PREDICTION OF HEART FAILURE

USING

LOGISTIC REGRESSION ANALYSIS:

By Group 8:

HINAL PANCHAL

JENY GEORGE

ASHWIN PANDEY

BUDDHADITYA RATH

INTRODUCTION:

As per the Centers for Disease Control and Prevention report, heart disease is the prime killer of both men and women in the United States and around the globe. Heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United States. One person dies every 36 seconds in the United States from cardiovascular disease. About 655,000 Americans die from heart disease each year—that is 1 in every 4 deaths. Heart disease costs the United States about $219 billion each year from 2014 to 2015. This includes the cost of health care services, medicines, and lost productivity due to death.

There are several data mining techniques that can be leveraged by researchers/ statisticians to help health care professionals determine heart disease and its potential causes. Some of the significant risk factors associated with heart disease are age, blood pressure, total cholesterol, diabetes, hypertension, family history of heart disease, obesity, lack of physical exercise, etc.

The objective of this project is to build a regression model and run statistical tests to assess how strongly are the clinical factors associated with heart disease and how it is related to the higher probability of getting a heart disease. We shall be implementing Multiple and Logistic Regression approaches together with the data. This project uses the Cleveland heart disease dataset.

DATASET:

Here is a glimpse of the dataset in hand:

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	class
63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
41	0	2	130	204	0	2	172	0	1.4	1	0	3	0

The dataset consists of **297** records with **14** variable columns.

Since the outcome variable ***class*** has more than 2 levels, we have created a new variable **heart disease:1** to represent binary 0/1 outcome where any **value > 0 shall be 1** and all **0 values will stay 0.**

Also, we renamed **sex** levels (originally Male and Female) as 1/ 0 for better clarity.

Source: https://archive.ics.uci.edu/ml/datasets/Heart+Disease/

DATA DICTIONARY:

Age: It is a *continuous* data type which describes the age of the person in years.

Sex: It is a *discrete* data type that describes the gender of the person. Here 0 = Female and 1 = Male.

Chest Pain type: It is a *discrete* data type that describes the chest pain type with following parameters- 1 = Typical angina; 2 = Atypical angina; 3 = Non-anginal pain; 4 = Asymptotic.

RestingBloodPressure: It is a *continuous* data type which describes resting blood pressure in mm Hg.

Cholesterol: It is a *continuous* data type that describes the serum cholesterol in mg/dl.

FastingBloodSugar: It is a *discrete* data type that compares the fasting blood sugar of the person with 120 mg/dl. If FBS >120 then 1 = true else 0 = false.

RestECG: It is a *discrete* data type that shows the resting ECG results where 0 = normal; 1 = ST-T wave abnormality; 2 = left ventricular hypertrophy.

MaxHeartRate: It is a continuous data type that describes the max heart rate achieved.

ExerciseAngina: It is a *discrete* data type where exercise induced angina is shown by 1 = Yes and 0 = No.

Oldpeak: It is a *continuous* data type that shows the depression induced by exercise relative to weight.

Slope: It is a *discrete* data type that shows us the slope of the peak exercise segment where 1= up-sloping; 2 = flat; 3 = down-sloping.

ColoredVessels: It is a *continuous* data type that shows us the number of major vessels coloured by fluoroscopy that ranges from 0 to 3.

Thal: It is a *discrete* data type that shows us Thalassemia where 3 = normal ; 6 = fixed defect ; 7 = reversible defect.

KEY VARIABLES:

	heart disease:1	*age*	*sex*
heart disease:1	1		
age	0.227075155	1	
sex	0.278466697	-0.092399479	1
chestpain	0.408944687	0.110470866	0.008908026
bloodpressure	0.153490026	0.290476262	-0.0663402
cholestrol	0.080284751	0.202643546	-0.198089063
bloodsugar	0.00316683	0.132061989	0.0388503
restecg	0.166343488	0.149916512	0.033896828
maxheartrate	-0.423817064	-0.394562881	-0.060496006
exerciseangina	0.42135549	0.096488805	0.14358125
oldpeak	0.424052057	0.197122616	0.106567243
slope	0.333049109	0.159404737	0.033344964
coloredvessel	0.463188625	0.362210343	0.0919248
thalassemia	0.526639576	0.126585998	0.383651748

Based on correlation outcomes we are going to focus on these variables:

Age: It is a *continuous* data type which describes the age of the person in years

Sex: It is a *discrete* data type that describes the gender of the person. Here 0 = Female and 1 = Male

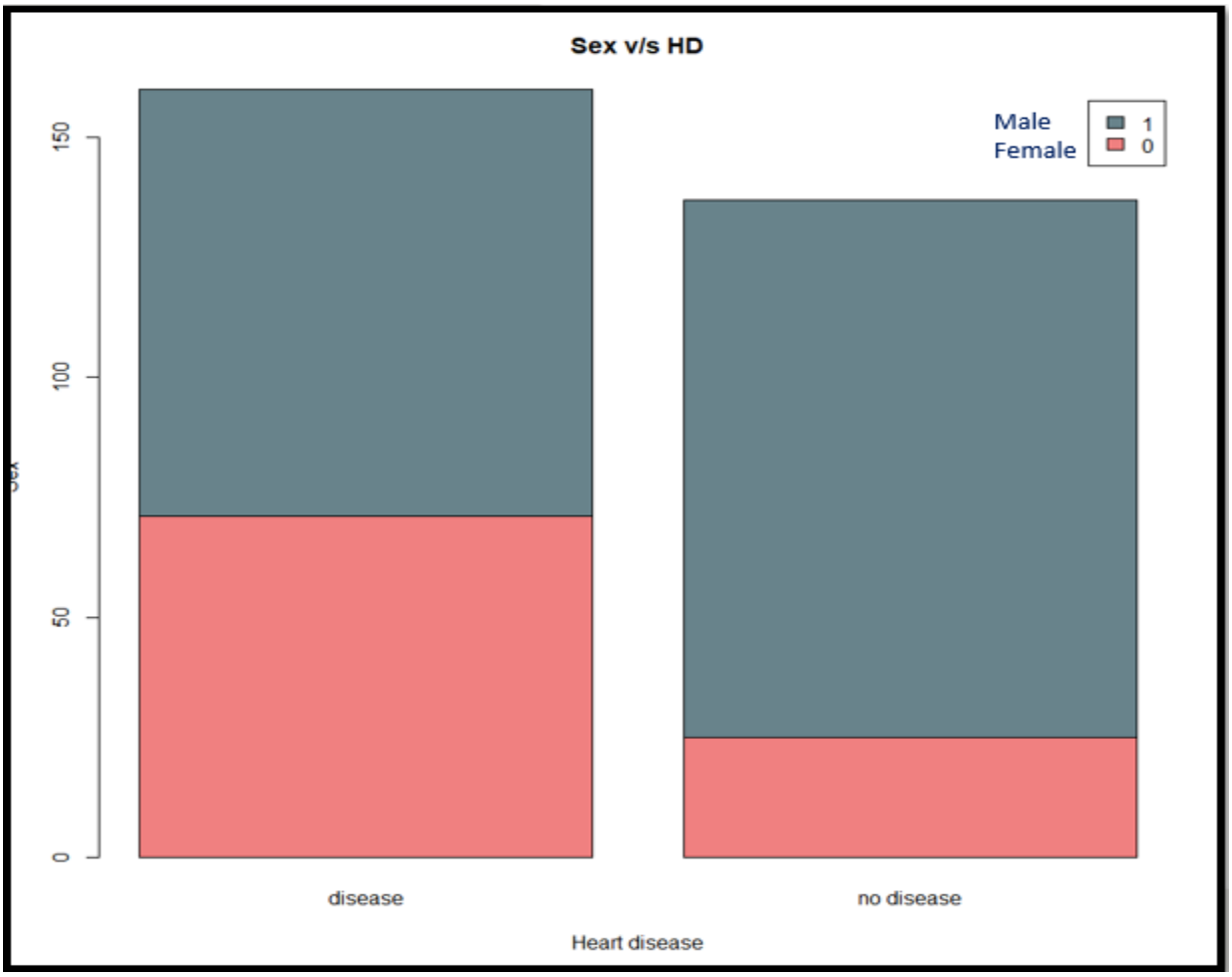
MaxHeartRate: It is a *discrete* data type that shows us Thalassemia where 3 = normal; 6 = fixed defect; 7 = reversible defect.

ExerciseAngina: It is a *discrete* data type where exercise induced angina is shown by 1 = Yes and 0 = No.

ANALYSIS AND INSIGHTS:

SEX: Since sex is a binary variable in this dataset, chi-squared test will be the appropriate test for this variable. Here is the output on using chi-square test to assess the relationship between sex and heart disease:1 (outcome variable).

Observation : CHISQ.TEST value = **1.59453305336596E-06 = p-value < 0.05** this means that we can conclude that, heart disease is dependent on Sex.



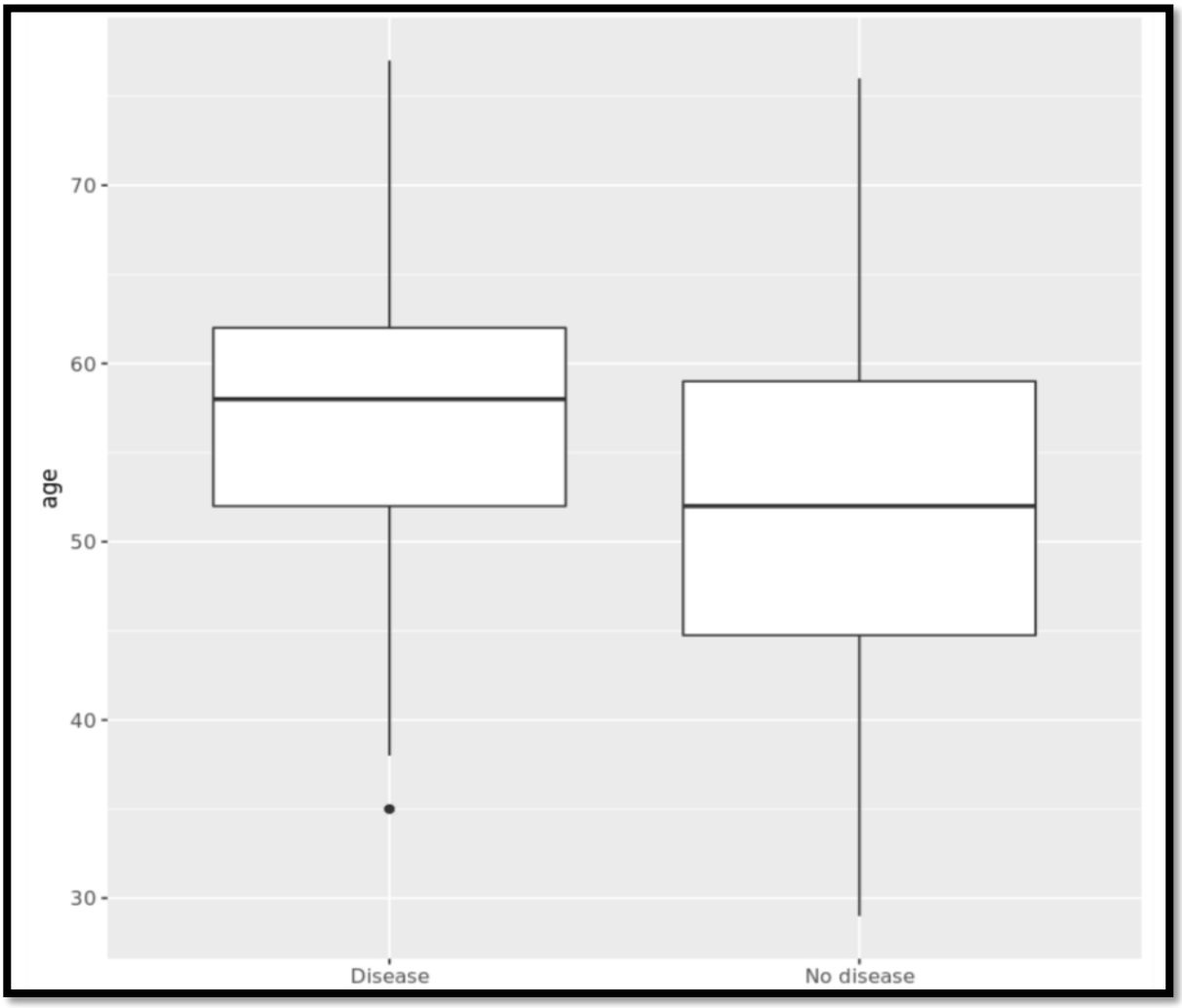
Sex v/s Heart Disease:1

The graphical plot above and statistical test clearly show us that the clinical variable Sex were chosen are significantly associated with our outcome since p-value < 0.05 for the test.

AGE: Since age is a continuous variable, we used t-test statistic to determine relationship between age and heart disease:1.

Observation: p-value = **4.1294E < 0.05**

We can infer that heart disease is dependent on Age.



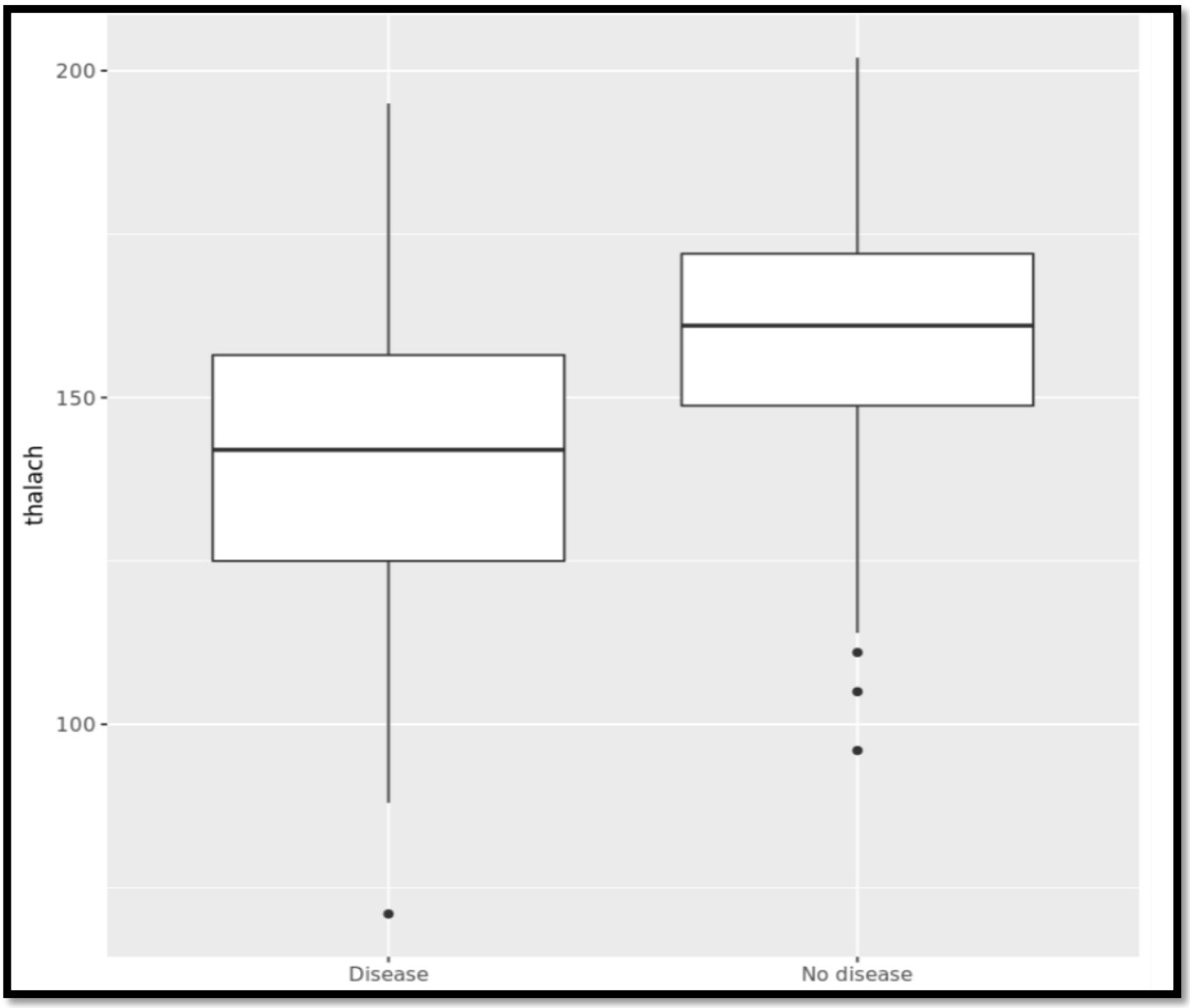
Age v/s Heart Disease:1

The graphical plot above and statistical test clearly show us that the clinical variable Age were chosen are significantly associated with our outcome since p-value < 0.05 for the test.

MAX HEART RATE: Using t-test statistic to assess relationship between thalach(Maximum Heart Rate) and heart disease:1.

Observation: p-value = **1.114E < 0.05**

We can infer that heart disease is dependent on MaxHeartRate.



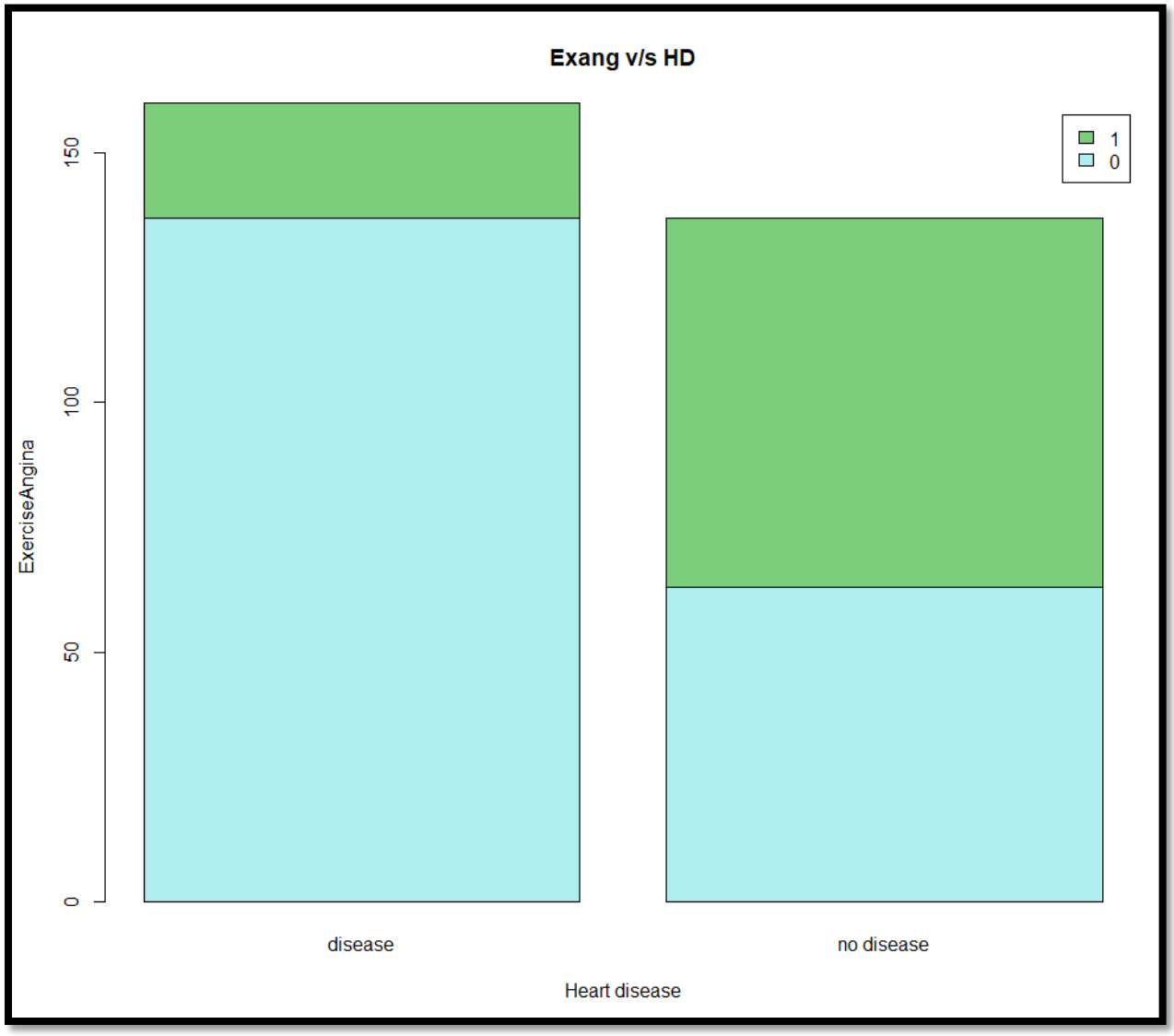
Thalach v/s Heart Disease:1

The graphical plot above and statistical test clearly show us that the clinical variable Thalach (Maximum Heart Rate) were chosen are significantly associated with our outcome since p-value < 0.05 for the test.

EXERCISEANGINA:

Observation : CHISQ.TEST value = **1.167E6-10= p-value < 0.05**

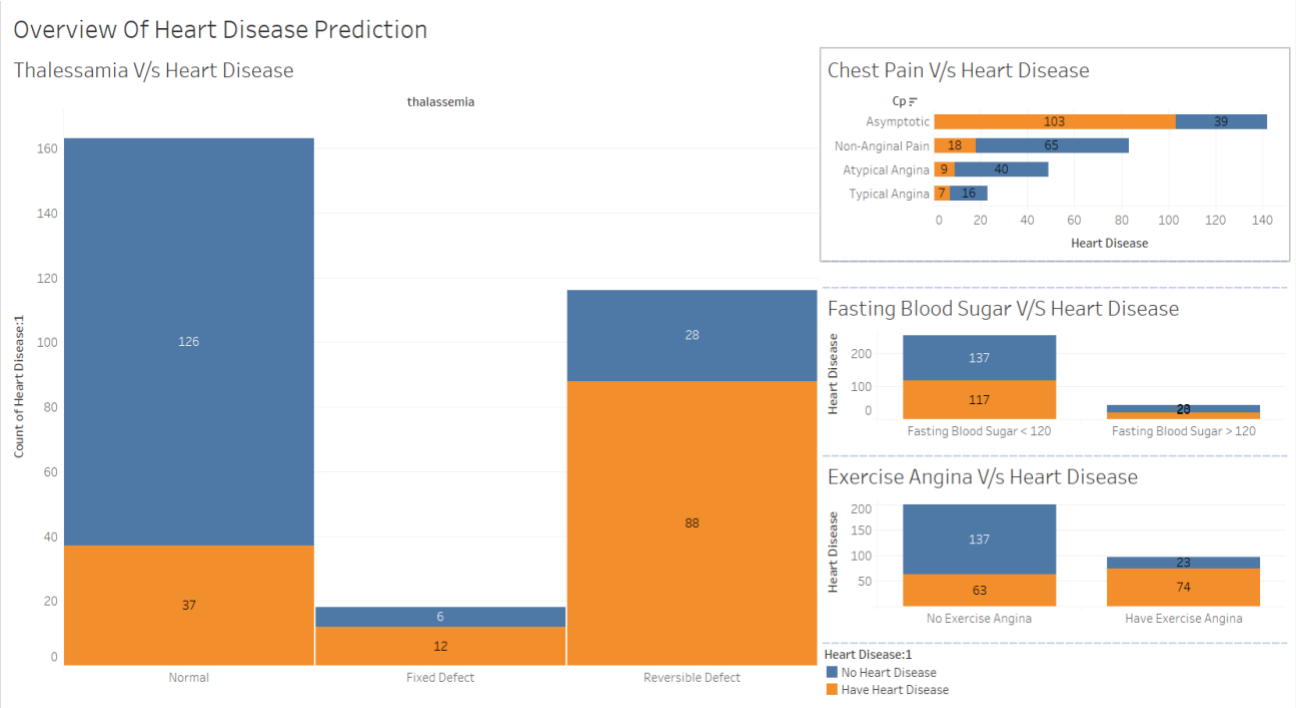
We can infer that heart disease is dependent on Exercised induced angina.



Exang v/s Heart Disease:1

INSIGHTS USING TABLEAU DASHBOARD

The following dashboard was created in Tableau considering the factors which would lead an individual to have a heart disease.

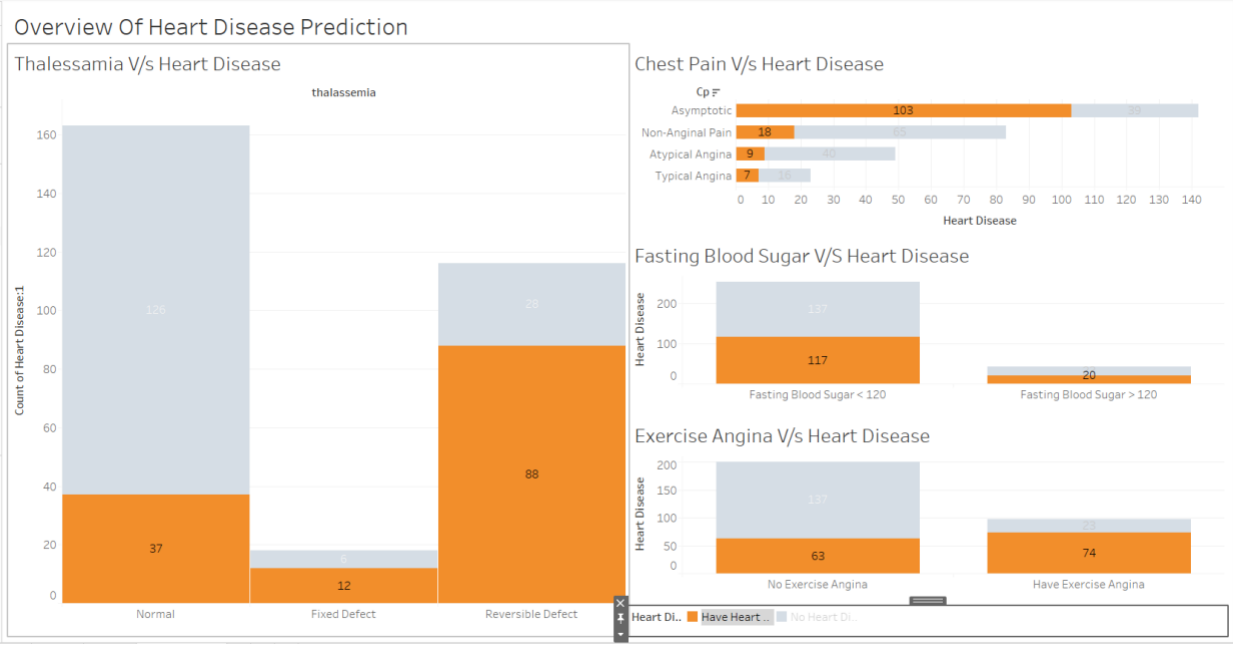


From the above dashboard, the factors which would lead the person having a heart disease are:

1. Thalessamia
2. Chest Pain
3. Fasting Blood Sugar
4. Exercise Angina

INSIGHTS FROM DASHBOARD IN TABLEAU:

Factors contributing for a person to have a heart disease:



From the above dashboard, we can infer that

Thalessamia V/S Heart Disease

1. 37 people with normal thalessamia had a heart disease.
2. 12 people with Fixed defect of thalessamia suffered from the heart disease
3. 88 people with reversible defect of thalessamia suffered from the heart disease

Therefore, people with Reversible Defect of Thalessamia are more prone to having a heart disease.

Chest Pain V/S Heart Disease

1. 103 people who had Asymptotic Chest Pain suffered from heart disease.
2. 18 people who had Non-Anginal Chest Pain suffered from heart disease.
3. 9 people who had Atypical Anginal Chest Pain suffered from heart disease.
4. 7 people who had Typical Anginal Chest Pain suffered from heart disease.

Overall, majority of the people with Asymptotic Chest Pain suffered from hear disease.

Fasting Sugar V/S Heart Disease

1. 117 people with Fasting Blood sugar less than 120 had heart disease
2. 20 people who had Fasting Blood Sugar more than 120 had heart disease.

Hence, people having fasting blood sugar less than 120 suffered from heart disease.

Exercise Angina V/S Heart Disease:

1. 63 people who had no Exercise Angina suffered from Heart Disease
2. 74 people who had Exercise Angina suffered from Heart Disease.

Finally, people with Exercise Angina are more prone to have a heart disease.

To conclude,

People with Reversible Defect Thalessamia, Asymptotic Chest pain, Fasting blood sugar less than 120 and having Exercise Angina are likely to have a heart disaease.

FITTING LOGISTIC REGRESSION MODEL WITH VARIABLES:

We have fitted a Logistic Regression model here since there are two predicting variables and one binary outcome variable. This model will help us determine the effect that a max heart rate, age, exercise angina and sex etc. can have on the likelihood that an individual will have a heart disease.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
Intercept	b1	b2	b3	b4	b5	b6	b7	b8	b9	b10	b11	b12	b13		cut-off	0.5
-7.365809669	-0.014171833	1.311664278	0.575664657	0.024036723	0.004993112	-1.022264933	0.245155621	-0.020683635	0.925926233	0.247236939	0.569568154	1.267552553	0.344063725			

heart disease:1	age	gender	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	Logit	Odds	Probability of Heart Failure	Pro
0	63	1	1	145	233	1	2	150	0	2.3	3	0	6	-1	0.362	0.265933514	
1	67	1	4	160	286	0	2	108	1	1.5	2	3	3	6.1	445.9	0.99776249	
1	67	1	4	120	229	0	2	129	1	2.6	2	2	7	4.8	121.6	0.991840522	
0	37	1	3	130	250	0	0	187	0	3.5	3	0	3	-0.7	0.477	0.322988426	
0	41	0	2	130	204	0	2	172	0	1.4	1	0	3	-3.8	0.023	0.022498758	
0	56	1	2	120	236	0	0	178	0	0.8	1	0	3	-3.5	0.03	0.028866127	

Logistic Model:

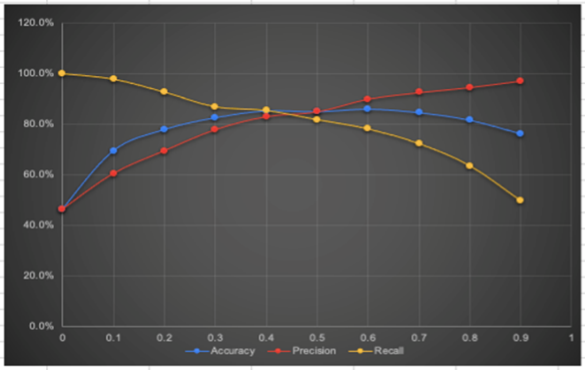
(7.3658+(0.0142)*age+(1.3117)*sex+(0.5757)*cp+(0.0240)trestbps+(0.0050)*chol+(1.0223)*fbs+(0.2452)*restecg +(-0.0207)*thalach+(0.9259)*exang+(0.2472)*oldpeak+(0.5696)*slope +(1.2676)*ca+(0.3441)*thal).

Accuracy Matrix:

Cutoff	Accuracy	Precision	Recall
	84.8%	84.8%	81.8%
0	0.461279461	0.461279461	1
0.1	0.693602694	0.603603604	0.97810219
0.2	0.777777778	0.693989071	0.927007299
0.3	0.824915825	0.777777778	0.868613139
0.4	0.851851852	0.829787234	0.854014599
0.5	0.848484848	0.848484848	0.817518248
0.6	0.858585859	0.899159664	0.781021898
0.7	0.845117845	0.925233645	0.722627737
0.8	0.814814815	0.945652174	0.635036496
0.9	0.760942761	0.971428571	0.496350365

Formula Bar

Cutoff	FP	TP
	12.5%	81.8%
0	1	1
0.1	0.55	0.97810219
0.2	0.35	0.927007299
0.3	0.2125	0.868613139
0.4	0.15	0.854014599
0.5	0.125	0.817518248
0.6	0.075	0.781021898
0.7	0.05	0.722627737
0.8	0.03125	0.635036496
0.9	0.0125	0.496350365



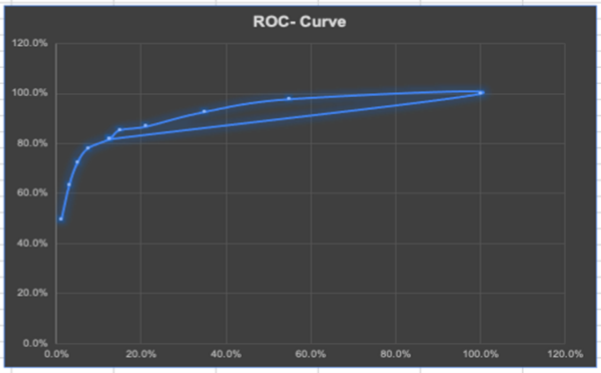


Chart outputs and ROC curve

			Predicted	Predicted	
			0	1	
Actual		0	140	20	160
Actual		1	25	112	137
	Total		165	132	297

Accuracy	84.8%
Precision	84.8%
Recall	81.8%
TP Rate	81.8%
FP Rate	12.5%

EXAMPLE:

Let us take an example of a 45-year-old female with a max heart rate of 150 and no chest pain:

Plugging this into the model we get the value= (-7.3658+(0.0142) *45+(1.3117) *0+(0.5757) *0+(0.024) *0+(0.005)* 0+(1.0223)* 0+(0.2452)* 0+(-0.0207)* 150+(0.9259)* 0+(0.2472)* 0+(0.5696)* 0+(1.2676)* 0+(0.3441)* 0)

=**0.177** which means very unlikely chances of heart failure.

INFERENCES:

- There is a direct relation between a chance of heart failure with parameters such as age, sex, different kinds of chest pains and your heart rate when you exercise or run.

- Individuals who have experienced chest pains of any kind should constantly get their vitals and bodily functions checked as they get older because they are very likely to have a heart failure.

- Hospitals can have a digital system to monitor and send out reminders to keep in check of patients who are at risk if predicted with high risk of heart failure, this will give doctors enough time to prevent or cure a heart disease saving lives. They can be prescribed with better diets and exercise regimen or surgeries if extreme in cases.

- We have also seen a higher no. of men having heart disease or issues as opposed to women so there should be many more awareness campaigns for the same in clinics and hospitals.

REFERENCES

- CDC Homepage:
  https://www.cdc.gov/heartdisease/heart_failure.html
- Kaggle:
  https://www.kaggle.com/dileep070/heart-disease-prediction-using-logistic-regression

THANK YOU!