Assignment 1 – Part B

KNN Classifier for **Breast Cancer Dataset**

**setwd("C:/Users/ASUS/Desktop/PR Lab/a1")**

**prc<-read.csv("Breast\_Cancer.csv",stringsAsFactors = FALSE)**

**str(prc)**

**prc<- prc[-1]**

**table(prc$diagnosis\_result)**

**prc$diagnosis<- factor(prc$diagnosis\_result, levels = c("B", "M"), labels = c("Benign", "Malignant"))**

**round(prop.table(table(prc$diagnosis)) \* 100, digits = 1)**

**normalize<- function(x) {**

**return ((x - min(x)) / (max(x) - min(x))) }**

**prc\_n<-as.data.frame(lapply(prc[2:9], normalize))**

**summary(prc\_n$radius)**

**prc\_train<- prc\_n[1:65,]**

**prc\_test<- prc\_n[66:100,]**

**prc\_train\_labels<- prc[1:65, 1]**

**prc\_test\_labels<- prc[66:100, 1]**

**install.packages(class)**

**library(class)**

**prc\_test\_pred<- knn(train = prc\_train, test = prc\_test,cl = prc\_train\_labels, k=10, prob=T)**

**#Evaluate model performance**

**library(gmodels)**

**CrossTable(x = prc\_test\_labels, y = prc\_test\_pred,prop.chisq=FALSE)**

**##Output:**

|  |
| --- |
| setwd("C:/Users/ASUS/Desktop/PR Lab/a1")  > prc<-read.csv("Breast\_Cancer.csv",stringsAsFactors = FALSE)  >  > str(prc)  'data.frame': 569 obs. of 32 variables:  $ id : int 842302 842517 84300903 84348301 84358402 843786 844359 84458202 844981 84501001 ...  $ diagnosis\_result : chr "M" "M" "M" "M" ...  $ radius\_mean : num 18 20.6 19.7 11.4 20.3 ...  $ texture\_mean : num 10.4 17.8 21.2 20.4 14.3 ...  $ perimeter\_mean : num 122.8 132.9 130 77.6 135.1 ...  $ area\_mean : num 1001 1326 1203 386 1297 ...  $ smoothness\_mean : num 0.1184 0.0847 0.1096 0.1425 0.1003 ...  $ compactness\_mean : num 0.2776 0.0786 0.1599 0.2839 0.1328 ...  $ concavity\_mean : num 0.3001 0.0869 0.1974 0.2414 0.198 ...  $ concave.points\_mean : num 0.1471 0.0702 0.1279 0.1052 0.1043 ...  $ symmetry\_mean : num 0.242 0.181 0.207 0.26 0.181 ...  $ fractal\_dimension\_mean : num 0.0787 0.0567 0.06 0.0974 0.0588 ...  $ radius\_se : num 1.095 0.543 0.746 0.496 0.757 ...  $ texture\_se : num 0.905 0.734 0.787 1.156 0.781 ...  $ perimeter\_se : num 8.59 3.4 4.58 3.44 5.44 ...  $ area\_se : num 153.4 74.1 94 27.2 94.4 ...  $ smoothness\_se : num 0.0064 0.00522 0.00615 0.00911 0.01149 ...  $ compactness\_se : num 0.049 0.0131 0.0401 0.0746 0.0246 ...  $ concavity\_se : num 0.0537 0.0186 0.0383 0.0566 0.0569 ...  $ concave.points\_se : num 0.0159 0.0134 0.0206 0.0187 0.0188 ...  $ symmetry\_se : num 0.03 0.0139 0.0225 0.0596 0.0176 ...  $ fractal\_dimension\_se : num 0.00619 0.00353 0.00457 0.00921 0.00511 ...  $ radius\_worst : num 25.4 25 23.6 14.9 22.5 ...  $ texture\_worst : num 17.3 23.4 25.5 26.5 16.7 ...  $ perimeter\_worst : num 184.6 158.8 152.5 98.9 152.2 ...  $ area\_worst : num 2019 1956 1709 568 1575 ...  $ smoothness\_worst : num 0.162 0.124 0.144 0.21 0.137 ...  $ compactness\_worst : num 0.666 0.187 0.424 0.866 0.205 ...  $ concavity\_worst : num 0.712 0.242 0.45 0.687 0.4 ...  $ concave.points\_worst : num 0.265 0.186 0.243 0.258 0.163 ...  $ symmetry\_worst : num 0.46 0.275 0.361 0.664 0.236 ...  $ fractal\_dimension\_worst: num 0.1189 0.089 0.0876 0.173 0.0768 ...  > prc<- prc[-1]  > table(prc$diagnosis\_result)  B M  357 212  >  > prc$diagnosis<- factor(prc$diagnosis\_result, levels = c("B", "M"), labels = c("Benign", "Malignant"))  > round(prop.table(table(prc$diagnosis)) \* 100, digits = 1)  Benign Malignant  62.7 37.3  >  > normalize<- function(x) {  + return ((x - min(x)) / (max(x) - min(x))) }  >  > prc\_n<-as.data.frame(lapply(prc[2:9], normalize))  > summary(prc\_n$radius)  Min. 1st Qu. Median Mean 3rd Qu. Max.  0.0000 0.2233 0.3024 0.3382 0.4164 1.0000  >  > prc\_train<- prc\_n[1:65,]  > prc\_test<- prc\_n[66:100,]  >  > prc\_train\_labels<- prc[1:65, 1]  > prc\_test\_labels<- prc[66:100, 1]  >  > install.packages(class)  Error in install.packages : 'match' requires vector arguments  > library(class)  >  > prc\_test\_pred<- knn(train = prc\_train, test = prc\_test,cl = prc\_train\_labels, k=10, prob=T)  >  > #Evaluate model performance  > library(gmodels)  > CrossTable(x = prc\_test\_labels, y = prc\_test\_pred,prop.chisq=FALSE)    Cell Contents  |-------------------------|  | N |  | N / Row Total |  | N / Col Total |  | N / Table Total |  |-------------------------|    Total Observations in Table: 35    | prc\_test\_pred  prc\_test\_labels | B | M | Row Total |  ----------------|-----------|-----------|-----------|  B | 13 | 6 | 19 |  | 0.684 | 0.316 | 0.543 |  | 1.000 | 0.273 | |  | 0.371 | 0.171 | |  ----------------|-----------|-----------|-----------|  M | 0 | 16 | 16 |  | 0.000 | 1.000 | 0.457 |  | 0.000 | 0.727 | |  | 0.000 | 0.457 | |  ----------------|-----------|-----------|-----------|  Column Total | 13 | 22 | 35 |  | 0.371 | 0.629 | |  ----------------|-----------|-----------|-----------| |
|  |
| |  | | --- | | > | |

**Assignment 1 – Part A**

Basics Of R Programming on **Big Mart Sales Dataset**

**#inserting data**

sales <- read.csv("C:/Users/ASUS/Desktop/PR Lab/a1/yearly\_sales.csv")

head(sales)

summary(sales)

**#plotting graph**

plot(sales$num\_of\_orders,sales$sales\_total,

main = "Number of Orders vs. Sales")

results <- lm(sales$sales\_total ~ sales$num\_of\_orders)

summary(results)

hist(results$residuals, breaks = 800)

**#writing to new file**

df <- read.csv("C:/Users/ASUS/Desktop/PR Lab/a1/yearly\_sales.csv")

print (df)

write.csv(df,"C:/Users/ASUS/Desktop/PR Lab/a1/write.csv", row.names = FALSE)

**#vector**

vector <- read.csv('C:/Users/ASUS/Desktop/PR Lab/a1/write.csv')

v1 <- vector[[1]] # by column number

v2 <- vector[["cust\_id"]] # by column name

v3 <- vector$cust\_id # by column name

print (v3)

**#Arrays**

result <- array(c(v3),dim = c(3,3,2))

print(result)

**#matrix**

mat <- as.matrix(df)

class(mat)

str(mat)

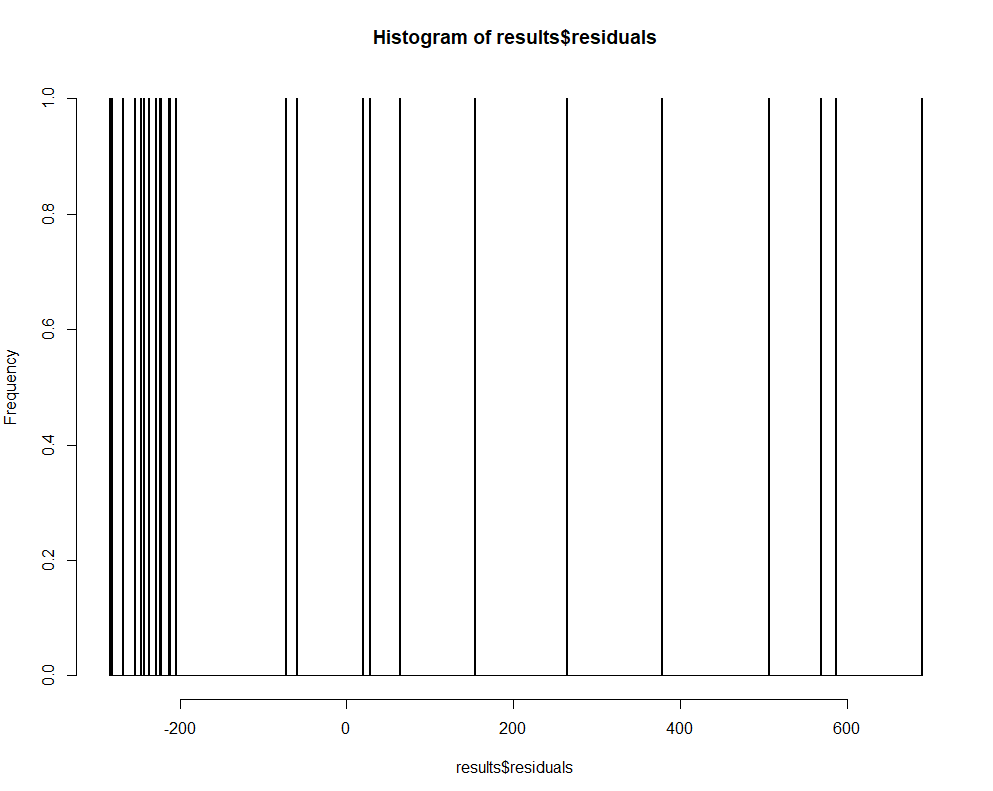
nrow(mat)

ncol(mat)

mat[1, 1]

mat[1:4, 1:2]

**##Output:**



##Output:

> #inserting data

>

> sales <- read.csv("C:/Users/ASUS/Desktop/PR Lab/a1/Big\_Mart\_Sales.csv")

> head(sales)

  cust\_id sales\_total num\_of\_orders gender

1  100001         800            56      M

2  100002         217            78      F

3  100003          75            34      F

4  100004         678            23      F

5  100005         324            44      M

6  100006         567            11      F

> summary(sales)

    cust\_id       sales\_total    num\_of\_orders    gender

 Min.   :1e+05   Min.   : 11.0   Min.   :  4.00   F:13

 1st Qu.:1e+05   1st Qu.: 56.0   1st Qu.: 17.00   M:12

 Median :1e+05   Median : 89.0   Median : 45.00

 Mean   :1e+05   Mean   :292.7   Mean   : 61.68

 3rd Qu.:1e+05   3rd Qu.:456.0   3rd Qu.: 66.00

 Max.   :1e+05   Max.   :985.0   Max.   :456.00

>

> #plotting graph

>

> plot(sales$num\_of\_orders,sales$sales\_total,

+      main = "Number of Orders vs. Sales")

>

> results <- lm(sales$sales\_total ~ sales$num\_of\_orders)

> summary(results)

Call:

lm(formula = sales$sales\_total ~ sales$num\_of\_orders)

Residuals:

   Min     1Q Median     3Q    Max

-284.9 -238.0 -204.8  154.0  689.1

Coefficients:

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

(Intercept)         304.5249    79.6654   3.823 0.000873 \*\*\*

sales$num\_of\_orders  -0.1920     0.7532  -0.255 0.801015

---

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

Residual standard error: 323.6 on 23 degrees of freedom

Multiple R-squared:  0.002818,  Adjusted R-squared:  -0.04054

F-statistic: 0.06501 on 1 and 23 DF,  p-value: 0.801

>

> hist(results$residuals, breaks = 800)

>

> #writing to new file

>

> df <- read.csv("C:/Users/ASUS/Desktop/PR Lab/a1/Big\_Mart\_Sales.csv")

> print (df)

   cust\_id sales\_total num\_of\_orders gender

1   100001         800            56      M

2   100002         217            78      F

3   100003          75            34      F

4   100004         678            23      F

5   100005         324            44      M

6   100006         567            11      F

7   100007         872             7      M

8   100008          44            89      F

9   100009          67            66      M

10  100010         234            55      M

11  100011         344           123      M

12  100012         236           456      F

13  100013         876            78      F

14  100014         985            45      M

15  100015          56             9      M

16  100016          89            10      F

17  100017          35            81      F

18  100018         456            13      M

19  100019          11            45      F

20  100020          56            55      F

21  100021          87            66      M

22  100022          88            21      M

23  100023          65            56      F

24  100024          33            17      M

25  100025          22             4      F

>

> write.csv(df,"C:/Users/ASUS/Desktop/PR Lab/a1/write.csv", row.names = FALSE)

>

> #vector

>

> vector <- read.csv('C:/Users/ASUS/Desktop/PR Lab/a1/write.csv')

>

> v1 <- vector[[1]]  # by column number

> v2 <- vector[["cust\_id"]]  # by column name

> v3 <- vector$cust\_id  # by column name

>

> print (v3)

 [1] 100001 100002 100003 100004 100005 100006 100007 100008 100009 100010 100011 100012 100013 100014

[15] 100015 100016 100017 100018 100019 100020 100021 100022 100023 100024 100025

>

> #Arrays

>

> result <- array(c(v3),dim = c(3,3,2))

> print(result)

, , 1

       [,1]   [,2]   [,3]

[1,] 100001 100004 100007

[2,] 100002 100005 100008

[3,] 100003 100006 100009

, , 2

       [,1]   [,2]   [,3]

[1,] 100010 100013 100016

[2,] 100011 100014 100017

[3,] 100012 100015 100018

>

>

> #matrix

>

> mat <- as.matrix(df)

> class(mat)

[1] "matrix"

> str(mat)

 chr [1:25, 1:4] "100001" "100002" "100003" "100004" "100005" "100006" "100007" "100008" "100009" ...

 - attr(\*, "dimnames")=List of 2

  ..$ : NULL

  ..$ : chr [1:4] "cust\_id" "sales\_total" "num\_of\_orders" "gender"

> nrow(mat)

[1] 25

> ncol(mat)

[1] 4

> mat[1, 1]

 cust\_id

"100001"

> mat[1:4, 1:2]

     cust\_id  sales\_total

[1,] "100001" "800"

[2,] "100002" "217"

[3,] "100003" " 75"

[4,] "100004" "678"