Northeastern University

Final Project

Final Report

Northeastern University - Toronto
ALY6015: Intermediate Analytics
Prof. Yvonne Leung

Adeyemi Adebimpe
Srinivas Shanmuga Ganesh Krishnaswamy
Utsav Desai

Introduction

The purpose of this project analysis is to conduct a study and investigate the prevalence of heart disease in individuals of different age groups. We are trying to answer the following questions:

1. What percentage of people in various age groups have heart disease?
2. Is there any association between the individual's gender and the different preliminary testing parameters? In other words, is the results of maximum heart rate, cholesterol, and resting blood pressure independent of the people's gender?
3. An individual's heartbeat count is measured to study the heart's functioning capacity. Does the age and gender of a person significantly impact the increase or decrease of the heart rate?
4. What are the various factors that influence the development of heart disease in humans? Can such conditions be predicted using a robust model?

Background:

According to World Health Organization (W.H.O), unhealthy diet, physical inactivity, tobacco use, and harmful use of alcohol increase the risk of heart disease. These behaviors may result in high blood pressure, raised blood lipids, overweight, and obesity. Factors like poverty, stress, and hereditary issues are also determinants of cardiovascular diseases. Statistics from the Canadian Government states that heart disease is the second leading cause of death in Canada. The Public Health Agency of Canada's Canadian Chronic Disease Surveillance System (CCDSS) data in 2012/13 reveal that 1 in 12 Canadian adults aged 20 and above live with a diagnosed heart condition. It also states that the death rate accrued to heart conditions is higher in adults aged 40 and above.

Through this project analysis, we are interested to gain a better understanding of the factors that influence the development of the disease in an individual and making accurate predictions. With a robust prediction model, we should be able to identify the types of chest pain, gender factors, level of blood pressure or cholesterol level, heart rate, or age associated with the severity of the heart disease. Like, chances of having heart disease for males with heaving high blood pressure and any type of chest pain? Chances of having heart disease for females with high cholesterol and heart rate? Etc.

The heart disease data from Kaggle will be used for this study. The data set contains 14 variables. The variable ‘target’ is the predicted output which serves as the dependent variable while the other 13 factors serve as the independent variables.

Methods:

Testing and feature selections methods such as Two-way ANOVA, and Chi-Square are used to investigate the cause and effect of certain factors and their relationship with the development of heart disease. Stepwise feature selection method is used for model optimization. The target variable is categorical. Thus, we are using the Generalized Linear modeling technique to make predictions.

Analysis

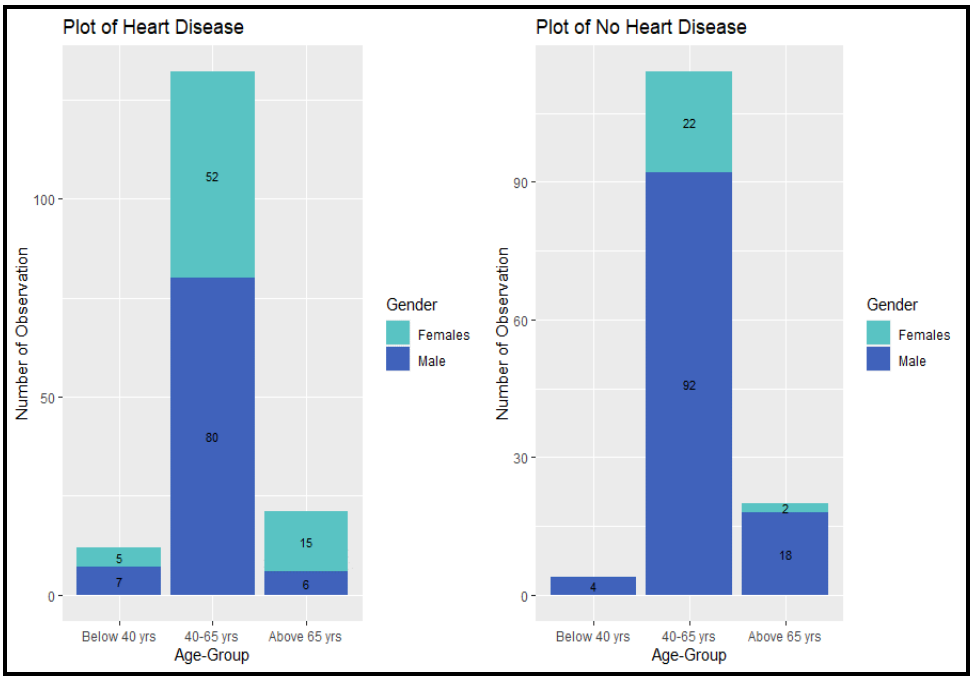


Fig.-1: Age-Group and Gender wise Heart Disease Observation

From the above graph, we can infer the frequency of individuals in various age groups having and not having heart disease. It is evident that people in the 40-65 year age group have a high risk of developing heart disease and that females in this age group are more venerable than males.

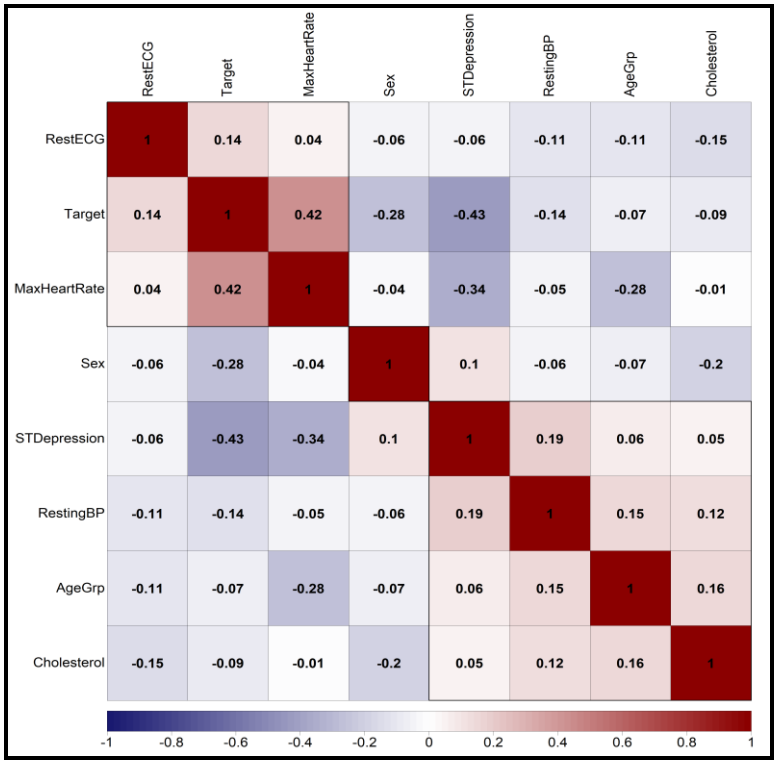


Fig.-2: Correlation Matrix

The above figure is a correlation plot showing the correlation values between the variables. The most related risk factors are used in the analysis. Figure 2 shows correlation values of each variable with each other and the color red indicates that variables are positively related with each other and color blue indicates that variables are negatively related with each other. We, therefore, identified that Maximum Heart Rate has a significant positive where ST Depression has a significant negative relationship with the target variable.

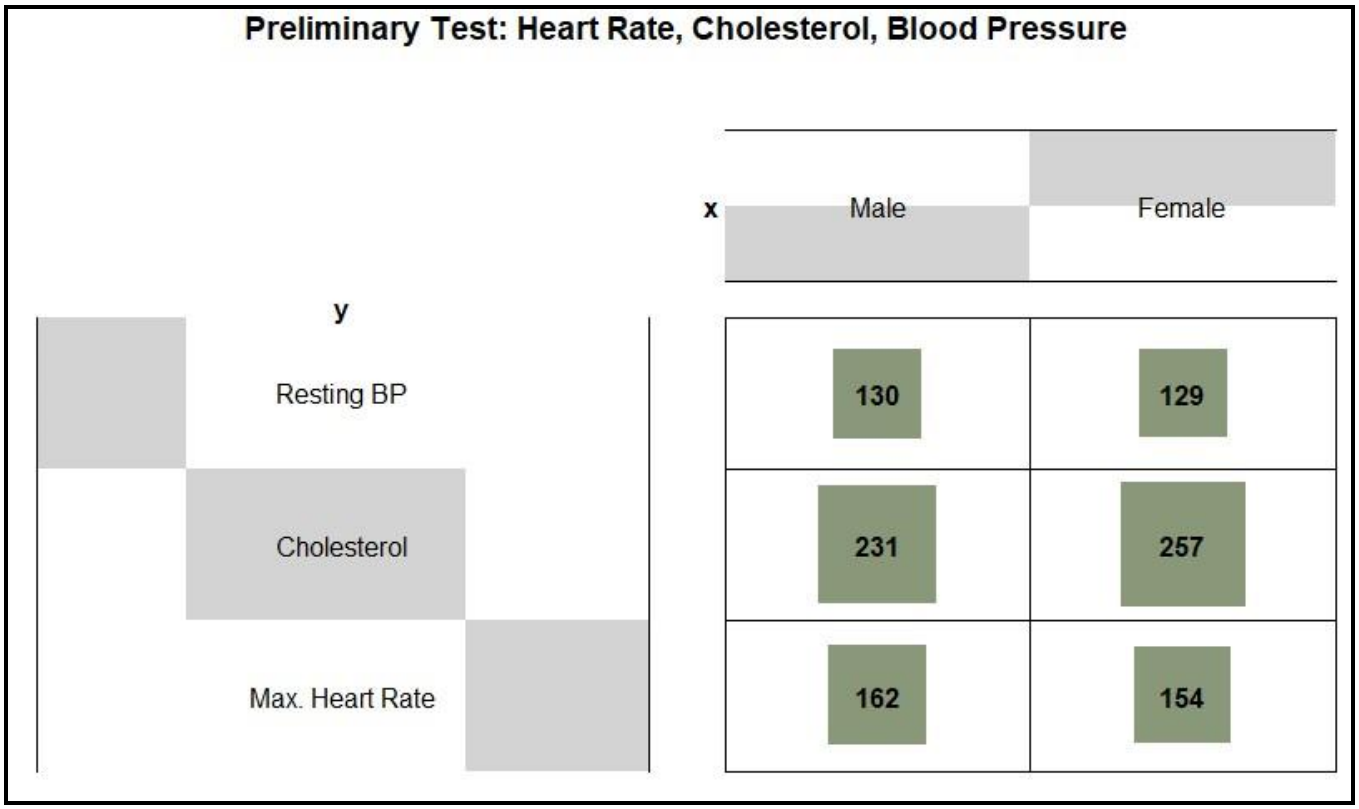


Fig.-3: Gender wise Resting BP, Cholesterol and Max. Heart Rate

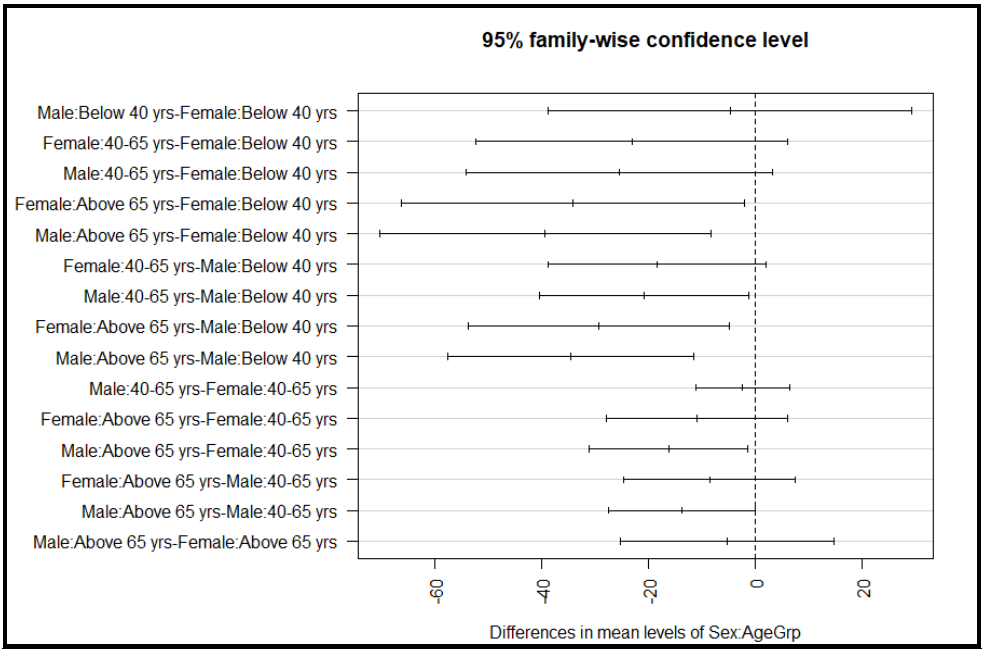
A test of independence is conducted to establish an association between gender and pre-diagnostic parameters such as maximum heart rate, cholesterol, and resting blood pressure. A balloon graph provides better visualization of the relationship between the group.

Chi-Square: Gender association with Heart Rate, Cholesterol, and Blood pressure

	Chi-Squared	Degrees of Freedom	p-value	Critical Value	Method
Values	1.29	2	0.523573807475011	5.991	Pearson's Chi-squared test

Fig.-4: Chi-Square results

The null hypothesis states that there is no relationship between the testing parameters and gender, whereas the alternative hypothesis states that there is a relationship between them. From the results, we infer that the chi-squared value is less than the critical value, as a result, we cannot reject the null hypothesis. To conclude, we don't have enough evidence to claim that there is an association between gender and the preliminary testing parameters such as Maximum Heart rate, Cholesterol, and Resting Blood Pressure.



Tukey Test: Two-way ANOVA: Heart Rate Vs Gender and Age Group

	differences	lower	upper	p-value
Male:Below 40 yrs-Female:Below 40 yrs	-4.727273	-38.81516	29.3606115	0.9987078
Female:40-65 yrs-Female:Below 40 yrs	-23.135135	-52.33858	6.0683097	0.2087372
Male:40-65 yrs-Female:Below 40 yrs	-25.546512	-54.21857	3.1255425	0.1117496
Female:Above 65 yrs-Female:Below 40 yrs	-34.117647	-66.27078	-1.9645155	0.0303108
Male:Above 65 yrs-Female:Below 40 yrs	-39.333333	-70.40251	-8.2641544	0.0044447
Female:40-65 yrs-Male:Below 40 yrs	-18.407862	-38.83082	2.0150964	0.1040649
Male:40-65 yrs-Male:Below 40 yrs	-20.819239	-40.47484	-1.1636356	0.0308016
Female:Above 65 yrs-Male:Below 40 yrs	-29.390374	-53.84605	-4.9347017	0.0084294
Male:Above 65 yrs-Male:Below 40 yrs	-34.606061	-57.61801	-11.5941086	0.0003136
Male:40-65 yrs-Female:40-65 yrs	-2.411376	-11.19773	6.3749810	0.9695475
Female:Above 65 yrs-Female:40-65 yrs	-10.982512	-27.98067	6.0156427	0.4332698
Male:Above 65 yrs-Female:40-65 yrs	-16.198198	-31.04432	-1.3520785	0.0234163
Female:Above 65 yrs-Male:40-65 yrs	-8.571135	-24.63920	7.4969310	0.6451056
Male:Above 65 yrs-Male:40-65 yrs	-13.786822	-27.55827	-0.0153742	0.0495603
Male:Above 65 yrs-Female:Above 65 yrs	-5.215686	-25.25040	14.8190323	0.9758147

Fig.-5: ANOVA Test Statistics

The two-way ANOVA test determines whether age and gender affect an individual's heart rate. Based on the findings, we can conclude that there is enough evidence to claim that both age and gender impact the maximum heart rate of an individual. A Tukey pairwise test is conducted to determine the difference between the groups. The p-value less than the significance level indicates that there is a significant difference between groups. There was no statistically significant difference in maximum heart rate between the male and female groups over the age of 65. However, the male group over the age of 65 did have a statistically significant difference in maximum heart rate compared to the female group under the age of 40.

Logistics Regression: Heart Disease

	Dependent variable:
	Target
SexMale	-1.526***
	(0.456)
AgeGrp40-65 yrs	0.093
	(0.918)
AgeGrpAbove 65 yrs	0.842
	(1.035)
MaxHeartRate	0.045***
	(0.011)
RestECG	0.631*
	(0.365)
STDepression	-0.722***
	(0.204)
Thalassemia	-0.852***
	(0.313)
FastingBSugar1	0.867*
	(0.504)
NumMajorVessels	-0.862***
	(0.208)
Constant	-3.031
	(2.048)
Observations	220
Log Likelihood	-88.079
Akaike Inf. Crit.	196.158
Significance levels	*p<0.1; **p<0.05; ***p<0.01

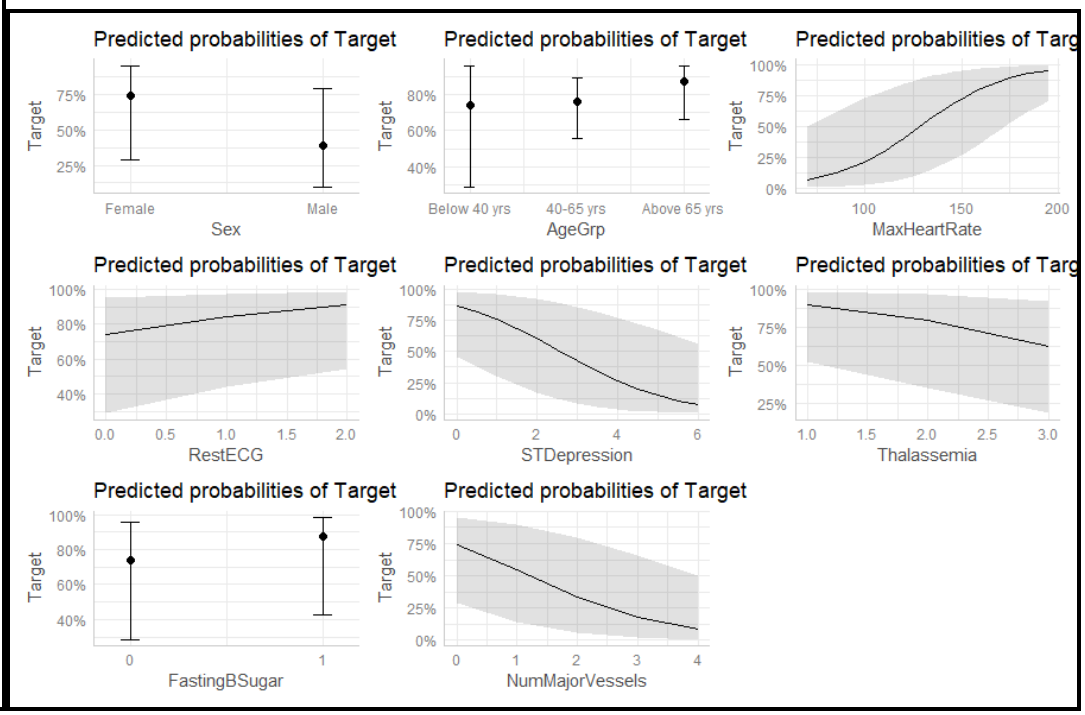
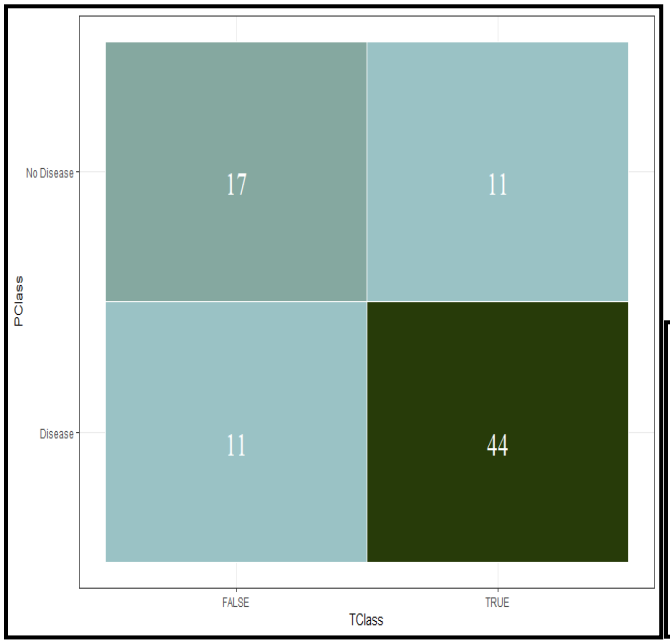


Fig.-6: GLM Statistics and variables Relation with Target variable

A model is developed to predict the prevalence of heart disease in people of a specific age group. The model was optimized using a stepwise feature selection method. The interaction of gender and age resulted in a high degree of collinearity among the independent variables. The interaction was removed from the model to resolve the multi-collinearity. The prevalence of the disease was influenced by factors such as maximum heart rate, gender, number of major vessels (cardiac fluoroscopy), and thalassemia.



Indicator	Training Data	Testing Data
Accuracy	82.72727	73.49398
Sensitivity	84.54545	80.00000
Specificity	80.90909	60.71429
Precision	81.57895	81.57895
Recall	84.54545	80.00000

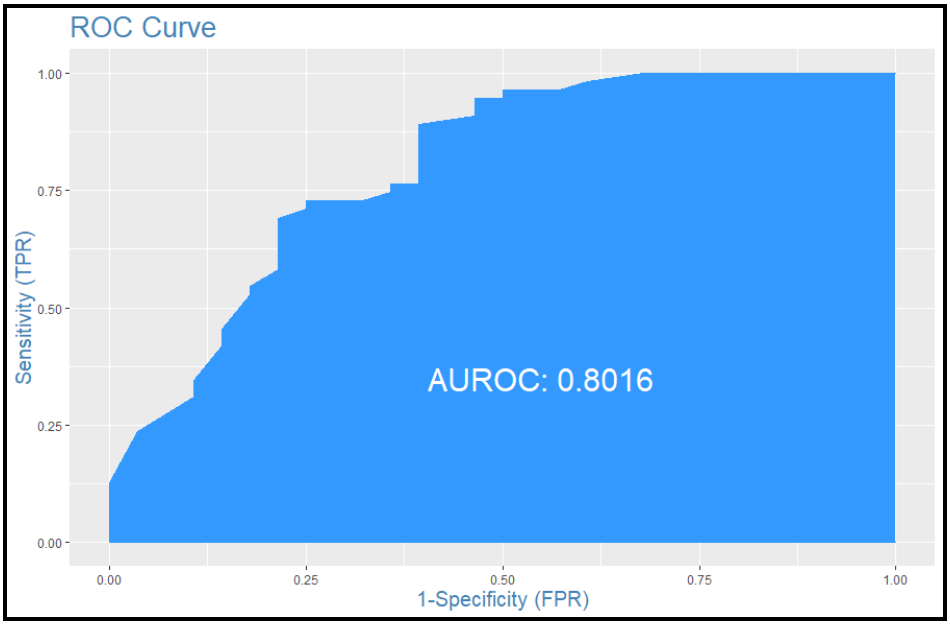


Fig.-7: GLM Confusion Matrix Statistics and ROC Curve

To assess the model's performance, a confusion matrix was created. For the testing sample, the model's sensitivity and specificity decrease. The model's accuracy is 82.7 percent for training data and 73.4 percent for testing samples, indicating that the model performs poorly. The ROC curve shows that the model performs poorly in classifying the data, with 80 percent of the area falling under the curve.

Conclusion/Interpretation

This project examined the prevalence of heart disease in different age groups of both males and females, taking into account various risk factors that may be associated. The project used a variety of analytic methods, including the Two-way ANOVA test, the Chi-square test, StepAIC, and GLM- Logistic regression.

The percentage of heart disease is more in men compared to women. In the Chi-Square test results, we can infer that the maximum heart rate, cholesterol, and resting blood pressure results are independent of a person’s gender. Through ANOVA analysis, we found that both age and gender influence the heart rate in individuals.

The model was subjected to various performance evaluation and validation processes, yielding a predictive model with 73.4% accuracy and 81% precision. The factors such as Resting Blood Pressure, Cholesterol, and interaction of sex and age group has no significance. Although the prediction model performed significantly well for the training sample, it underperforms with the testing sample. As a result, we need to investigate more risk factors that eventually helps in predicting heart disease in human.

By changing one's lifestyle and eating habits, except for Thalassemia (a hereditary disorder), all other risk factors can be rectified and thus, minimize the risk of a heart condition.

References

- Heart Disease UCI | Kaggle
- Table 1 | A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method (hindawi.com)
- https://www.canada.ca/en/public-health/services/publications/diseases-conditions/heart-disease-canada.html
- https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)#:~:text=The%20most%20important%20behavioural%20risk,lipids%2C%20and%20over
- https://www.canada.ca/en/public-health/services/publications/diseases-conditions/heart-disease-canada.html
- https://ieeexplore.ieee.org/abstract/document/8740989

Appendix (R Code)

```
# --- Final Project : Heart Disease --- #

# install.packages(c('gapminder', 'gridExtra', 'data.table', 'gginference', 'yarrr', 'dplyr', 'ggplot2', 'plotly', 'corrplot', 'psych', 'kableExtra'))

# load packages

install.packages('ggstatsplot')

library(ggstatsplot)

library(psych)

library(dplyr)

library(gridExtra)

library(gapminder)

library(yarrr)

library(ggplot2)

#library(plotly)

library(corrplot)
```

```
library(data.table)

library(gginference)

library(kableExtra)

rm(list = ls())

getwd()

setwd("C://Users//91956//Srinivas Shanmuga G//NEU//Q2//ALY 6015 Intermediate Analytics//Assignments//Final Project")

heartDF <- read.csv("heart.csv")

str(heartDF)

summary(heartDF)

psych::describe(heartDF)

dplyr::glimpse(heartDF)

colnames(heartDF) <- c('Age', 'Sex', 'ChestPainType', 'RestingBP', 'Cholesterol', 'FastingBSugar', 'RestECG', 'MaxHeartRate',

                       'ExerciseInducedPain', 'STDepression', 'STSlope', 'NumMajorVessels', 'Thalassemia', 'Target')

rapply(heartDF,function(x)length(unique(x)))

heartDF <- subset(heartDF, select = c('Age', 'Sex', 'RestingBP', 'Cholesterol', 'RestECG', 'FastingBSugar', 'NumMajorVessels', 'MaxHeartRate', 'STDepression','Thalassemia', 'Target'))



heartDF$Sex <- as.factor(heartDF$Sex)

heartDF$Target <- as.factor(heartDF$Target)

heartDF$FastingBSugar <- as.factor(heartDF$FastingBSugar)

heartDF$Thalassemia[which(heartDF$Thalassemia == 0)] <- 1
```

```
heartDF$Thalassemia[which(heartDF$Thalassemia == 0 | heartDF$Thalassemia == 1 )] <-
'normal'

heartDF$Thalassemia[which(heartDF$Thalassemia == 2)] <- 'fixed defect'

heartDF$Thalassemia[which(heartDF$Thalassemia == 3)] <- 'reversable defect'

AgeGrp <- cut(heartDF$Age,

        breaks = c(-Inf,40,65,Inf),

        labels = c('1','2','3'),

        right = FALSE)

heartDF <- data.frame(heartDF, AgeGrp)

colnames(heartDF)

#heartDF <- subset(heartDF, select = c('Target', 'Sex', 'AgeGrp', 'MaxHeartRate', 'STDepression',
'RestingBP', 'Cholesterol', 'RestECG'))

heartDF <- subset(heartDF, select = c('AgeGrp', 'Sex', 'RestingBP', 'Cholesterol', 'RestECG',
'FastingBSugar', 'NumMajorVessels', 'MaxHeartRate', 'STDepression','Thalassemia', 'Target'))

heart_group <- heartDF %>% group_by(Sex, AgeGrp, Target) %>% tally()

heart_disease_group <- subset(heart_group, heart_group$Target == 1)

not_heart_disease_group <- subset(heart_group, heart_group$Target == 0)

levels(heartDF$AgeGrp) <- c('Below 40 yrs','40-65 yrs','Above 65 yrs')

levels(heartDF$Sex) <- c('Female','Male')

plot1 <- ggplot2::ggplot(heart_disease_group,

          aes(x = AgeGrp, y = n, fill = Sex, label = n)) +

  xlab('Age-Group') + ylab('Number of Observation') +

  scale_x_discrete(labels = c('Below 40 yrs', '40-65 yrs', 'Above 65 yrs')) +

  scale_fill_manual('Gender', values = c('#59C3C3', '#4062BB'), labels=c('Males', 'Female')) +
```

```
  ggtitle('Plot of Heart Disease') +
  geom_bar(position='stack', stat='identity') +
  geom_text(size = 3, position = position_stack(vjust = 0.5))
plot2 <- ggplot2::ggplot(not_heart_disease_group,
                aes(x = AgeGrp, y = n, fill = Sex, label = n)) +
  xlab('Age-Group') +
  ylab('Number of Observation') +
  scale_x_discrete(labels = c('Below 40 yrs', '40-65 yrs', 'Above 65 yrs')) +
  scale_fill_manual('Gender', values = c('#59C3C3', '#4062BB'), labels=c('Males', 'Female')) +
  ggtitle('Plot of No Heart Disease') +
  geom_bar(position='stack', stat='identity') +
  geom_text(size = 3, position = position_stack(vjust = 0.5))
grid.arrange(plot1, plot2, ncol= 2)
plot3 <- ggplot2::ggplot(data = heartDF,
                         mapping = aes(x = AgeGrp, y = MaxHeartRate, fill = Target)) +
  xlab('Age-Group') +
  ylab('Maximum Heart Rate') +
  ggtitle('Box-Plot of Heart Disease') +
  scale_x_discrete(labels = c('Below 40 yrs', '40-65 yrs', 'Above 65 yrs')) +
  scale_fill_manual('Heart\nDisease', values = c('#FC766AFF', '#5B84B1FF'), labels=c('Yes', 'No'))
+
  ylim(c(110, 180)) +
  geom_boxplot() +
```

```
  geom_violin(trim = FALSE, alpha = 0.2)

plot4 <- ggplot2::ggplot(data = heartDF,

                         mapping = aes(x = Sex, y = MaxHeartRate, fill = Target)) +

  xlab('Age-Group') +

  ylab('Maximum Heart Rate') +

  ggtitle('Box-Plot of Heart Disease') +

  scale_x_discrete(labels = c('Males', 'Females')) +

  scale_fill_manual('Heart\nDisease', values = c('#FC766AFF', '#5B84B1FF'), labels=c('Yes', 'No'))
+

  ylim(c(110, 180)) +

  geom_boxplot()  +

  geom_violin(trim = FALSE, alpha = 0.2)

grid.arrange(plot3, plot4, ncol= 2)

set.seed(1)

heartDFCor <- cor(heartDF %>% type.convert(as.is=TRUE))

heartDFCor[is.na(heartDFCor)] = 0

heartDFCor <- round(heartDFCor, 2)

png(file='corr.png', res=150, width=10000, height=4500)

corrplot(as.matrix(heartDFCor), tl.cex = 3, tl.col = 'black', method = 'color',

         outline = T,  order='hclust',

         addCoef.col = 'black', number.digits = 2, number.cex = 3,

         cl.pos = 'b', cl.cex = 3, addrect = 3, rect.lwd = 3,

         col = colorRampPalette(c('midnightblue', 'white','darkred'))(100))
```

```
dev.off()

heart_below_40 <- subset(heartDF, Target == 1 | AgeGrp == 1,

                         select = c('MaxHeartRate'))

heart_below_40 <- data.frame(heart_below_40, group = 'heart_below_40')

heart_40_to_65 <- subset(heartDF, Target == 1 | AgeGrp == 2,

                         select = c('MaxHeartRate'))

heart_40_to_65 <- data.frame(heart_40_to_65, group = 'heart_40_to_65')

heart_above_65 <- subset(heartDF, Target == 1 | AgeGrp == 3,

                         select = c('MaxHeartRate'))

heart_above_65 <- data.frame(heart_above_65, group = 'heart_above_65')

# H0: There is no difference in max heart rate with heart disease in the three groups of age
(claim).

# H1: There is a difference in max heart rate with heart disease in the three groups of age.

alpha <- 0.05

data <- rbind(heart_below_40, heart_40_to_65, heart_above_65)

result <- kruskal.test(MaxHeartRate ~ group, data = data)

result.data <- data.frame(

  Values = c(result$statistic, '5.991', result$p.value, alpha, result$parameter),

  stringsAsFactors = FALSE

)

tidyr::as_tibble(result.data)

t(result.data) %>% kable(col.names = c('Kruskal-Wallis chi-squared', 'Critical Value', 'p-value',
'alpha', 'Degree of freedom'),
```

```
                caption = 'The Kruskal-Wallis Test') %>%

  kable_styling(bootstrap_options = c('striped', 'hover', 'condensed'),

                html_font = 'Calibri',

                font_size = 20)

# --- CHI-SQUARE --- #

preLimTDF <- subset(heartDF, select = c('AgeGrp', 'Sex',
'RestingBP','Cholesterol','MaxHeartRate','Target'))

blw40M <- preLimTDF[sample(which ( preLimTDF$Sex == 'Male' & preLimTDF$AgeGrp %in%
c('Below 40 yrs') ) ,1), ]

bt65M <- preLimTDF[sample(which ( preLimTDF$Sex == 'Male' & preLimTDF$AgeGrp %in% c('40-
65 yrs') ) ,1), ]

ab65M <- preLimTDF[sample(which ( preLimTDF$Sex == 'Male' & preLimTDF$AgeGrp %in%
c('Above 65 yrs') ) ,1), ]

blw40F <- preLimTDF[sample(which ( preLimTDF$Sex == 'Female' & preLimTDF$AgeGrp %in%
c('Below 40 yrs') ) ,5), ]

bt65F <- preLimTDF[sample(which ( preLimTDF$Sex == 'Female' & preLimTDF$AgeGrp %in%
c('40-65 yrs') ) ,5), ]

ab65F <- preLimTDF[sample(which ( preLimTDF$Sex == 'Female' & preLimTDF$AgeGrp %in%
c('Above 65 yrs') ) ,5), ]

mean(preLimTDF$RestingBP[which( preLimTDF$Sex == 'Male' & preLimTDF$Target == 1)])

c1 <- c(129.74,230.98,161.90)

c2 <- c(128.73,256.75,154.02)

prelTestMatrix <- matrix(c(c1,c2),ncol = 2)

rownames(prelTestMatrix) <- c("Resting BP","Cholesterol","Max. Heart Rate")

colnames(prelTestMatrix) <- c("Male","Female")

compltPreDF <- rbind(blw40M,bt65M,ab65M,blw40F,bt65F,ab65F)
```

```
data.matrix(t(compltPreDF[,2:5]))

chiSqTest <- chisq.test(prelTestMatrix)

dev.off()

balloonplot(

  t(as.table(prelTestMatrix)),

  main = "Preliminary Test: Heart Rate, Cholesterol, Blood Pressure",

  dotchar = 15,

  dotcolor = "#899878",

  label=TRUE,

  label.digits=2,

  label.size=1,

  show.margins = FALSE

)

Chitest.data <- data.frame(

  #Output = c("Chi-Squared", "Degrees of Freedom","p-value", "Critical Value", "Method"),

  Values = c(round(chiSqTest$statistic,2), chiSqTest$parameter, '3.559e-10', 7.82,
chiSqTest$method),

  stringsAsFactors = FALSE

)

t(Chitest.data) %>% kable(col.names = c('Chi-Squared','Degrees of Freedom','p-value','Critical
Value','Method'),caption = "Chi-Square: Heart Disease in different Age groups") %>%

  kable_styling(bootstrap_options = c("striped", "hover", "condensed"),html_font = "Calibri",
font_size = 20)

# --- TWO- WAY ANOVA Test --- #
```

```
table(heartDF$Sex,heartDF$AgeGrp)

par(mar=c(9, 4.1, 4.1, 2.1))

#create boxplots

boxplot(MaxHeartRate ~ Sex + AgeGrp,

        data = heartDF,

        main = "Max Heart Rate by Gender and Age Group",

        xlab = "Group",

        ylab = "Max Heart Rate",

        col = "steelblue",

        border = "black",

        las = 2 #make x-axis labels perpendicular

)

anovaModel <- aov(MaxHeartRate ~ Sex + AgeGrp, data = heartDF)

summary(anovaModel)

tukTest <- TukeyHSD(anovaModel,conf.level=.95)

par(mar=c(4.1, 13, 4.1, 2.1))

plot(tukTest, las = 2)



test <- as.data.frame(rbind(tukTest$Sex,tukTest$AgeGrp))

test %>% kable(col.names = c('differences','lower','upper','p-value'),caption = "Tukey Test:
Analysis of Variance: Age Vs Types of Angina") %>%

  kable_styling(bootstrap_options = c("striped", "hover", "condensed"),html_font = "Calibri",
font_size = 20)
```

```
hist(anovaModel$residuals, main = "Histogram of Residuals", xlab = "Residuals", col = "steelblue")

# --- Equal Variance --- #

car::leveneTest(MaxHeartRate ~ Sex * AgeGrp, data = heartDF)

# --- create Train and Test Data --- #

# --- Standardized Data --- #

heartDFStd <- heartDF %>%

  mutate_if(is.numeric, funs(as.numeric(scale(.))))

# Create Training Data

input_ones <- heartDF[which(heartDF$Target == 1), ]

input_zeros <- heartDF[which(heartDF$Target == 0), ]

set.seed(123)  # for repeatability of samples

input_zeros_training_rows <- sample(1:nrow(input_zeros), 0.8*nrow(input_zeros))

input_ones_training_rows <- sample(1:nrow(input_ones), 0.8*nrow(input_zeros))

training_ones <- input_ones[input_ones_training_rows, ]

training_zeros <- input_zeros[input_zeros_training_rows, ]

trainingData <- rbind(training_ones, training_zeros)

nrow(training_ones)

nrow(training_zeros)

# Create Testing Data

testing_ones <- input_ones[-input_ones_training_rows, ]

testing_zeros <- input_zeros[-input_zeros_training_rows, ]

testingData <- rbind(testing_ones, testing_zeros)
```

```
nrow(testing_ones)

nrow(testing_zeros)

# --- Logistic Regression --- #

# --- Best Fit Model --- #

bstFitModel <- glm(formula = Target ~ Sex + AgeGrp + MaxHeartRate + RestECG  +

                     STDepression + Thalassemia + FastingBSugar + NumMajorVessels,data =
trainingData, family = binomial(link="logit"))

summary(bstFitModel)

coefficients(bstFitModel)

factorNames <- c("Sex","AgeGrp","MaxHeartRate",

                 "RestECG","STDepression","Thalassemia","FastingBSugar",

                 "NumMajorVessels")

# --- Graph --- #

plts = lapply(factorNames,function(i){

  return(plot(ggpredict(bstFitModel,i)))

})

patchwork::wrap_plots(plts)

# --- Multi-collinearity --- #

car::vif(bstFitModel)

# --- McFadden's R2 --- #

pscl::pR2(bstFitModel)["McFadden"]

# --- Probability Train--- #

bestFitMTrain <- predict(bstFitModel,trainingData, type="response")
```

```
predictedTr.classes <- as.factor(ifelse(bestFitMTrain > 0.5, 1, 0))

head(predictedTr.classes)

mean(predictedTr.classes == trainingData$Target)

# --- Probability Test--- #

bestFitMTest <- predict(bstFitModel, testingData, type="response")

predictedTe.classes <- ifelse(bestFitMTest > 0.5, 1, 0)

head(predictedTe.classes)

mean(predictedTe.classes == testingData$Target)



#find optimal cutoff probability to use to maximize accuracy

optimal <- optimalCutoff(testingData$Target, bestFitMTest,optimiseFor='Ones')

optimal

# --- Confusion Matrix--- #

caret::confusionMatrix(as.factor(trainingData$Target), as.factor(predictedTr.classes), positive =
"1")

caret::confusionMatrix(as.factor(testingData$Target), as.factor(predictedTe.classes), positive =
"1")

table_mat_tr <- table(trainingData$Target, bestFitMTrain > 0.5)

table_mat_tr

table_mat_te <- table(testingData$Target, bestFitMTest > 0.5)

table_mat_te

rownames(table_mat_te) <- c('No Disease','Disease')

graphics::fourfoldplot(table_mat_te, color = c("#CC6666", "#99CC99"),
```

```
    conf.level = 0.95, margin = 1, main = "Confusion Matrix : Test Data")


TClass <- factor(c('FALSE', 'FALSE', 'TRUE', 'TRUE'))

PClass <- factor(c('No Disease', 'Disease', 'No Disease', 'Disease'))

Y      <- c(17, 11, 11, 44)

df <- data.frame(TClass, PClass, Y)

ggplot(data =  df, mapping = aes(x = TClass, y = PClass)) +

  geom_tile(aes(fill = Y), colour = "white") +

  geom_text(aes(label = sprintf("%1.0f", Y)), vjust = 1, colour = "White",family = "serif",size = 8) +

    scale_fill_gradient(low = "#9AC2C5", high = "#273B09") +

  theme_bw() + theme(legend.position = "none")

#calculate sensitivity

sensi_tr <- sensitivity(trainingData$Target, bestFitMTrain)

sensi_te <- sensitivity(testingData$Target, bestFitMTest)

#calculate specificity

speci_tr <- specificity(trainingData$Target, bestFitMTrain)

speci_te <- specificity(testingData$Target, bestFitMTest)

#calculate total misclassification error rate

misClassError(testingData$Target, bestFitMTest, threshold=optimal)

# --- Accuracy --- #

accuracy_Test_Tr <- sum(diag(table_mat_tr)) / sum(table_mat_tr)

accuracy_Test_Tr
```

```
accuracy_Test_Te <- sum(diag(table_mat_te)) / sum(table_mat_te)

accuracy_Test_Te

# --- Variable Importance --- #

caret::varImp(bstFitModel)

precision <- function(matrix) {

  # True positive

  tp <- matrix[2, 2]

  # false positive

  fp <- matrix[1, 2]

  return (tp / (tp + fp))

}

prec_te <- precision(table_mat_te)

prec_te

prec_tr <- precision(table_mat_tr)

prec_tr

# --- Recall --- #

recall <- function(matrix) {

  # true positive

  tp <- matrix[2, 2]

  # false positive

  fn <- matrix[2, 1]

  return (tp / (tp + fn))
```

```
}

rec_te <- recall(table_mat_te)

rec_te

rec_tr <- recall(table_mat_tr)

rec_tr

# --- F1 Score --- #

f1 <- 2 * ((prec_te * rec_te) / (prec_te + rec_te))

f1

dev.off()

#plot the ROC curve

plotROC(testingData$Target, bestFitMTest)

tablDt <- data.frame()

tablDt <- cbind(c(accuracy_Test_Tr * 100,sensi_tr * 100,speci_tr * 100,prec_tr * 100,rec_tr *
100),c(accuracy_Test_Te * 100,sensi_te * 100,speci_te * 100,prec_tr * 100,rec_te * 100))

rownames(tablDt) <- c('Accuracy','Sensitivity','Specificity','Precision','Recall')

colnames(tablDt) <- c('Training Data','Testing Data')

stargazer(bstFitModel,

        title = "Logistics Regression: Heart Disease",

        #dep.var.caption ="D : Weight" ,

        #covariate.labels ="Height (inches)",

        #column.labels = c("Male","Female"),

        notes.label = "Significance levels",

        type = "html",
```

```
        out = "C:/Users/91956/Srinivas Shanmuga G/NEU/Q2/ALY 6015 Intermediate
Analytics/Assignments/Final Project/final_1.htm")

customGreen0 = "#A1D2CE"

customGreen1 = "#62A8AC"

customRed = "#ff7f7f"

tbl_data <- cbind("Indicator" = rownames(tablDt), as.data.frame(tablDt))

rownames(tbl_data) <- c()

formattable(

  tbl_data,

  align = c("l","r"),

  list(`Indicator` = formatter("span", style = ~ style(color = "grey", font.weight = "bold")),

       `Training Data` = color_tile(customGreen0, customGreen1),

       `Testing Data` = color_tile(customGreen0, customGreen1)

  )

)
```