Data Description

The Cardiovascular Disease Dataset is downloaded from Kaggle which is a collection of preprocessed data from 1000 patients who visited a multispecialty hospital in India for heart-related issues. The original dataset contains 12 variables that measure various aspects of the patients' health and heart condition, but we only selected a few of them as follows:

1. Gender (categorical)
2. Chest pain type (categorical)
3. Exercise induced angina (categorical)
4. Age (Numeric)
5. Resting blood pressure (Numeric)
6. Maximum heart rate achieved (Numeric)
7. Serum cholesterol (Numeric)
8. Number of major vessels (Numeric)
9. Classification (categorical) (Response variable)

The dataset aims to provide a comprehensive and diverse set of features that can be used for exploring the risk factors and symptoms of heart disease, as well as for developing and evaluating predictive models that can diagnose heart disease at an early stage using suitable procedure and models.

We will use this data to perform and apply logistic regression model to predict(classify) whether a certain patient will suffer from heart disease or not as well as applying suitable machine learning algorithms.

Cardiovascular Disease Dataset Description

S.No	Attribute	Assigned Code	Unit	Type of the Data
1	Age	age	In Years	Numeric
2	Gender	gender	1,0(0= female, 1 = male)	Binary
3	Chest pain type	chestpain	0,1,2,3 (Value 0: typical angina Value 1: atypical angina Value 2: non-anginal pain Value 3: asymptomatic)	Nominal
4	Resting blood pressure	restingBP	94-200 (in mm HG)	Numeric
5	Serum cholesterol	serumcholestrol	126-564 (in mg/dl)	Numeric
6	Maximum heart rate achieved	maxheartrate	71-202	Numeric
7	Exercise induced angina	exerciseangia	0,1 (0 = no, 1 = yes)	Binary
8	Number of major vessels	noofmajorvessels	0,1,2,3	Numeric
9	Classification	target	0,1 (0= Absence of Heart Disease, 1= Presence of Heart Disease)	Binary

Descriptive analysis

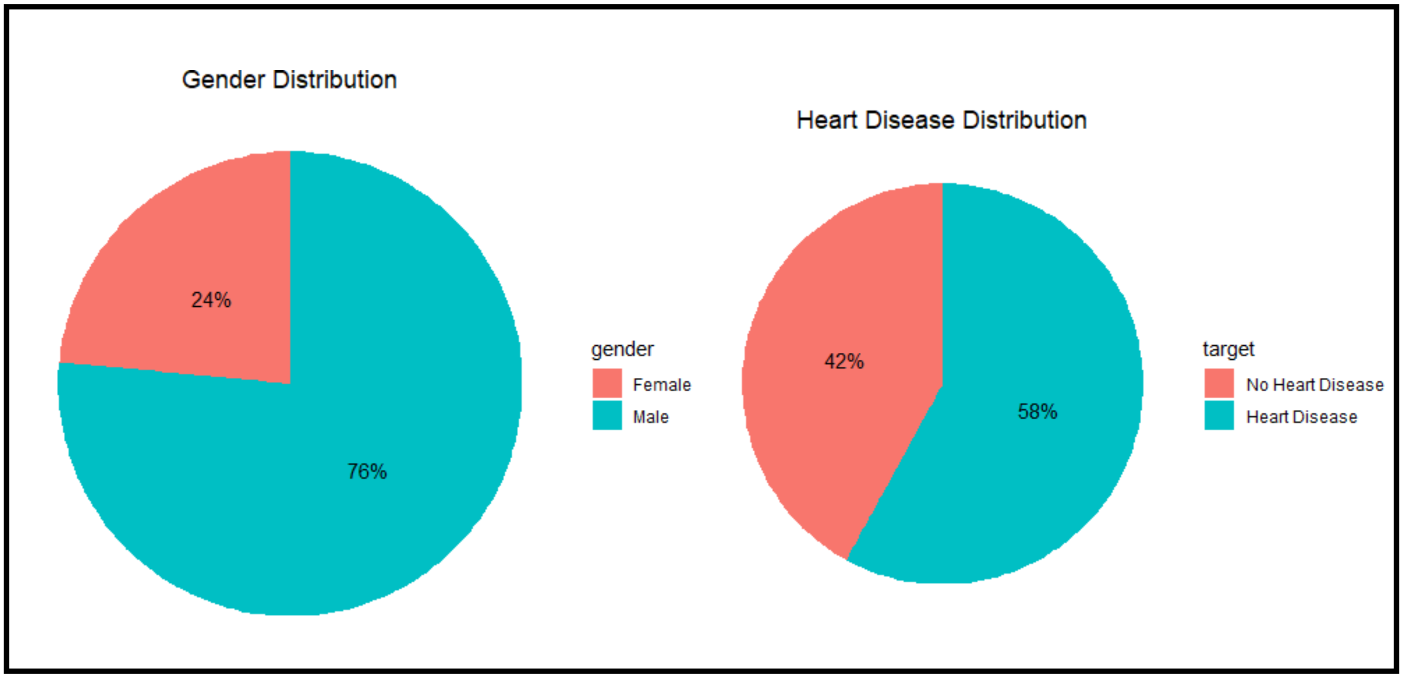


Figure 1:Gender & Hear disease Pie charts

The figure above shows the Gender distribution in our data where 76% of the patients were male while only 24% were female. 58% of our patients suffered from heart disease while the other 42% didn’t suffer from it which shows that the majority of our patients suffer from heart disease.

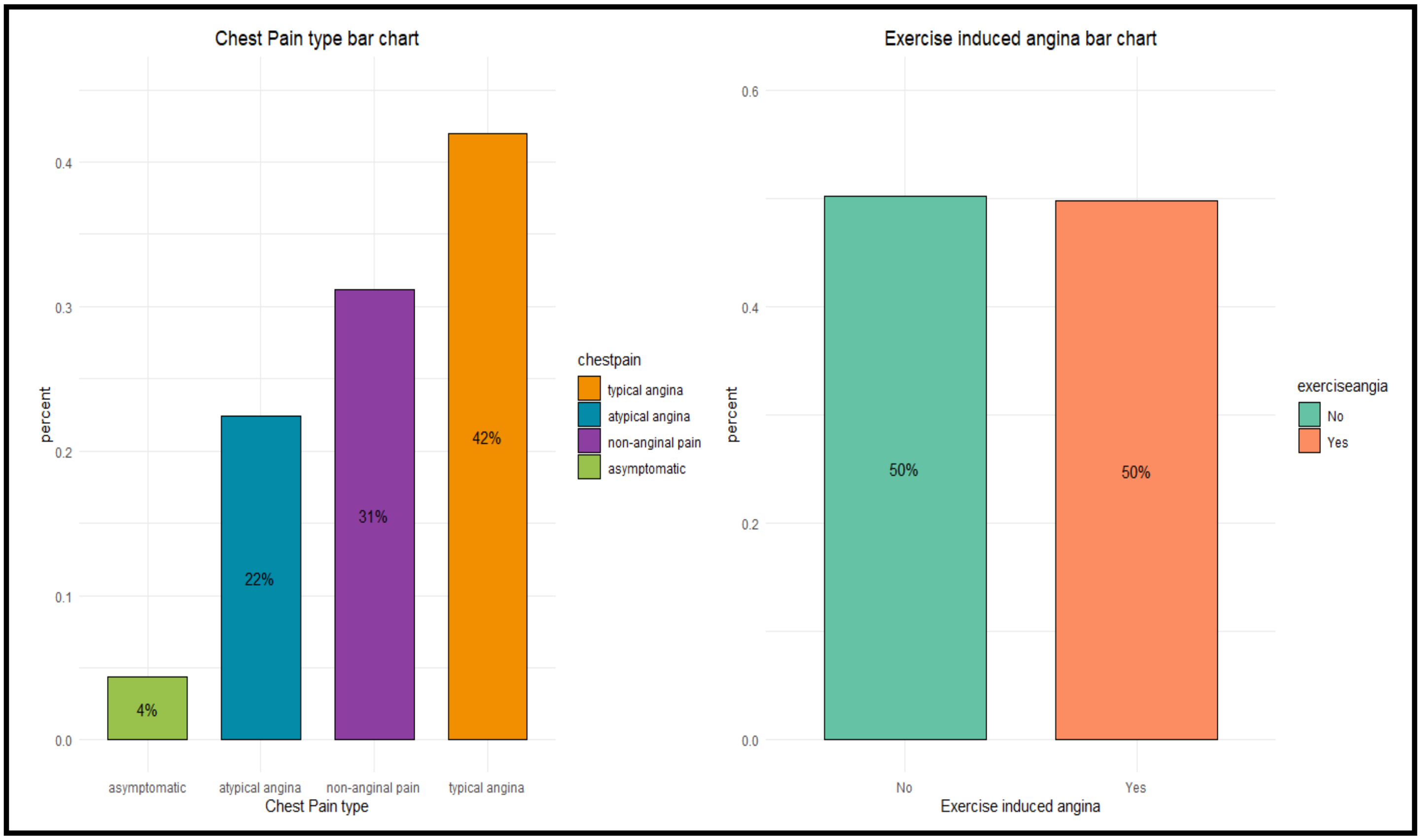
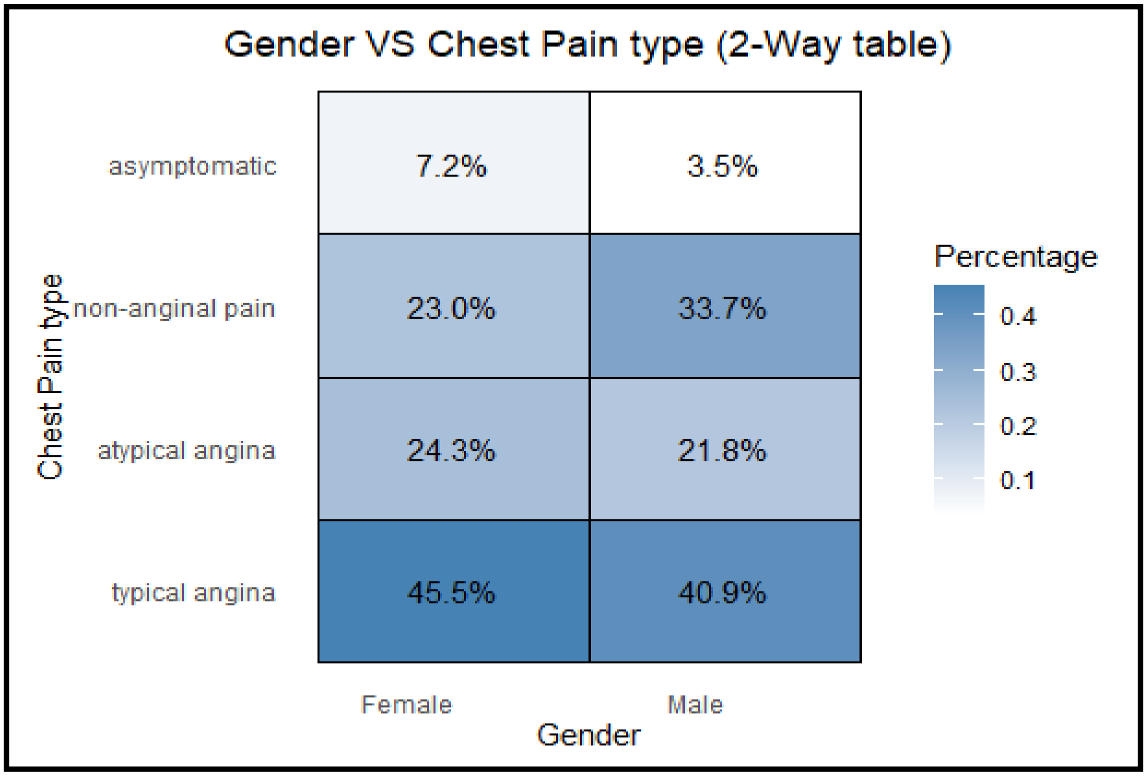


Figure 2:Chest pain type & Exercise angina Bar charts

For the plot on the left42% of the chest pain types felt were typical angina pain followed by non-anginal pain with 31% while asymptomatic with the least chest pain felt among our patients with only 4%.

For the plot on the right, we notice that there's almost a balance between whether exercising included angina or not since total number of patients who felt angina was 498 while the patients who didn't feel it was 502.

Table 1: Joint Distribution for gender & Chest pain type



Chest Pain type	Female	Male
asymptomatic	7.2%	3.5%
non-anginal pain	23.0%	33.7%
atypical angina	24.3%	21.8%
typical angina	45.5%	40.9%

The table above shows the distribution of chest pain type among female & male patients, 45.5% of the females experienced the typical angina compared with the 40.9% of the males who experienced the same type of chest pain.

It's also worth to mention that despite that 45.5% of females had a typical angina, 24.3% of them had atypical angina compared with the 21.8% in males which is lower than the females.

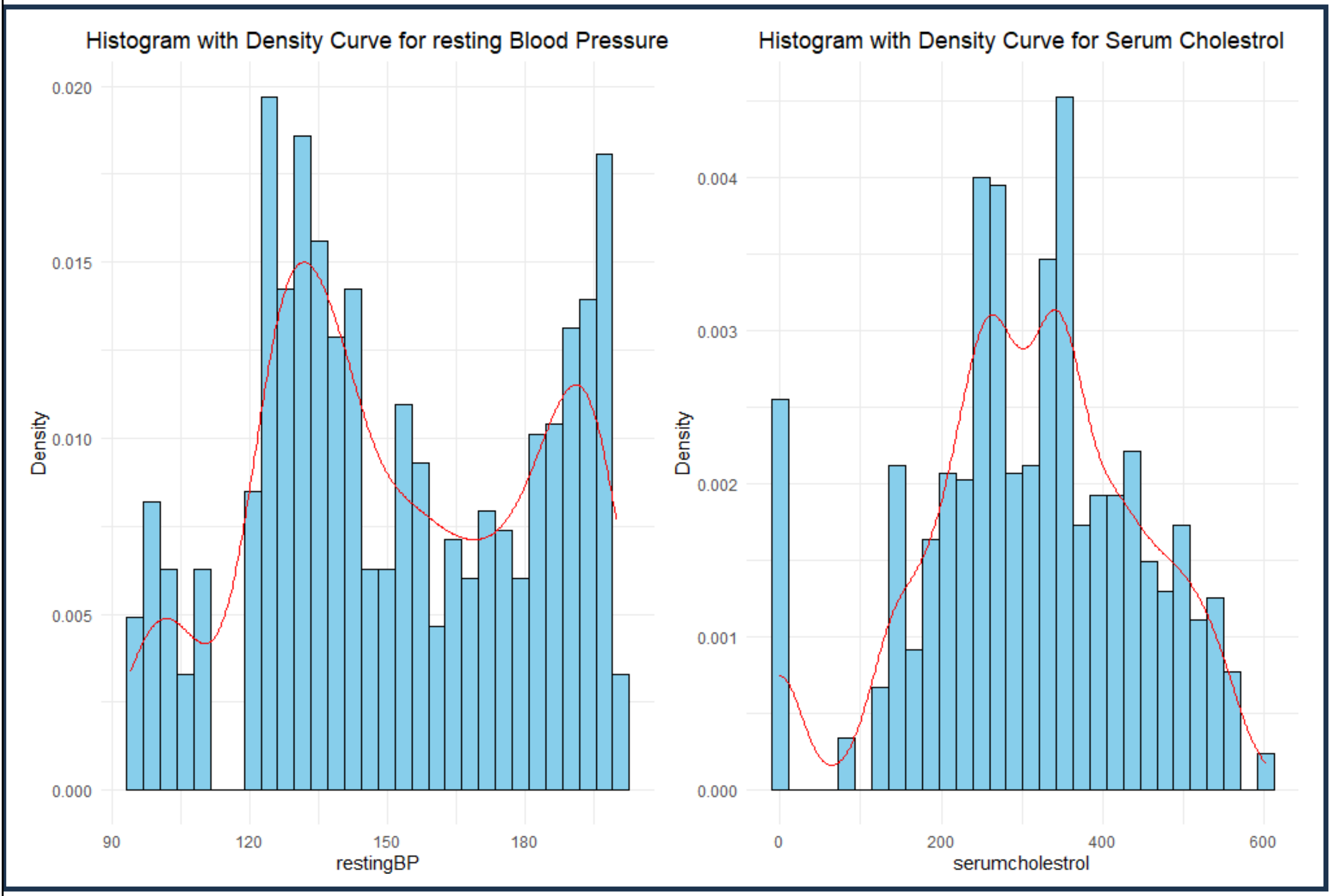


Figure 3: Histograms for resting blood pressure and serum cholestrol.

Table 2: Resting blood pressure & serum Cholestrol summary statistics

	min	Q1	median	Q3	Max	Mean	skewness
RestingBP	94	129	147	181	200	151.75	0.02
serumcholestrol	0	235.75	318	404.25	602	311.45	-0.31

The distribution of RestingBP is almost symmetric with skewness coefficient 0.02 unlike serumcholestrol which is almost negatively skewed with skewness coefficient -0.31.

For 25% of the patients, They experienced 129 for resting blood pressure while the mean resting blood pressure for the patients was 151.75.

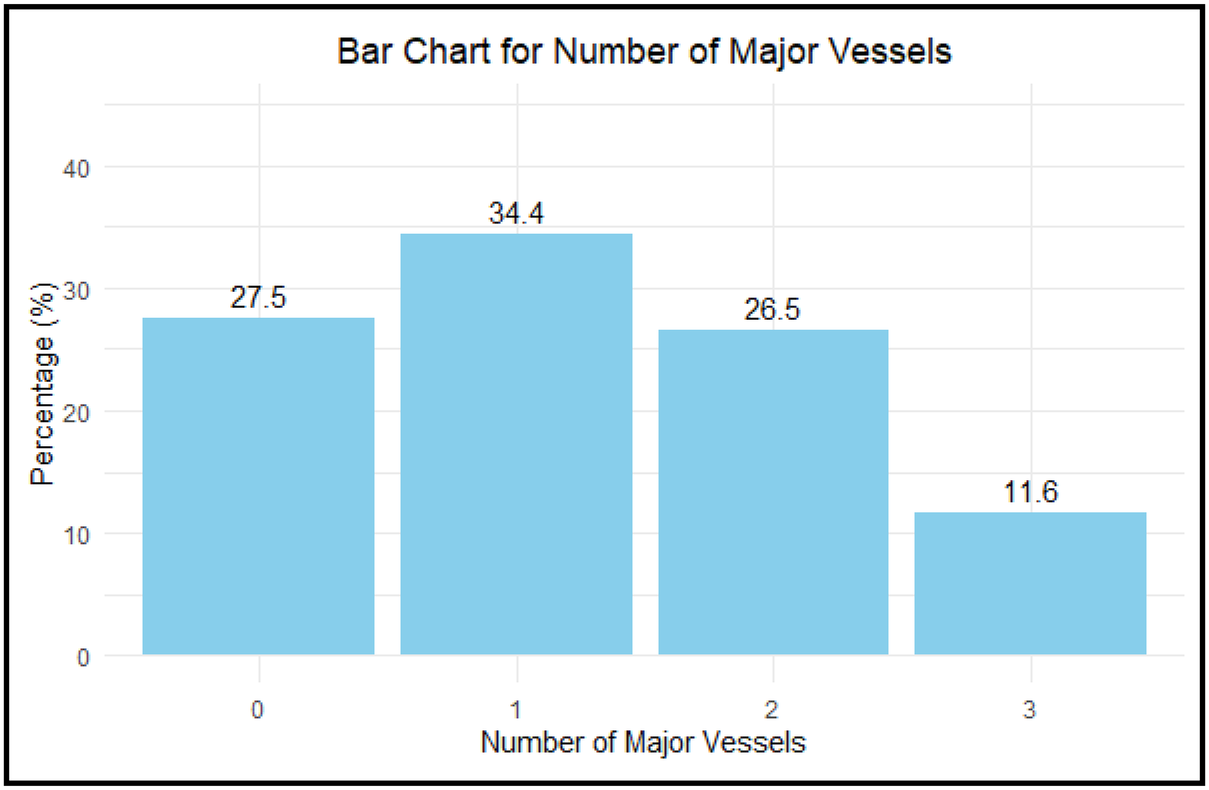


Figure 4: Bar chart for number of major vessels.

Number of major vessels (1) is the most frequent category in our patients since we have 34.4% of the patients who have only one major vessel, followed by 27.5% of the patients who do not have major vessels (0) then comes in the third place 26.5% of the patients who have 2 major vessels, finally, 11.6% of the patients who have 3 major vessels.

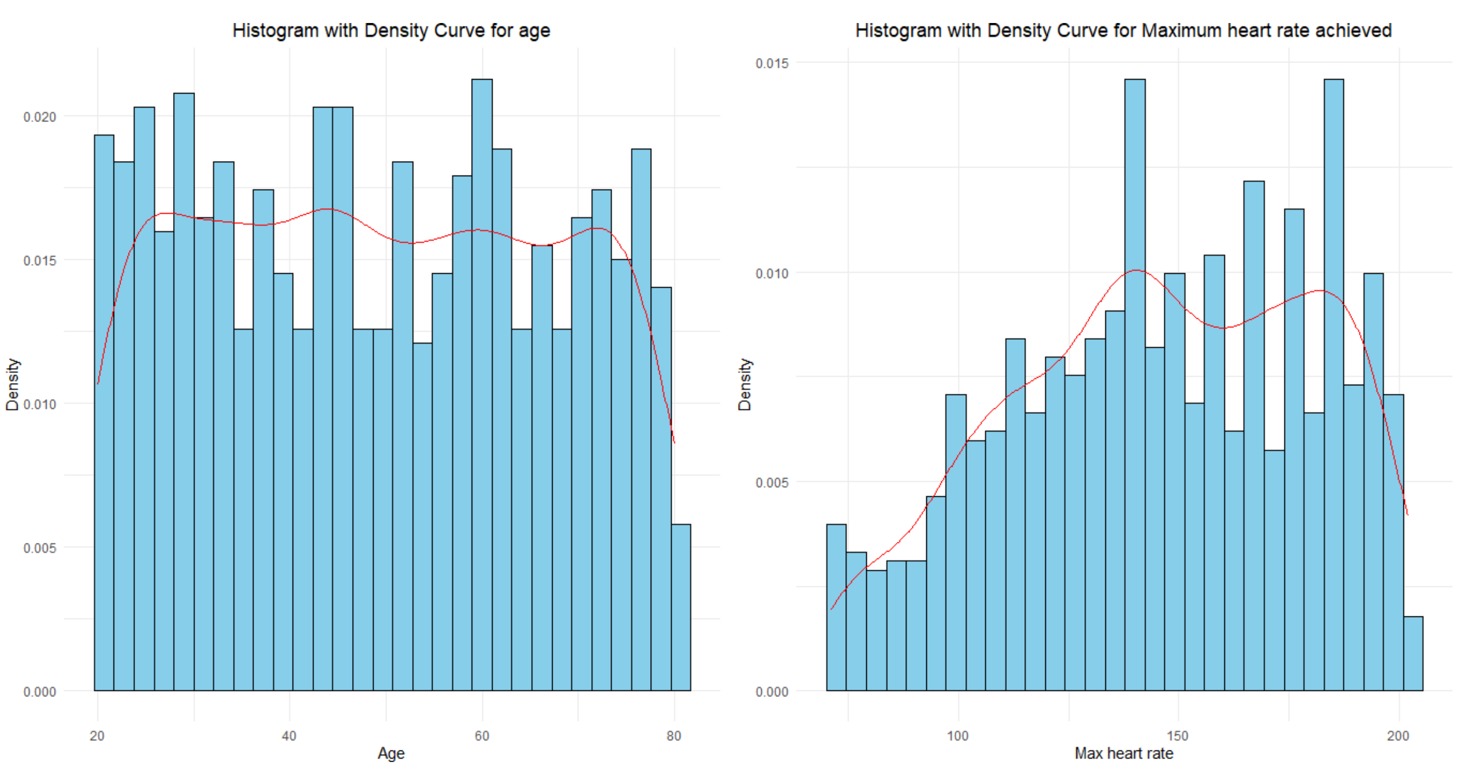


Figure 5: Histograms for age & maximum heart rate achieved.

Table 3: Age & max heart rate summary statistics

	min	Q1	median	Q3	Max	Mean	skewness
Age	20	34	49	64	80	49	0.03
Max heart rate	71	120	146	175	202	145.48	-0.25

The distribution of age in figure 3 is almost symmetric which goes along with the statistics in table 2 since median age = mean age = 49 as well as the skewness coefficient which is around zero. The youngest patient we have in the data is 20 years old while the oldest patient is 80 years old with average patient age of 49.

The distribution of max heart rate in figure 3 is barely negatively skewed with the skewness coefficient (-0.25) which mean that most of the heart rates happened at the larger values.

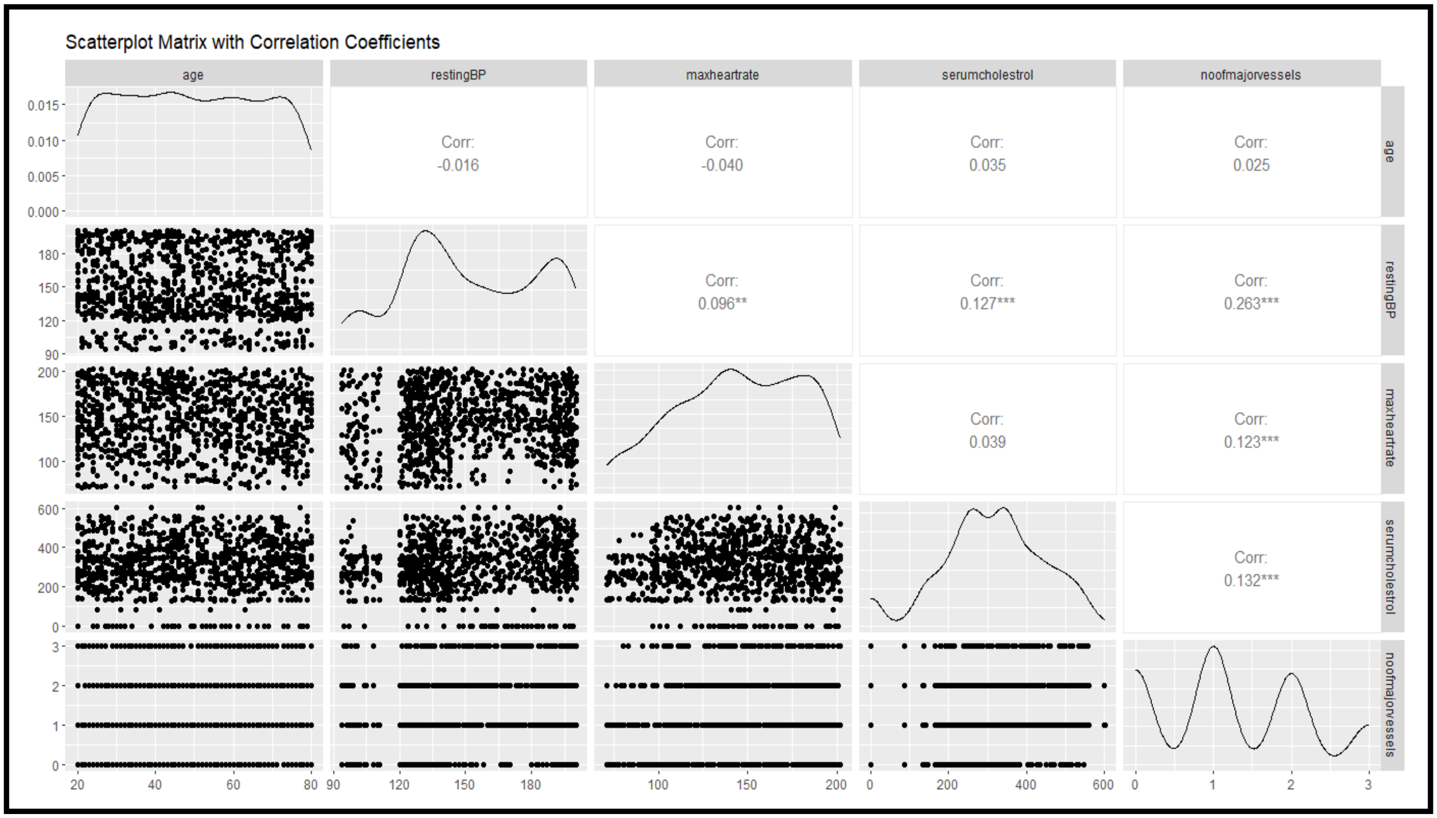


Figure 6: Scatterplot matrix for quantitative variables

The scatterplots above show scattered points for almost all combinations of variables since the highest correlation value is for Resting blood pressure number of major vessels which shows a weak positive linear relationship between the 2 variables.

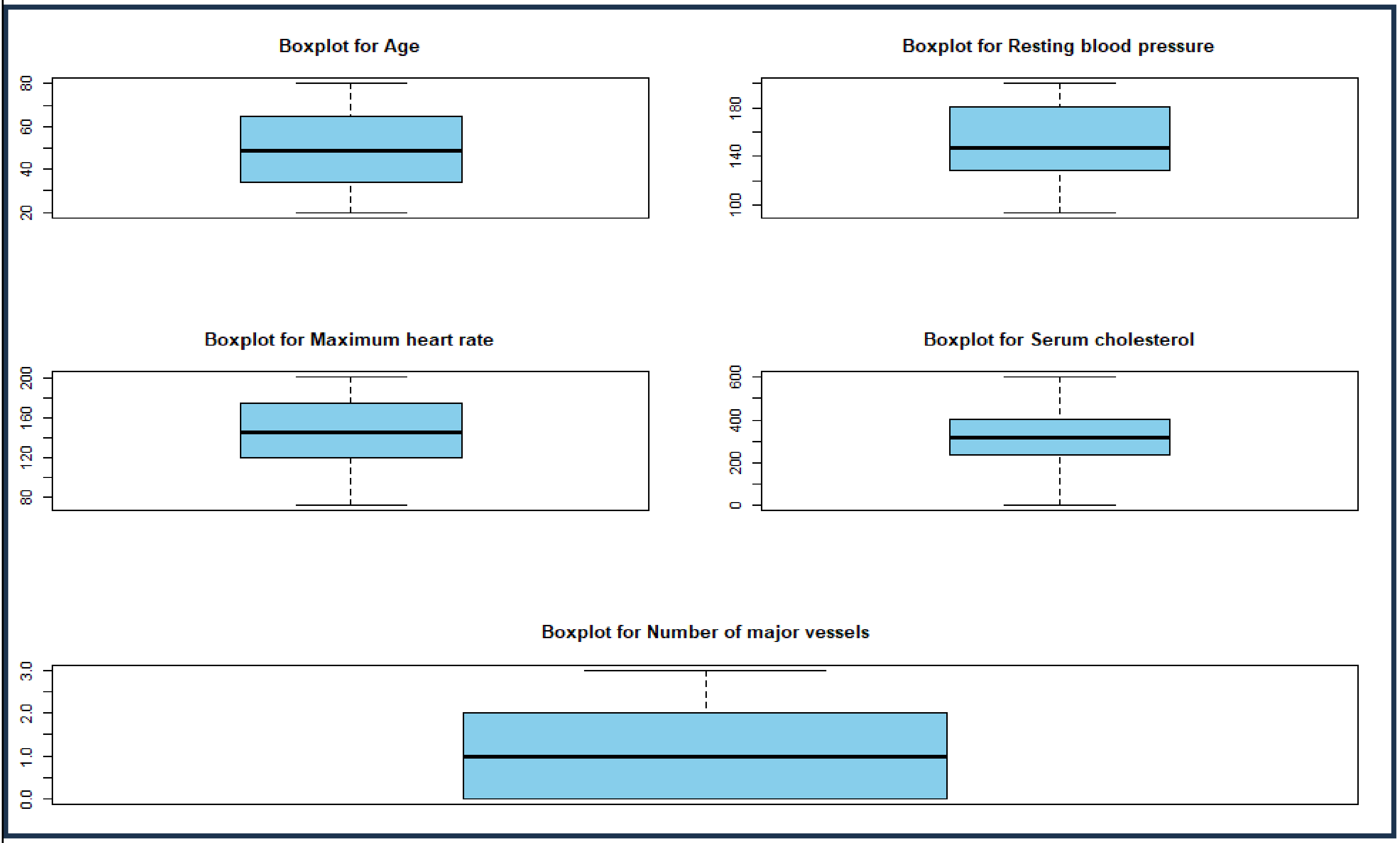


Figure 7: Boxplots for quantitative variables

Figure 7 shows that we don’t have outliers for any of the 5 quantitative variables since the data is already preprocessed.

Logistic regression model:

At the first we will split the data into training (70%) and testing (30%).

To predict (classify) whether a certain patient will suffer from heart disease or not, we will use logistic regression because we have the response variable is categorical variables.

Binary logistic regression : between Classification a binary response variable and a set of explanatory variables (Gender, Chest pain, Exercise induced angina ,Age, Resting blood pressure, Maximum heart rate achieved, Serum cholesterol, Number of major vessels).

$$\log\left(\frac{\hat{\text{Target}}}{1-\hat{\text{Target}}}\right) = \hat{\beta}_0 + \hat{\beta}_1 Gender + \hat{\beta}_2 \text{Chest pain} + \hat{\beta}_3 \text{Exercise induced angina} + \hat{\beta}_4 \text{Age} + \hat{\beta}_5 \text{Resting blood pressure} + \hat{\beta}_6 \text{ Maximum heart rate achieved} + \hat{\beta}_7 \text{ Serum cholesterol} + \hat{\beta}_8 \text{ Number of major vessels}$$

Step (1): Model Building using Stepwise procedure on training data

We will use this procedure to choose a better model with fewer variables.

The left variables after this procedure are Gender, Exercise induced angina, Age, Serum cholesterol.

We will look at the estimated coefficients of this model

target		**Odds ratio**	**estimated**	**Std. Err**	**z**	**p>z**
Chest pain	Atypical angina	3.05325123	1.116207	0.292890	3.811	0.000138
	Non angina	20.1233131	3.001879	0.292725	10.255	0.0000000225
	Asymptomatic	11.5656442	2.448039	0.632326	3.871	0.000108
Resting blood pressure		1.04273075	0.041843	0.004473	9.355	0.0000000225
Maximum heart rate achieved		1.01080798	0.010750	0.003425	3.138	0.001699
Number of major vessels		2.69535251	0.991529	0.126463	7.840	0.0000013735
Constant		.0000498462	-9.906567	0.933365	-10.614	0.0000000225

Interpret

From this table, we will interpret their odds ratio:

The estimated odds ratio for being target is higher for atypical angina chest pain than the typical angina chest pain by 205.3% holding other variables constant.

For Non angina: The estimated odds ratio for being target for Non angina chest pain is 19.12 times that the estimated odds ratio for the typical angina chest holding other variable constant.

For Asymptomatic: The estimated odds ratio for being target for Asymptomatic chest pain is 10.56 times that the estimated odds ratio for the typical angina chest holding other variables constant.

The estimated odds ratio for being target increases with a fraction of 4.273% when Resting blood pressure increases by one unit holding other variables constant.

The estimated odds ratio for being target increases with a fraction of 1.08% when Maximum heart rate achieved increases by one unit holding other variables constant.

The estimated odds ratio for being target increases with a fraction of 1.69% when Number of major vessels increases by one unit holding other variables constant.

e^{B0}=.0000498462 the estimated odds of presence for heart disease when typical angina chest pain, Resting blood pressure equal zero, Maximum heart rate achieved equal zero& Number of major vessels equal zero.

Chosen model:

$$\log\left(\frac{\hat{\text{Target}}}{1-\hat{\text{Target}}}\right) = -9.906567+.\ 1.116207\ \text{Chest pain1}+3.001879\ \text{Chest pain2}$$
$$+\ 2.448039\ \text{Chest pain3} +0.041843\ \text{Resting blood pressures}$$
$$+0.010750\ \text{Maximum heart rate achieved} +0.991529\ \textit{Number of major Vessels}$$

Check for multicollinearity

$$\text{VIF} = \frac{1}{1-R^2}$$

By using VIF we find there is no multicollinearity.

Goodness of fit (at the test data)

Since we have a continuous variable, then we will use Hosmer-Lem show test, because this test groups the data using the quantiles of the predicted probabilities before running the person test statistic

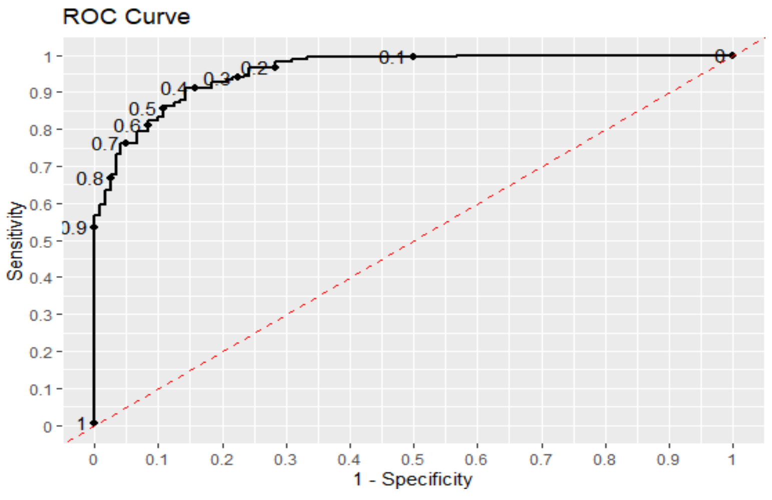
<u>Hosmer-Lem show test</u>:

H_0: insignificant model **H_1**: significant model

	Degree of freedom	Deviance Resid.	Df Resid	Dev	Pr(>chi)
Null			699	956.07	
Chest pain	3	229.713	696	726.36	2.2e-16
Resting blood pressure	1	1480326	695	578.03	2.2e-16
Maximum heart rate achieved	1	10.489	694	567.54	0.001201
Number of major vessels	1	71.496	693	496.05	2.2e-16

All the variable is a significant, then the model is significant as a whole.

<u>Roc curve</u>



Area under the curve AUC= 0.956

The AUC (Area under the ROC curve) is 0.956which is higher than 0.5 indicating that this model is better than the model with intercept only.

We find that the best cutoff point nearly occurs at 0.403329, we will recalculate the classification table based on the best cutoff point

classification table

TRUE	**Classified**		
	1	0	Total
1	164	16	180
0	17	103	120
Total	181	119	300

Sensitivity: 91.11% **Specificity**: 85.83% **Accuracy**: 89%

From the classification table, we can see:

1) Among the 180 true Y=1, 164 were classified as $\hat{Y}$=1; a 91.11% correct classification percentage.

2) Among the120 true Y=0,103 were classified as $\hat{Y}$=0; a 85.83% correct classification percentage

3) Overall; 283(164+103) out of 300 were correctly classified; the overall percentage of correct classification is 89%, which is reasonably high.

Machine learning Algorithms:

We will apply one of the common machine learning algorithms, which is **Decision tree** to classify whether the patient has heart disease or not.

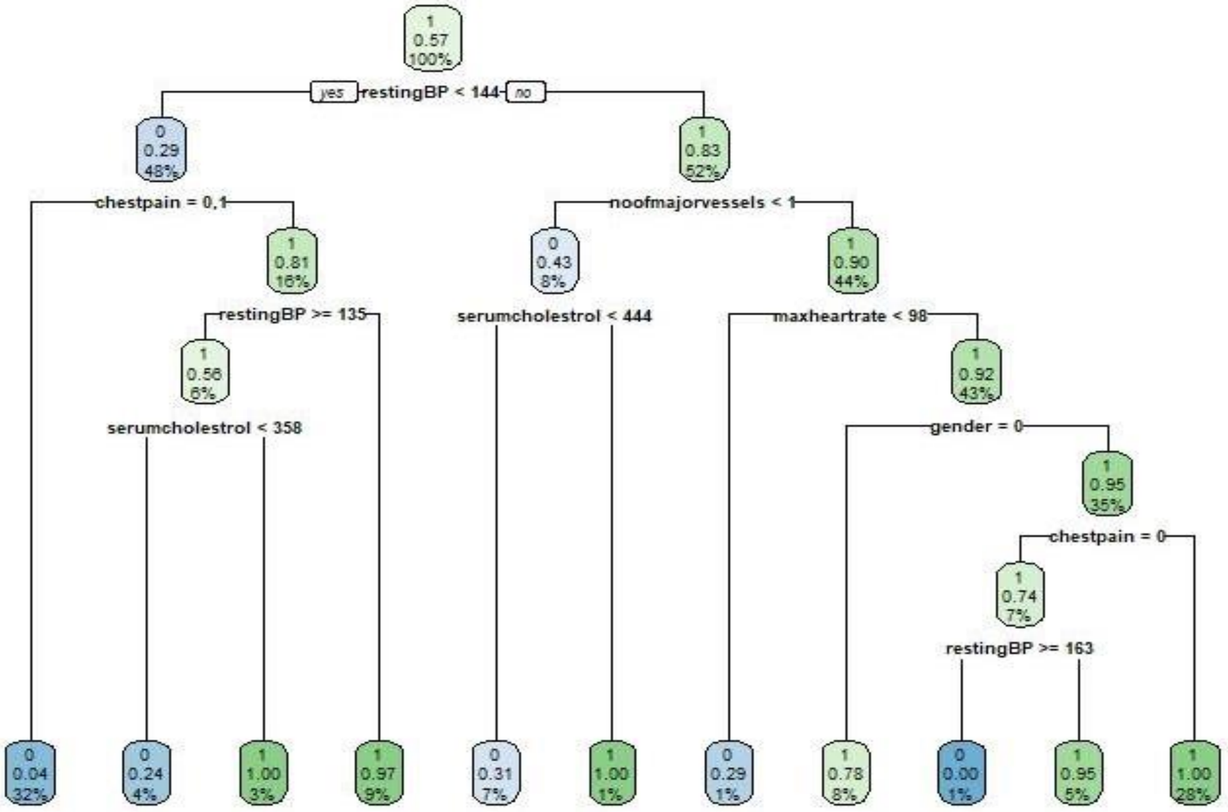


Figure 8: Decision tree for the model

From ***Figure 8***, we can conclude that:

- the patients who have restring blood pressure between 135 and less than 144 mm GH, with non-anginal pain and asymptomatic chest pain and with Serum cholesterol greater than or equal 358 mg/dl are exactly carrying the heart disease.
- the patients who have restring blood pressure greater than or equal 144 mm GH, with number of major vessels less than 1 and with Serum cholesterol greater than or equal 444 mg/dl are exactly carrying the heart disease.
- the male patients who have restring blood pressure greater than or equal 144 mm GH, with number of major vessels greater than or equal 1, with maximum heart rate greater than or equal 98, with typical angina chest pain and with resting blood pressure greater than or equal 163 mm GH are not exactly carrying the heart attack.
- If the resting blood pressure is less than 144 mm HG then 29% will not have heart

disease and 71% will not have the heart disease, then if the chest pain is typical angina or atypical angina then only 4% will have the disease and 96% will not. But if the chest pain is non-anginal pain or asymptomatic, then 81% of the patient will have the disease than 19% will not have, then if the resting blood pressure is greater than or equal 135 mm HG then 56% of the patients will have the disease than 44% will not have and otherwise, 97% of the patients will have the heart disease, then if serum cholesterol is less than 358 mg/dl then 24% of the patients will not have the disease and otherwise 100% will have the disease.

- If the resting blood pressure is greater than or equal 144 mm HG then 83% of the patients will have the disease than 17% will not have, but if number of major vessels is less than 1 then 43% will not have the disease than 57% will have the disease and if serum cholesterol is less than 444 mg/dl then 31% of the patients will not have the disease but otherwise 100% of the patients will have the disease. in the other hand, the number of major vessels is greater than or equal to 1, then 90% of the patients will have the disease than 10% will not. If the maximum heart rate is less than 98, then 29% of the patients will not have the disease but if the maximum heart rate is greater than or equal to 98 then 92% of the patients will have the disease but if the patient is woman then 78% of them will have the disease than 22% will not but if the patient is man then 95% of them will carry the disease but if they have the typical angina, then 74% of them will have the disease and if their resting blood pressure is greater than or equal 163 mm HG then no one of them will have the disease and otherwise 95% of them will have the disease and if the male patients suffer from atypical angina , non-anginal pain or asymptomatic pain, then 100% of them will have the disease.

Table 4: importance of the variables in splitting the data.

Variable	Overall
chestpain	201.171143
restingBP	149.399552
noofmajorvessels	128.426935
serumcholestrol	113.793858
maxheartrate	77.312427
gender	13.915016
age	3.072845
exerciseangia	1.736645

From *Table 4,* we can conclude that:

- The most important variable in splitting the patients to whether have heart disease or not is chest pain.
- The lowest important variable in splitting the patients to whether have heart disease or not is Exercise included angina.

Table 5: confusion matrix.

Predicted	**True**		
	1	0	Total
1	115	5	120
0	24	156	180
Total	139	161	300

Table 6: accuracy measures

measure	value
Accuracy	0.9033
Sensitivity	0.8273
Specificity	0.9689

From ***table 6***, we can conclude that:

- Overall, the model is correctly classified the patients into carrying the heart disease or not by a proportion 90.33%.
- 82.73% of the patients who have the heart disease, the model could classify them exactly correct.
- 96.89% of the patients who do not have the heart disease, the model could classify them exactly correct.

The main results and conclusion:

- Most of the patients have the heart disease than those who do not have.
- Most of the patients suffer from the chest pain.
- The age of the patients is ranged between 20 and 49 years old.
- The highest two correlated variables are resting blood pressure and the number of major vessels.
- The most significantly variables affect whether the patients are carrying the heart disease or not are chest Pain, resting blood pressure, maximum heart rate and number of major vessels.
- Our logistic model we used to fit the data is good fit the data based on the significance, Hosmer – Lem show test, Roc curve and the classification table.
- Using the Decision tree, we found that the most variable affects in determining whether the patient is carrying the disease or not is chest Pain. In other hand, the lowest variable affects in determining whether the patient is carrying the disease or not is exercise included angina.
- Using the Decision tree, the accuracy of prediction whether the patient has heart disease or not, Specificity and Sensitivity have been improved than using the logistic model in predictions.
- Finally, we can say that women are more exposure to have heart attack than men and the patients with highly Serum cholesterol are also more exposure to have heart attack than those they are with low Serum cholesterol.