CANCER PREDICATION

1)Initial Variables

```
age

ethnicity

ER

PR

RT

CT

HT

N

tumorStage

tumorSize

grade
```

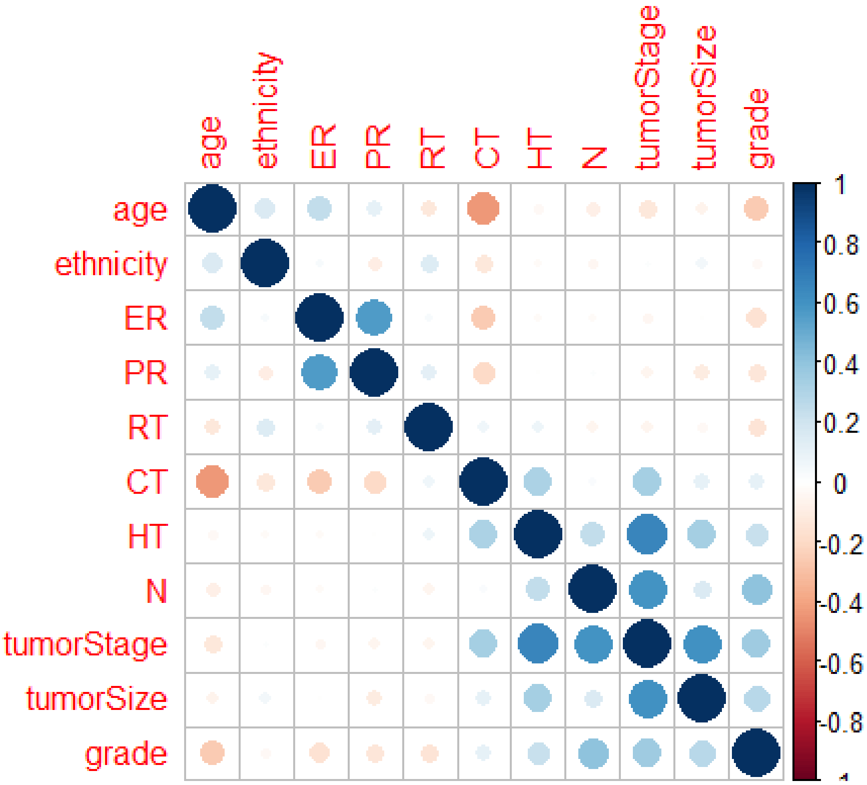
2)Pre-processing

1) Checked the summary of data

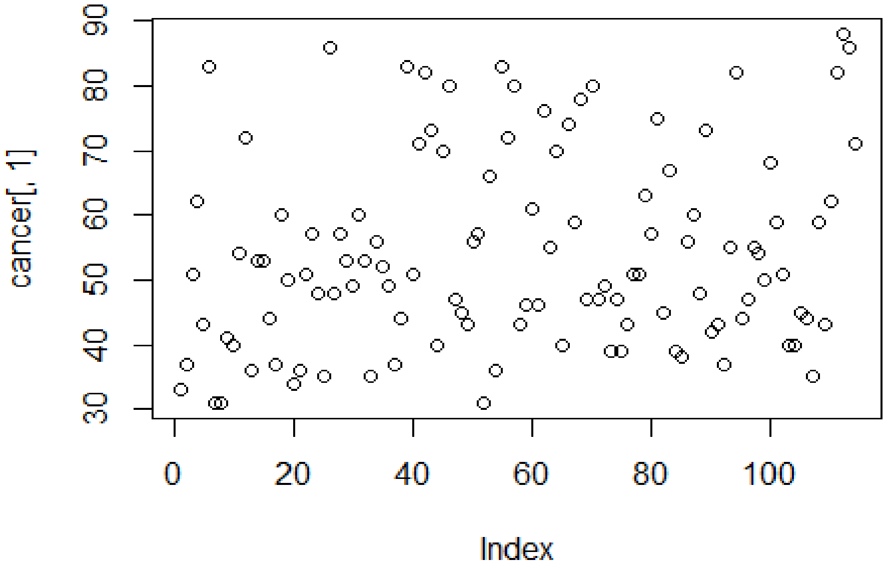
```
     age          ethnicity          ER               PR
Min.   :31.00   Min.   :0.00   Min.   :0.0000   Min.   :0.0000
1st Qu.:43.00   1st Qu.:1.00   1st Qu.:0.0000   1st Qu.:0.0000
Median :51.00   Median :2.00   Median :1.0000   Median :1.0000
Mean   :54.04   Mean   :1.57   Mean   :0.6579   Mean   :0.5702
3rd Qu.:62.00   3rd Qu.:2.00   3rd Qu.:1.0000   3rd Qu.:1.0000
Max.   :88.00   Max.   :2.00   Max.   :1.0000   Max.   :1.0000


    RT               CT               HT               N
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
Median :1.0000   Median :1.0000   Median :1.0000   Median :0.00000
Mean   :0.5351   Mean   :0.5439   Mean   :0.5702   Mean   :0.04386
3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000
Max.   :1.0000   Max.   :1.0000   Max.   :2.0000   Max.   :1.00000


  tumorStage      tumorSize         grade
Min.   :1.000   Min.   :0.000   Min.   :0.0000
1st Qu.:2.000   1st Qu.:1.625   1st Qu.:0.0000
Median :2.000   Median :2.300   Median :0.0000
Mean   :1.956   Mean   :2.689   Mean   :0.2193
3rd Qu.:2.000   3rd Qu.:3.475   3rd Qu.:0.0000
Max.   :4.000   Max.   :7.500   Max.   :1.0000
```

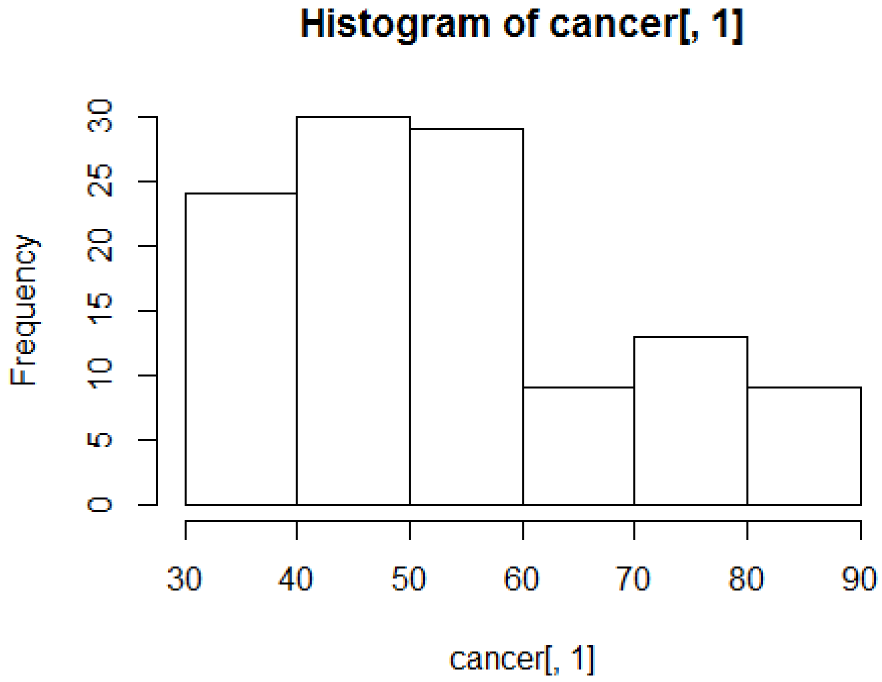
2) Checked the correlation of data

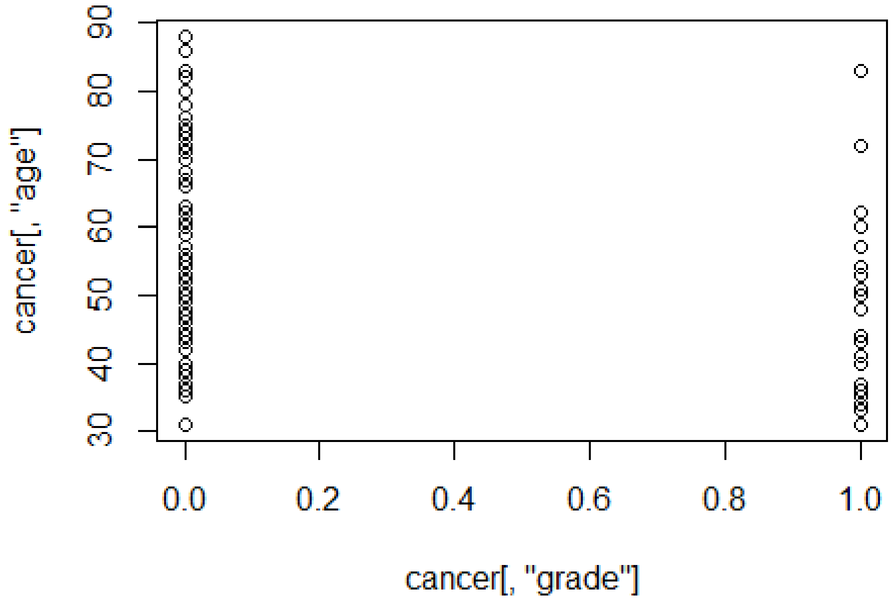
3) Some Variables are scaled for Normalization of data

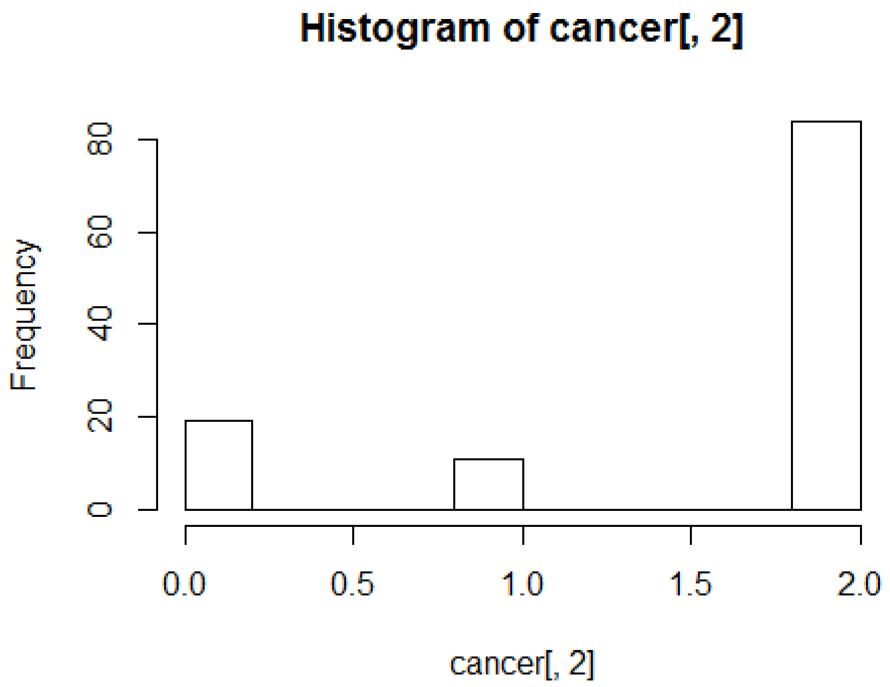


4) Observed the data by all means




Histogram of cancer[, 1]
Frequency
0
5
10
15
20
25
30
30
40
50
60
70
80
90
cancer[, 1]


cancer[, "age"]
30
40
50
60
70
80
90
0.0
0.2
0.4
0.6
0.8
1.0
cancer[, "grade"]



5) Checked for missing values

6) Outlier treatment is not done since most of them were categorical variables

3)Feature Selection

1) First of all, all the variables are taken into consideration

2) Then according step() function number of variables are reduced

4)Model Development

1) Generalized Linear Model (GLM) algorithm have been run

5)Model Testing and Accuracy Checking

1) RMSE is checked for developed model

2) Accuracy also is calculated

6)Codes

```
#Set working directory where my dataset resides
setwd("C:/Users/DELL/Desktop/Aegis/Machine Learning/Data")



#Read the .csv file of cancer dataset
cancer=read.csv("cancer.csv")

#checking the structure of data
head(cancer)

#checking summary of data
summary(cancer)

#Checking the Correlation of the cancer dataset's features
library("corrplot")
corrplot(cor(cancer))

#Ploting the first column i.e age of cancer dataset
plot(cancer[ ,1])

#Histogram of age column
hist(cancer[ ,1])

#plot grade vs age column of cancer dataset
plot(cancer[ ,"grade"],cancer[,"age"])

#Correlation of age and grade of cancer dataset
cor(cancer[ ,"grade"],cancer[,"age"])

#Histogram of ethnicity column of cancer dataset
hist(cancer[ ,2])
```

```
#correlation of ethnicity and grade of cancer dataset

cor(cancer[ ,2],cancer[,"grade"])


#checking is there any NA data in row

anyNA(cancer)


#if any NA then its row and column number i.e its position

which(is.na(cancer),T)


#Divide the grade 1 and grade 0 data

class1=subset(cancer,grade==1)

class0=subset(cancer,grade==0)


#taking sample of 70% grade 1 and grade 0 data

ind0=sample(1:nrow(class0),round(0.70*(nrow(class0))))

ind1=sample(1:nrow(class1),round(0.70*(nrow(class1))))


train1=class1[ind1,]

train0=class0[ind0,]

test1=class1[-ind1,]

test0=class0[-ind0,]


#creating final train and test data

train=rbind(train1,train0)

test=rbind(test1,test0)


#logistic model of cancer data

fit1=glm(grade ~ sqrt(age) + ethnicity + ER + PR + RT + CT + HT + factor(tumor
Stage) + tumorSize ,family=binomial("logit"),train)#71.651


#step function applied on fit1 model

step(fit1)
```

```
#created the new logistic model according to step() function
fit=glm(grade ~ sqrt(age) + RT + factor(tumorStage) + tumorSize ,family=binomial("logit"),train)#63.269

#plot of logistic model
plot(fit)

#checking is there any multicolinearity in logistic model
library("car")
vif(fit)

#predicting the grade of our test data
out=predict(fit,test,type="response")

#checking summary of our logistic model
summary(fit)

#Rounding of the grade vaue
out=ifelse(out>0.5,1,0)
out

#checking the accuracy
count=0
accuracy=0
for(i in 1:nrow(test)){
  if(out[i]==test[i,11]){
    count=count+1
  }
accuracy=c(accuracy,count/nrow(test))
}
accuracy

#checking root mean square error
RMSE=sqrt(mean((out-test["grade"])^2))
RMSE
```