833333	0.8928571	0.87500	0.877193
Model 3 - KNN	0.6833333	0.5714286	0.78125	0.627451

Here is the summary of the KNN Validation Results

- This accuracy is slightly better than the Novice model
- Sensitivity of Novice models better than KNN
- Logistic regression model is way better than KNN in accuracy, sensitivity and specificity.

Poor performance of the KNN model in our case is due to large of number of dimensions.

Lets look at couple more models and see whether we could beat the logistic regression.

Note that we can not perform LDA and QDA on this dataset as the some of the variable are factors and LDA and QDA works on numerical data and when multinational distribution is assumed.

5.6 Model 4 - Classification Tree Heart Disease Model

Classification trees, or decision trees, are used in prediction problems where the outcome is categorical so we will use Classification trees to predict the heart disease.

In Classification Tree Model we estimate the below conditional probabilities by which class is the most common among the training set observations within the partition

$$p(x_1, x_2) = \Pr(Y = \text{"}Unhealthy\text{"} \mid X).$$

We are going to use CART package in R to train the algorithm, and CART uses *Gini Index* to choose the partition, here is the definition of the Gini Index.

$$\text{Gini}(j) = \sum_{k=1}^{K} \hat{p}_{j,k}(1 - \hat{p}_{j,k})$$

In a perfect scenario, the outcomes in each of our partitions are all of the same category since this will permit perfect accuracy. The *Gini Index* is going to be 0 in this scenario, and become larger the more we deviate from this scenario. The partitions that produce minimum *Gini Index* would be chosen.

5.6.1 Train & Optimize Model

Lets train and optimize Heart Disease Classification Model using cross validation with 10 folds and use *complexity parameter* (cp) as tuning parameter with values 0, 0.0041667, 0.0083333, 0.0125, 0.0166667, 0.0208333, 0.025, 0.0291667, 0.0333333, 0.0375, 0.0416667, 0.0458333, 0.05, 0.0541667, 0.0583333, 0.0625, 0.0666667, 0.0708333, 0.075, 0.0791667, 0.0833333, 0.0875, 0.0916667, 0.0958333, 0.1. Lets keep the mtry at default value which is the square root of number of parameters, and minsplit at 0 so that the algorithm gets the flexibility during training.

Here is the performance of the Heart Disease Classification Tree for various values of Complexity Parameter

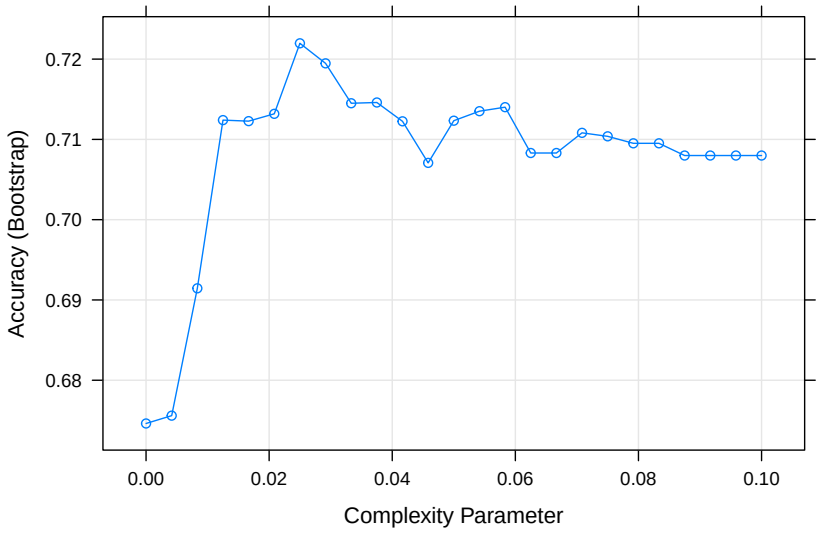


Figure 15: Results of Cross Valtion

As we can see from the above graph the best value of Complexity Parameter (cp) is 0.025 which has got an accuracy of 0.7219706 during training.

5.6.2 Validate Model

As we got the parameters that optimized the model lets retrain the algorithm using the optimal parameters obtained during the above cross training on the entire train_set. Here is the Heart Disease classification tree after the retraining on the entire train_set,

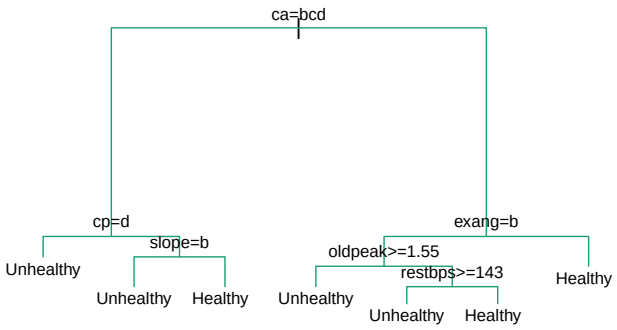


Figure 16: Heart Disease Classfication Tree

Here is the variable importance based on the fit

```
##         ca      exang         cp    oldpeak    thalach        age      slope
## 28.1289710 13.5945707 12.9004495 11.3479312 10.4788429 10.3560891  6.1920869
##    restbps       chol       thal        sex        fbs
##  4.2058000  2.7516068  1.1445503  0.6627643  0.5925926
```

Lets use the trained model and predict the Heart Disease outcome.

Here is the Confusion Matrix based on the predicted outcomes.

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction  Unhealthy Healthy
##   Unhealthy        22       3
##   Healthy           6      29
##
##                Accuracy : 0.85
```

```
##                  95% CI : (0.7343, 0.929)
##     No Information Rate : 0.5333
##     P-Value [Acc > NIR] : 2.293e-07
##
##                   Kappa : 0.6966
##
##  Mcnemar's Test P-Value : 0.505
##
##             Sensitivity : 0.7857
##             Specificity : 0.9062
##          Pos Pred Value : 0.8800
##          Neg Pred Value : 0.8286
##              Prevalence : 0.4667
##          Detection Rate : 0.3667
##    Detection Prevalence : 0.4167
##       Balanced Accuracy : 0.8460
##
##        'Positive' Class : Unhealthy
##
```

As seen from the above results, overall accuracy of the algorithm is 0.85, where as the sensitivity is 0.7857143, that is 78.5714286 percent of patients who are Unhealthy will be predicted as Unhealthy, and its specificity is 0.90625, that is 90.625 percent of patients who are Healthy will be predicted as Healthy.

Here is the model Comparison table

Table 16: Model Comparison Table

method	accuracy	Sensitivity	Specificity	F1
Model 1 - Guessing	0.5500000	0.6071429	0.50000	0.5573770
Model 2 - Logistic Regression	0.8833333	0.8928571	0.87500	0.8771930
Model 3 - KNN	0.6833333	0.5714286	0.78125	0.6274510
Model 4 - Classification Trees	0.8500000	0.7857143	0.90625	0.8301887

As we see from the above Classification Tree Model has the height specificity across all built so far , but the sensitivity is still less than the Logistic Regression Model.

As we know that Random Forest improves the performance of the Classification trees by building large number of random trees. Lets build Random Forest Model in the next section.

5.7 Model 5 - Random Forest Heart Disease Model

Random forests are a **very popular** machine learning approach that addresses the shortcomings of decision trees using a clever idea. The goal is to improve prediction performance and reduce instability by *averaging* multiple decision trees (a forest of trees constructed with randomness). Lets use Random Forest to model Heart Disease predictions.

5.7.1 Train & Optimize Model

Lets train and optimize Random Forest Heart Disease Model using cross validation with 10 folds and use the following turning parameters,

```
1) mtry:  1, 2, 3, 4, 5, 6, 7, 8, 9, 10

2) nodesize: 1, 11, 21, 31, 41, 51
```

Here are the results of the cross validation

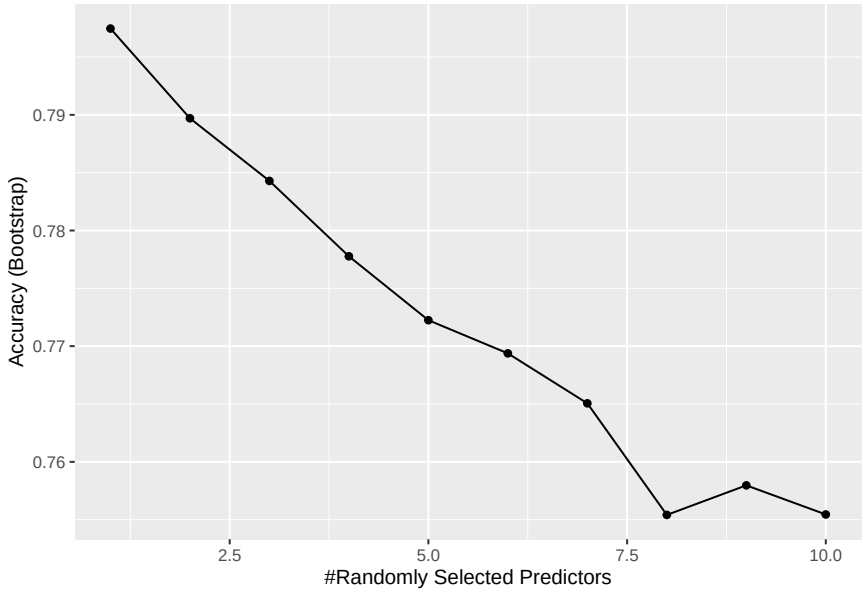
Table 17: Accuracy of the fit for various values of mtr and nodesize

nodesize	mtry	accuracy
1	1	0.7901373
11	1	0.7987261
21	1	0.8039397
31	2	0.7843756
41	1	0.7995353
51	1	0.7959951

Above table shows the best nodesize and mtry in each fold of the cross validation along with thier accuracy.

We can see that optimal values for which the max accuracy has achieved, that is 0.8039397 , are mtry = 1 and nodesize = 21.

Here is a plot of the accuracy when we set the optimal nodesize, 21, and fit the model with various values of mtry, and it clearly shows that the maximum accuracy achived when mtry = 1.



5.7.2 Validate Model

We need to retrain the algorithm on the train_set using the parameters that optimized the model in the previous section so that the model is exposed to the entire train_set.

Here is the Confusion Matrix from the predictions of the test_set by using the Random Forest Heart Disease Model

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction  Unhealthy Healthy
##   Unhealthy        24       3
##   Healthy           4      29
##
##                Accuracy : 0.8833
##                  95% CI : (0.7743, 0.9518)
##     No Information Rate : 0.5333
##     P-Value [Acc > NIR] : 7.387e-09
##
##                   Kappa : 0.7651
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.8571
##             Specificity : 0.9062
##          Pos Pred Value : 0.8889
##          Neg Pred Value : 0.8788
##              Prevalence : 0.4667
##          Detection Rate : 0.4000
##    Detection Prevalence : 0.4500
##       Balanced Accuracy : 0.8817
##
##        'Positive' Class : Unhealthy
##
```

As seen from the above results, overall accuracy of the algorithm is 0.8833333, where as the sensitivity is 0.8571429, that is 85.7142857 percent of patients who are Unhealthy will be predicted as Unhealthy, and its specificity is 0.90625, that is 90.625 percent of patients who are Healthy will be predicted as Healthy.

Here is the Random Forest Model comparison with other models.

Table 18: Model Comparison Table

method	accuracy	Sensitivity	Specificity	F1
Model 1 - Guessing	0.5500000	0.6071429	0.50000	0.5573770
Model 2 - Logistic Regression	0.8833333	0.8928571	0.87500	0.8771930
Model 3 - KNN	0.6833333	0.5714286	0.78125	0.6274510
Model 4 - Classification Trees	0.8500000	0.7857143	0.90625	0.8301887
Model 5 - Random Forest	0.8833333	0.8571429	0.90625	0.8727273

As we see from the above Random Forest Model has performed better than Classification Trees , but the sensitivity is still less than the Logistic Regression Model.

5.8 Model 6 - Ensemble Model

Lets combine the models we have already built in the previous sections to build an Ensemble Model which could improve the overall performance of the predictions.

The approach that we follow to select the models for Ensemble Model is as follows

- Calculate the Training Performance of all the models that we built so far except the Novice Model
- Calculate the combined average of the all the model
- Select the models that performed greater than or equal to the combined average

See below for their training accuracy of individual models and their combined average

Logistic Regression Heart Disease Model Training Accuracy : 0.8607595

KNN Heart Disease Model Training Accuracy : 0.7046414

Heart Disease Classification Tree Model Training Accuracy : 0.8565401

Random Forest Heart Disease Model Training Accuracy : 0.8607595

Combined Accuracy of all the above Models : 0.8607595

We see that Logistic Regression and Random Forest Models only have their Training Accuracy greater than or equal to the combined accuracy, so we can consider combining these models to improve the accuracy of the predictions.

Here is the approach we take for predicting the outcome

- For every row in test_set, set the Random Forest vote to 1 if Random Forest predicts the outcome as “Unhealthy” otherwise 0.

- For every row in test_set, set the Logistic Regression vote to 1 if Logistic Regression predicts the outcome as "Unhealthy" otherwise 0.
- For every row in the test_set, calculate the average vote, by taking the average of Random Forest vote and the Logistic Regression vote
- If the average vote > 0.5 then predict the outcome as "Unhealthy" otherwise "Healthy"

Here are the results of the prediction after combining Logistic Regression Heart Disease Model and Random Forest Heart Disease Models

```
## [1] 0.9166667
```

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction  Unhealthy Healthy
##   Unhealthy        24       1
##   Healthy           4      31
##
##                Accuracy : 0.9167
##                  95% CI : (0.8161, 0.9724)
##     No Information Rate : 0.5333
##     P-Value [Acc > NIR] : 1.297e-10
##
##                   Kappa : 0.8315
##
##  Mcnemar's Test P-Value : 0.3711
##
##             Sensitivity : 0.8571
##             Specificity : 0.9688
##          Pos Pred Value : 0.9600
##          Neg Pred Value : 0.8857
##              Prevalence : 0.4667
##          Detection Rate : 0.4000
##    Detection Prevalence : 0.4167
##       Balanced Accuracy : 0.9129
##
##        'Positive' Class : Unhealthy
##
```

As seen from the above results, overall accuracy of the algorithm is 0.9166667, where as the sensitivity is 0.8571429, that is 85.7142857 percent of patients who are Unhealthy will be predicted as Unhealthy, and its specificity is 0.96875, that is 96.875 percent of patients who are Healthy will be predicted as Healthy.

Here is the Ensemble Model comparison with other models.

Table 19: Model Comparison Table

method	accuracy	Sensitivity	Specificity	F1
Model 1 - Guessing	0.5500000	0.6071429	0.50000	0.5573770
Model 2 - Logistic Regression	0.8833333	0.8928571	0.87500	0.8771930
Model 3 - KNN	0.6833333	0.5714286	0.78125	0.6274510
Model 4 - Classification Trees	0.8500000	0.7857143	0.90625	0.8301887
Model 5 - Random Forest	0.8833333	0.8571429	0.90625	0.8727273
Model 6 - Ensemble Model	0.9166667	0.8571429	0.96875	0.9056604

As we see from the above Ensemble Model has the height accuracy and specificity across all the models but the sensitivity is still less than the Logistic Regression Model

6 Results

Now that we have completed building the Models for Hear Disease Prediction System, lets look at all Model performances and analyze the results.

Here is the table which summarizes the performance of all the models

Table 20: Model Comparison Table

method	accuracy	Sensitivity	Specificity	F1
Model 1 - Guessing	0.5500000	0.6071429	0.50000	0.5573770
Model 2 - Logistic Regression	0.8833333	0.8928571	0.87500	0.8771930
Model 3 - KNN	0.6833333	0.5714286	0.78125	0.6274510
Model 4 - Classification Trees	0.8500000	0.7857143	0.90625	0.8301887
Model 5 - Random Forest	0.8833333	0.8571429	0.90625	0.8727273
Model 6 - Ensemble Model	0.9166667	0.8571429	0.96875	0.9056604

We started with Model 1 - Guessing as a novice model which has not considered any of the independent variables, and this model is only use to give us a base line estimate. This models accuracy and specificity are same as flipping a coin, sensitivity is little high, that is 0.6071429 is due to prevalence of the Heart Disease.

Model 2 - Logistic Regression has second height accuracy after Model 6 - Ensemble Model, and height sensitivity among all the models, that is 0.8928571. But this model's specificity, which is 0.875, is 0.09375 less than the model with the highest specificity. As it is a linear model it is simple and easy to interpret. This model could be chosen to predict the Heart Disease if sensitivity is more important than specificity, that is predicting the patient with Heart Disease as Heart Disease is more important than predicting the patient who is Healthy as Healthy by taking a little compromise in specificity. Slight compromise in sensitivity would slightly increase the chance of predicting Unhealthy patients as Healthy, and this incorrect diagnosis could put a patient in danger, where as a slight compromise in specificity would slightly increase the chance of predicting the patient who is Healthy as Unhealthy and in which case the patient's incorrect diagnosis adverse impact is less.

Based on the above argument we could recommend Logistic Regression when sensitivity is more important than the specificity while predicting the Heart Disease.

Model 3 - KNN Models accuracy and specificity are higher than the Novice Model but sensitivity is less than the Novice Model. Sensitivity of this model is 0.5714286, this sensitivity is not even close to flipping a coin and hence this model would not be recommended for predicting the Heart Disease. Poor performance of this model is due the the number of dimensions, the total number of dimensions after hot encoding, that is by adding the dummy variables for each of the categories, is 20, these are high number of dimensions, and due to the curse of dimensional, that is if you want to inclue 10% of the data in a neighborhood in a 20 dimensional space, then each dimension space would have to be 89.12509, that is around 90% of each dimension is taken away for just 10% of the data, and hence the neighborhood is no longer local and hence the poor performance.

Model 4 - Classification Trees Model accuracy, 0.85 , is better than Novice and KNN models, but still less than the logistic regression model, specificity is 0.90625 , and is better than Logistic Regression Model but not better than the Ensemble Model. Its sensitivity, 0.7857143 is 0.1071429 less than the Logistic Regression Model. The main advantage of this model is its interpretation, very easy to interpret and even so that the logistic regression model. As there is a significant difference in the sensitivity with the best model and moreover this model's variance would be high as it would most frequently get over-fitted with the training data, we would not recommend this for the prediction. However we could use this models variable importance to understand how individual variables contribution to the output.

Model 5 - Random Forest Model accuracy, 0.8833333 , is better than Novice and KNN models, and same as the logistic regression model, specificity is 0.90625 is better than Logistic Regression Model but not better than the Ensemble Model. Its specificity is 0.0714286 more than the Logistic Regression Model which put this model at advantage if specificity is more than sensitivity. Due to the randomness introduced during the training from random selection of variables during model fitting and the random samples from bootstrapping reduces the variance in the predictions.Its main drawback of this model is that we lose the model interpretation.

Model 6 - Ensemble Model accuracy, 0.9166667 , is higer than Novice, KNN , logistic regression model and Classification Trees models. Its specificity, 0.90625 is higher than all other models, and in specific it is 0.0714286 more than the Logistic Regression Model which put this model at advantage if specificity is more than sensitivity. The main downside of the model is loss of interpretation due to the average of all the models.

In summary we could choose Model 2 - Logistic Regression if sensitivity is more important than the specificity. If specificity and overall accuracy is more important than sensitivity we could choose the Model 6 - Ensemble Model as it has got higher values of both accuracy and specificity.

7 Conclusion

We have started the project with data download and clean up, and later used data visualization to present the analysis of all the independent variables and gained insights into their effects on the output. We built various machine learning models to predict whether the patient has Heart Disease in Methods section and compared all the models in the Results section and in the same section we discussed individual models pro's and con's, and recommended $Model2 - LogisticRegression$ if sensitivity is more important than the specificity, and if specificity and overall accuracy is more

important than sensitivity we can recommended $Model6 - EnsembleModel$. The goal of this project has been achieved.

The following are the limitations of the model,

1) Size of the input data set is only 303 rows and hence the algorithm's might not had enough exposure to the real world data
2) As noted in the analysis section under Heart Disease Distribution sub section, prevalence of Heart Disease in Men in the dataset is significantly more than the CDC heart disease prevalence and this could impact the predictions.

Following are the future considerations for the Heart Disease Prediction System

- There are 3 more dataset are present in the UCI machine learning repository for 3 other counties and those datasets needs to be analyzed and combined incorporated in the algorithm to improve the prediction power.
- As per CDC Race or Ethnicity seems to play a role in Heart Disease but this variable is not present in the processed dataset we have taken so if this information is available consider including it in the algorithms. This is one of many factors which could explain the presence of Heart Disease so need more data exploration to find out other factors which are missing in the dataset and include them if data is available.
- There are around 2 percent of the data has missing values, that is 6 rows, and they were removed as the percent of missing values are low, but to improve the performance we could employ missing value imputation techniques such as random forest to impute the values.
- The best algorithm that gave us the better sensitivity is Logistic Regression, and R Sqaured for this model is 0.5172917, which is only around half of the variance in the output and the remaining variance is still unexplained so we could do a better job on the predictions by trying other algorithms.
- Deep learning algorithms and other machine learning algorithms needs to be considered in futher improving the sensitivity and specificity of the predicted outcome.