



PROBABILITY AND STATISTICS PROJECT

TOPIC OF PROJECT: HEART DISEASE PREDICTION

SOFTWARE
USED:

RStudio







TEAM MEMBERS



REG. NO.	NAME
21BCE5724	SUGANDHI NINAD NILESH
21BCE5726	DUNNA SRAVAN KUMAR
21BCE5733	AMAN ANAND
21BCE5871	RAJEEV RANJAN



01 Basic Statistical Analysis

Consists of basic analysis functions such as mean, median, mode, etc.

Data
Visualization

Visualization of data using plots, histogram, etc

Correlation
Correlation between members of data items.

Multiple Linear
Regression
Using Im function for predicting output

Logistic
Regression
Using glm function for predicting output

Prediction Using
Plots

Using plots function for plotting between prediction and actual values





LINK OF UPLOADED DATASET TAKEN FROM KAGGLE PASTED BELOW

https://drive.google.com/file/d /1tdbow15rjD-@egVGK4IrgBsk@810Z_Xn/vi ew?usp=sharing

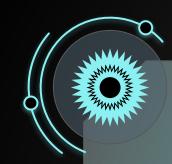




SCREENSHOT OF DATASET

A1	A1 • : X • fx index														
4	Α	В	С	D	Е	F	G	Н	1	J	K	L	М	N	0
1	index	Age	Sex	Chest_pain_type	BP	Cholesterol	FBS_over_120	EKG_results	Max_HR	Exercise_angina	ST_depression	Slope_of_ST	Number_of_vessels_	Thallium	Heart_Disease
2	0	7	0 1	. 4	130	322	0	2	109	0	2.4	2	3	3	1
3	1	6	57 (3	115	564	0	2	160	0	1.6	2	0	7	0
4	2	5	7 1	. 2	124	261	0	0	141	0	0.3	1	0	7	1
5	3	6	14 1	. 4	128	263	0	0	105	1	0.2	2	1	7	0
6	4	7	'4 (2	120	269	0	2	121	1	0.2	1	1	3	0
7	5	6	5 1	. 4	120	177	0	0	140	0	0.4	1	0	7	0
8	6	5	6 1	. 3	130	256	1	2	142	1	0.6	2	1	6	1
9	7	5	9 1	. 4	110	239	0	2	142	1	1.2	2	1	7	1
10	8	6	0 1	. 4		293	0			0	1.2	2	2		
11	9	6	i3 (4	150	407	0	2	154	0	4	2	3		
12	10	5	9 1	. 4	135	234	0	0	161	0	0.5	2	0	7	0
13	11		i3 1	. 4	142	226	0			1	0	1	0		0
14	12	4	4 1	. 3	140	235	0	2	180	0			0		0
15	13		1 1	. 1	134	234	0			0	2.6	2	2		1
16	14		7 (303	0					_			0
17	15		1 (4		149	0			0					0
18	16		6 1			311	0			1					
19	17		i3 1	. 4	140	203	1			1		3	0		_
20	18		14 1			211	0			1					
21	19		10 1			199	0								-
22	20		i7 1			229	0			1					
23	21		8 1			245	0	_						_	0
24	22		3 1			303	0			0					0
25	23		7 1			204	0			0		1	0	_	0
26	24		4 (288	1			1	0	-			0
27	25		8 (275	0			0		1	0		0
28	26		6 (138	243	0	2	152	1	0	2	0	3	0
4		Heart_D	isease_Prec	iction (+)							1	4			





BASIC STATISTICAL ANALYSIS



000

MEAN

```
attach(Heart_Disease_Prediction)
  mean(Age)
   [1] 54.43333
mean(Chest_pain_type)
   [1] 3.174074
mean(BP)
   [1] 131.3444
mean(Cholesterol)
   [1] 249.6593
mean(FBS_over_120)
   [1] 0.1481481
> mean(Max_HR)
  [1] 149.6778
> mean(ST_depression)
  [1] 1.05
mean(Slope_of_ST)
   [1] 1.585185
> mean(Thallium)
```

[1] 4.696296



000

MEDIAN

- median(Age)
 [1] 55
- median(BP)
 [1] 130
- median(Cholesterol)
 [1] 245
- median(Max_HR)
 [1] 153.5
- median(ST_depression)
 [1] 0.8
- median(Thallium)
 [1] 3





[1] 234

```
my_mode <- function(x) {</pre>
                                                     reate mode function
      unique_x <- unique(x)
      tabulate_x <- tabulate(match(x, unique_x))
      unique_x[tabulate_x == max(tabulate_x)]
my_mode(Age)
[1] 54
my_mode(Sex)
\lceil 1 \rceil 1
print("This defines that number of men is not equal to number of women")
[1] "This defines that number of men is not equal to number of women"
my_mode(Chest_pain_type)
\lceil 1 \rceil 4
print("Most of the people have the chest pain of 4th type")
[1] "Most of the people have the chest pain of 4th type"
my_mode(BP)
[1] 120
>print("120 is the most repeated among the recorded BP in the dataset")
 [1] "120 is the most repeated among the recorded BP in the dataset"
>my_mode(Cholesterol)
```



MODE (CONTINUED)

print("234 is the most repeated value of cholestrol recorded among the patients")
[1] "234 is the most repeated value of cholestrol recorded among the patients"
> my_mode(Max_HR)
[1] 162
> my_mode(Thallium)
[1] 3
> print("3 units of thallium is present in most of the patients body")
[1] "3 units of thallium is present in most of the patients body">
my_mode(Heart_Disease)
[1] 0
> print("Number of Patients having Heart Disease is less than number of patients not suffering from any Heart Disease")
[1] "Number of Patients having Heart Disease is less than number of patients not

suffering from any Heart Disease"

SUMMARY

```
summary(Age)
 Min. 1st Qu.
               Median
                          Mean 3rd Qu.
                                          Max.
         48.00
                 55.00
 29.00
                          54.43
                                  61.00
                                          77.00
summary(Sex)
                         Mean 3rd Qu.
 Min. 1st Qu. Median
                                          Max.
 0.0000 0.0000 1.0000
                         0.6778 1.0000
                                         1.0000
summary(Chest_pain_type)
 Min. 1st Qu. Median
                          Mean 3rd Qu.
                                          Max.
         3.000
                 3.000
 1.000
                          3.174
                                  4.000
                                          4.000
summary(BP)
 Min. 1st Qu.
               Median
                         Mean 3rd Qu.
                                          Max.
 94.0
        120.0
                130.0
                         131.3
                                 140.0
                                         200.0
summary(Cholesterol)
 Min. 1st Qu. Median
                          Mean 3rd Qu.
                                          Max.
 126.0
         213.0
                 245.0
                          249.7
                                  280.0
                                          564.0
```

SUMMARY CONTINUED

```
summary(FBS_over_120)
Min. 1st Ou. Median Mean 3rd Ou.
                                      Max.
0.0000
       0.0000 \quad 0.0000 \quad 0.1481 \quad 0.0000
                                     1.0000
summary(EKG_results)
Min. 1st Ou. Median
                     Mean 3rd Qu.
                                      Max.
0.000
       0.000
             2.000
                      1.022 2.000
                                      2.000
summary(Max_HR)
Min. 1st Qu. Median Mean 3rd Qu.
                                      Max.
71.0
      133.0 153.5 149.7 166.0
                                     202.0
summary(Exercise_angina)
Min. 1st Qu. Median Mean 3rd Qu.
                                      Max.
0.0000
      0.0000 0.0000 0.3296 1.0000
                                     1.0000
```

SUMMARY CONTINUED

```
summary(ST_depression)
  Length Class
                      Mode
       1 character character
  summary(Slope_of_ST)
  Min. 1st Qu. Median Mean 3rd Qu.
                                       Max.
 1.000
        1.000
              2.000 1.585 2.000
                                      3.000
summary(Number_of_vessels_fluro)
  Min. 1st Qu. Median Mean 3rd Qu.
                                       Max.
0.0000 0.0000 0.0000 0.6704 1.0000
                                     3.0000
  summary(Thallium)
  Min. 1st Qu. Median Mean 3rd Qu.
                                       Max.
```

4.696

7.000

7.000

3.000

3.000

3.000

VARIANCE & STANDARD DEVIATION

- var(Age)
- [1] 82.97509
- sd(Age)
- [1] 9.109067
- var(BP)
- [1] 319.0371
- > sd(BP)
- [1] 17.86161
- var(Cholesterol)
- [1] 2671.467
- > sd(Cholesterol)
- [1] 51.68624
- var(Max_HR)
- [1] 536.6504
- sd(Max_HR)
- [1] 23.16572

COVARIANCE

- \triangleright cov1 = cov(Age, BP)
- Cov1[1] 44.42639
- cov2 = cov(Sex, BP)
- ➤ Cov2
 - [1] -0.5242875
- cov3 = cov(BP, Cholesterol)
- Cov3[1] 159.7312
- cov(Age, Heart_Disease) [1] 0.9628253
- cov(Sex, Heart_Disease) [1] 0.06939281
- cov(Chest_pain_type, Heart_Disease) [1] 0.1974391
- cov(BP, Heart_Disease)
 - [1] 1.38166
- cov(Cholesterol, Heart_Disease)

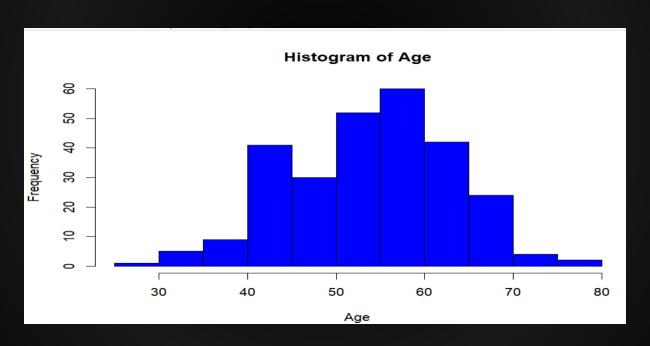
COVARIANCE CONTINUED

- [1] 3.036762
- cov(FBS_over_120, Heart_Disease)
 - [1] -0.002891367
- cov(EKG_results, Heart_Disease) [1] 0.09045849
- cov(Max_HR, Heart_Disease) [1] -4.826518
- cov(Exercise_angina, Heart_Disease) [1] 0.09830648
- cov(ST_depression, Heart_Disease) [1] 0.23829
- cov(Slope_of_ST, Heart_Disease) [1] 0.1032631
- cov(Number_of_vessels_fluro, Heart_Disease) [1] 0.2139612
- cov(Thallium, Heart_Disease) [1] 0.5072284

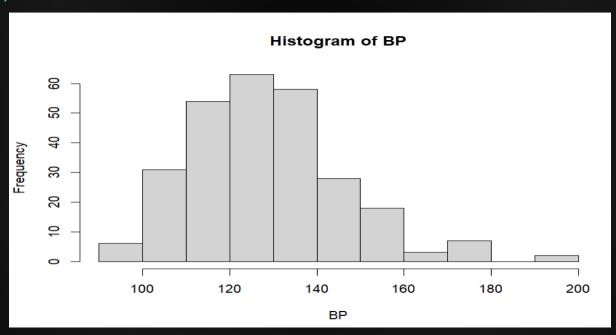


HISTOGRAM

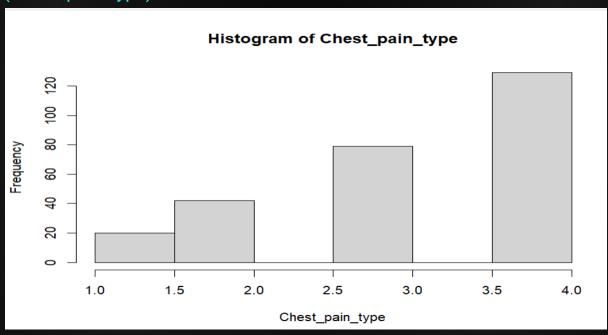
> hist(Age, col = 'blue')



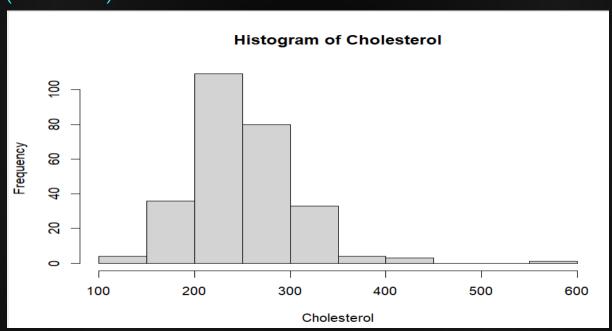
> hist(BP)



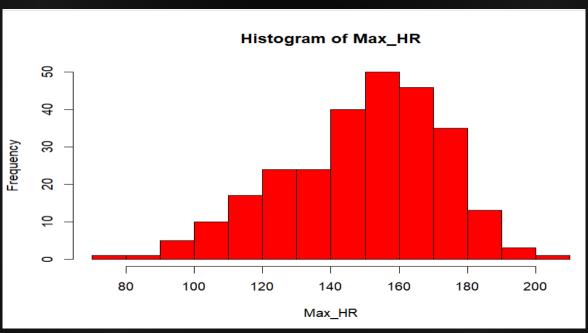
> hist(Chest_pain_type)



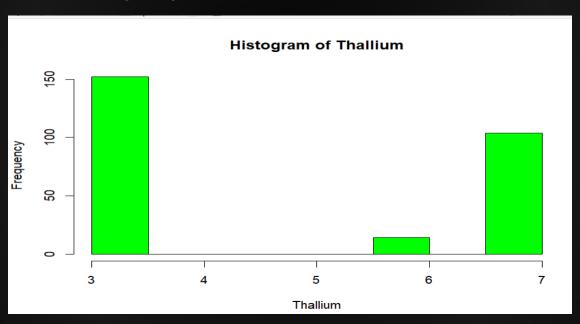
> hist(Cholesterol)



> hist(Max_HR, col = 'red')

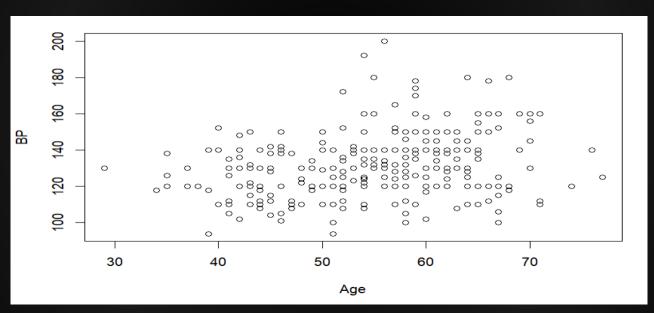


> hist(Thallium, col = 'green')



PLOT

> plot(Age,BP)





- cor.test(Age, Heart_Disease)
- Pearson's product-moment correlation
- data: Age and Heart_Disease
- t = 3.557, df = 268, p-value = 0.0004435
- alternative hypothesis: true correlation is not equal to 0
- 95 percent confidence interval: 0.09536339 0.32349863
- sample estimates:
- cor 0.2123222
- cor.test(Sex, Heart_Disease)
- Pearson's product-moment correlation
- data: Sex and Heart_Disease t = 5.1054, df = 268, p-value = 6.267e-07
- alternative hypothesis: true correlation is not equal to 0
- 95 percent confidence interval:
- 0.1849169 0.4027815

cor

- sample estimates:
 - 0.2977208

- cor.test(Chest_pain_type, Heart_Disease) Pearson's product-moment correlation data: Chest_pain_type and Heart_Disease t = 7.5203, df = 268, p-value = 8.262e-13 alternative hypothesis: true correlation is not equal to 0 95 percent confidence interval:
- 0.3136922 0.5113315

t = 2.575, df = 268, p-value = 0.01056

- sample estimates: cor 0.4174362
- cor.test(BP, Heart_Disease)
- Pearson's product-moment correlation data: BP and Heart Disease
- alternative hypothesis: true correlation is not equal to 0 95 percent confidence interval:
- 0.03668728 0.26975488
- sample estimates:
- cor 0.1553827

- cor.test(Cholesterol, Heart_Disease) Pearson's product-moment correlation data: Cholesterol and Heart_Disease t = 1.9457, df = 268, p-value = 0.05274 alternative hypothesis: true correlation is not equal to 0 95 percent confidence interval: -0.001374728 0.234098256 sample estimates: cor 0.1180205 cor.test(FBS_over_120, Heart_Disease) Pearson's product-moment correlation
- data: FBS_over_120 and Heart_Disease t = -0.26719, df = 268, p-value = 0.7895 alternative hypothesis: true correlation is not equal to 0 95 percent confidence interval:
- -0.1354309 0.1032582 sample estimates:
- cor -0.01631883

- cor.test(EKG_results, Heart_Disease) Pearson's product-moment correlation data: EKG_results and Heart_Disease t = 3.0316, df = 268, p-value = 0.00267 alternative hypothesis: true correlation is not equal to 0 95 percent confidence interval: 0.06410841 0.29505301 sample estimates: cor 0.1820908 cor.test(Max_HR, Heart_Disease) Pearson's product-moment correlation
- data: Max_HR and Heart_Disease t = -7.5438, df = 268, p-value = 7.12e-13 alternative hypothesis: true correlation is not equal to 0 95 percent confidence interval: -0.5122953 -0.3148691
- sample estimates: cor
- -0.418514

- cor.test(Exercise_angina, Heart_Disease)
 Pearson's product-moment correlation
 data: Exercise_angina and Heart_Disease
 t = 7.5611, df = 268, p-value = 6.383e-13
 alternative hypothesis: true correlation is not equal to 0
 95 percent confidence interval:
 0.3157306 0.5130005
 sample estimates:
 cor
 0.4193027
- cor.test(ST_depression, Heart_Disease)
 Pearson's product-moment correlation
 data: ST_depression and Heart_Disease
 t = 7.5319, df = 268, p-value = 7.678e-13
 alternative hypothesis: true correlation is not equal to 0
 95 percent confidence interval:
- sample estimates: Cor

0.3142722 0.5118066

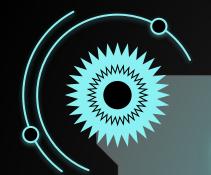
0.4179674

```
cor.test(Slope_of_ST, Heart_Disease)
Pearson's product-moment correlation
data: Slope_of_ST and Heart_Disease
t = 5.8718, df = 268, p-value = 1.272e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.2274052 0.4392872
sample estimates:
cor
0.337616
cor.test(Number_of_vessels_fluro, Heart_Disease)
Pearson's product-moment correlation
data: Number_of_vessels_fluro and Heart_Disease
t = 8.3725, df = 268, p-value = 3.173e-15
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.3552718 0.5450837
sample estimates:
```

cor

0.4553365

- cor.test(Thallium, Heart_Disease)
- Pearson's product-moment correlation
- data: Thallium and Heart_Disease
- t = 10.099, df = 268, p-value < 2.2e-16
- alternative hypothesis: true correlation is not equal to 0
- 95 percent confidence interval:
- 0.4327681 0.6063908
- sample estimates:
- cor
- 0.5250203



Multiple Linear Regression

04



- fit1 = Im(Heart_Disease~BP + Age + Sex + Thallium)
- > fit1
- Call:
- Im(formula = Heart_Disease ~ BP + Age + Sex + Thallium)
- Coefficients:
- (Intercept) BP Age Sex Thallium -0.900582 0.001716 0.008852 0.152465 0.113803

```
fit1 = Im(BP ~ Age + Sex + Chest_pain_type + FBS_over_120 + EKG_results + Max_HR + Exercise_angina + ST_depression + Slope_of_ST + Number_of_vessels_fluro + Thallium)
fit1
```

Call:

(Intercept)

Im(formula = BP ~ Age + Sex + Chest_pain_type + FBS_over_120 + EKG_results + Max_HR + Exercise_angina + ST_depression + Slope_of_ST + Number_of_vessels_fluro + Thallium)

Sex

Coefficients:

83.8148 0.5267 -4.1680
Chest_pain_type FBS_over_120 EKG_results
-1.7281 5.9246 1.3346
Max_HR Exercise_angina ST_depression
0.1108 2.7272 3.1902
Slope_of_ST Number_of_vessels_fluro Thallium

Age

-0.5168 -1.1262 1.2081 reached 'max' / getOption("max.print") -- omitted 204 rows]



```
    fit2 = glm(Heart_Disease ~ Age + Sex + Chest_pain_type + BP + Cholesterol + FBS_over_120 + EKG_results + Max_HR + Exercise_angina + ST_depression + Slope_of_ST + Number_of_vessels_fluro + Thallium, data = data, family = binomial(link = 'logit'))
    summary(fit2)
    Call: glm(formula = Heart_Disease ~ Age + Sex + Chest_pain_type + BP + Cholesterol + FBS_over_120 + EKG_results + Max_HR + Exercise_angina + ST_depression +
```

Slope_of_ST + Number_of_vessels_fluro + Thallium, family = binomial(link = "logit"), data

Deviance Residuals:

Min 1Q Median 3Q Max
-2.6123 -0.5120 -0.1739 0.4023 2.4210

= data)

Coefficients:

Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.446417 3.088097 -2.735 0.00624 **

Age -0.017477 0.025720 -0.680 0.49681

Sex 1.542109 0.540762 2.852 0.00435 **

```
Chest_pain_type 0.700895 0.215280 3.256 0.00113 **

BP 0.025216 0.011450 2.202 0.02765 *

Cholesterol 0.007228 0.004077 1.773 0.07628 .

FBS_over_120 -0.794811 0.574662 -1.383 0.16664

EKG_results 0.301668 0.197838 1.525 0.12730

Max_HR -0.021045 0.010579 -1.989 0.04666 *

Exercise_angina 0.829386 0.431091 1.924 0.05436 .

ST_depression 0.343690 0.227068 1.514 0.13013

Slope_of_ST 0.442276 0.391077 1.131 0.25809

Number_of_vessels_fluro 1.165271 0.269283 4.327 1.51e-05 ***

Thallium 0.341384 0.106066 3.219 0.00129 **
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 370.96 on 269 degrees of freedom

Residual deviance: 179.60 on 256 degrees of freedom

AIC: 207.6

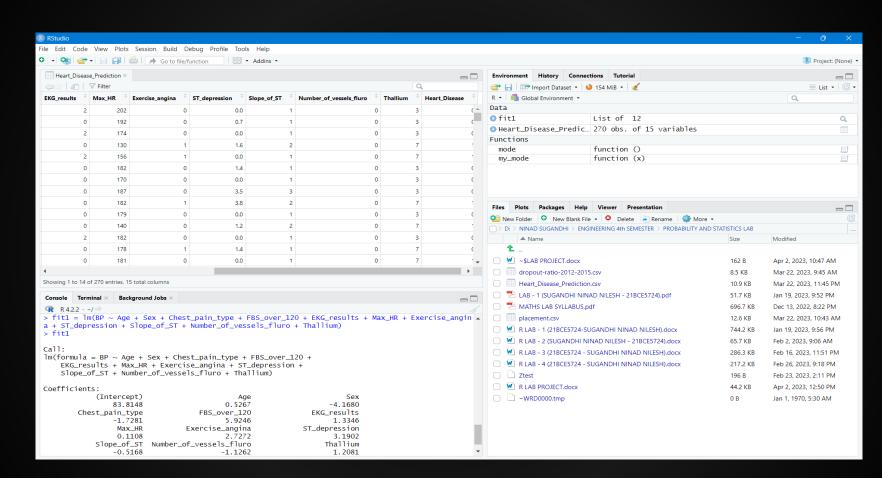
Number of Fisher Scoring iterations: 6

```
> predict_regression = predict(fit2, data, type = 'response')
> predict_regression
      2 3
                  4 5
0.991998619 0.627598571 0.179790533 0.954226458 0.144175093
0.287328023 0.786154238
            10
                  11
                      12 13
0.934964448 0.971845881 0.995878630 0.505348858 0.892430729
0.113024750 0.470539572
       16
              17
                    18
                        19
                               20
0.208746590 0.040909142 0.995507390 0.893626135 0.109723983
0.176905386 0.990297874
       23
              24
                    25
                        26
                                         28
0.075866594 0.222628921 0.106010724 0.065691610 0.035331466
0.255772844 0.042899619
                    32
                        33
29
       30
             31
                               34
                                         35
0.750430000 0.039501130 0.751857127 0.803520602 0.010753215
0 992479545 0 921771902
36
       37
              38
                    39
                           40
                                  41
                                         42
```

0.944	402505	0.95427	4292 U.	1433563	80 0.06	61/83/9	0.108//18	369
0.417	657619	0.01707	4168					
43	44	45	46	47	48	49		
0.125	400301	0.02707	5692 0.	5397864	37 0.04	9383376	0.8296652	235
0.238	294968	0.97994	9088					
50	51	52	53	5/	55	56		

0.981066996 0.932240375 0.007371878 0.178814730 0.040073054

0.021324125 0.010087727 0.932523472 0.011790340 0.319069745 0.861428506 0.186275875 0.616813096 0.047361787 0.006059465 0.241091018 0.983049517 0.356493518 0.128239900 0.087324301 0.053363024





Prediction Using Plots



plot(BP, predict_regression, main = "Blood Pressure", col = rgb(red = 0.5, blue = 0.2, green = 0.7)plot(Cholesterol, predict_regression, main = "Cholestrol", col = rgb(red = 0.4, blue = 0.8, green = 0.2) plot(FBS_over_120, predict_regression, main = "FBS over 120", col = rgb(red = 0.2, blue = 0.2) 0.6, green = 0.1) plot(EKG_results, predict_regression, main = "EKG results", col = rgb(red = 0.2, blue = 0.6, green = 0.1) plot(Max_HR, predict_regression, main = "Max HR", col = rgb(red = 0.5, blue = 0.6, green = 0.6)plot(Exercise_angina, predict_regression, main = "Exercise Angina", col = rgb(red = 0.5, blue = 0.5, green = 0.5)) plot(ST_depression, predict_regression, main = "ST depression", col = rgb(red = 0.8, blue = 0.2, green = 0.2) plot(Number_of_vessels_fluro, predict_regression, main = "Number of vessels of fluro", col = rgb(red = 0.1, blue = 0.8, green = 0.4))plot(Thallium, predict_regression, main = "Thallium", col = rgb(red = 0.1, blue = 0.8, green = (8.0)

plot(Age, predict_regression, main = "Age", col = rgb(red = 0.5, blue = 0.2, green = 0.8))
 plot(Sex, predict_regression, main = "Sex", col = rgb(red = 0.6, blue = 0.3, green = 0.2))

plot(Chest_pain_type, predict_regression, main = "Chest pain type", col = rgb(red = 0.2,

 \rightarrow par(mfrow = c(3,4))

blue = 0.4, green = 0.6)

