**Mini Project (Supervised Learning)**

**Regression**

Consider ‘Heart’ data set from ISLR library.

Explain the data set

What could be the target variable?

Whether this is a classification problem or regression problem?

If it is a regression problem, get least RMSE or RSE and best R^2

If it is a classification problem, get best precision (true positives/ (true positives + false positives)) and recall (true positives/ (true positives + false negatives))

Technologies to be used: R and Python (Jupyter note book)

First Step understand the Heart dataset.

In Dataset first column is unnamed which is looks like primary keys / patients sl.no. / Primary keys .

Data set Contains 303 numbers of rows or observation and 14 attributes apart from unnamed columns. If we consider this columns also then there will be total 15 columns.

Description of the dataset

1. Attribute : Age

Age in years. Type – int.

1. Attribute : Sex

2 values {0: female, 1: male}

Sex can be Female or Male . Type – Binary or Boolean.

1. Attribute : Chest Pain

Chest pain type (typical angina, atypical angina, non-angina, or asymptomatic angina)

Type : Categorical

1. Attribute : RestBP ( resting blood pressure on admission to hospital)

Resting blood pressure (mm Hg)

Type : Continuous

1. Attribute : Chol

Serum cholesterol (mg/dl)

Type : Continuous

1. Attribute : Fbs

Fasting blood sugar (< 120 mg/dl or > 120 mg/dl)

2 values {0: <= 120 mg/dl, 1: > 120 mg/dl}

Type : Binary or Boolean

1. Attribute : RestECG

Resting electrocardiography results (normal, ST-T wave abnormality, or left ventricular hypertrophy)

{0: normal, 1: ST-T wave abnormality, 2: left ventricular hypertrophy}

Type : Categorical

1. Attribute : MaxHR

Max. Heart rate achieved during thalium stress test

Type : Continous

1. Attribute : ExAng

Exercise induced angina (yes or no)

2 values {0: no, 1: yes}

Type : Binary or Boolean

1. Attribute : Oldpeak

ST depression induced by exercise relative to rest.

Type : Continuous

1. Attribute : Slope

Slope of peak exercise ST segment (unsloping, flat, or downsloping)

3 values {1: upsloping, 2: flat, 3: downsloping}

Type : Categorical

1. Attribute : Ca

Number of major vessels colored by fluoroscopy

Type : Discrete

Discrete (0,1,2,3,4)

And having 4 NA’s in this Ca column.

1. Attribute : Thal

Thalium stress test result (normal, fixed defect, or reversible defect)

Type : Categorical

There is two NA values is there in Thal columns.

1. Attribute : AHD stands for ( Atherosclerotic Heart Disease )

[Atherosclerosis](https://www.webmd.com/heart-disease/what-is-atherosclerosis) -- sometimes called hardening of the arteries -- can slowly narrow the [arteries](https://www.webmd.com/heart/picture-of-the-arteries) throughout your body. When atherosclerosis affects arteries that carry blood to the [heart](https://www.webmd.com/heart/picture-of-the-heart) muscle, it’s called [coronary artery disease](https://www.webmd.com/heart-disease/guide/heart-disease-coronary-artery-disease).

Atherosclerosis can create life-threatening blockages.

So, AHD is the response variable. 2 values {0: no (no heart disease) , 1: yes(yes there is heart disease)}

Note : Type column indicates whether an attribute is binary , integer , categorical, or continuous.

This is a problem of Classification and will used Logistics regression in that.

1 (Age) + 1 (Sex) + 3 (ChestPain) + 1 (RestBP) + 1 (Chol) + 1 (Fba) + 2 (RestECG) +

1 (MaxHR) + 1 (ExAng) + 1 (Oldpeak) + 2 (Slope) + 1 (Ca) + 2 (Thal) = 18

The actual number of feature variables (after converting categorical variables to dummy ones) is 18.

> summary(hd)

id Age Sex

Min. : 1.0 Min. :29.00 Min. :0.0000

1st Qu.: 76.5 1st Qu.:48.00 1st Qu.:0.0000

Median :152.0 Median :56.00 Median :1.0000

Mean :152.0 Mean :54.44 Mean :0.6799

3rd Qu.:227.5 3rd Qu.:61.00 3rd Qu.:1.0000

Max. :303.0 Max. :77.00 Max. :1.0000

ChestPain RestBP Chol

asymptomatic:144 Min. : 94.0 Min. :126.0

nonanginal : 86 1st Qu.:120.0 1st Qu.:211.0

nontypical : 50 Median :130.0 Median :241.0

typical : 23 Mean :131.7 Mean :246.7

3rd Qu.:140.0 3rd Qu.:275.0

Max. :200.0 Max. :564.0

Fbs RestECG MaxHR

Min. :0.0000 Min. :0.0000 Min. : 71.0

1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:133.5

Median :0.0000 Median :1.0000 Median :153.0

Mean :0.1485 Mean :0.9901 Mean :149.6

3rd Qu.:0.0000 3rd Qu.:2.0000 3rd Qu.:166.0

Max. :1.0000 Max. :2.0000 Max. :202.0

ExAng Oldpeak Slope

Min. :0.0000 Min. :0.00 Min. :1.000

1st Qu.:0.0000 1st Qu.:0.00 1st Qu.:1.000

Median :0.0000 Median :0.80 Median :2.000

Mean :0.3267 Mean :1.04 Mean :1.601

3rd Qu.:1.0000 3rd Qu.:1.60 3rd Qu.:2.000

Max. :1.0000 Max. :6.20 Max. :3.000

Ca Thal AHD

Min. :0.0000 fixed : 18 No :164

1st Qu.:0.0000 normal :166 Yes:139

Median :0.0000 reversable:117

Mean :0.6722 NA's : 2

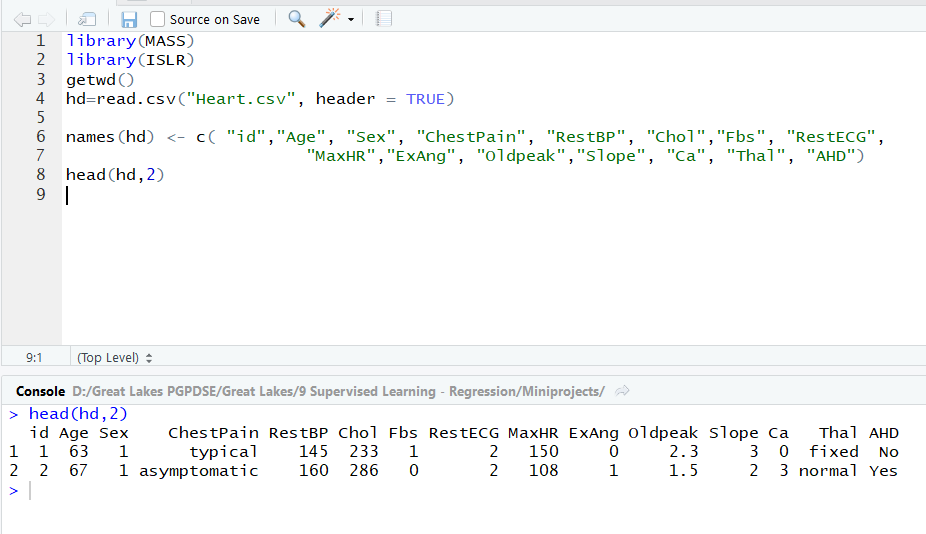
3rd Qu.:1.0000

Max. :3.0000

NA's :4

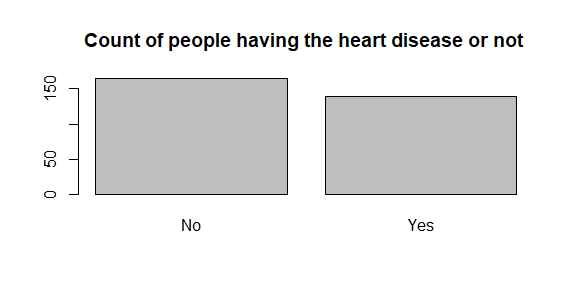
Ca and Thal two column is having the NA column.

Importing of Dataset with the head () first two rows is showed.



In the initial data id column name is not given I have added the id column name can be consider as Patient id.

Count of people having the Heart Disease or not referring to the last column AHD.

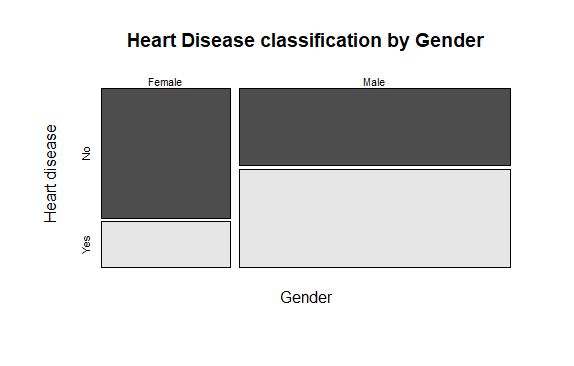


> summary(hd$AHD)

No Yes

164 139

164 patients is not having the heart disease and 139 patients is having the heart disease.



table(hd$Sex,hd$AHD)

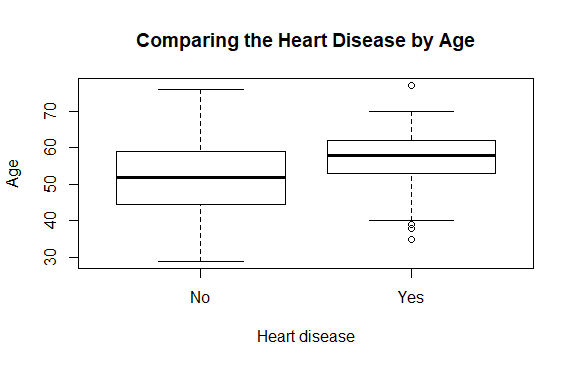
No Yes

Female 71 25

Male 89 112

In terms of % , 73.96% of Female have not heart disease and 44.28% of Male have not heart disease .

By comparing this female is having more chance of not having the heart disease.



There is a few outlier in case of having the heart disease one on the upper side of the whiskers and three on the lower side of the whiskers which implies that in higher and lower age there is also a chance of having the heart disease.

> correlationMatrix <- cor(hd[,c(2,5,6,9,11)])

> print(correlationMatrix)

Age RestBP Chol MaxHR Oldpeak

Age 1.0000000 0.29047626 2.026435e-01 -3.945629e-01 0.19712262

RestBP 0.2904763 1.00000000 1.315357e-01 -4.910766e-02 0.19124314

Chol 0.2026435 0.13153571 1.000000e+00 -7.456799e-05 0.03859579

MaxHR -0.3945629 -0.04910766 -7.456799e-05 1.000000e+00 -0.34763997

Oldpeak 0.1971226 0.19124314 3.859579e-02 -3.476400e-01 1.00000000

There is no correlation between all numerical variable . So all variable are independent .

> highlyCorrelated <- findCorrelation(correlationMatrix, cutoff=0.5)

> print(highlyCorrelated)

integer(0)

So we are considering all numeric value as there is no correlation between the any numerical value.

> modlog<- glm(AHD~.,data=trainData,family = "binomial",na.action = na.omit)

> summary(modlog)

Call:

glm(formula = AHD ~ ., family = "binomial", data = trainData,

na.action = na.omit)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.93227 -0.40718 -0.07394 0.30314 3.00730

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.448757 3.656153 -0.943 0.345540

id 0.005198 0.003031 1.715 0.086321 .

Age -0.034490 0.030375 -1.135 0.256172

SexMale 2.130378 0.762934 2.792 0.005233 \*\*

ChestPainnonanginal -2.154639 0.663404 -3.248 0.001163 \*\*

ChestPainnontypical -1.179831 0.740921 -1.592 0.111299

ChestPaintypical -1.833831 0.947188 -1.936 0.052858 .

RestBP 0.016567 0.016237 1.020 0.307585

Chol 0.004453 0.004979 0.894 0.371151

FbsHigh -0.179086 0.795180 -0.225 0.821812

RestECGleft ventricular 0.306277 0.498633 0.614 0.539061

MaxHR -0.022301 0.014922 -1.494 0.135057

ExAngyes 0.549184 0.565153 0.972 0.331177

Oldpeak 0.294427 0.290177 1.015 0.310274

Slopeflat 1.785774 0.631770 2.827 0.004704 \*\*

Slopedown 1.133205 1.105272 1.025 0.305234

Ca1 2.699516 0.708470 3.810 0.000139 \*\*\*

Ca2 4.063026 1.034707 3.927 8.61e-05 \*\*\*

Ca3 2.153260 0.985054 2.186 0.028821 \*

Thalnormal 0.647029 1.029308 0.629 0.529607

Thalreversable 2.212115 1.000266 2.212 0.026999 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 287.12 on 207 degrees of freedom

Residual deviance: 118.73 on 187 degrees of freedom

AIC: 160.73

Number of Fisher Scoring iterations: 6

> logRegConfMat <- confusionMatrix(logRegPrediction, testData[,"AHD"])

> logRegConfMat

Confusion Matrix and Statistics

Reference

Prediction No Yes

No 41 5

Yes 7 36

Accuracy : 0.8652

95% CI : (0.7763, 0.9283)

No Information Rate : 0.5393

P-Value [Acc > NIR] : 5.93e-11

Kappa : 0.7296

Mcnemar's Test P-Value : 0.7728

Sensitivity : 0.8542

Specificity: 0.8780

Pos Pred Value : 0.8913

Neg Pred Value : 0.8372

Prevalence : 0.5393

Detection Rate : 0.4607

Detection Prevalence : 0.5169

Balanced Accuracy : 0.8661

'Positive' Class : No

> AUC$logReg <- roc(as.numeric(testData$AHD),as.numeric(as.matrix((logRegPredictionprob))))$auc

> AUC$logReg

Area under the curve: 0.9157

> Accuracy$logReg <- logRegConfMat$overall['Accuracy']

> Accuracy$logReg

Accuracy

0.8651685

> # Confusion Matrix

> **table(Actualvalues = testData$AHD, Predictedvalues = logRegPredictionprob> 0.63**)

Predictedvalues

Actualvalues FALSE TRUE

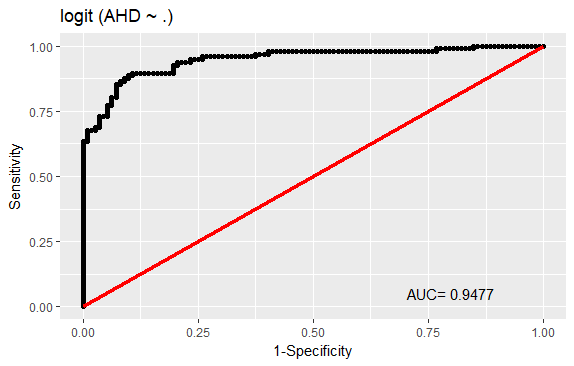
No 44 4

Yes 6 35

Confusion Matrix shows that it has greater accuracy with comparable Sensitivity and Sensitivity.

library(Deducer)

rocplot(modlog)



Conditional Probability Plot.

> # Conditional Probability plot

> cdplot(testData$AHD~logRegPrediction,data = testData, main = "Conditional Probability Density Curve")

