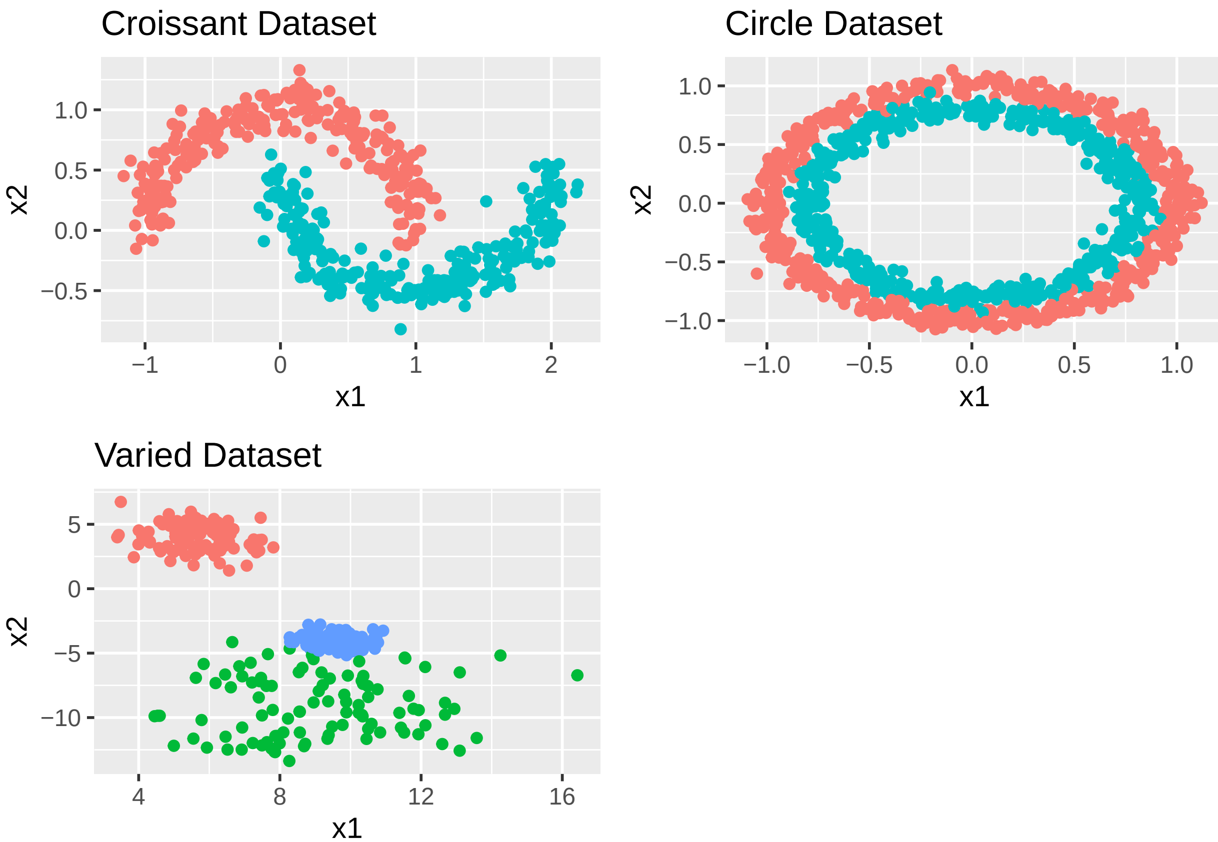
Assignment 3

Mitheysh Asokan

Question 1: Classification

1.1. Preprocess and Plot

```
croissant <- read.csv('croissant.csv')

circles <- read.csv('circles.csv')
varied <- read.csv('varied.csv')

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

g1 <- ggplot(croissant, aes(x1,x2,colour=y)) +
  geom_point() +
  ggtitle("Croissant Dataset") +
  theme(legend.position = "none")

g2 <- ggplot(circles, aes(x1,x2,colour=y)) +
  geom_point() +
  ggtitle("Circle Dataset") +
  theme(legend.position = "none")

g3 <- ggplot(varied, aes(x1,x2,colour=y)) +
  geom_point() +
  ggtitle("Varied Dataset") +
  theme(legend.position = "none")

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



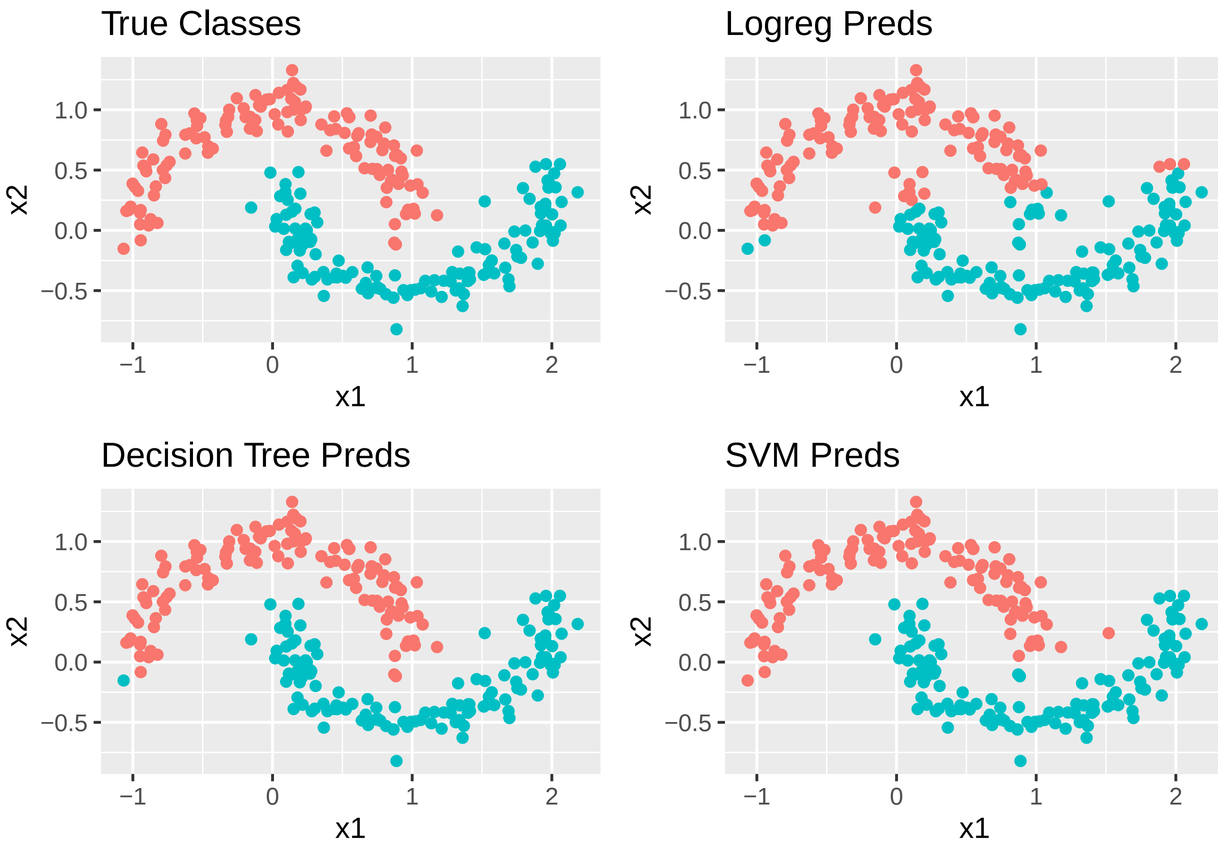
1.2. Train / Test split

```
set.seed(112)

dat <- croissant
train_ind <- sample(1:nrow(dat), floor(0.5*nrow(dat)))

train.croissant <- dat[ train_ind,]
test.croissant  <- dat[-train_ind,]

dat <- circles
train_ind <- sample(1:nrow(dat), floor(0.5*nrow(dat)))

train.circles <- dat[ train_ind,]
test.circles <- dat[-train_ind,]

dat <- varied
train_ind <- sample(1:nrow(dat), floor(0.5*nrow(dat)))

train.varied <- dat[ train_ind,]
test.varied <- dat[-train_ind,]
```

1.3. Train and Test

Croissants

```r
logreg.croissant <- glm(y ~ x1+x2, data= train.croissant, family="binomial")
tree.croissant <- tree(y~x1+x2, data=train.croissant)
svmfit.croissant <-  svm(y ~ x1+x2, data=train.croissant , kernel ="radial",
                         cost =1,gamma =1,scale =FALSE)

preds.logreg <- predict(logreg.croissant,test.croissant,type = "response") > 0.5
preds.tree <- predict(tree.croissant,test.croissant, type="class")
preds.svm <- predict(svmfit.croissant,test.croissant, type="class")

g1 <- ggplot(test.croissant, aes(x1,x2,colour=y)) +
  geom_point() +
  ggtitle("True Classes") +
  theme(legend.position = "none")
g2 <- ggplot(test.croissant, aes(x1,x2,colour=preds.logreg)) +
  geom_point() +
  ggtitle("Logreg Preds") +
  theme(legend.position = "none")
g3 <- ggplot(test.croissant, aes(x1,x2,colour=preds.tree)) +
  geom_point() +
  ggtitle("Decision Tree Preds") +
  theme(legend.position = "none")
g4 <- ggplot(test.croissant, aes(x1,x2,colour=preds.svm)) +
  geom_point() +
  ggtitle("SVM Preds") +
  theme(legend.position = "none")

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




```
Accuracy(preds.logreg, test.croissant$y==1)
```

```
## [1] 0.904
```

```
ConfusionMatrix(preds.tree, test.croissant$y)
```

```
##          y_pred
## y_true    0   1
##        0 123   1
##        1   0 126
```

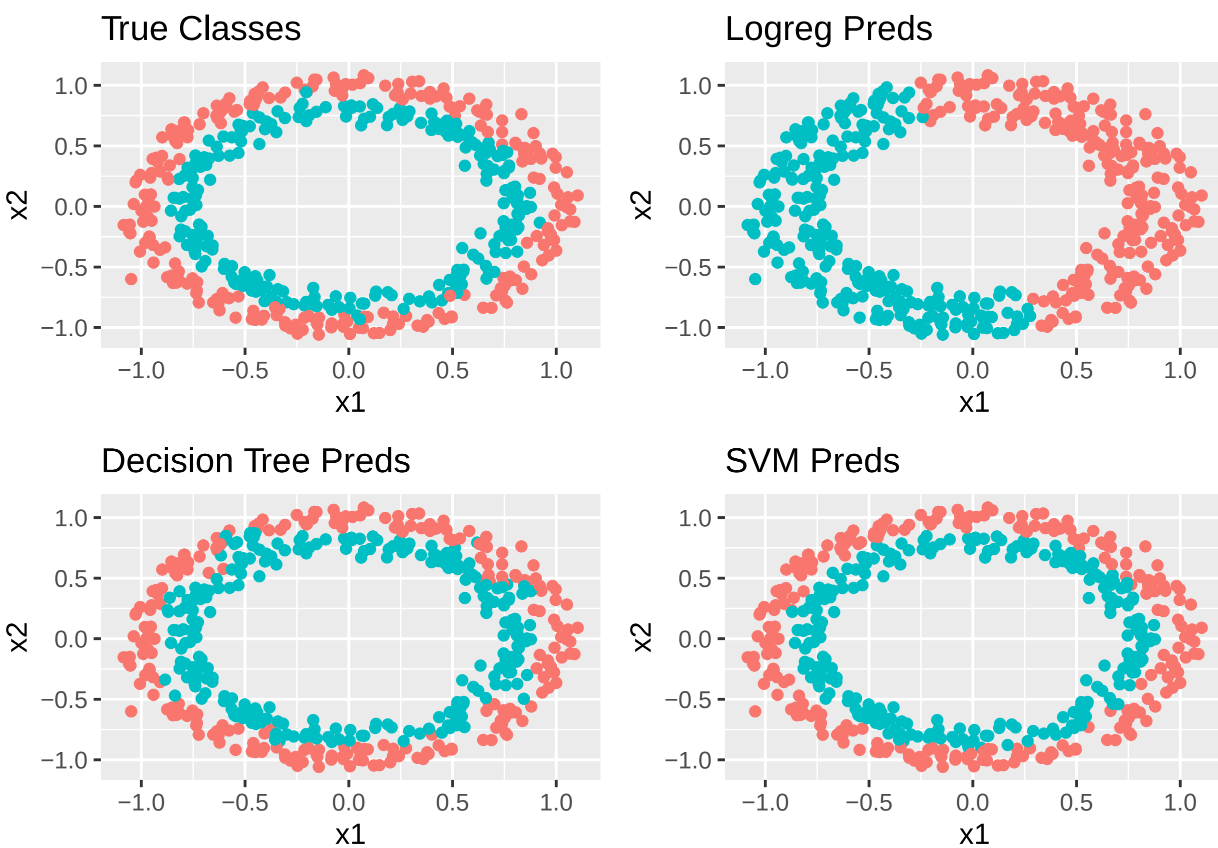
```
table(predict=preds.svm,actual=(test.croissant$y==1))
```

```
##          actual
## predict FALSE TRUE
##       0   122    1
##       1     2  125
```

Based on the results, the decision tree produced the highest accuracy.

Circles

```r
logreg.circles <- glm(y ~ x1+x2, data= train.circles, family="binomial")
tree.circles <- tree(y~x1+x2, data=train.circles)
svmfit.circles <-  svm(y ~ x1+x2, data=train.circles , kernel ="radial",
                       cost =1, gamma=1,scale =FALSE)

preds.logreg <- predict(logreg.circles,test.circles,type = "response") > 0.5
preds.tree <- predict(tree.circles,test.circles, type="class")
preds.svm <- predict(svmfit.circles,test.circles, type="class")

g1 <- ggplot(test.circles, aes(x1,x2,colour=y)) +
  geom_point() +
  ggtitle("True Classes") +
  theme(legend.position = "none")
g2 <- ggplot(test.circles, aes(x1,x2,colour=preds.logreg)) +
  geom_point() +
  ggtitle("Logreg Preds") +
  theme(legend.position = "none")
g3 <- ggplot(test.circles, aes(x1,x2,colour=preds.tree)) +
  geom_point() +
  ggtitle("Decision Tree Preds") +
  theme(legend.position = "none")
g4 <- ggplot(test.circles, aes(x1,x2,colour=preds.svm)) +
  geom_point() +
  ggtitle("SVM Preds") +
  theme(legend.position = "none")

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




```
Accuracy(preds.logreg, test.circles$y==1)
```

```
## [1] 0.468
```

```
ConfusionMatrix(preds.tree, test.circles$y)
```

```
##        y_pred
## y_true   0   1
##       0 222  28
##       1  16 234
```

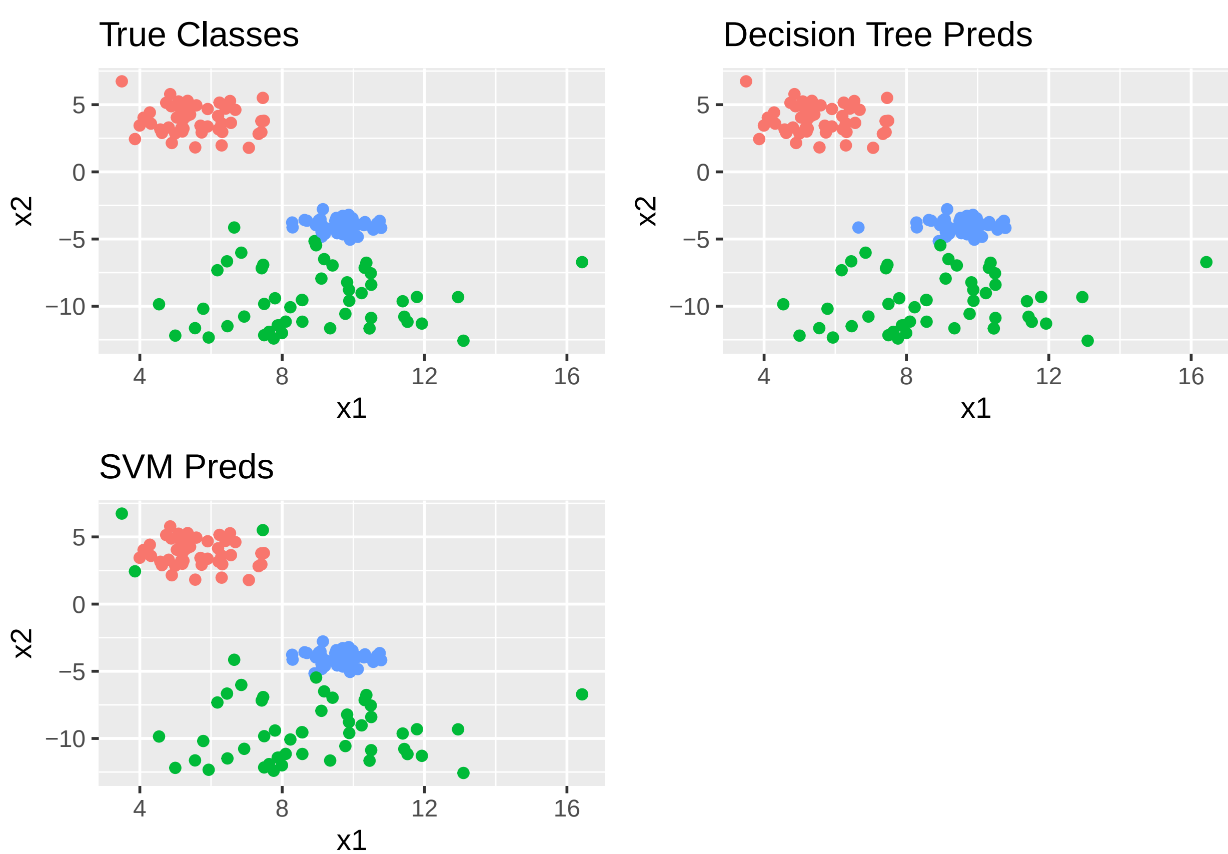
```
table(predict=preds.svm,actual=(test.circles$y==1))
```

```
##        actual
## predict FALSE TRUE
##       0   244    6
##       1     6  244
```

Based on the results, the SVM model produced the highest accuracy.

Varied

```r
tree.varied <- tree(y~x1+x2, data=train.varied)
svmfit.varied <-  svm(y ~ x1+x2, data=train.varied , kernel ="radial",
                      cost =1, gamma=1,scale =FALSE)

preds.tree <- predict(tree.varied,test.varied, type="class")
preds.svm <- predict(svmfit.varied,test.varied, type="class")

g1 <- ggplot(test.varied, aes(x1,x2,colour=y)) +
  geom_point() +
  ggtitle("True Classes") +
  theme(legend.position = "none")
g2 <- ggplot(test.varied, aes(x1,x2,colour=preds.tree)) +
  geom_point() +
  ggtitle("Decision Tree Preds") +
  theme(legend.position = "none")
g3 <- ggplot(test.varied, aes(x1,x2,colour=preds.svm)) +
  geom_point() +
  ggtitle("SVM Preds") +
  theme(legend.position = "none")

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




```r
ConfusionMatrix(preds.tree, test.varied$y)
```

```
##        y_pred
```

```
## y_true  0  1  2
##      0 51  0  0
##      1  0 49  2
##      2  0  0 48
```

```
table(predict=preds.svm,actual=(test.varied$y==1))
```

```
##        actual
## predict FALSE TRUE
##       0    48    0
##       1     3   50
##       2    48    1
```

Based on the results, the Decision tree model produced the highest accuracy.

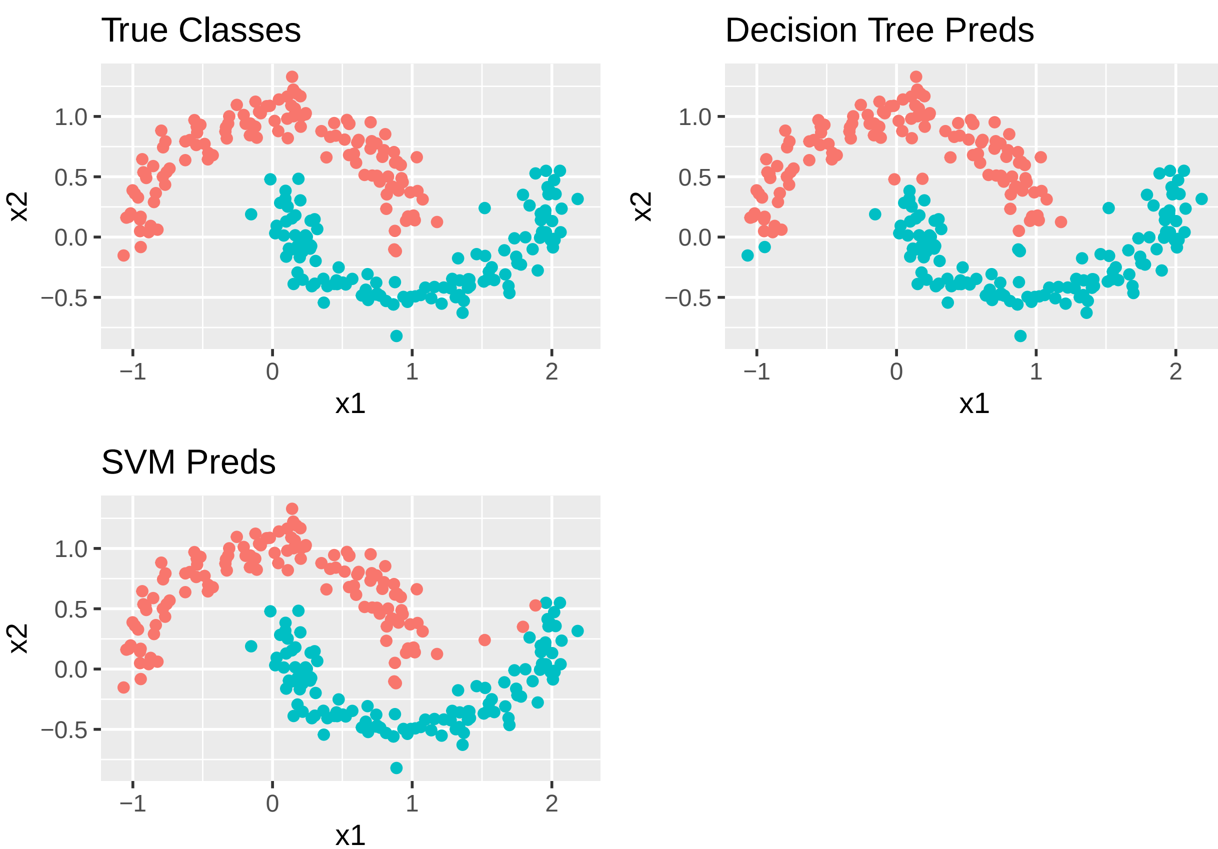
1.4. Cross Validation

Croissant

```
train_control <- trainControl(method = "cv", number = 10)
logreg.croissant <- train(y ~ x1+x2, data= train.croissant,
                          trControl = train_control,method = "glm",
                          family=binomial())

tree.croissant <- rpart(y~x1+x2, data=train.croissant)

svmfit.croissant <-  tune(svm ,y ~ x1+x2,data=train.croissant ,
                          kernel ="radial",scale =FALSE,
                          ranges =list(cost=c(0.01, 0.05, .1 ,1 ,10 ,100 ,1000),
                                       gamma=c(0.5,1,2,3,4)))

preds.logreg <- predict(logreg.croissant,test.croissant,type = "prob") > 0.5
preds.tree <- predict(tree.croissant,test.croissant, type="class")
preds.svm <- predict(svmfit.croissant$best.model,test.croissant, type="class")

g1 <- ggplot(test.croissant, aes(x1,x2,colour=y)) +
  geom_point() +
  ggtitle("True Classes") +
  theme(legend.position = "none")

summary(preds.logreg)
```

```
##      0                1
##  Mode :logical   Mode :logical
##  FALSE:126       FALSE:124
##  TRUE :124       TRUE :126
```

```r
g3 <- ggplot(test.croissant, aes(x1,x2,colour=preds.tree)) +
  geom_point() +
  ggtitle("Decision Tree Preds") +
  theme(legend.position = "none")
g4 <- ggplot(test.croissant, aes(x1,x2,colour=preds.svm)) +
  geom_point() +
  ggtitle("SVM Preds") +
  theme(legend.position = "none")

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




```r
Accuracy(preds.logreg, test.croissant$y==1)
```

```
## [1] 0.5
```

```r
ConfusionMatrix(preds.tree, test.croissant$y)
```

```
##        y_pred
## y_true   0   1
##      0 120   4
##      1   2 124
```

```
table(predict=preds.svm,actual=(test.croissant$y==1))
```

```
##        actual
## predict FALSE TRUE
##       0   124    3
##       1     0  123
```

Based on the results, the SVM model produced the highest accuracy. But, it is only slightly better than Decision tree.

Circles

```
train_control <- trainControl(method = "cv", number = 10)
logreg.circles <- train(y ~ x1+x2, data= train.circles,
                        trControl = train_control,method = "glm",
                        family=binomial())

tree.circles <- rpart(y~x1+x2, data=train.circles)

svmfit.circles <-  tune(svm ,y ~ x1+x2,data=train.circles ,kernel ="radial",
                        scale =FALSE,
                        ranges =list(cost=c(0.01, 0.05, .1 ,1 ,10 ,100 ,1000),
                                     gamma=c(0.5,1,2,3,4)))

preds.logreg <- predict(logreg.circles,test.circles,type = "prob") > 0.5
preds.tree <- predict(tree.circles,test.circles, type="class")
preds.svm <- predict(svmfit.circles$best.model,test.circles, type="class")

g1 <- ggplot(test.circles, aes(x1,x2,colour=y)) +
  geom_point() +
  ggtitle("True Classes") +
  theme(legend.position = "none")

summary(preds.logreg)
```

```
##      0                1
##  Mode :logical   Mode :logical
##  FALSE:258       FALSE:242
##  TRUE :242       TRUE :258
```

```
g3 <- ggplot(test.circles, aes(x1,x2,colour=preds.tree)) +
  geom_point() +
  ggtitle("Decision Tree Preds") +
  theme(legend.position = "none")
g4 <- ggplot(test.circles, aes(x1,x2,colour=preds.svm)) +
  geom_point() +
  ggtitle("SVM Preds") +
```

```r
  theme(legend.position = "none")


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


```r
Accuracy(preds.logreg, test.circles$y==1)
```

```
## [1] 0.5
```

```r
ConfusionMatrix(preds.tree, test.circles$y)
```

```
##       y_pred
## y_true   0   1
##      0 235  15
##      1  32 218
```

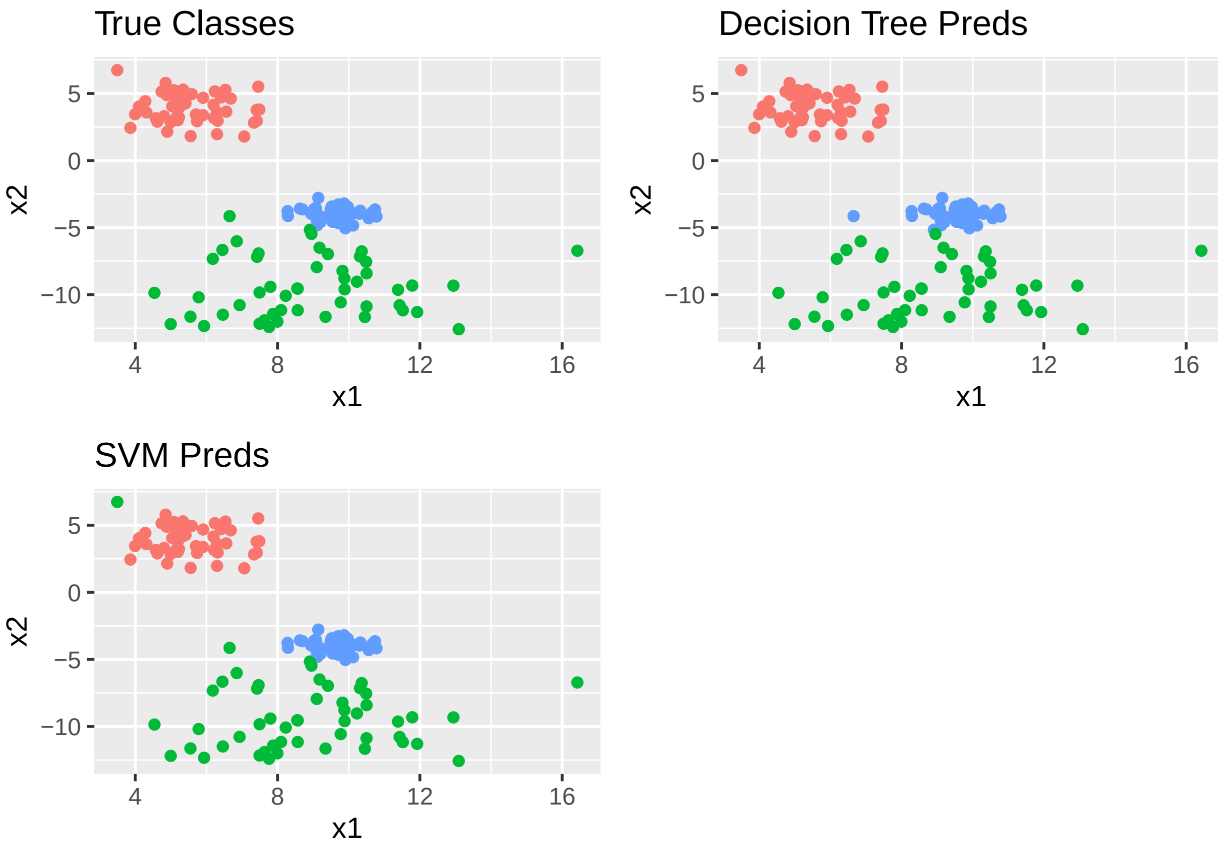
```r
table(predict=preds.svm,actual=(test.circles$y==1))
```

```
##        actual
## predict FALSE TRUE
##       0   242    4
##       1     8  246
```

Based on the results, the SVM model produced the highest accuracy.

Varied

```
tree.varied <- rpart(y~x1+x2, data=train.varied)

svmfit.varied <-  tune(svm ,y ~ x1+x2,data=train.varied ,kernel ="radial",
                       scale =FALSE,
                       ranges =list(cost=c(0.01, 0.05, .1 ,1 ,10 ,100 ,1000),
                                    gamma=c(0.5,1,2,3,4)))

preds.tree <- predict(tree.varied,test.varied, type="class")
preds.svm <- predict(svmfit.varied$best.model,test.varied, type="class")

g1 <- ggplot(test.varied, aes(x1,x2,colour=y)) +
  geom_point() +
  ggtitle("True Classes") +
  theme(legend.position = "none")
g2 <- ggplot(test.varied, aes(x1,x2,colour=preds.tree)) +
  geom_point() +
  ggtitle("Decision Tree Preds") +
  theme(legend.position = "none")
g3 <- ggplot(test.varied, aes(x1,x2,colour=preds.svm)) +
  geom_point() +
  ggtitle("SVM Preds") +
  theme(legend.position = "none")

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




```
ConfusionMatrix(preds.tree, test.varied$y)
```

```
##        y_pred
## y_true  0  1  2
##       0 51  0  0
##       1  0 49  2
##       2  0  0 48
```

```
table(predict=preds.svm,actual=(test.varied$y==1))
```

```
##          actual
## predict FALSE TRUE
##       0    50    0
##       1     1   51
##       2    48    0
```

Based on the results, the decision tree produced the highest accuracy.

Question 2: Tree-based methods

2.1 Preprocess

2.2 Decision Trees for Regression

2.3. Decision Trees for Classification

2.4. Bagging: Regression

2.5. Bagging: Classification

2.6. Random Forest: Regression