Predicting factors affecting presence or absence of Heart Disease and determining the gender at risk.

Guided by
Professor
Meeta Pradhan

T.A
Hardik Patel

GROUP 3

Monica Yamsani

Rasagna Vallabh

Yash Raj

Sai Sravanthi Kilari

IUPUI

INTRODUCTION

- Cardiovascular diseases are one of the leading cause of death all globally. About one third of death occur due to this diseases worldwide.
- Heart diseases refers to any condition that affects structure and function of heart.
- In most of heart diseases, a plaque builds inside the arteries thereby narrowing the blood vessel. This leads to blockage of blood flow.
- According to WHO, it is estimated that 17.9 million people died because of cardiovascular diseases in 2019, which is 32% of all global deaths.

AIM

AIM 1

The aim of our study is to determine the factors affecting occurrence of heart disease in the Cleveland population.

AIM 2

Which gender is highly affected by heart diseases in the Cleveland population.

OBJECTIVE

- The objective of the study is to identify the which factor has maximum influence in the occurrence of heart disease
- Which gender is at a higher risk of developing heart disease and by what percentage.

HYPOTHESIS

Null Hypothesis (H0): There is no association between the factors and occurrence of heart disease.

Alternative Hypothesis(H1): There is association between factors and occurrence of heart disease.

DATASET

- The dataset we used for our project is Cleveland Heart disease taken from the UCI repository.
- Cleveland heart disease data includes 303 individual's data of age > 29 years.
- Data set includes 14 columns age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, num.

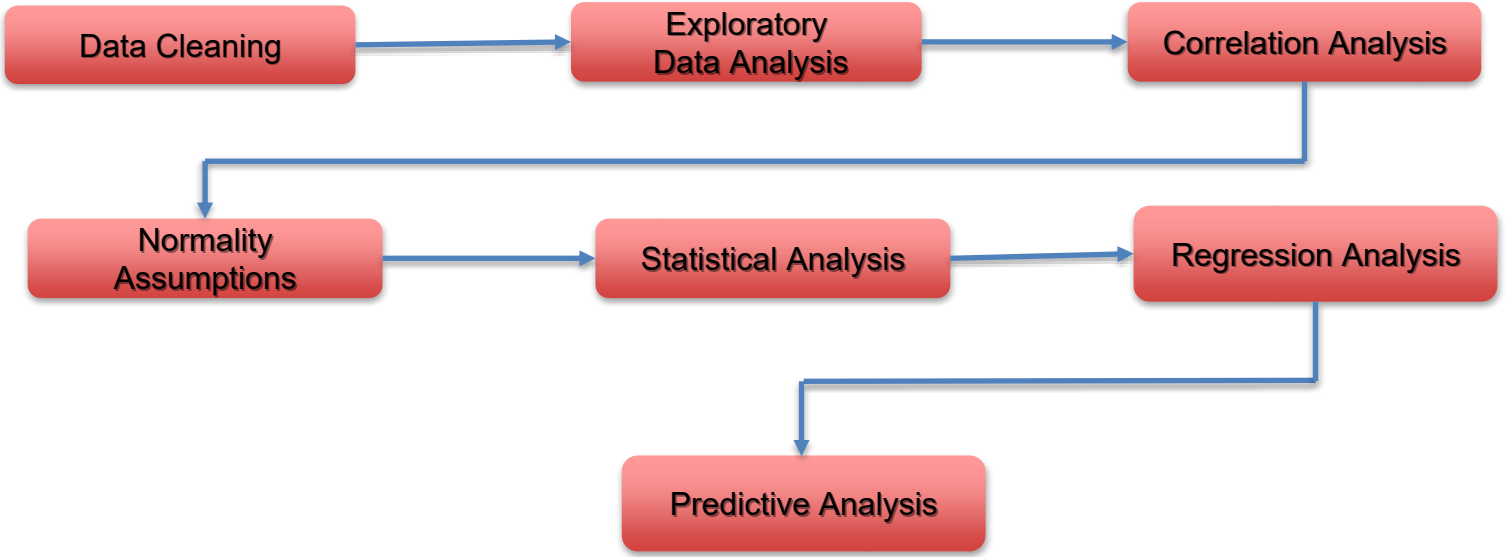
Link:

Index of /ml/machine-learning-databases/heart-disease (uci.edu)

```
##Attribute information:##
###age - age in years

##sex - (1 = male; 0 = female)

###cp - chest pain type
 ### 0: Typical angina: chest pain related decrease blood       supply to the heart
  ###1: Atypical angina: chest pain not related to heart
  ###2: Non-anginal pain: typically esophageal spasms (non      heart related)
  ###3: Asymptomatic: chest pain not showing signs of           disease

###trestbps - resting blood pressure (in mm Hg on admission to the hospital)
### anything above 130-140 is typically cause for concern

###chol - serum cholestoral in mg/dl
###serum = LDL + HDL + .2 * triglycerides
###above 200 is cause for concern

###fbs - (fasting blood sugar > 120 mg/dl)
       ###(1 = true; 0 = false)
       ###'>126' mg/dL signals diabetes

###restecg - resting electrocardiographic results
 ### 0: Nothing to note
  ###1: ST-T Wave abnormality
  ###can range from mild symptoms to severe problems
  ###signals non-normal heart beat
 ###2: Possible or definite left ventricular hypertrophy
 ###Enlarged heart's main pumping chamber

###thalach - maximum heart rate achieved

###exang - exercise induced angina (1 = yes; 0 = no)
```

```
##exang - exercise induced angina (1 = yes; 0 = no)

##oldpeak - ST depression induced by exercise relative to    rest
 ###looks at stress of heart during excercise
 ###unhealthy heart will stress more

##slope - the slope of the peak exercise ST segment
  ###0: Upsloping: better heart rate with excercise            (uncommon)
  ###1: Flatsloping: minimal change (typical healthy heart)
  ###2: Downslopins: signs of unhealthy heart

##ca - number of major vessels (0-3) colored by flourosopy
###colored vessel means the doctor can see the blood passing  through the more blood movement the better (no clots)

##thal - thalium stress result
  ###1,3: normal
  ###6: fixed defect: used to be defect but ok now
  ###7: reversable defect: no proper blood movement when       excercising

##target - have disease or not (1=yes, 0=no) (= the predicted attribute)
```

METHODOLOGY



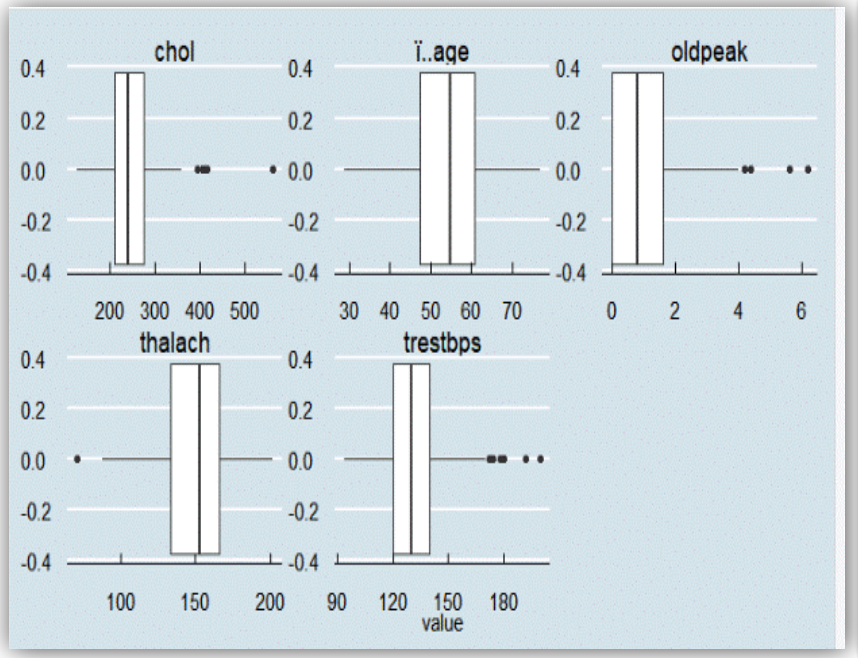
DATA CLEANING

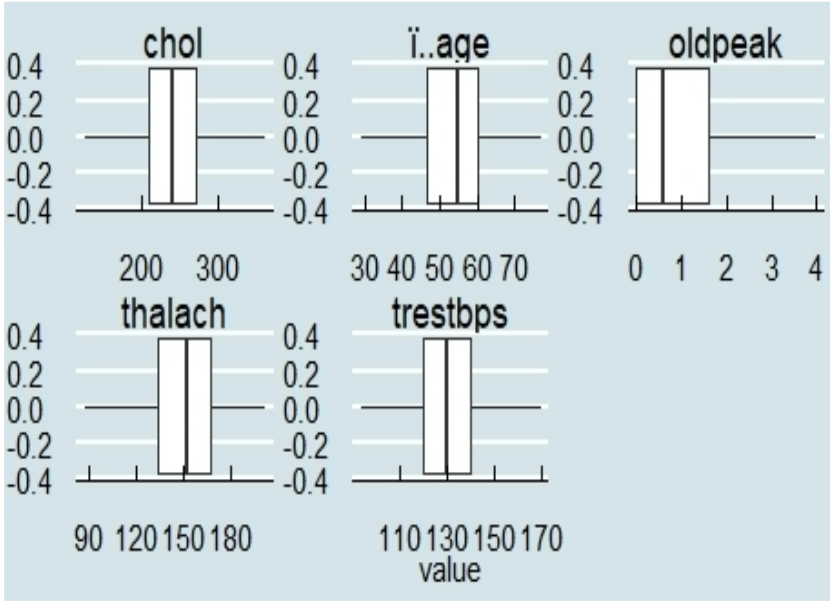
- We assigned labels to categorical variables

```
#sex column
cleve_data$sex[cleve_data$sex == 0] = "female"
cleve_data$sex[cleve_data$sex == 1] = "male"

#cp
cleve_data$cp[cleve_data$cp==0] = "typical angina"
cleve_data$cp[cleve_data$cp==1] = "atypical angina"
cleve_data$cp[cleve_data$cp==2] = "non-anginal pain"
cleve_data$cp[cleve_data$cp==3] = "asymptotic"

#fbs
cleve_data$fbs[cleve_data$fbs==0] = "false"
cleve_data$fbs[cleve_data$fbs==1] = "true"

#exang
cleve_data$exang[cleve_data$exang == 1] = "yes"
cleve_data$exang[cleve_data$exang == 0] = "no"

#restecg
cleve_data$restecg[cleve_data$restecg==0] = "Nothing to note"
cleve_data$restecg[cleve_data$restecg==1] = "ST-T Wave abnormality"
cleve_data$restecg[cleve_data$restecg==2] = " Definite left ventricular hypertrophy"

#slope
cleve_data$slope[cleve_data$slope == 0] = "upsloping"
cleve_data$slope[cleve_data$slope == 1] = "flat"
cleve_data$slope[cleve_data$slope == 2] = "downsloping"

#thal
cleve_data$thal[cleve_data$thal == 1] = "normal"
cleve_data$thal[cleve_data$thal == 2] = "fixed defect"
cleve_data$thal[cleve_data$thal == 3] = "reversible defect"
```

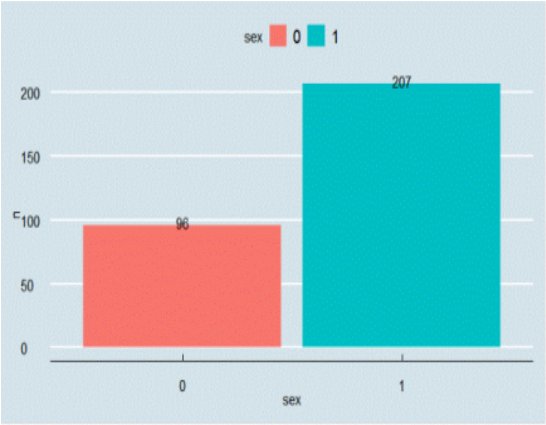
- Removed the outliers

EXPLORATORY DATA ANALYSIS



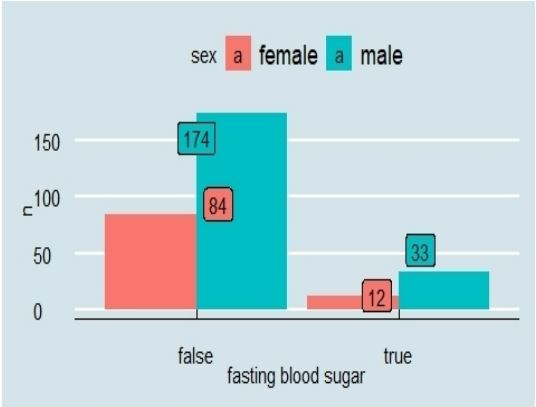


Box Plots of all variables

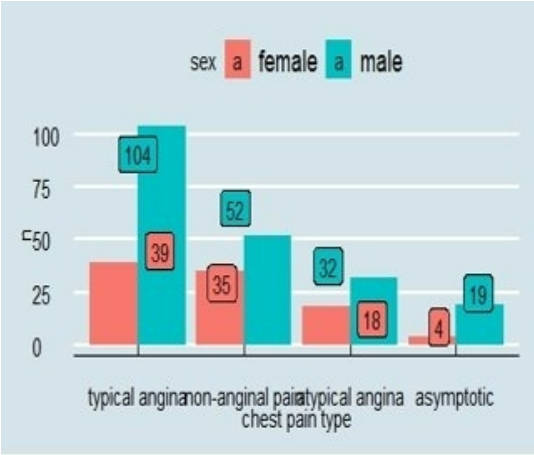


Total males and females

Fbs distribution in males and females

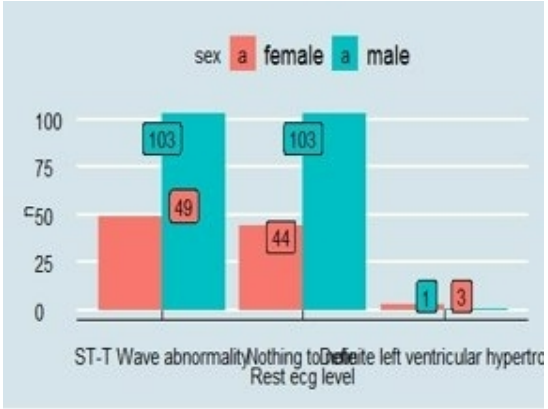


Bar plots of categorical variables



Types of chest pain among males and females

ECG slope among males and females

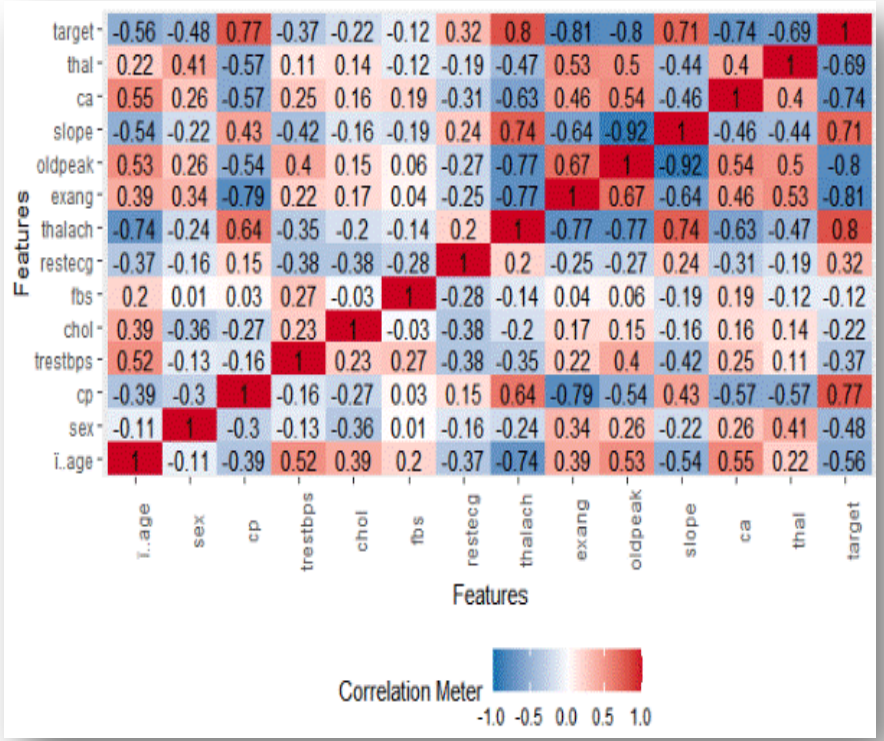


Data distribution

Characteristic	female, N = 85[7]	male, N = 199[7]
ï..age	54.94 (9.65)	53.59 (8.94)
cp		
asymptotic	4 / 85 (4.7%)	18 / 199 (9.0%)
atypical angina	18 / 85 (21%)	31 / 199 (16%)
non-anginal pain	33 / 85 (39%)	50 / 199 (25%)
typical angina	30 / 85 (35%)	100 / 199 (50%)
trestbps	129.71 (15.60)	130.00 (15.31)
chol	250.74 (48.39)	238.38 (42.70)
fbs		
false	76 / 85 (89%)	168 / 199 (84%)
true	9 / 85 (11%)	31 / 199 (16%)
restecg		
Definite left ventricular hypertrophy	2 / 85 (2.4%)	0 / 199 (0%)
Nothing to note	37 / 85 (44%)	100 / 199 (50%)
ST-T Wave abnormality	46 / 85 (54%)	99 / 199 (50%)
thalach	151.45 (20.76)	149.44 (23.49)
exang	16 / 85 (19%)	74 / 199 (37%)
oldpeak	0.71 (0.84)	1.05 (1.08)
slope		
downsloping	44 / 85 (52%)	94 / 199 (47%)
flat	38 / 85 (45%)	92 / 199 (46%)
upsloping	3 / 85 (3.5%)	13 / 199 (6.5%)
ca		
0	59 / 85 (69%)	106 / 199 (53%)
1	14 / 85 (16%)	49 / 199 (25%)
2	10 / 85 (12%)	25 / 199 (13%)
3	2 / 85 (2.4%)	14 / 199 (7.0%)
thal		
0	1 / 85 (1.2%)	1 / 199 (0.5%)
fixed defect	74 / 85 (87%)	86 / 199 (43%)
normal	1 / 85 (1.2%)	16 / 199 (8.0%)
reversible defect	9 / 85 (11%)	96 / 199 (48%)
heart disease		
0	17 / 85 (20%)	108 / 199 (54%)
1	68 / 85 (80%)	91 / 199 (46%)

Means of each variable among males and females

CORRELATION ANALYSIS



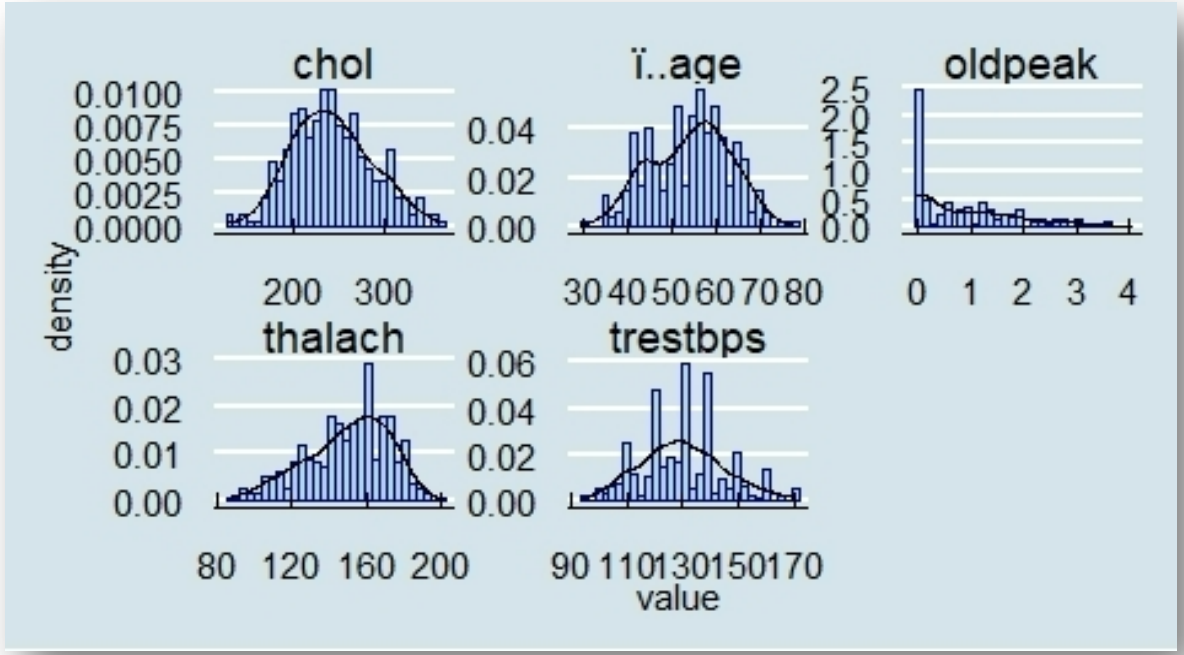
- Target is our dependent variable.
- Highly positive correlated variables are cp (0.77), thalach (0.8), slope (0.71)
- Highly negative correlated variables are exang (-0.81), oldpeak (-0.8), ca (-0.74) and thal (-0.69)

NORMALITY ASSUMPTION

Shapiro Wilk Test

- All the variable p values are < 0.05 which indicates that the data did not pass the normalcy assumptions.
- Based on normality assumptions, we can say that the Shapiro tests are not statistically significant.
- Therefore, there is an evidence for rejecting null hypothesis (i.e., data is normally distributed).

Variables	P values
chol	5.365 e-19
age	0.05
ca	< 2.2 e-16
thal	< 2.2 e-16
sex	< 2.2 e-16
cp	< 2.2 e-16
thalach	< 3.76 e-17
oldpeak	<2.2 e-16
Slope	<2.2 e -16
exang	<2.2 e-16
trestbps	< 1.458e -06
restecg	< 2.2e-16
target	<2.2 e-16

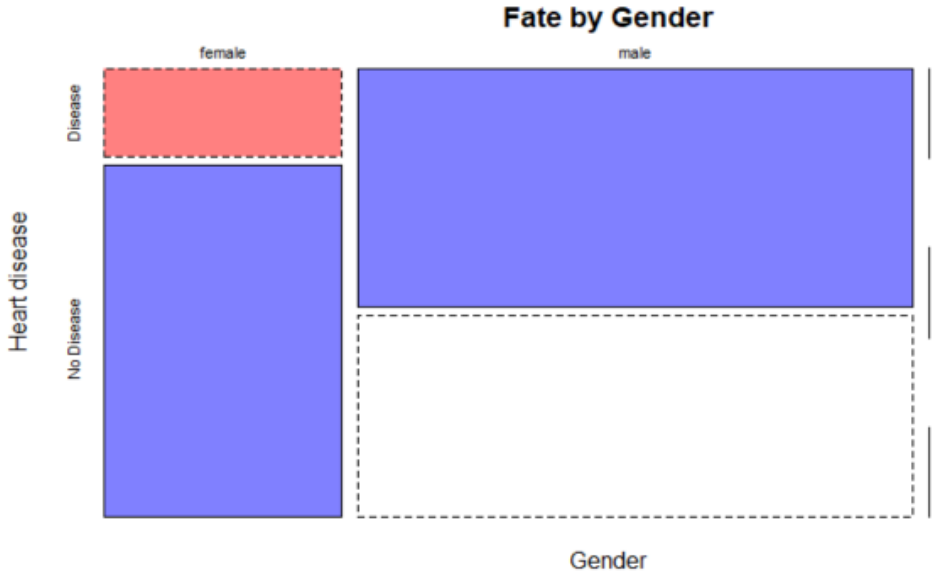


Histograms depicting the data distribution

STATISTICAL ANALYSIS

Chi square test

- We performed chi square to determine the relationship between the target and sex.
- p value is < 0.05 which indicates that there is statistical association between sex and target .
- Therefore, we reject the null hypothesis (i.e., there is no association)



##Chi square test don ebetween the target and sex to determine thier assocoiation

```
chisq<-chisq.test(cleve_data$sex,cleve_data$target)
chisq
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  cleve_data$sex and cleve_data$target
## X-squared = 27.015, df = 1, p-value = 2.019e-07
```

ODDS RATIO AND RISK RATIO

```
##
##      0   1
##   0  24  72
##   1 114  93
```

```
epi.2by2(tab,method="cohort.count")
```

```
##             Outcome +    Outcome -      Total        Inc risk *        Odds
## Exposed +          24           72         96              25.0       0.333
## Exposed -         114           93        207              55.1       1.226
## Total             138          165        303              45.5       0.836
##
## Point estimates and 95% CIs:
## -------------------------------------------------------------------
## Inc risk ratio                                 0.45 (0.31, 0.66)
## Odds ratio                                     0.27 (0.16, 0.47)
## Attrib risk in the exposed *                   -30.07 (-41.07, -19.07)
## Attrib fraction in the exposed (%)            -120.29 (-218.18, -52.52)
## Attrib risk in the population *                -9.53 (-18.32, -0.73)
## Attrib fraction in the population (%)         -20.92 (-30.04, -12.44)
## -------------------------------------------------------------------
## Uncorrected chi2 test that OR = 1: chi2(1) = 23.914 Pr>chi2 = <0.001
## Fisher exact test that OR = 1: Pr>chi2 = <0.001
##  Wald confidence limits
##  CI: confidence interval
##  * Outcomes per 100 population units
```

The odds of male having heart disease is 73% more than females

Males are 55% at more risk of developing heart disease

Kruskal Wallis Rank Sum Test

- All the variables p values are < 0.05 indicating that the not all group medians are equal.

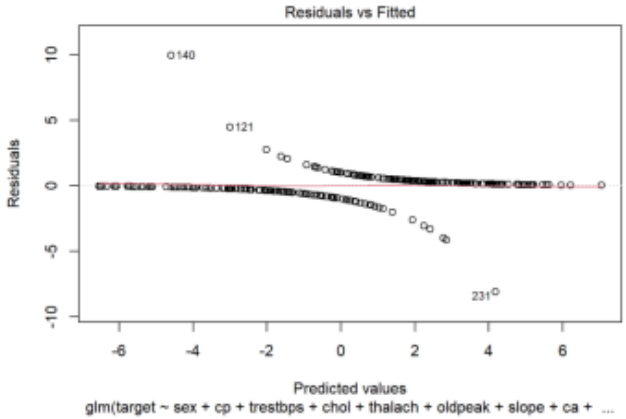
- Therefore, there is enough evidence to reject the null hypothesis

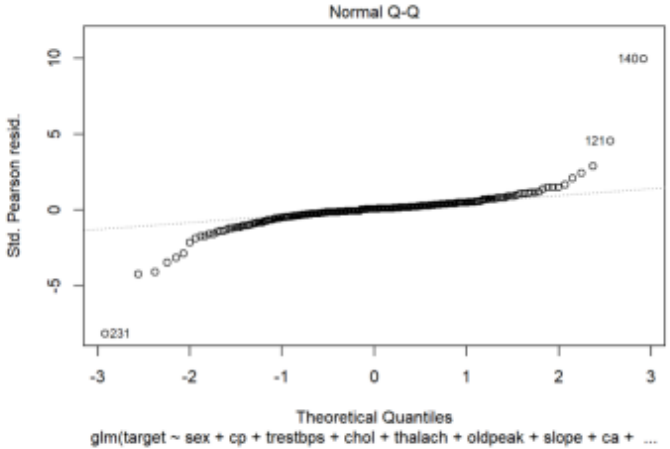
Post Hoc Test

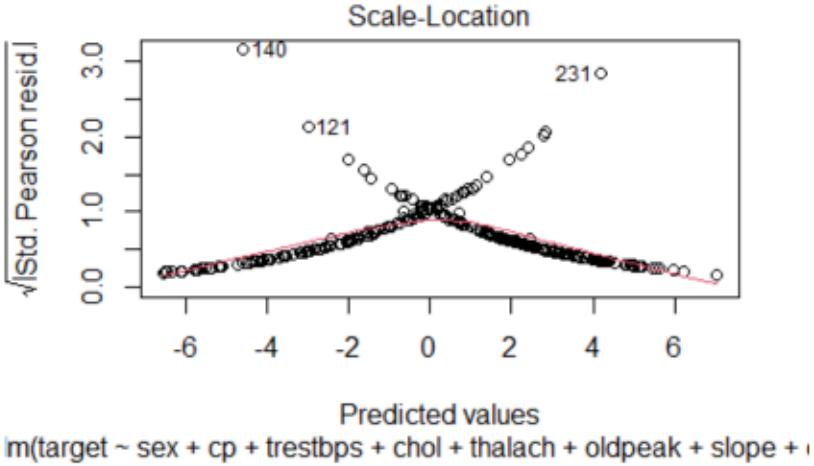
- We performed Dunn- Kruskal Wallis pair comparison test to determine if there are any differences within the groups and p values adjusted with the Benjamini-Hochberg method for two groups.

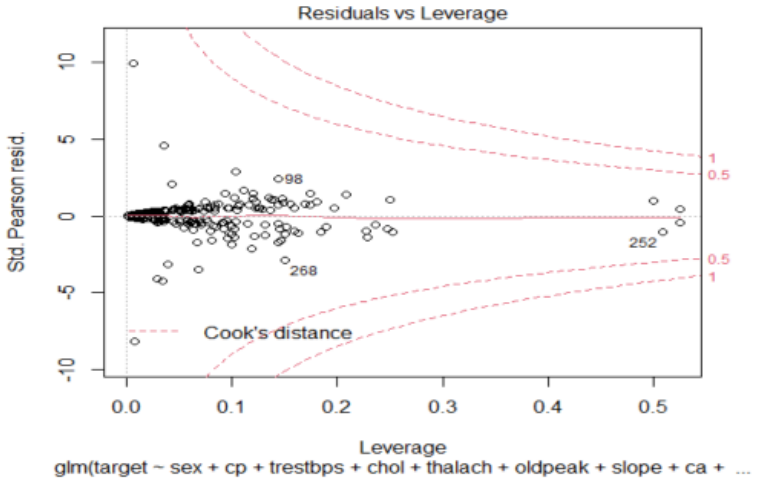
Variables	P values
age	3.429 e-05
ca	1.83 e -15
cp	1.157 e-15
thal	2.407 e-12
oldpeak	2.395 e-13
trestbps	0.0346
chol	0.03566
sex	1.049 e-06
restecg	0.009806
slope	1.08 e -10
exang	3.198 e-14
thalach	9.748 e- 14

Logistic regression
With step wise AIC backward regression






Scale-Location
√|Std. Pearson resid.|
140
231
121
0.0
1.0
2.0
3.0
-6
-4
-2
0
2
4
6
Predicted values
m(target ~ sex + cp + trestbps + chol + thalach + oldpeak + slope +


Residuals vs Leverage
Std. Pearson resid.
10
5
0
-5
-10
98
268
252
Cook's distance
1
0.5
0.5
1
0.0
0.1
0.2
0.3
0.4
0.5
Leverage
glm(target ~ sex + cp + trestbps + chol + thalach + oldpeak + slope + ca + ...

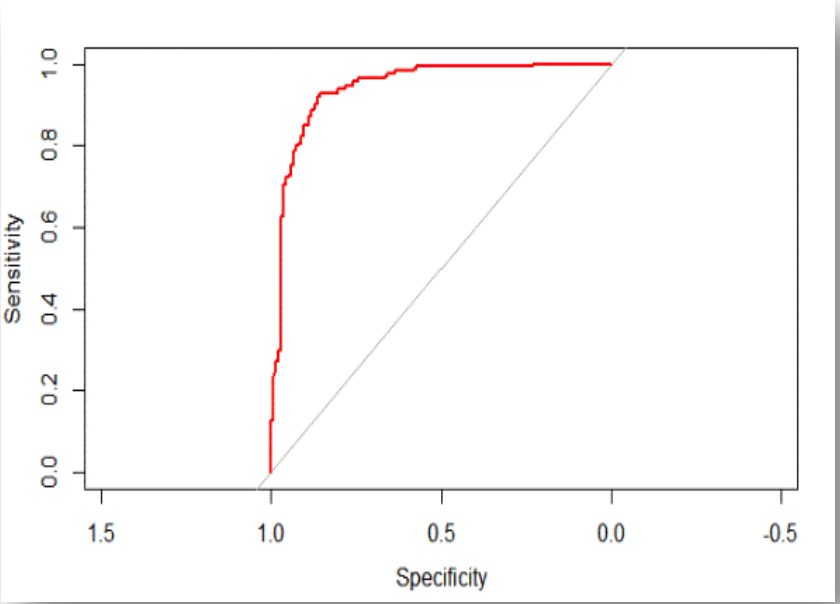
RESULTS

Logistic regression

Characteristic	N	log(OR)[1]	95% CI[1]	p-value
ï..age	303	0.03	-0.02, 0.08	0.3
sex	303			<0.001
0		—	—	
1		-1.9	-3.0, -0.78	
cp	303			<0.001
0		—	—	
1		0.86	-0.24, 2.0	
2		2.0	1.0, 3.1	
3		2.4	1.1, 3.9	
trestbps	303	-0.03	-0.05, 0.00	0.025
chol	303	0.00	-0.01, 0.00	0.3
fbs	303			0.4
0		—	—	
1		0.45	-0.69, 1.6	
restecg	303			0.5
0		—	—	
1		0.46	-0.32, 1.3	
2		-0.71	-5.5, 3.4	
thalach	303	0.02	0.00, 0.04	0.083
exang	303			0.085
0		—	—	
1		-0.78	-1.7, 0.11	
oldpeak	303	-0.40	-0.89, 0.07	0.094
slope	303			0.009
0		—	—	
1		-0.78	-2.6, 0.92	
2		0.69	-1.2, 2.5	
ca	303			<0.001
0		—	—	
1		-2.3	-3.4, -1.3	
2		-3.5	-5.2, -2.0	
3		-2.2	-4.3, -0.55	
4		1.3	-2.0, 5.0	
thal	303			0.003
0		—	—	
1		2.6	-2.0, 7.6	
2		2.4	-2.2, 7.1	
3		0.92	-3.7, 5.6	

[1] OR = Odds Ratio, CI = Confidence Interval

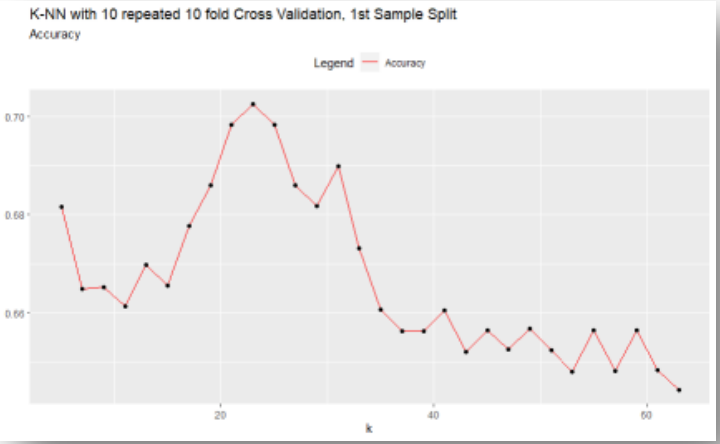
ROC curve



- Logistic regression model achieved a very high AUC score of 0.9401.
- That means our model was able to correctly predict 0 classes as 0 and 1 classes as 1.
- The Higher the AUC, the better the model is at distinguishing between patients with the disease and no disease.

KNN MODEL

k	accuracy	kappa
5	0.6743333	0.3402848
7	0.6743333	0.3402099
9	0.6618333	0.3159125



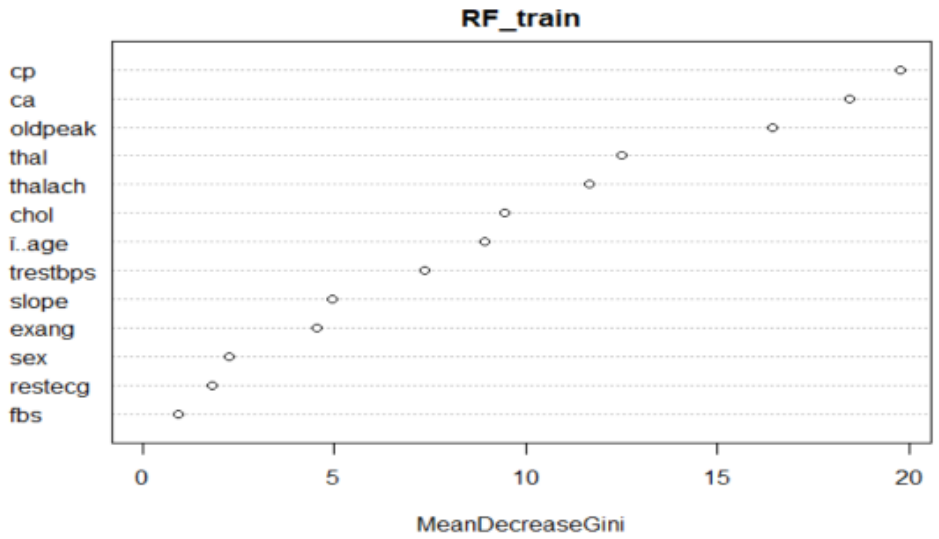
- When K-NN with 10 repeated 10-fold Cross Validation was performed
- The accuracy reached to 0.703 at k = 23
- Hence, the k=23 is the optimal value of k for the model

RANDOM FOREST MODEL

```
          Reference
Prediction  1  0
         1 27  5
         0  6 23

               Accuracy : 0.8197
                 95% CI : (0.7002, 0.9064)
    No Information Rate : 0.541
    P-Value [Acc > NIR] : 4.82e-06

                  Kappa : 0.6379

 Mcnemar's Test P-Value : 1
```

The model is 81% Accurate

Variable importance from Random Forest



	Logistic regression	KNN	Random Forest
Accuracy	0.94	0.67(K=5) 0.70(K=23)	0.803
Kappa	-	0.33	0.6

SUMMARY

	Shapiro Wilk test	Kruskal Wallis Test	Chi Square	Logistic Regression
Significance value	0.05	0.05	0.05	0.05
P value	<0.05 for all the variables	<0.05 for all the variables	2.091e	<0.05

Conclusion

- Significant association was found among the variables and the target (which is the presence or absence of heart disease
- The variable with highest association is chest pain obtained from random forest
- Chi square results indicate that males are at a higher risk than females
- Among the machine learning models logistic regression gave the highest accuracy value of 0.94.

Discussion

- Most of the variables of the data were categorical in nature.
- With correlation analysis, the highest positive correlation was found among thalach but random forest variable importance has shown that chest pain had the highest influence in presence of heart disease.
- Among all the machine learning models the logistic regression has the highest accuracy of 0.94 and KNN gave the least (0.67 with K=5)

Critique

- The data did not have other important variables that could result in occurrence of heart disease such as BMI underlying comorbid conditions, diet, adverse habits.
- We have used only Cleveland data set. However similar data was available for California, Switzerland, South Africa and Hungary.

References

- Detrano, R., Yiannikas, J., Salcedo, E., Rincon, G., Go, R., Williams, G., & Leatherman, J. (1984). Bayesian probability analysis: a prospective demonstration of its clinical utility in diagnosing coronary disease. *Circulation*, *69*(3), 541-547. doi: 10.1161/01.cir.69.3.541

- Cardiovascular diseases (CVDs). (2021). Retrieved 9 December 2021, from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

- gtsummary: Presentation-Ready Data Summary and Analytic Result Tables (r-project.org)

- Heart disease dataset (r-project.org)