CSE – 3020

Data Visualization

Lab DA – 4

Name	**:**	**Anish Desai**
Reg. No.	**:**	**20BCE0461**
Slot	**:**	**L39 + L40**
Guided by	**:**	**Prof. Jyotismita Chaki**

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Question – 1 :

Create a multivariate linear regression model. Interpret the result in terms of the important features (density, block, fertilizer) needed to increase the "yield" amount. With some dummy data predict the value of "yield".

Code :

Data Viz   Lab DA-4.

Anish Desai
20BCE0461

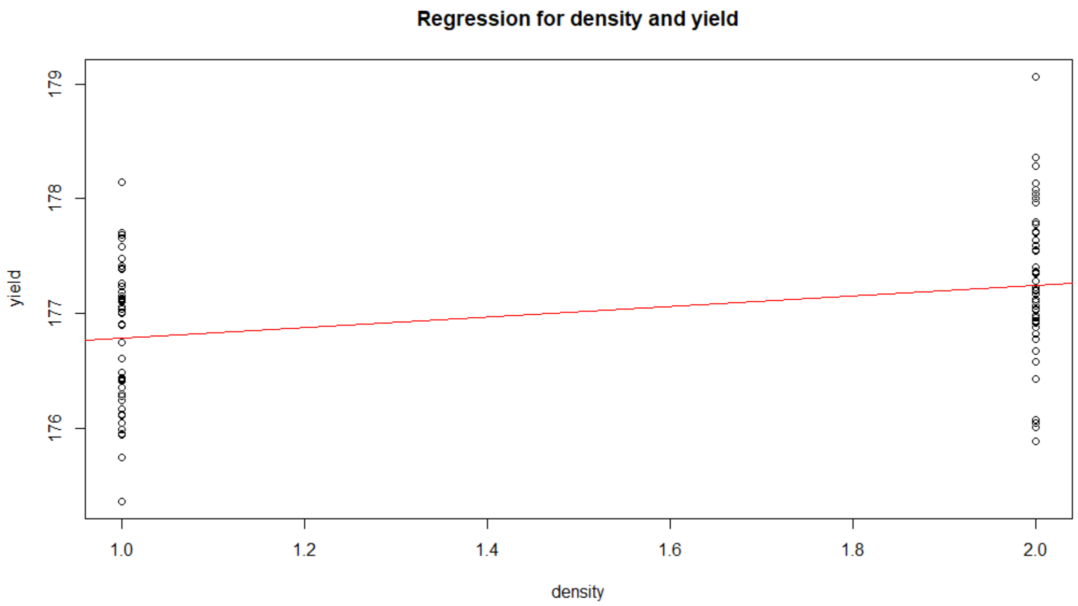
```
#Q1
library(tidyverse)
library(caret)
theme_set(theme_bw())
#Loading the dataset
harvest = read.csv("C:/Users/ANISHDESAI/Documents/Lab_DataViz
                                      /harvest.csv")

#View the dataset
view(harvest)

#Split the data into training and test set
#To make it reproduceable - same sample test for every run
set.seed(123)
#yield is the dependent variable
training.samples <- harvest$yield %>%
                 #80% training and 20% testing sample
                 createDataPartition(p=0.8, list = FALSE)
train.data <- harvest[training.samples, ]
test.data <- harvest[-training.samples, ]
```

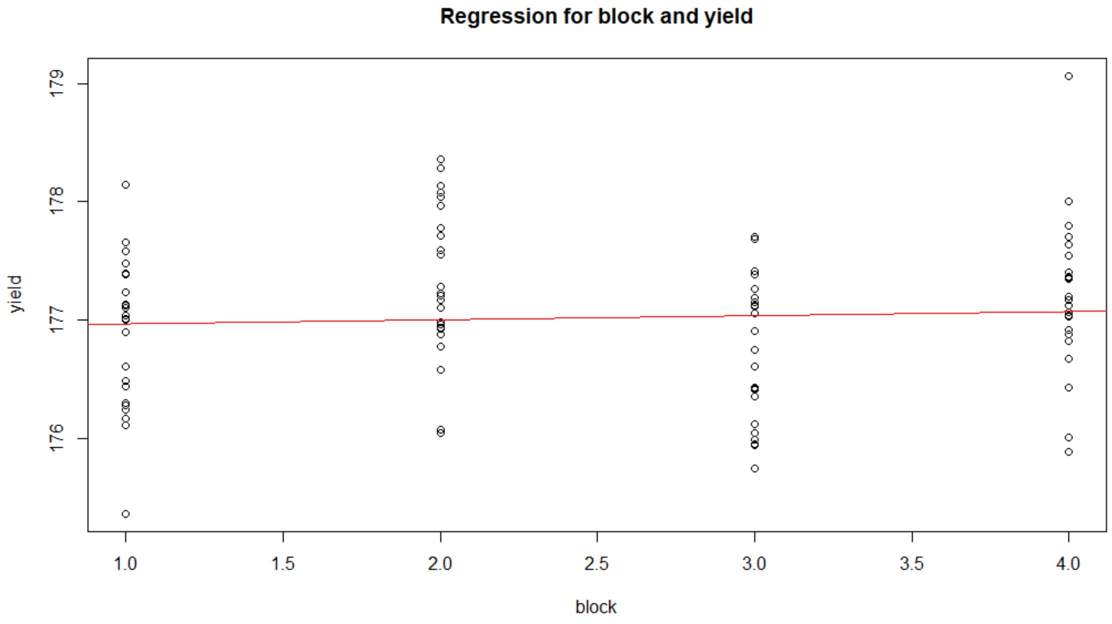
```
#Build the model
model <- lm(yield ~. , data=train.data)
# Summarize the model
summary(model)

#Plot LR
plot(harvest$density, harvest$yield, main = "Regression
        for density and yield", xlab = 'density',
                                ylab = 'yield')

abline (lm(yield~ density, data=harvest), col = 'red')

plot (harvest$block, harvest$yield, main = 'Regression
        for block and yield', xlab = 'block',
                              ylab = 'yield')
abline (lm(yield ~block, data=harvest), col='red')

plot (harvest$fertilizer , harvest$yield, main = 'Regression
        for fertilizer and yield', xlab = 'fertilizer',
                                   ylab = 'yield')
abline(lm(yield~fertilizer, data=harvest), col ='red')

#Predict value using LR
density =2
block =4
fertilizer =3
```

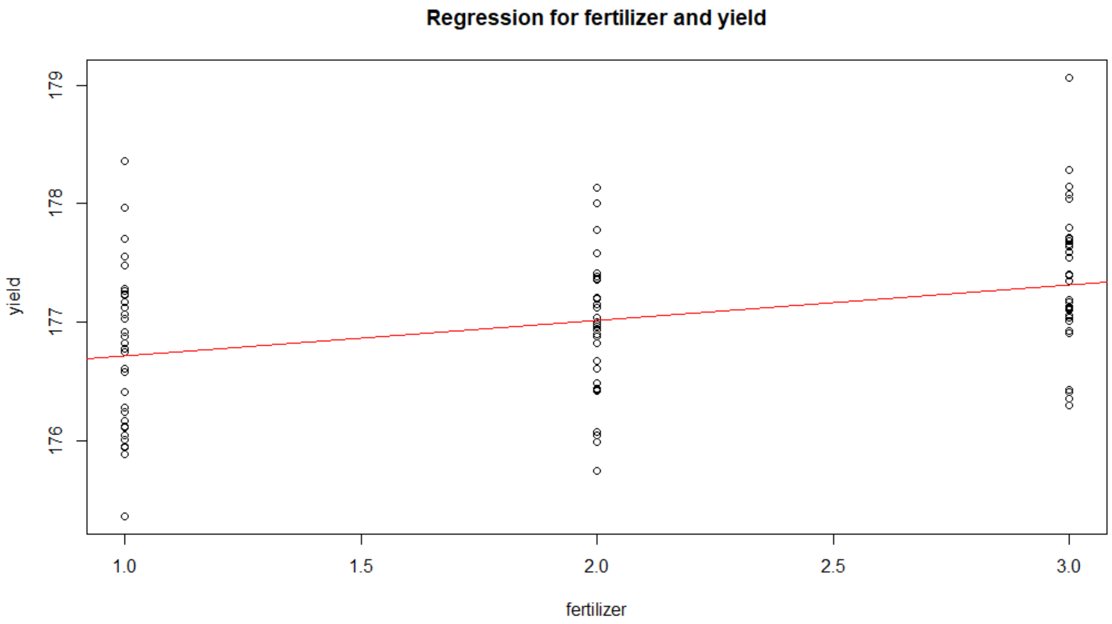
```
data_harvest = data.frame(density, block, fertilizer)
data_harvest
prediction <- predict(model, data_harvest)
prediction

# Load Library
library(hydroGOF)
predictYlinregress <- predict(model, test.data)
RMSE(train.data$yield, predictYlinregress)
```

Output :

```
> summary(model)

Call:
lm(formula = yield ~ ., data = train.data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.30515 -0.39533 -0.04853  0.39115  1.59464 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 175.80469    0.28063 626.469  < 2e-16 ***
density       0.56056    0.14288   3.923 0.000190 ***
block        -0.09507    0.06576  -1.446 0.152383    
fertilizer    0.30691    0.08150   3.766 0.000325 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5863 on 76 degrees of freedom
Multiple R-squared:  0.2672,	Adjusted R-squared:  0.2383 
F-statistic: 9.239 on 3 and 76 DF,  p-value: 2.774e-05
```








```
> data_harvest
  density block fertilizer
1       2     4          3
> prediction <- predict(model, data_harvest)
> prediction
       1
177.4663

> RMSE(train.data$yield, predictYlinregress)
[1] 0.7683697
```

Predicted value for dummy data and RMSE value of the model

Interpretation :

Interpretation #Q1

Using the Linear regression, we have found the relationship between the dependent variable 'yield' and independent variables 'density', 'block' and 'fertilizer'.

From the co-efficients section, we can infer that the LR is of the form

yield = 0.56 * density + 0.31 * fertilizer − 0.095 * block + 175.804

This implies that density and fertilizer are more important and significant to determine yield while 'block' is not an important factor

Another important parameter is t-value. The larger values of density and fertilizer implies greater association with outcome variable yield and thus block is of lesser significance and can be removed for a better model

Residual standard error is 0.5863 which is low, hence better model.

For dummy values of density = 2, block = 4 and fertilizer = 3, the predicted value of 'yield' is 177.4663

RMSE is a measure of performance of regression models.

For this model, RMSE = 0.768 which is quite less and hence can infer that our model is good.

Question – 2 :

Create a multivariate logistic regression model. Interpret the result in terms of the important features (density, block, fertilizer) needed to increase the "yield" amount. With some dummy data predict the value of "yield".

Code :

```
#Q2
# loading the library
library(rattle.data)
# Dataset 'harvest' already loaded
# Checking the structure of harvest dataset
str(harvest)
```

```
#Prep training and test data
library (dplyr)
#Using sample_frac to create 70-30 split into test
                                                and train

train <- sample_frac (harvest, 0.7)
sample_id <- as.numeric(rownames(train))
```

```
test <- harvest[-sample_id,]
require(nnet)
#Training the multinomial model
# 'yield' is the dependent variable
multinom.fit <- multinom(yield ~., data=train)

#Checking the model
summary(multinom.fit)

#Predicting values for train dataset
train$predicted <- predict(multinom.fit, newdata =
                                          train, "class")

#Building classification table
ctable <- table(train$yield, train$predicted)
#Calculating accuracy
round(sum(diag(ctable))/sum(ctable))*100, 2)

#Predicting values for test dataset
test$predicted <- predict(multinom.fit, newdata=
                                          test, "class")
#Building classification table
ctable <- table(test$yield, test$predicted)
#Calculating accuracy
round(sum(diag(ctable))/sum(ctable))*100,2)
```

```
#Predict value using LogR
density = 2
block = 4
fertilizer=3
data_harvest = data.frame(density, block, fertilizer)
data_harvest
prediction <- predict(multinom.fit, data_harvest)
prediction
```

Output :

```
Residual Deviance: 236.8636
AIC: 764.8636
```

```
> #Predicting the values for train dataset
> train$precticed <- predict(multinom.fit, newdata = train, "class")
> #Building classification table
> ctable <- table(train$yield, train$precticed)
> #Calculating accuracy - sum of diagonal elements divided by total obs
> round((sum(diag(ctable))/sum(ctable))*100,2)
[1] 14.93
```

Accuracy for training dataset

```
> #Predicting the values for test dataset
> test$precticed <- predict(multinom.fit, newdata = test, "class")
> #Building classification table
> ctable <- table(test$yield, test$precticed)
> #Calculating accuracy - sum of diagonal elements divided by total obs
> round((sum(diag(ctable))/sum(ctable))*100,2)
[1] 0
```

Accuracy for testing dataset

```
> data_harvest
  density block fertilizer
1       2     4          3
> prediction <- predict(multinom.fit, data_harvest)
> prediction
[1] 176.430830126686
```

Predicted value for dummy data

Interpretation :

Interpretation #Q2

The residual deviance of the model is 236.8636

Greater the value of residual deviance, poorer is the model

Hence, we can infer from this that our model is somewhat bad.

We can also see that the while the accuracy of the model for training dataset is 14.93%, that for the testing dataset is 0%.

Thus, we can conclude that Logistic Regression model for the dataset is very poor and is almost inaccurate.

Question – 3 :

Create a support vector regression model. Interpret the result in terms of the important features (density, block, fertilizer) needed to increase the "yield" amount. With some dummy data predict the value of "yield".

Code :

```
#Q3

#Load required libraries
library(e1071)
library(hydroGOF)

#Plot
makePlot <- function(x,y) {
        plot(x, y, col = "black", pch=5, lwd=1)
        lines(x, y, lty=2, lwd=2)
        grid() }

#Predict value using SVM
density = 2
block = 4
fertilizer = 3
data_harvest = data.frame(density, block, fertilizer)
```

```
#SVM model 1
svm1 <- svm(yield ~ density, harvest)
#predicted values
predictedYsvm1 <- predict(svm1, harvest)
#Viz comparison
makePlot(harvest$density, harvest$yield)
title("Original data + SVR Model")
points(harvest$density, predictYsvm1, col='blue',
                                         pch=4)
points(harvest$density, predictYsvm1, col='blue',
                                         type='l')
```

```
#Checking the model
summary(svm1)
# Predicting result for some value of x
pred_svm1 <- predict (svm1, data_harvest $ density)
pred_svm1
#Comparing result with LR
lrm1 <- lm (yield ~ density, data =harvest)
predict Y linregress1 <- predict (lrm1, harvest)
RMSE (harvest $ yield, predict Y linregress1)
RMSE (harvest $ yield, predictYsvm1)
```

```
# SVM model 2
svm2 <- svm (yield ~ block, harvest)
#predicted values
predictedYsvm2 <- predict (svm2, harvest)
#Viz comparison
make Plot (harvest $ block, harvest $ yield)
title ("Original data + SVR Model")
points (harvest $ block, predictYsvm2, col = 'blue',
                                        pch=4)
points (harvest $ block, predict Ysvm2, col = 'blue',
                                        type = 'l')
```

```
#Checking the model
summary(svm2)
#Predicting result for some value of x
pred_svm2 <- predict(svm2, data_harvest $ block)
pred_svm2
#Comparing result with LR
lrm2 <- lm(yield ~ block, data=harvest)
predictYlinregress2 <- predict(lrm2, harvest)
RMSE(harvest$yield, predictYlinregress2)
RMSE(harvest$yield, predictYsvm2)
```

```
#SVM model 3
Svm3 <- svm(yield~fertilizer, harvest)
#predicted values
predictYsvm3 <- predict(svm3, harvest)
#Viz comparison
makePlot(harvest$fertilizer, harvest$yield)
title("Original data + SVR Model")
points(harvest$fertilizer, predictYsvm3, col='blue', pch=4)
points(harvest$fertilizer, predictYsvm3, col='blue', type='l')
```

```
#Checking the model
summary(svm3)
#Predicting result for some value of x
pred_svm3 <- predict(svm3, data_harvest$ fertilizer)
pred_svm3
#Comparing the result with LR
lrm3 <- lm(yield ~ fertilizer, data = harvest)
predictYlinregress3 <- predict(lrm3, harvest)
RMSE(harvest$yield, predictYlinregress3)
RMSE(harvest$yield, predictYsvm3)
```

Output :

```
> summary(svm1)

Call:
svm(formula = yield ~ density, data = harvest)


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1
      gamma:  1
    epsilon:  0.1


Number of Support Vectors:  88
```

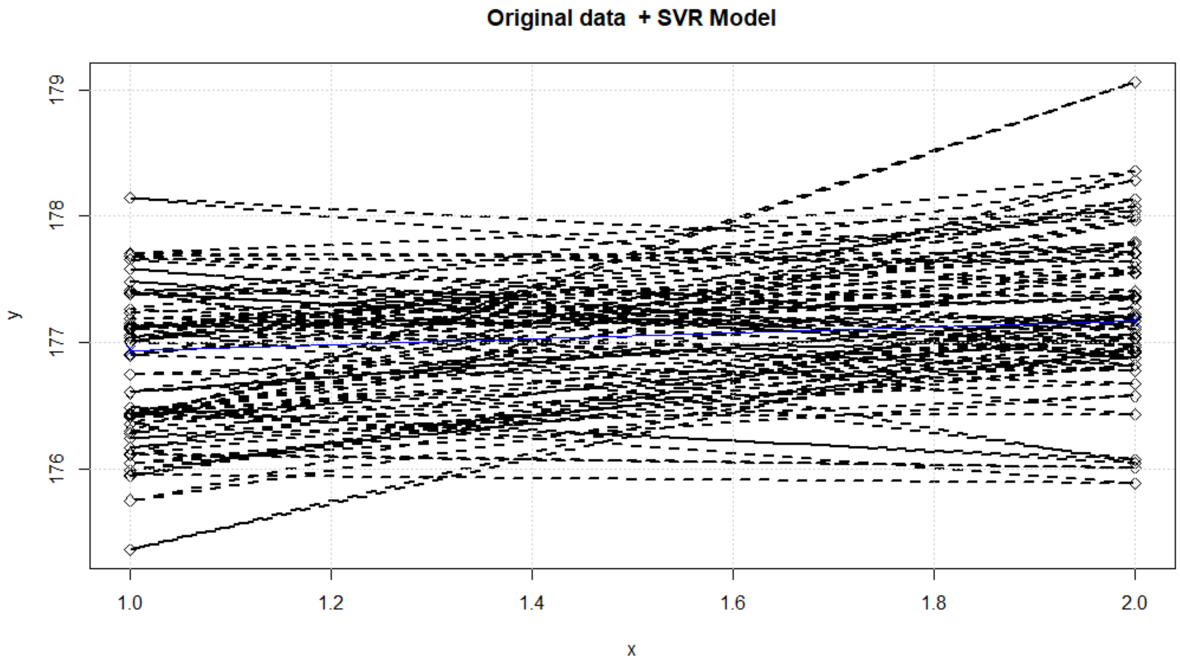
Summary of SVM Model 1 <u>(density vs yield)</u>

```
> #Predicting result for some value of x
> pred_svm1 <- predict(svm1,data_harvest$density)
> pred_svm1
       1
177.1669
```

Predicted value for dummy data

```
> RMSE(harvest$yield, predictYlinregress1)
[1] 0.619413
> RMSE(harvest$yield, predictYsvm1)
[1] 0.6307202
```

RMSE value of Linear Regression Model 1 and SVM Model 1



Plot of SVM Model 1

```
> summary(svm2)

Call:
svm(formula = yield ~ block, data = harvest)


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1
      gamma:  1
    epsilon:  0.1


Number of Support Vectors:  88
```

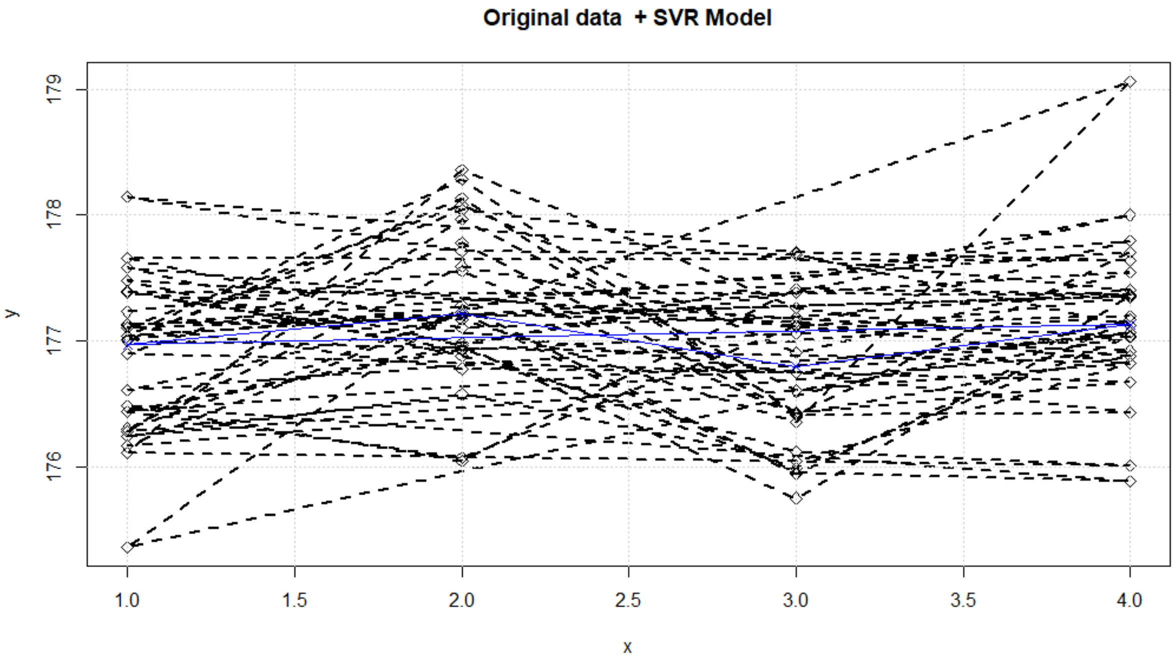
Summary of SVM Model 2 <u>(block vs yield)</u>

```
> #Predicting result for some value of x
> pred_svm2 <- predict(svm2,data_harvest$block)
> pred_svm2
       1
177.1277
```

Predicted value for dummy data

```
> RMSE(harvest$yield, predictYlinregress2)
[1] 0.6598871
> RMSE(harvest$yield, predictYsvm2)
[1] 0.6219541
```

RMSE value of Linear Regression Model 2 and SVM 2



Plot of SVM Model 2

```
> summary(svm3)

Call:
svm(formula = yield ~ fertilizer, data = harvest)


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1
      gamma:  1
    epsilon:  0.1


Number of Support Vectors:  89
```

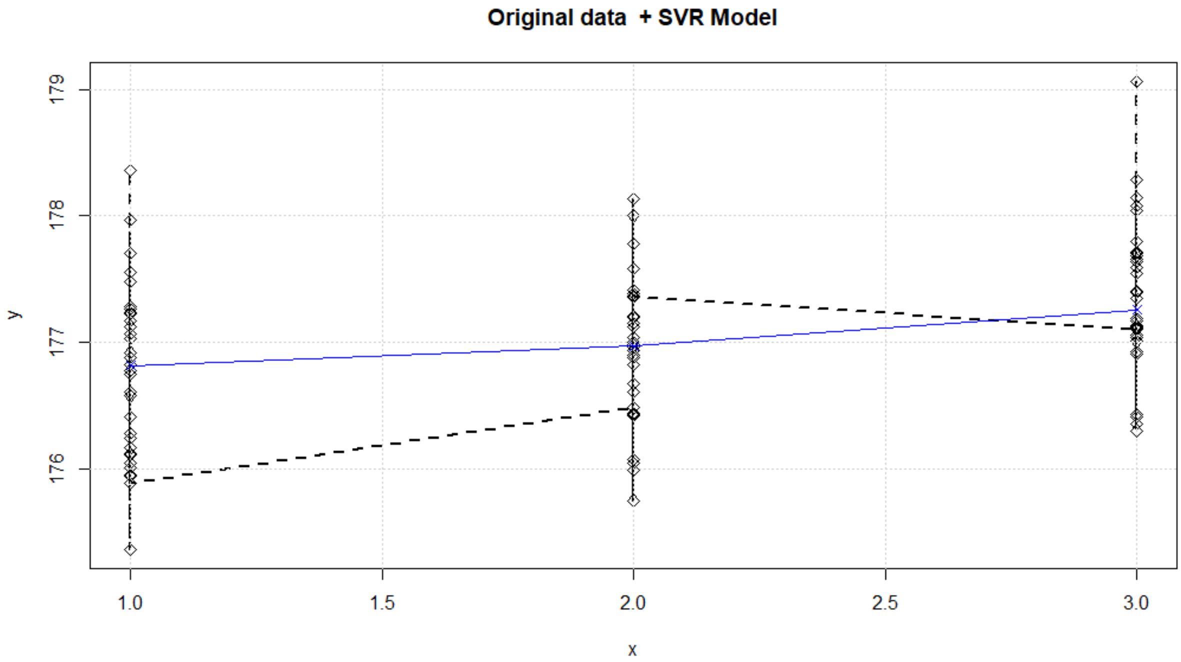
Summary of SVM Model 3 **<u>(fertilizer vs yield)</u>**

```
> #Predicting result for some value of x
> pred_svm3 <- predict(svm3,data_harvest$fertilizer)
> pred_svm3
       1
177.2543
```

Predicted value for dummy data

```
> RMSE(harvest$yield, predictYlinregress3)
[1] 0.6141645
> RMSE(harvest$yield, predictYsvm3)
[1] 0.6153963
```

RMSE value for Linear Regression Model 3 and SVM 3



Plot of SVM Model 3

Interpretation :

Interpretation #Q3

The first SVM Model - SVM Model 1 - is plotted in between 'density' and 'yield'. The RMSE value of SVM 1 is 0.63 while that of LR model for same variables is 0.62. This implies the LR model is slightly better than SVM model 1 for the given variables

For the SVM Model 2,
plot is in between 'block' and 'yield'.
RMSE of SVM 2 = 0.62
while that of LR 2 = 0.66
This implies the SVM Model 2 is better than LR Model for the given variables

For the SVM Model 3,
plot is in between 'fertilizer' and 'yield'.
RMSE of SVM 3 = 0.615
while that of LR 3 = 0.614
We can thus infer that both SVM and LR are equally good for the given variables

Corresponding plots and predictions for dummy data has been made.

Question – 4 :

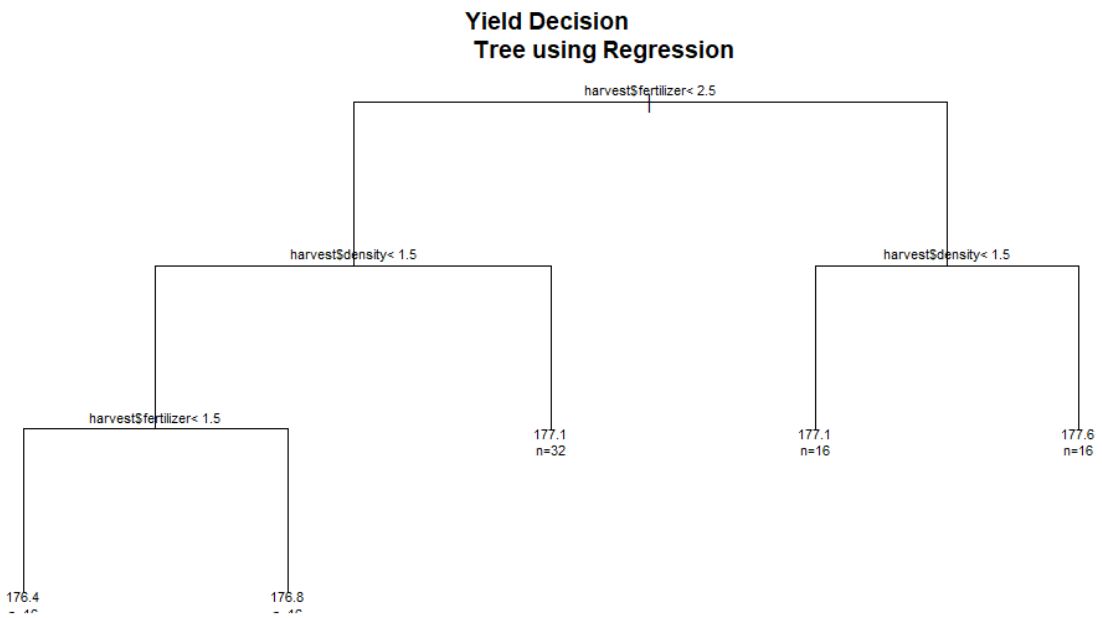
Create a decision tree regression model. Interpret the result. With some dummy data predict the value of "yield".

Code :

```
# Q4
# Load the Package
library (rpart)

# Create decision tree using regression
# For regression, method = 'anova'
# Predict yield using density, block and fertilizer
fit <- rpart(harvest $ yield ~ harvest$density +
                    harvest $ block + harvest $ fertilizer,
                    method = 'anova', data = harrest)

#Plot
plot (fit, uniform = TRUE, main = "Yield Decision
                    Tree using Regression")

text (fit, use.n = TRUE, cex = .7)

#Print model
print(fit)

#Create test data
df_dtr <- data.frame (density = 2, block = 4,
                    fertilizer = 3)
```

```
# Predicting yield
# using testing data and model
predict (fit, df_dtr, method = 'anova')

# Checking Performance
pred_dtr <- predict(fit, harvest, method = 'anova')
# Building classification table
ctable <- table(harvest $ yield, pred_dtr)
# Calculating accuracy
round((sum(diag(ctable))/sum(ctable))*100, 2)
```

Output :



Decision Tree Model

```
> #Print model
> print(fit)
n= 96

node), split, n, deviance, yval
      * denotes terminal node

1) root 96 41.954230 177.0155
  2) harvest$fertilizer< 2.5 64 25.255410 176.8451
    4) harvest$density< 1.5 32 10.467060 176.6089
      8) harvest$fertilizer< 1.5 16  5.152701 176.4396 *
      9) harvest$fertilizer>=1.5 16  4.396675 176.7783 *
    5) harvest$density>=1.5 32 11.218180 177.0813 *
  3) harvest$fertilizer>=2.5 32 11.127340 177.3562
    6) harvest$density< 1.5 16  3.905502 177.1356 *
    7) harvest$density>=1.5 16  5.665117 177.5767 *
```

Summary of Decision Tree Model

```
> #using testing data and model
> predict(fit, df_dtr, method = "anova")
       1        2        3        4        5 
176.4396 177.0813 176.4396 177.0813 176.4396 
```

Predicted value for dummy data

```
> #Checking Performance
> pred_dtr <- predict(fit,harvest,method="anova")
> #Building classification table
> ctable <- table(harvest$yield, pred_dtr)
> #Calculating accuracy - sum of diagonal elements divided by total obs
> round((sum(diag(ctable))/sum(ctable))*100,2)
[1] 3.12
```

Accuracy of the model

Interpretation :

Interpretation #Q4

From the decision tree model graph and the corresponding summary of the model, we can infer that :

(i) fertilizer < 2.5 values are the root node. There are 64 such observations.

(ii) If the density < 1.5, it further checks if the fertilizer < 1.5, the Value outcome is 176.4 or else 176.8

If density >= 1.5, then the outcome is 177.1 and there are 32 such observations

Else
(iii) If fertilizer >= 2.5, it goes to the right of root node. (32 such observations)

It further checks if density < 1.5, then outcome is 177.1

There are 16 such observations (n=16)

If density >= 1.5, then outcome is 177.6 (n=16 observations)

This is the way DT traverses to obtain the outcome.

The accuracy of the decision tree model is obtained as 3.12 %, which implies the model is very poor.

Question – 5 :

Compare the linear regression model, logistic regression model and support vector regression model and state which model is the best one for this dataset along with proper logic.

Solution :

Q5

As discussed earlier from the outcomes, the accuracy rate for logistic regression model is 14.93 % for training dataset and ~0 % for testing dataset. This accuracy rate is the least among all the 3 and is very less, thus logistic regression model is not selected.

When we compare Linear Regression model and SVM model, the overall RMSE of LR model was 0.768 which is quite low.

Comparing feature-wise, we obtained

	LR	SVM
(i) density vs yield	0.62	0.63
(ii) block vs yield	0.66	0.62
(iii) fertilizer vs yield	0.614	0.615

'block' is a less significant feature and is least associated with 'yield'. Thus, considering the other two variables, RMSE of LR model is less than SVM model for both. Lesser the RMSE, better the model.

Thus, Linear Regression model is the best among the three models for the given dataset.