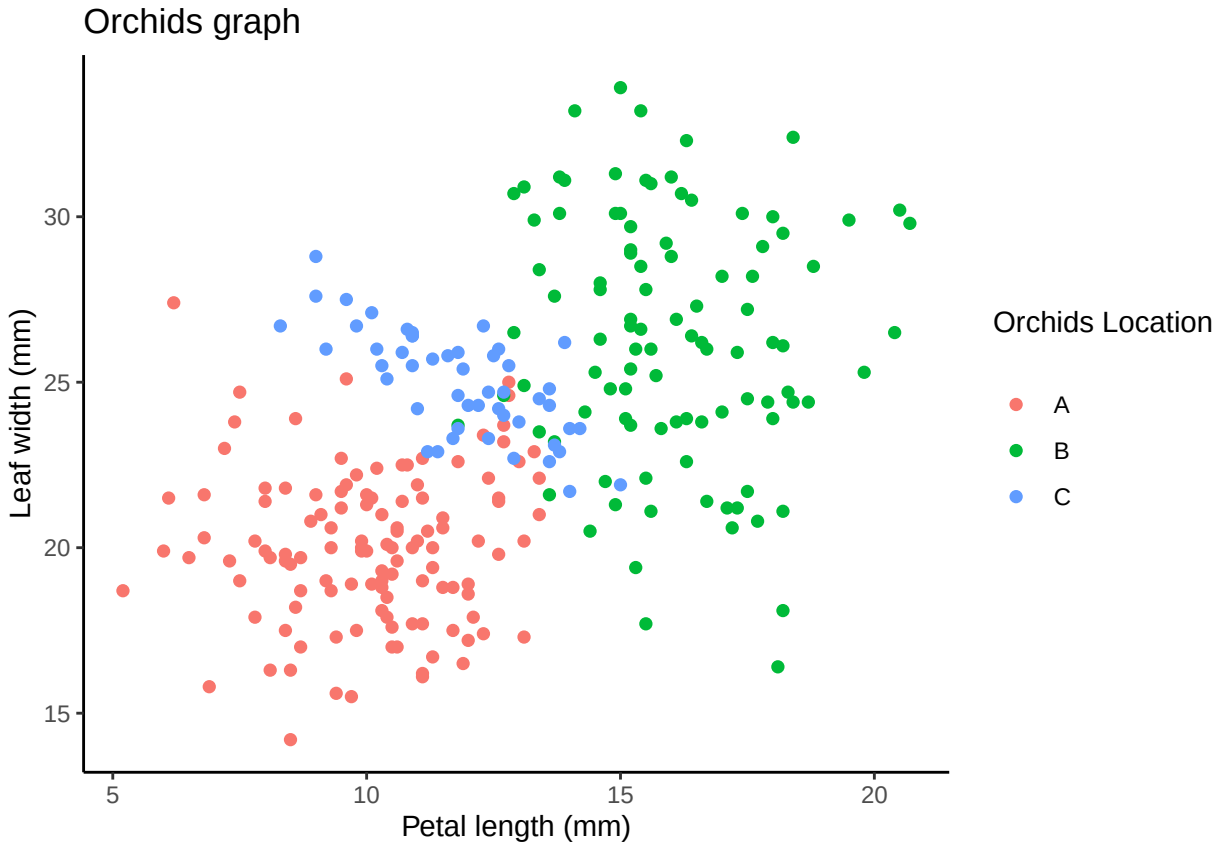
3.1 Machine Learning Task

By: Aderibigbe Olugbenga & Paul Hazell

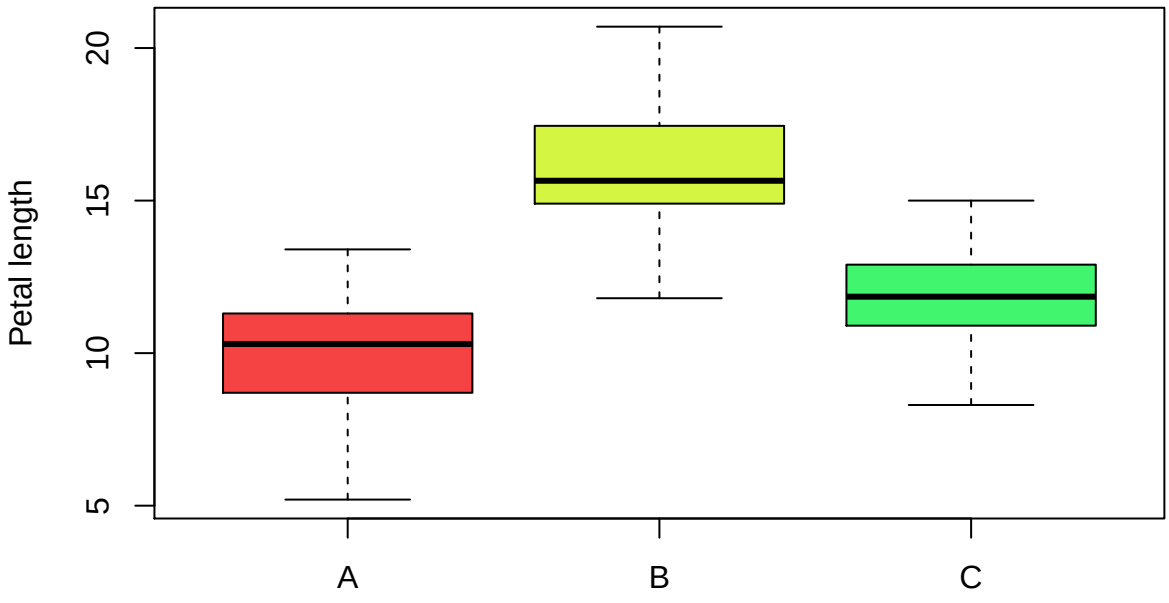
18/04/2020

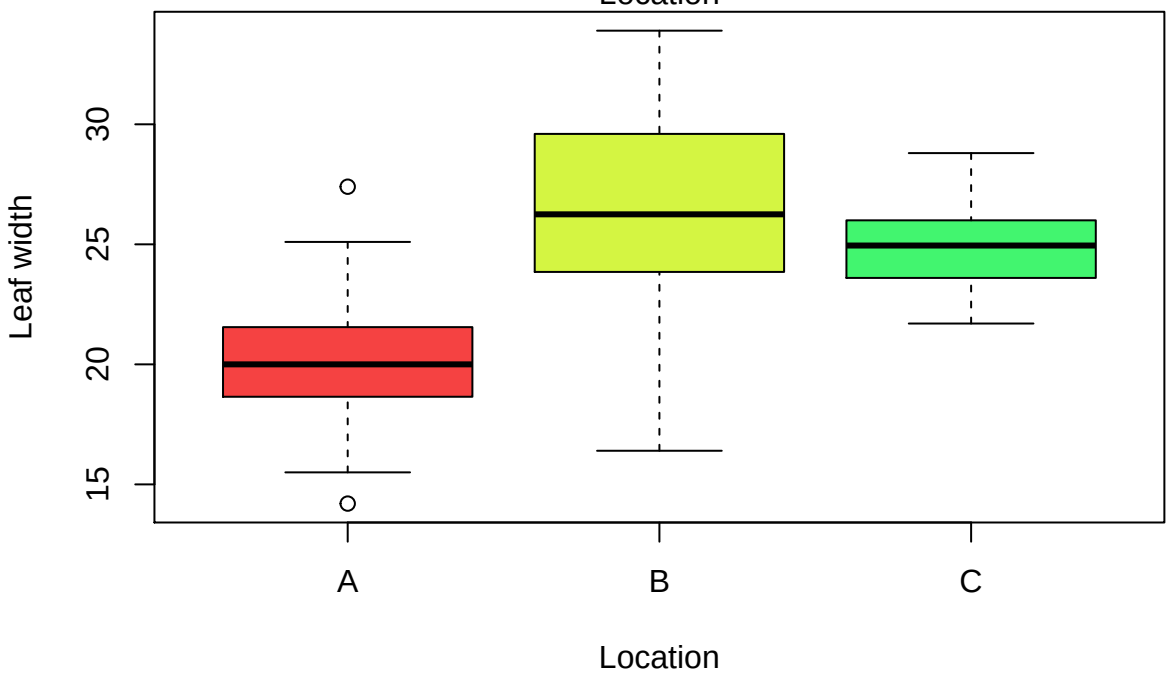
Machine Learning Part (a)

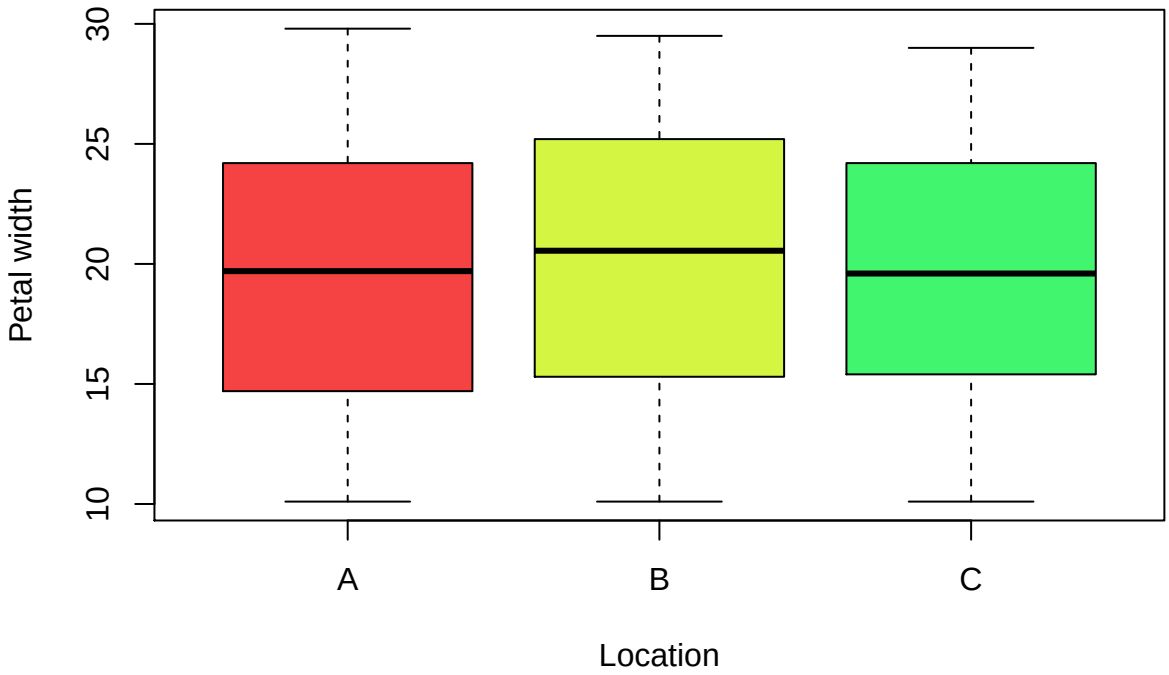
Graph of bivariate scatter plots to
distinguish between the three locations of Orchids



Boxplots of the data to choose two characteristcs
that should be used as predictors for orchids' locations.


Petal length
5
10
15
20
A
B
C
Location


Leaf width
15
20
25
30
A
B
C
Location



From the graph it can be seen that there is a considerable difference between the mean location of the Petal length (X1) and Leaf width (X2).

```
##   loc       X1
## 1   A 10.05917
## 2   B 16.01400
## 3   C 11.84800
```

```
##   loc       X2
## 1   A 20.01917
## 2   B 26.30500
## 3   C 24.95600
```

Whereas the difference between the mean Petal width (X3) with different locations is not much.

```
##   loc       X3
## 1   A 19.48333
## 2   B 20.04700
## 3   C 19.71400
```

So the Petal length and Leaf width data are used as predictors for the orchids' locations.

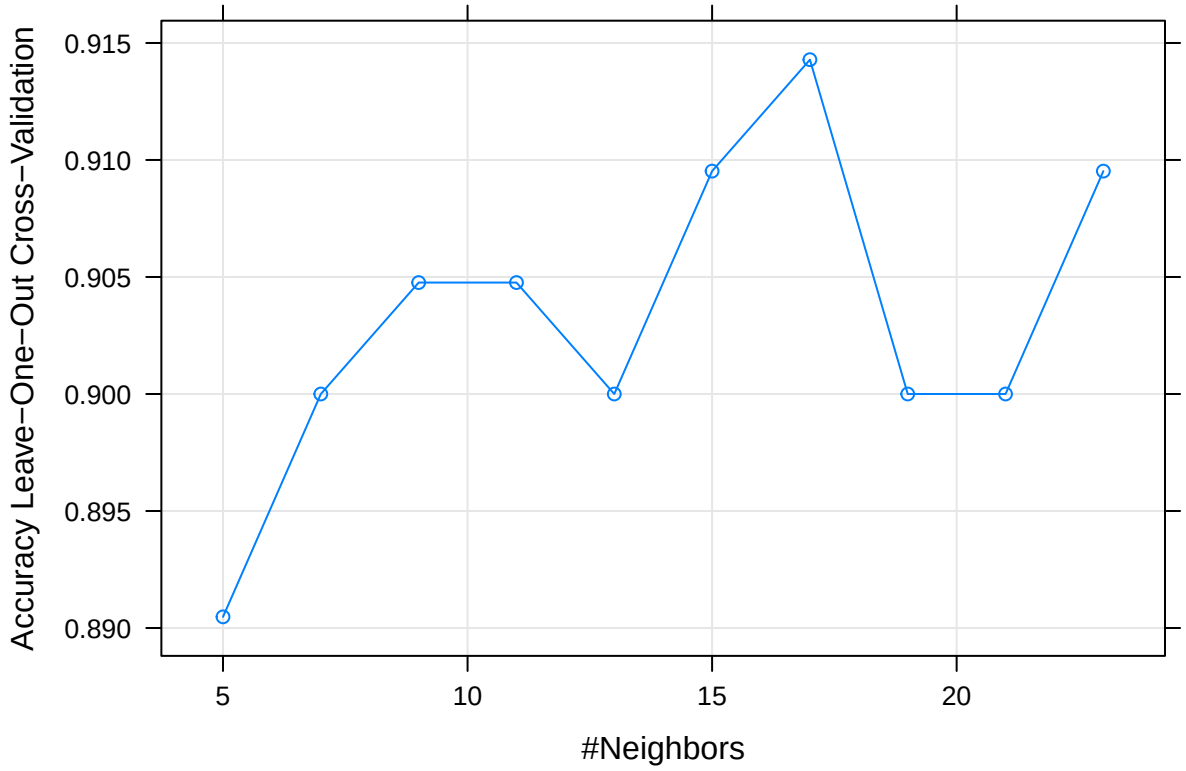
Machine learning task B

Creating a training set 210 randomly chosen data points and a test set of 60 data points.

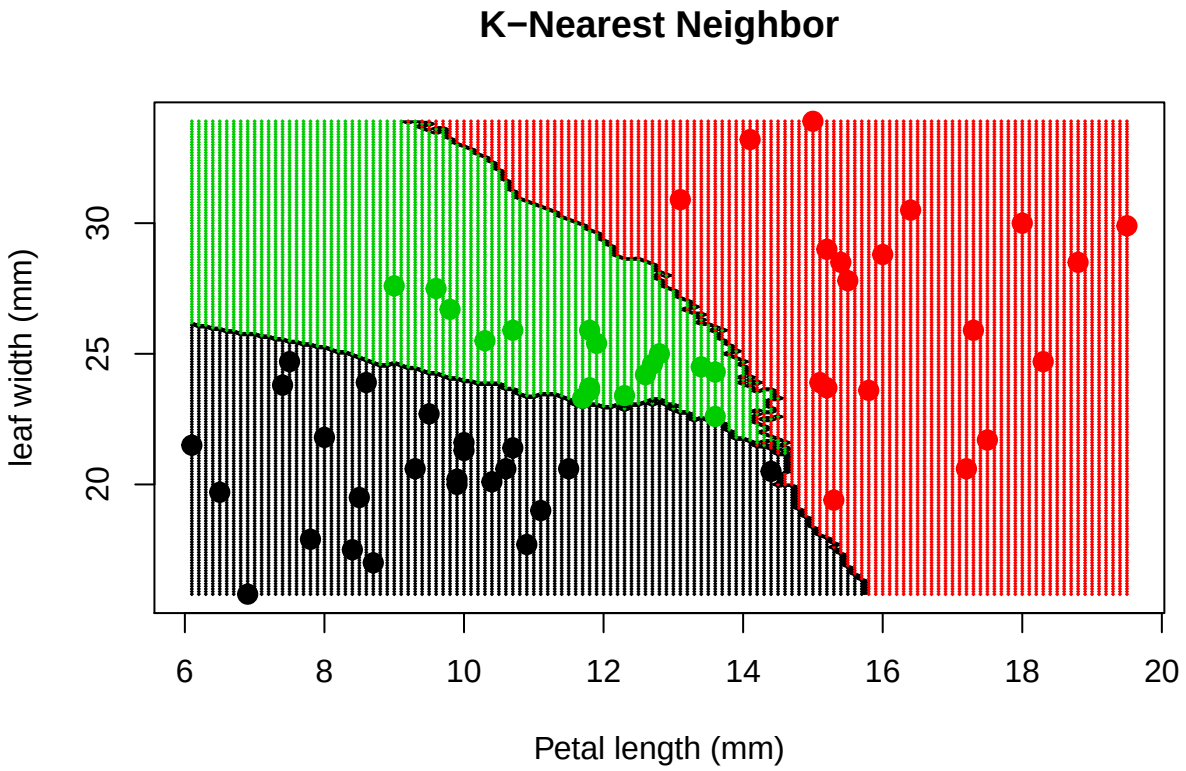
```
set.seed(1)
data.subset <- sample(270, 210)
model.train <- orchid[data.subset,]
model.test <- orchid[-data.subset,]
```

Machine learning task C KNN Method

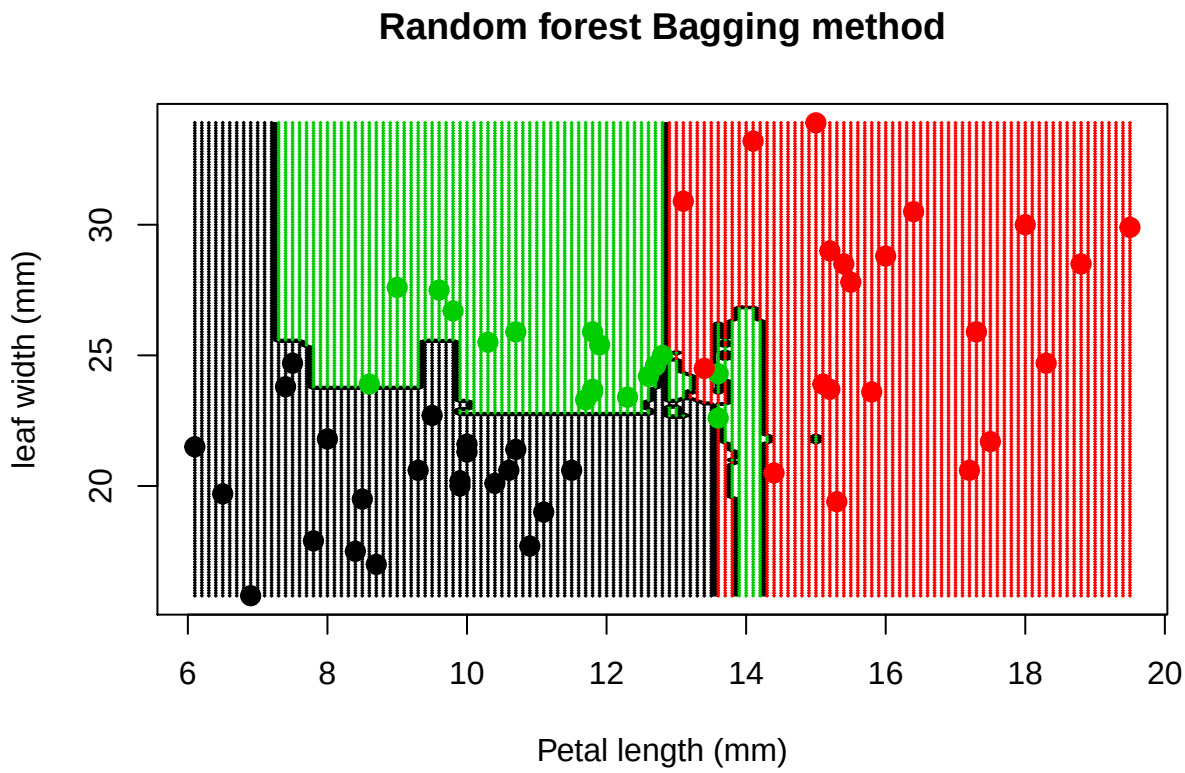
Graph showing the accuracy of K



The best optimal k value is 17



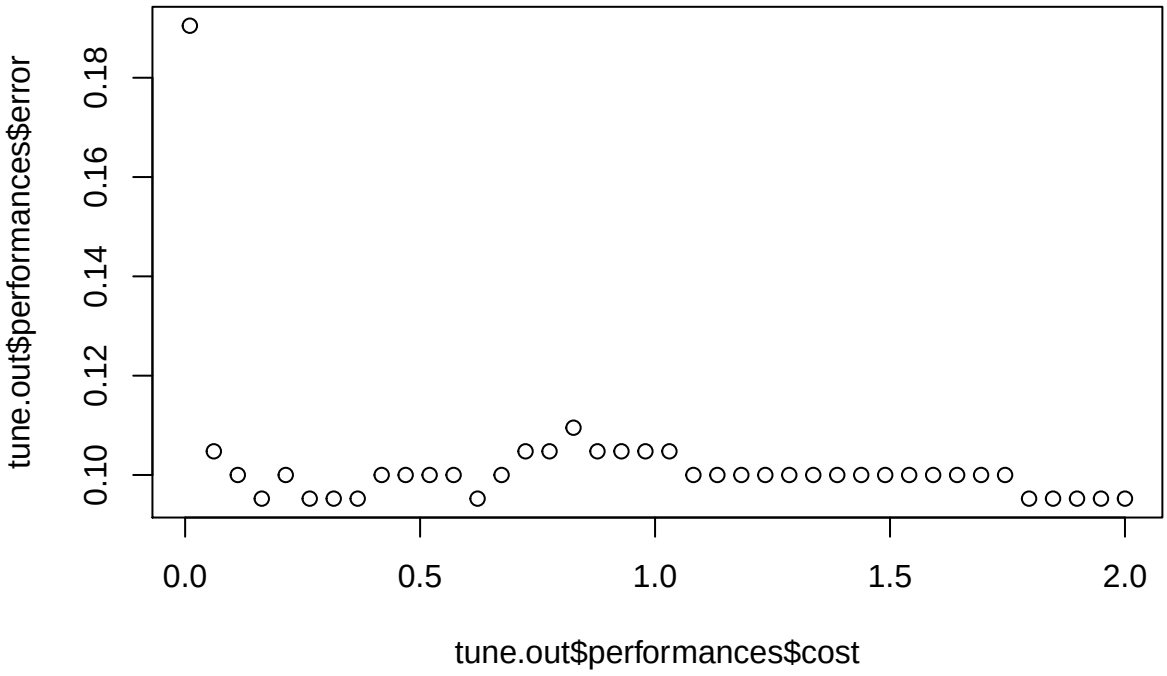
Machine learning task D Random Forest Bagging method



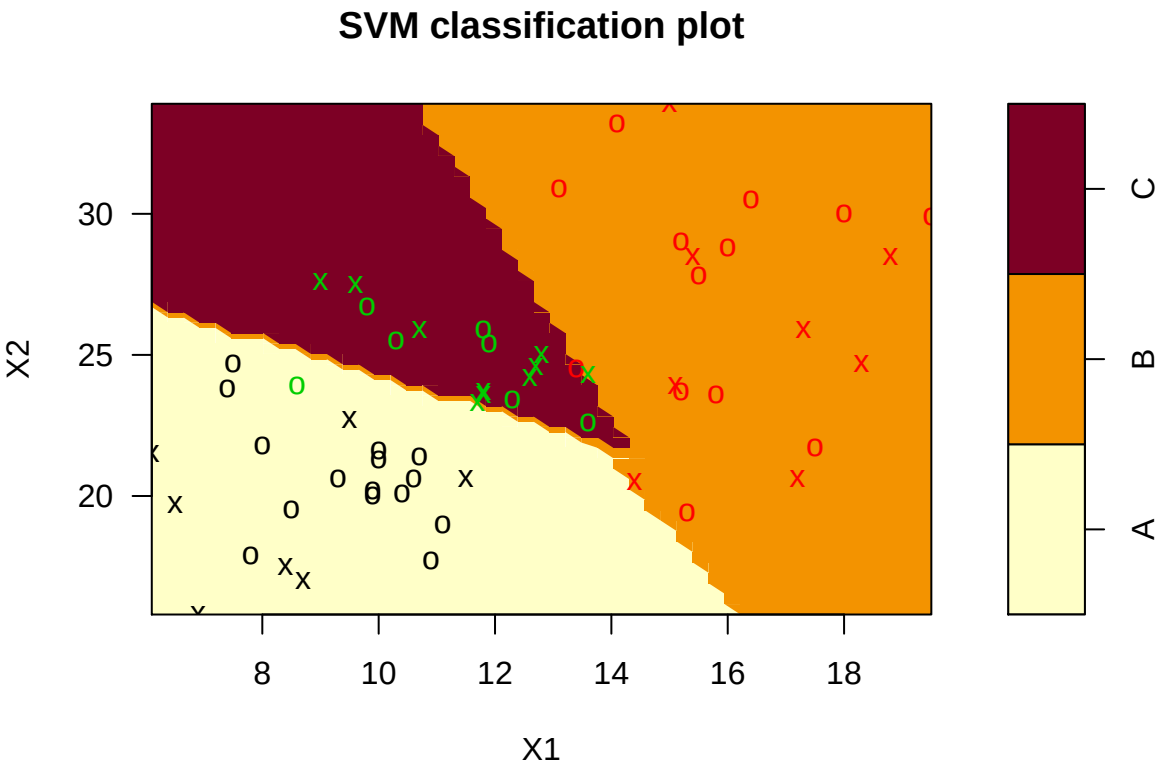
Machine learning task E Support Vector Machines

Linear kernel

Choosing the best cost parameter

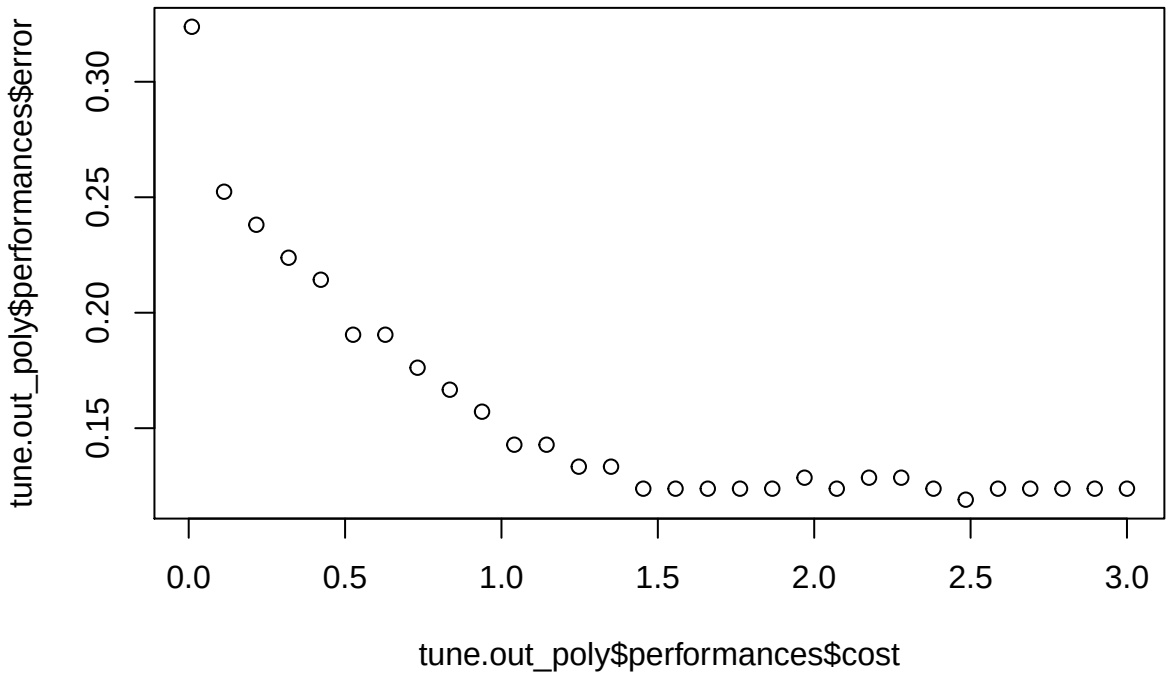


The best cost parameter for the linear kernel is 0.1630769

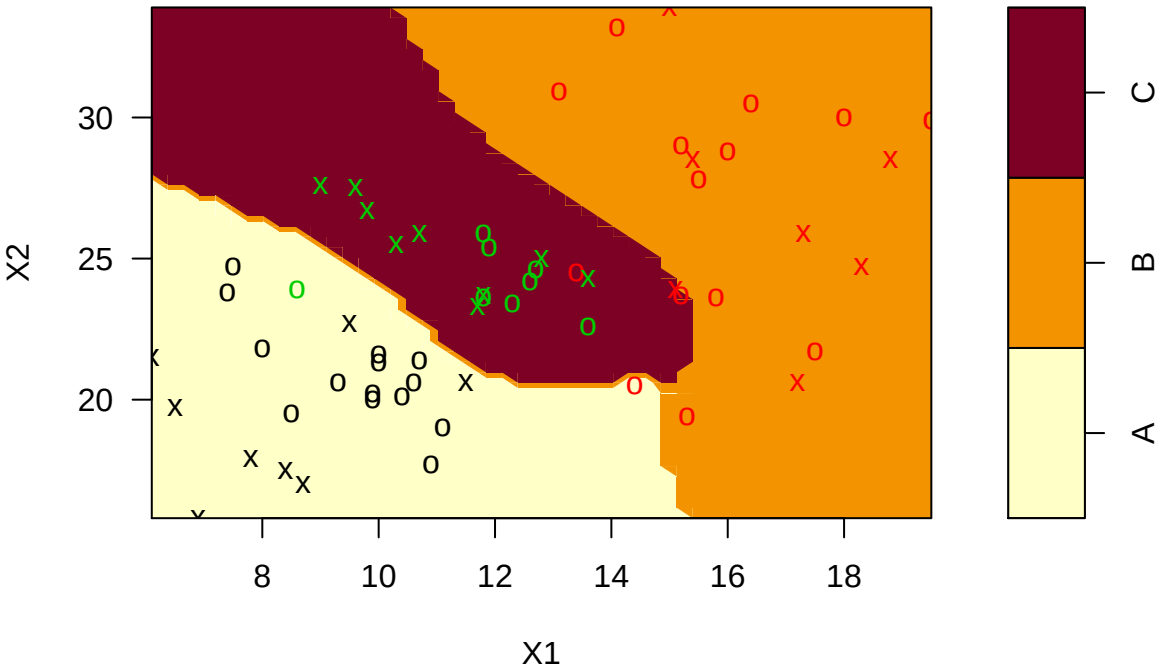


Machine learning task E Support Vector Machines Polynomial Kernel

Choosing the best cost parameter



The best cost parameter for the Polynomial kernel is 2.4844828



Test Accuracy for KNN Method

```
##
## predict.knn  1  2  3
##           1 22  1  1
##           2  0 19  0
##           3  0  1 16
```

```
## [1] 0.95
```

Test Accuracy for Random Forest Bagging Method

```
##
## bag_predict  1  2  3
##            A 22  0  0
##            B  0 21  0
##            C  0  0 17
```

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

Test Accuracy for Linear Kernel Support Vector

```
##
## predict.knn  1  2  3
##           1 22  1  1
##           2  0 19  0
##           3  0  1 16
```

```
## [1] 0.9666667
```

Test Accuracy for Polynomial Kernel Support Vector

```
##
## ypred  1  2  3
##     A 22  1  1
```

```
##      B  0 17  0
##      C  0  3 16
```

```
## [1] 0.9166667
```

From the results gotten it can be seen that the model with the best accuracy is the Random Forest Bagging method.
This method came out as the best because the Bagging method helps in increasing the accuracy of the Random Forest method and also helps reduce variance of the model to help in predicting accurate results.