Date 08/12/2022

No. Ex:- 09 Page No.

Experiment - 09

AIM: Analysing Boston Data.

Commands:-

```
> install.packages("tree")
> library(MASS)
> options(sciper=4)
> view(Boston)
```

O/P:-

It shows 1 to 11 of 506 enteries, 14 total columns ...n Indus cha nen rm age...

```
> set.seed(1)
> train <- sample(1:nrow(Boston), nrow((Boston)/2)
```

O/P

Values:-

train int [1:253] 505 234 167 129 ...

```
> tree.boston <- tree(med ~., Boston, subset=train)
> install.packages("tree")
> tree.boston <- tree(med ~.Boston, subset=train)
```

O/P

$ frame data.frame 130 obs of 5 variables

```
> summary(tree.boston)
```

O/P

Regression tree.

tree (formula = medv ~ ., Boston, subset=train)

Variables actually used in tree construction

[1] "rm" "lstat" "crim" "age"

Number of terminal nodes: 7

Teacher's Signature

Date

Expt. No. _____ Page No. _____

Residual mean deviance: 10.38 = 2555/246

Distribution of Residuals

Min	1st Qu	Median	Mean	3rd Qu	Max
-10.180	-1.77	-0.1775	0.0000	1.9230	16.580

→ Run cv to find best level at which to prune.

> cv.boston <- cv.tree(tree.Boston)

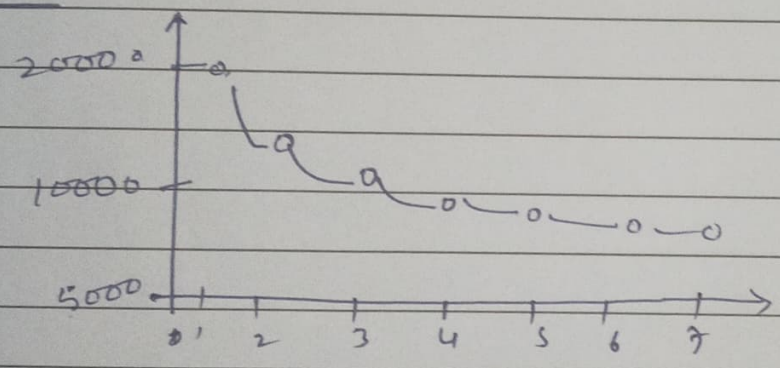
O/P

cv.boston list of 4

→ Construct a plot (dev = MSE on y-axis)

plot (cv.boston $ size, cv.boston $ dev, type='b')

O/P :-



> prune.boston <- prune.tree(tree.boston, best=5)

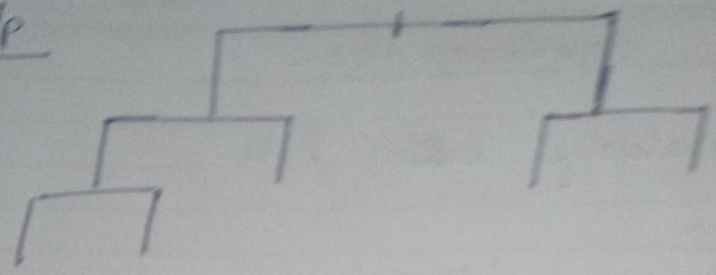
O/P

prune.boston list of 6

> plot (prune boston)

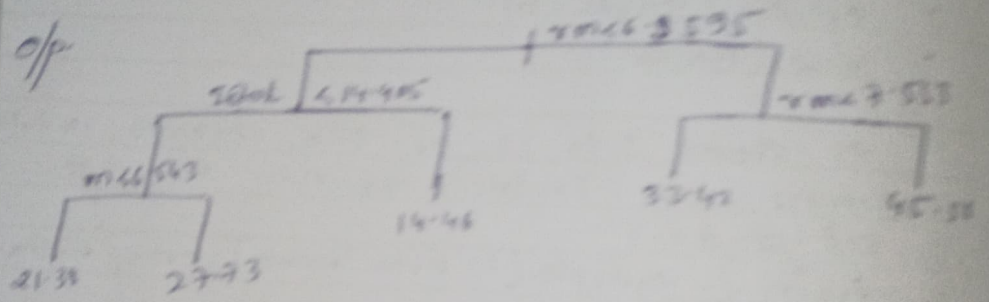
Teacher's Signature _____

o/p



> text (pruned_boston, pretty=0)

o/p



Date

Expt. No. Ex: 10

Page No.

Experiment - 10

AIM:- To given dataset (electrical consumption of an organization) It contains the monthly electrical consumption & annual average for various years

Program:-

Plot the various across the years. Is the consumption pattern comparable across the years?

```
> df <- read.csv("C:\\Users\\Rahul\\Desktop")
  print(df)
> Install.package("ggplot2")
> x = df[,]
> y = df[2,]
  plot(x)
  plot(y)
> x = df[1:5,]
> y = df[1,4]
  plot(x)
> plot(x, y, col='blue', xlab="years", ylab="Jan")
> df = (("Jan, Feb, march, April, May, June, July, August,
      September, October, November, December"))
> plot(x, df, col="red")
> cv.boston <- cv.tree tree.Boston)
> plot(cv.boston <- prone.tree(tree.boston, best=5)
> Plot(Prune, boston)
> text(prune.Boston, pretty=0)
```

Teacher's Signature

Date

Expt. No. ~~Ex. 0.11~~ 11 Page No.

Experiment-11

AIM:- Classification on Boston Dataset

Program:-

```
> Library(MASS)
> Installed.packages("tree")
> library(tree)
> options (scipen=4)
> view (Boston)
> set.seed (1)
> train <- sample(1:nrow(Boston), nrow(Boston)/2)
> tree.Boston <- tree (medv ~., Boston, subset = train)
> summary (tree Boston)
```

O/P

Regression tree

tree (formula = medv ~., data = Boston, subset = train)

variables actually used in tree construction

[1] "rm" "lstat" "crim" "age"

No. of terminal nodes: 7

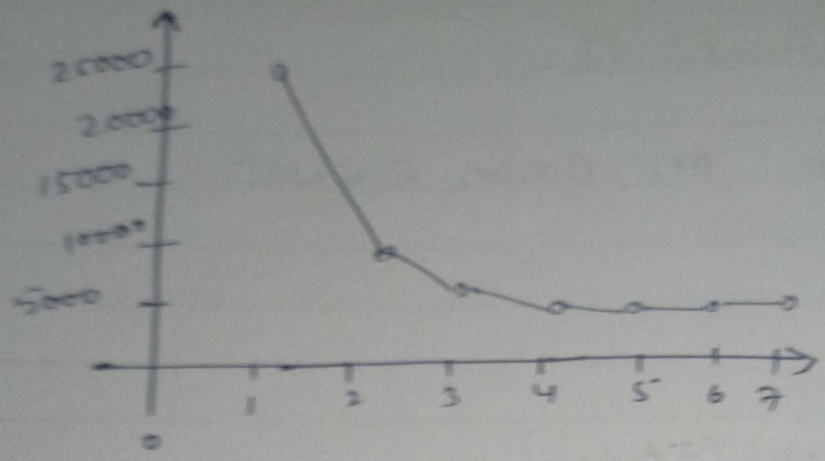
Residual mean deviance: 10.38 - 2555/246

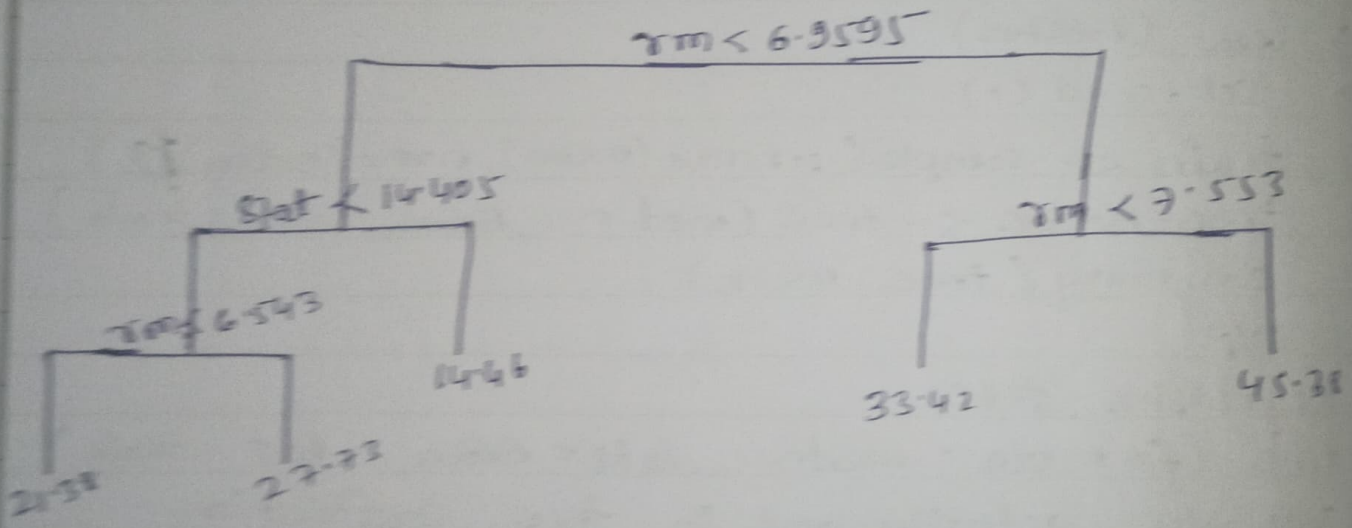
Distribution of residuals.

Min	1st Qu	Median	Mean	3rd Qu	Max
-10.080	-1.770	-0.1775	0.0000	1.9830	16.580

→ Run CV (cross validation) to final best level at which to prune.

Teacher's Signature


20000
15000
5000
0
1
2
3
4
5
6
7


rm < 6.9595
lstat < 14.405
rm < 7.553
rm < 6.543
14.46
33.42
45.38
21.38
23.22

Date

Expt. No.

Page No.

Km out list of 9

> table (km.out & cluster, class, lb)

	1	2
1	2	2
2	4	23

Teacher's Signature

Date

Expt. No. 12

Page No.

Experimental - 12

AIM: To learn & practice more ~~product~~ about K-means clustering.

Programs:

```
> library ("ggplot2")
> set.seed(2)
> x <- matrix (nor(50*2), ncol=2)
> gplot (x= x[,1], y=x[,2]
> x[1:25,1] <- x[1:25,1]+1
> x[1:25,2] <- x[1:25,2]-1
> gplot (x= x[1], y= x[,2]
```

Shift the 1st 25 pts to have mean (1,-1) instead of (0,0)

```
> x[1:25,1] <- x[1:25,1]+4
> x[1:25,2] <- x[1:25,2]-4
> gplot (x= x[,1], y= x[,2])
> class.lbl <- as.factor ((rep(1,25), rep(2,25)))

class.lbl factor w/2 levels "1","2": 1111...

> gplot (x= [,1], y= x[,2], color = class.lbl.
      size = I(3)) + theme.b/w()
> km.out <- kmeans (x,2, nstart=20)
```

Teacher's Signature

Date

Ex. No. 13

Page No.

Experiment - 13

AIM :- Hadoop Implementation on Single Node on Windows

Program :-

Install Java, Hadoop :-

Step 1 :

→ Verify the Java installed :- $ Javac -version

→ Step 2 :- Extract Hadoop at C:\Hadoop

→ Step 3 :- setting up the Hadoop_Home variable by using Windows environment variable. setting for Hadoop Path setting.

Step 4 :- Set Java - Home variable.
Use Windows enviornment variable setting for Java Path setting.

Step 5 :- Set Hadoop & Java bin directory Path. Copy the path & of the file & create a new in edit environment variable.

Step 6 :- Hadoop configuration
for Hadoop configuration we need to modify six files that are below :-

(1) Core-site.xml (2) Mapred-site.xml
(3) Hdfs-site.xml (4) YARN-site.xml.

Teacher's Signature

Date

Expt. No.

Page No.

Hadoop-env.cmd
Create two folder datanode & namenode.

Step 7:- Format the namenode folder:-
open command window & give the code
$ hdfs namenode -format
→ this will show that we have successfully install Hadoop.

Step 8:- Testing The setup:- (Running the Hadoop)
~~$ start-all.cmd~~
$ start-dfs.cmd and $ start-yarn.cmd.

~~Step 9:-~~

output:- localhost active:-
Started: Tue Jan 12:46:15 ICT 2023
Version: 2.7.1
Compiled: 2023-01-17 by Rahul Yadav.
cluster ID: CID 9.2636
Block Pool ID: BP-27.

Step 8:- Stop Hadoop:-
$ stop-dfs.cmd.

Teacher's Signature

Date

Expt. No. 14

Page No.

Experiment-14

AIM: Visualization & modification on Diabetes dataset.

Program:

① Determine the structure of the data
diabetes = read.csv ("diabetes.csv")
summary (diabetes)
str (diabetes) & Head (diabetes)
Mean_Glucose ← mean (diabetes $ Glucose)
median_Glucose ← median (diabetes & Glucose)

② Sd_Glucose ← sd (diabetes & Glucose)
Var_Glucose ← var (diabetes & Glucose)

~~③ Numeric.var ← sapply~~

③ Perform corelation against each other (install corrplot)
→ install.packages ('corrplot')
→ library ('corrplot')
→ numeric.var <- sapply (diabetes, is.numeric)
→ Corr.matrix <- cor (diabetes [,numeric.var])
→ corrplot (corr.matrix, main = "\n\n Correlation Plot for
→ "Numerical variables", order = "h.clust", tl.col = "black"
→ , tl.srt = 45, tl.cex = 0.8, cl.cex = 0.8)
→ box (which = "outer", lty = "solid")

Decision Tree: library (diabetes)
model1 ← rpt (outcome ~ Pressure + Glucose + BMI
data = tree, anova = "class")

Teacher's Signature

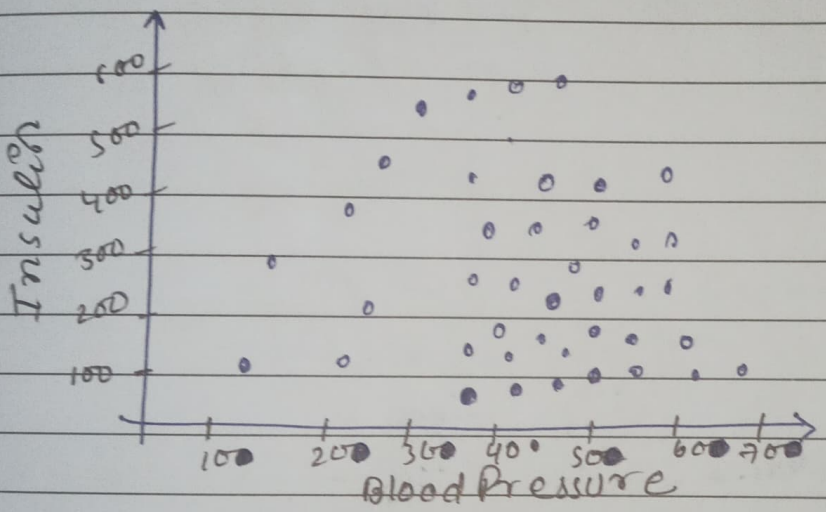
Date

Expt. No. Page No.

~~Decision Tree~~ :- plot (model1, vm → true)
box (which = 'outer', lty = 'solid').

Scatterplot :

```
sns.set_style('darkgrid')
grid = sns.JointGrid(x = "Blood Pressure", y = "Insulin",
                     data = diabetes);
grid.plot(sns.scatterplot)
```



OUTPUT :-

74.56 % of accuracy in decision Tree

The data has various parameter & decision tree is used because it can consider all these parameters as target & its contribution tell us which comes literally off diabetes in Glt class.

Teacher's Signature

Expt. No. 15 Date

Page No.

Experiment-15

AIM:- Visualize & Analyze on cat data set in R.

Program :-

```
library (MASS)
data(cats, package = 'MASS)
str(cats)
head (cats)
library(lattice)

xyplot (Hwt ~ Bwt | Sex, data = cats, style = c("p", "g", "l"),
        xlab = "Baby cat (Kg)", glof = "Heart by ()")

xyplot (Hwt ~ Bwt, gb = sex, data = cats, style = c("p", "g", "l"),
        xlab = "using(keyword", plot = "But (grou), as = True).
→ install.packages("e1071") . → library('e1071")
m ← svm (sex ~, dat = cats)
fcat(m, data = cats)
summary (m)
data.frame (plot(m), cats $ sex)

mt ← svm (sex., data = cats, gama = 0.5", cost = 678)
plot (sex1, data = cats)
table (true = cats $ sex, future = predict (x2))
```

Teacher's Signature

Date

Expt. No. Page No.

OUTPUT:-

→ The initial unproved model gave a function accuracy of about 53%.

→ The proved model with Gamma = 0.5 & cost = 67 but a particular of 98.9% accuracy.

→ No SVM is not suitable for these costs, sex departs the weights & height weight & size. Therefore we need to use a multiple regression model which assist ~~could~~ less outliers.

OUTPUT	Prediction	
TRUE	F	M
F	33	14
M	11	86.

Teacher's Signature