Lab Report 3

By: Abhishek Shah, Ravi Seth, Tanya Ralliaram

```
# Insert necessary packages
library('glmnet')
library('caret')
library('ISLR')
library('plotly')
library('gridExtra')
library('tree')
```

```
## Warning: package 'tree' was built under R version 4.0.4
```

```
library('rpart')
```

```
## Warning: package 'rpart' was built under R version 4.0.4
```

```
library('MLmetrics')
```

```
## Warning: package 'MLmetrics' was built under R version 4.0.4
```

```
library('e1071')
```

```
## Warning: package 'e1071' was built under R version 4.0.4
```

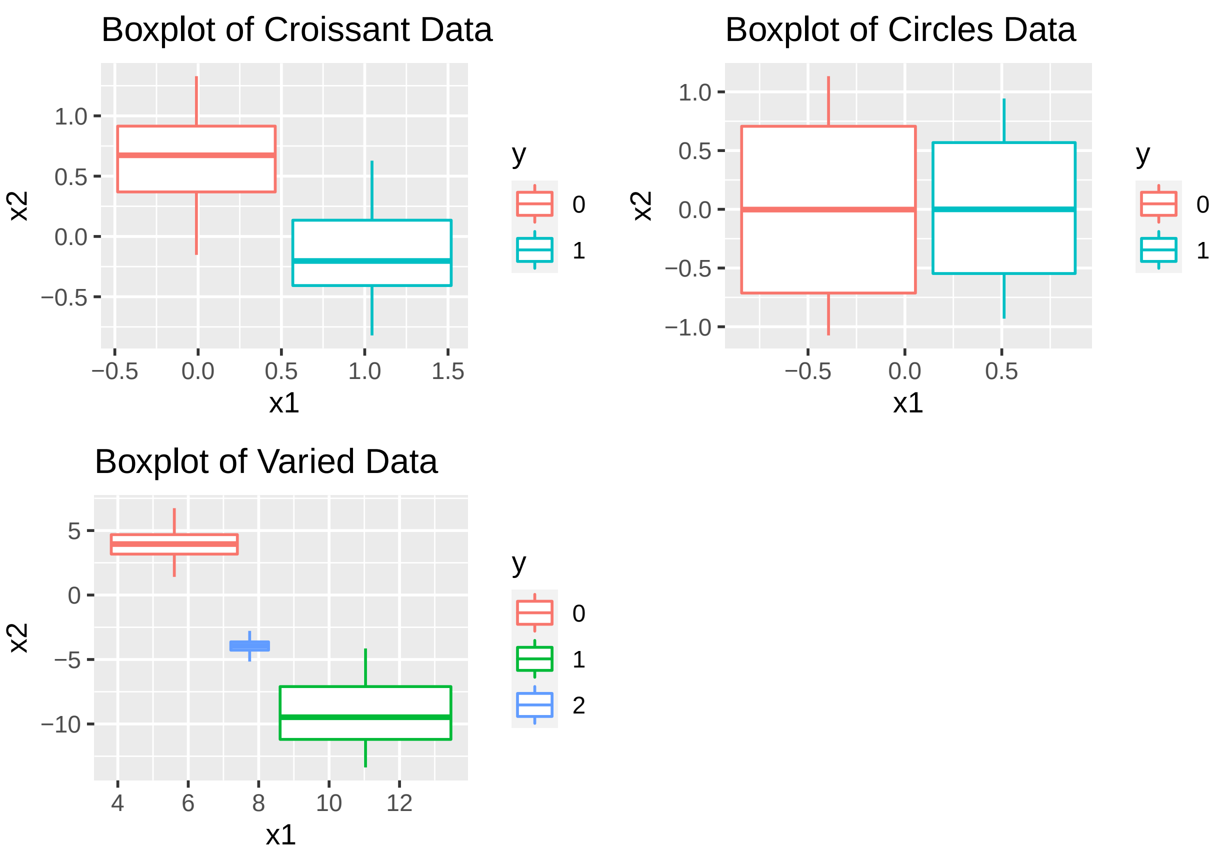
Question 1: Classification

```
# Read in data
croissant <- read.csv("data/croissant.csv")[,-1]
circles <- read.csv("data/circles.csv")[,-1]
varied <- read.csv("data/varied.csv")[,-1]
```

Question 1.1: Preprocess and Plot

```
croissant$y <- as.factor(croissant$y)
circles$y <- as.factor(circles$y)
varied$y <- as.factor(varied$y)
```

```
cro <- ggplot(data = croissant) +
  geom_boxplot(aes(x = x1, y=x2, colour=y)) +
  ggtitle("Boxplot of Croissant Data")
cir <- ggplot(data = circles) +
  geom_boxplot(aes(x = x1, y=x2, colour=y)) +
  ggtitle("Boxplot of Circles Data")
var <- ggplot(data = varied) +
  geom_boxplot(aes(x = x1, y=x2, colour=y)) +
  ggtitle("Boxplot of Varied Data")

grid.arrange(cro, cir, var, ncol=2)
```



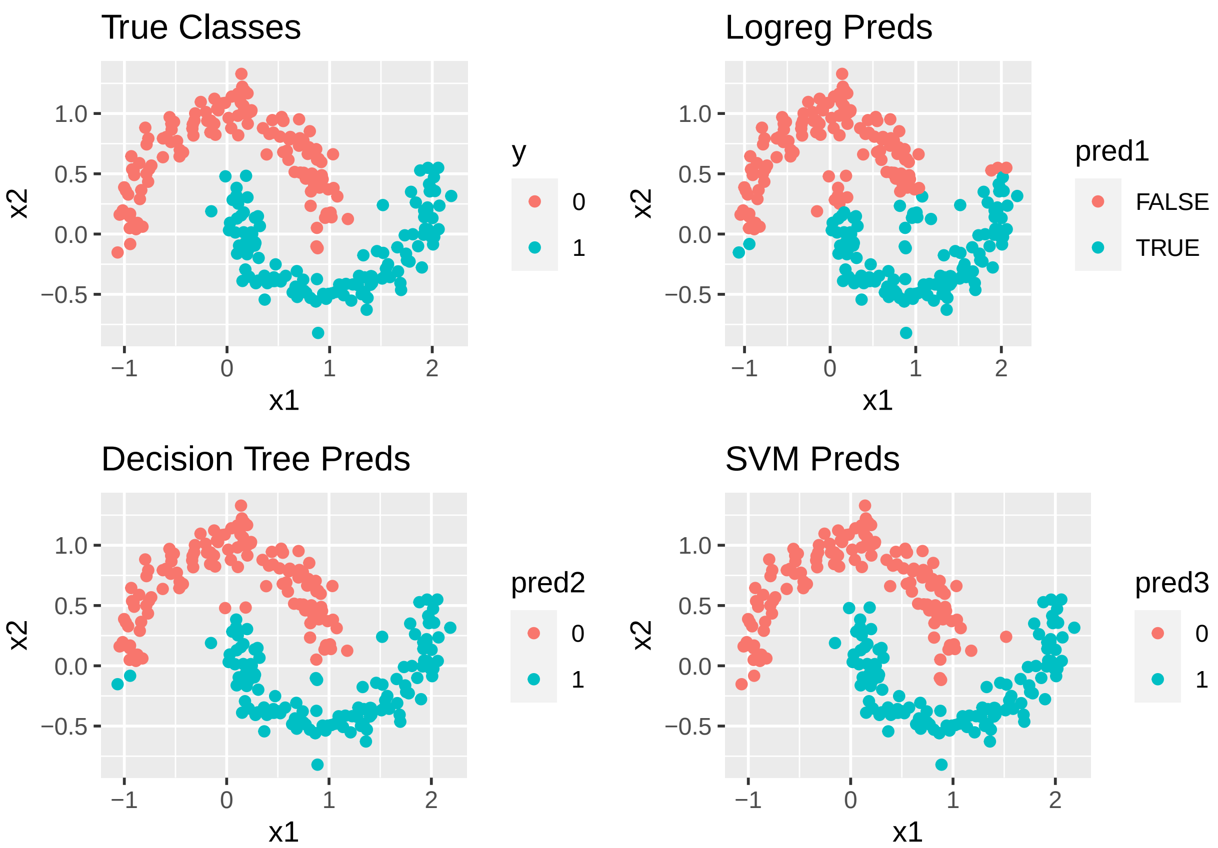
Questions 1.2-1.4 for Croissant Data

```
## Question 1.2
set.seed(112)

train_inds <- sample(1:nrow(croissant), floor(nrow(croissant)*0.5))
train <- croissant[ train_inds, ]
test  <- croissant[-train_inds, ]

y.train <- train$y
x.train <- model.matrix(y ~ .,train)[,-1]
x.test  <- model.matrix(y ~ .,test)[,-1]

## Question 1.3

# Logistic Regression
lreg <- glm(y ~ ., data=train, family = "binomial")
pred1 <- predict(lreg, newdata=as.data.frame(x.test), type = "response") > 0.5
lreg_acc <- mean(pred1 == (test$y==1))
lreg_con <- table(predict=pred1,actual=(test$y))

# Decision Tree
dtree <- rpart(y~., data=train)
pred2 <- predict(dtree, as.data.frame(x.test), type = "class")
dtree_acc <- Accuracy(pred2,test$y)
dtree_con <- table(predict=pred2,actual=(test$y))

# SVM
svmfit <- svm(y~.,data=train, kernel ="radial", gamma=1,cost=1)
pred3 <- predict(svmfit,as.data.frame(x.test), type="class")
print('We chose radial as the kernel as it best fits the shape of the data
            and thus should lead to a better prediction.')
```

```
## [1] "We chose radial as the kernel as it best fits the shape of the data \n            and thus
```

```
svm_acc <- Accuracy(pred3,test$y)
svm_con <- table(predict=pred3,actual=(test$y))

g1 <- ggplot(test, aes(x1,x2,colour=y)) +
  geom_point() +
  ggtitle("True Classes")
g2 <- ggplot(test, aes(x1,x2,colour=pred1)) +
  geom_point() +
  ggtitle("Logreg Preds")
g3 <- ggplot(test, aes(x1,x2,colour=pred2)) +
  geom_point() +
  ggtitle("Decision Tree Preds")
g4 <- ggplot(test, aes(x1,x2,colour=pred3)) +
  geom_point() +
```

```
  ggtitle("SVM Preds")

grid.arrange(g1,g2,g3,g4,ncol=2)
```




```
print('Looking at the four plots, we can see that Logisitc Regression has
      the most misclassifications and that Decision Tree performs pretty well,
      but not as well as SVM which seems to perform the best.')
```

```
## [1] "Looking at the four plots, we can see that Logisitc Regression has \n      the most miscla
```

```
sprintf("Logisitc Regression Accuracy: %f", lreg_acc)
```

```
## [1] "Logisitc Regression Accuracy: 0.904000"
```

```
sprintf("Decision Tree Accuracy: %f", dtree_acc)
```

```
## [1] "Decision Tree Accuracy: 0.976000"
```

```
sprintf("SVM Accuracy: %f", svm_acc)
```

```
## [1] "SVM Accuracy: 0.996000"
```

```r
print("In terms of accuracy, SVM has the highest and Decision Tree was second highest.
      Logistic Regression has the lowest out of the three.")
```

```
## [1] "In terms of accuracy, SVM has the highest and Decision Tree was second highest. \n      Lo
```

```r
lreg_con
```

```
##        actual
## predict   0   1
##   FALSE 112  12
##   TRUE   12 114
```

```r
dtree_con
```

```
##        actual
## predict   0   1
##       0 120   2
##       1   4 124
```

```r
svm_con
```

```
##        actual
## predict   0   1
##       0 124   1
##       1   0 125
```

```r
print("SVM was the least biased out of the three as it had zero False Positive (FP) and
      one False Negatives (FN). Decision Tree has 4 FP and 2 FN and Logistics Regression
      has 12 FP and FN.")
```

```
## [1] "SVM was the least biased out of the three as it had zero False Positive (FP) and \n      o
```

```r
## Question 1.4 for Croissant Data
set.seed(112)
```

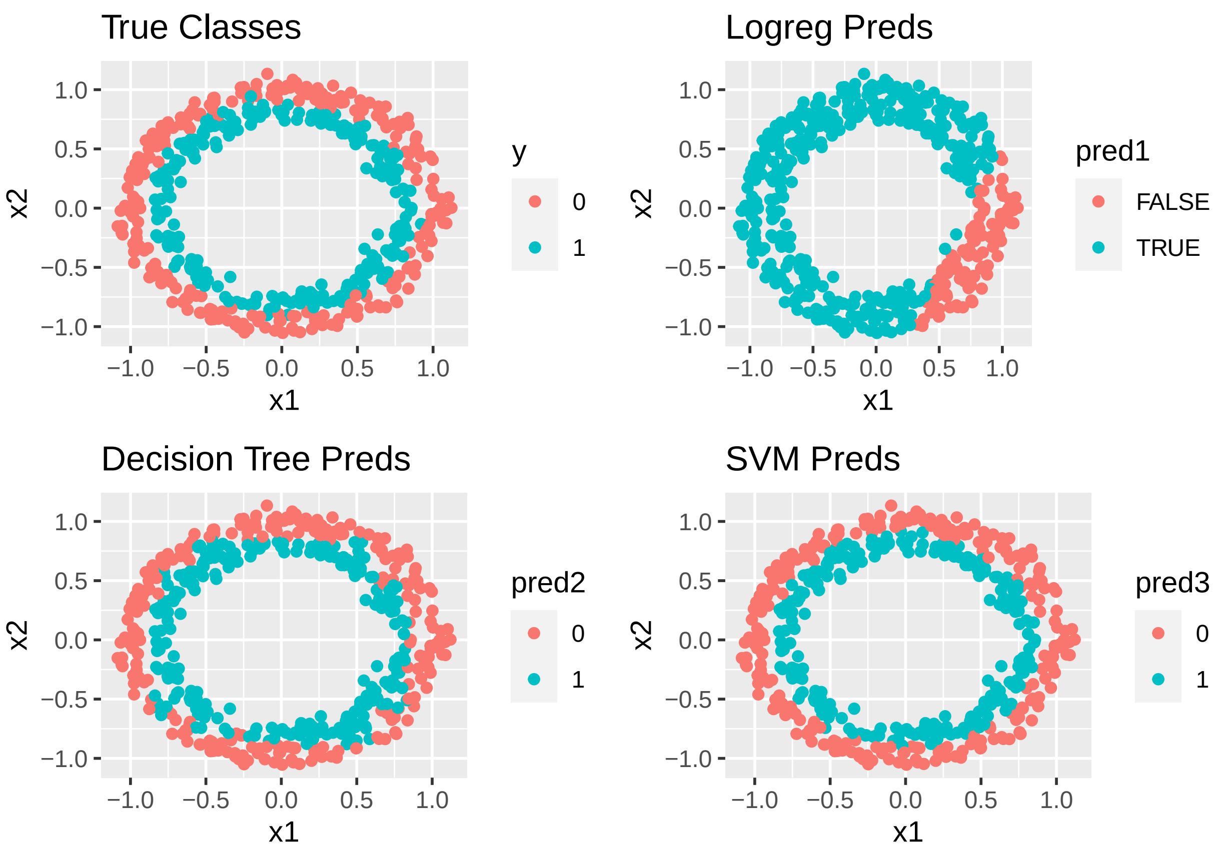
Questions 1.2-1.4 for Circle Data

```
## Question 1.2
set.seed(112)

train_inds <- sample(1:nrow(circles), floor(nrow(circles)*0.5))
train <- circles[ train_inds, ]
test  <- circles[-train_inds, ]

y.train <- train$y
x.train <- model.matrix(y ~ .,train)[,-1]
x.test  <- model.matrix(y ~ .,test)[,-1]

## Question 1.3

# Logistic Regression
lreg <- glm(y ~ ., data=train, family = "binomial")
pred1 <- predict(lreg, newdata=as.data.frame(x.test), type = "response") > 0.5
lreg_acc <- mean(pred1 == (test$y==1))
lreg_con <- table(predict=pred1,actual=(test$y))

# Decision Tree
dtree <- rpart(y~., data=train)
pred2 <- predict(dtree, as.data.frame(x.test), type = "class")
dtree_acc <- Accuracy(pred2,test$y)
dtree_con <- table(predict=pred2,actual=(test$y))

# SVM
svmfit <- svm(y~.,data=train, kernel ="radial", gamma=1,cost=1)
pred3 <- predict(svmfit,as.data.frame(x.test), type="class")
print('We chose radial as the kernel as it best fits the shape of the data
            and thus should lead to a better prediction.')
```

```
## [1] "We chose radial as the kernel as it best fits the shape of the data \n            and thus
```

```
svm_acc <- Accuracy(pred3,test$y)
svm_con <- table(predict=pred3,actual=(test$y))

g1 <- ggplot(test, aes(x1,x2,colour=y)) +
  geom_point() +
  ggtitle("True Classes")
g2 <- ggplot(test, aes(x1,x2,colour=pred1)) +
  geom_point() +
  ggtitle("Logreg Preds")
g3 <- ggplot(test, aes(x1,x2,colour=pred2)) +
  geom_point() +
  ggtitle("Decision Tree Preds")
g4 <- ggplot(test, aes(x1,x2,colour=pred3)) +
  geom_point() +
```

```r
  ggtitle("SVM Preds")

grid.arrange(g1,g2,g3,g4,ncol=2)
```




```r
print('Looking at the four plots, we can see that Logisitc Regression has
      a lot misclassifications and that Decision Tree performs well, but has noticeable
      misflassications. Again SVM seems to perform the best.')
```

```
## [1] "Looking at the four plots, we can see that Logisitc Regression has \n      a lot misclass
```

```r
sprintf("Logisitc Regression Accuracy: %f", lreg_acc)
```

```
## [1] "Logisitc Regression Accuracy: 0.506000"
```

```r
sprintf("Decision Tree Accuracy: %f", dtree_acc)
```

```
## [1] "Decision Tree Accuracy: 0.910000"
```

```r
sprintf("SVM Accuracy: %f", svm_acc)
```

```
## [1] "SVM Accuracy: 0.972000"
```

```r
print("In terms of accuracy, SVM has the highest and Decision Tree was second highest.
      Logistic Regression has the lowest out of the three.")
```

```
## [1] "In terms of accuracy, SVM has the highest and Decision Tree was second highest. \n      Lo
```

```r
lreg_con
```

```
##        actual
## predict   0   1
##   FALSE  58  45
##   TRUE  202 195
```

```r
dtree_con
```

```
##        actual
## predict   0   1
##       0 233  18
##       1  27 222
```

```r
svm_con
```

```
##        actual
## predict   0   1
##       0 253   7
##       1   7 233
```

```r
print("SVM was the least biased out of the three as it had 7 False Positives (FP) and
      7 False Negatives (FN). Decision Tree has 27 FP and 18 FN and Logistic Regression
      has 202 FP and 45 FN.")
```

```
## [1] "SVM was the least biased out of the three as it had 7 False Positives (FP) and \n      7 F
```

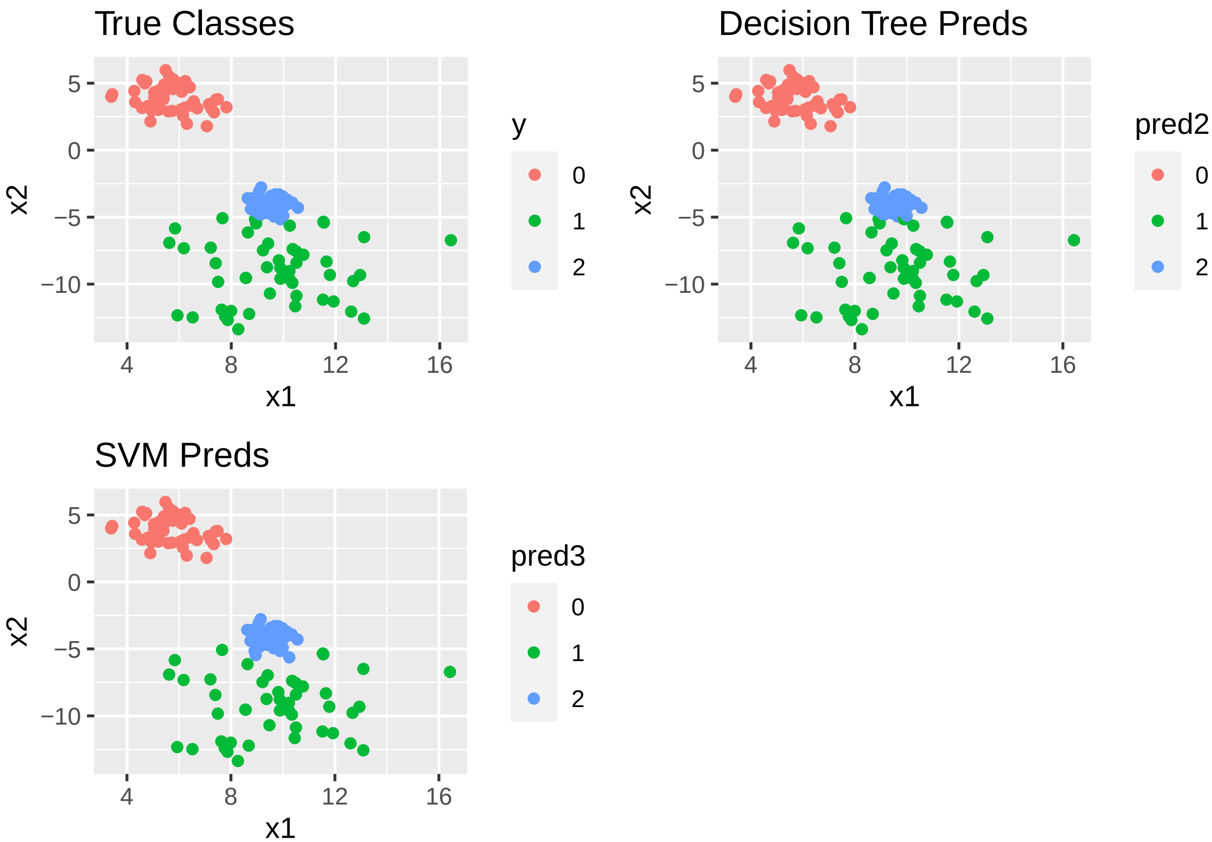
```r
## Question 1.4 for Circles Data
set.seed(112)
```

Questions 1.2-1.4 for Varied Data

```r
## Question 1.2
set.seed(112)

train_inds <- sample(1:nrow(varied), floor(nrow(varied)*0.5))
train <- varied[ train_inds, ]
test  <- varied[-train_inds, ]

y.train <- train$y
x.train <- model.matrix(y ~ .,train)[,-1]
x.test  <- model.matrix(y ~ .,test)[,-1]

## Question 1.3

# Decision Tree
dtree <- rpart(y~., data=train)
pred2 <- predict(dtree, as.data.frame(x.test), type = "class")
dtree_acc <- Accuracy(pred2,test$y)
dtree_con <- table(predict=pred2,actual=(test$y))

# SVM
svmfit <- svm(y~.,data=train, kernel ="radial", gamma=1,cost=1)
pred3 <- predict(svmfit,as.data.frame(x.test), type="class")
print('We chose radial as the kernel as it best fits the shape of the data
            and thus should lead to a better prediction.')
```

```
## [1] "We chose radial as the kernel as it best fits the shape of the data \n            and thus
```

```r
svm_acc <- Accuracy(pred3,test$y)
svm_con <- table(predict=pred3,actual=(test$y))

g1 <- ggplot(test, aes(x1,x2,colour=y)) +
  geom_point() +
  ggtitle("True Classes")
g3 <- ggplot(test, aes(x1,x2,colour=pred2)) +
  geom_point() +
  ggtitle("Decision Tree Preds")
g4 <- ggplot(test, aes(x1,x2,colour=pred3)) +
  geom_point() +
  ggtitle("SVM Preds")

grid.arrange(g1,g3,g4,ncol=2)
```




```r
print('Looking at the three plots, we can see that Logisitc Regression has
      a lot misclassifications and that Decision Tree performs well, but has noticeable
      misflassications. Again SVM seems to perform the best.')
```

```
## [1] "Looking at the three plots, we can see that Logisitc Regression has \n      a lot misclass
```

```r
sprintf("Decision Tree Accuracy: %f", dtree_acc)
```

```
## [1] "Decision Tree Accuracy: 0.986667"
```

```r
sprintf("SVM Accuracy: %f", svm_acc)
```

```
## [1] "SVM Accuracy: 0.980000"
```

```r
print("In terms of accuracy, Decision Tree performed a little bit better than SVM.")
```

```
## [1] "In terms of accuracy, Decision Tree performed a little bit better than SVM."
```

```r
dtree_con
```

```
##        actual
## predict  0  1  2
##       0 49  0  0
##       1  0 49  2
##       2  0  0 50
```

```
svm_con
```

```
##        actual
## predict  0  1  2
##       0 49  0  0
##       1  0 46  0
##       2  0  3 52
```

```
print("Decision Tree was the least biased has it had two false negatives for Class 2.
      SVM had 3 false positives for Class 1.")
```

```
## [1] "Decision Tree was the least biased has it had two false negatives for Class 2.\n      SVM
```

```
## Question 1.4 for Varied Data
set.seed(112)
```

Question 2: Tree-based methods

2.1. Preprocess

```
# 1
library("ISLR")
hitters <- Hitters
hitters$Salary <- log(hitters$Salary) # Q2 (Converted to log before dataset is split)
heart <- read.csv("data/Heart.csv")[-1] # Q3 (removed row identifier)

set.seed(112)
train_inds <- sample(1:nrow(hitters), floor(nrow(hitters)*0.7))
train.hitters <- hitters[ train_inds, ]
test.hitters  <- hitters[-train_inds, ]

train_inds <- sample(1:nrow(heart), floor(nrow(heart)*0.7))
train.heart <- heart[ train_inds, ]
test.heart  <- heart[-train_inds, ]
```

2.2. Decision Trees for Regression

```
# 1
```

2.3. Decision Trees for Classification

```
# 1
```

2.4. Bagging: Regression

```
# 1
```

Question 2.5. Bagging: Classification

```
# 1
```

Question 2.6. Random Forest: Regression

```
# 1
```