**CME 4403 Hw3 - Due Date: November 11th 2022, 23:50**

*Please send your source codes and results (figures, replies etc.) in one text file (e.g., YourName.docx / YourName.pdf), then submit it to the SAKAI.*

*Group homework is not allowed!*

1. You will code a **decision tree** (DT) that will be similar with the one given in the lab examples. The input data is provided in the text file “wheat\_types.txt”. The target feature is the “type” column. You will train and test the DT by using all features except the “type”. Apply the following steps and reply the questions in your report.

* You can use a different decision tree library apart from “tree”.
* Use 75% of samples for training and 25% of them for test purposes.
* Perform this sub-sampling process (cross-validation) for “100” iterations.

**library(tree)**

**wheatTypes <- read.delim(file='C:\\Users\\irfan\\Desktop\\R\\wheat\_types.txt',header = T,sep = ';')**

**set.seed(4132311)**

**dt\_acc <- numeric()**

**for(i in 1:100){**

**sub <- sample(1:nrow(wheatTypes), size=nrow(wheatTypes)\*0.75)**

**train <- tree(type ~ ., data = wheatTypes, subset = sub, split = "gini")**

**test\_predict <- table(predict(train, wheatTypes[-sub, ]), wheatTypes[-sub, "type"])**

**dt\_acc <- c(dt\_acc, sum(diag(test\_predict)) / sum(test\_predict))**

**if(max\_acc < sum(diag(test\_predict)) / sum(test\_predict)){**

**max\_tree <- train**

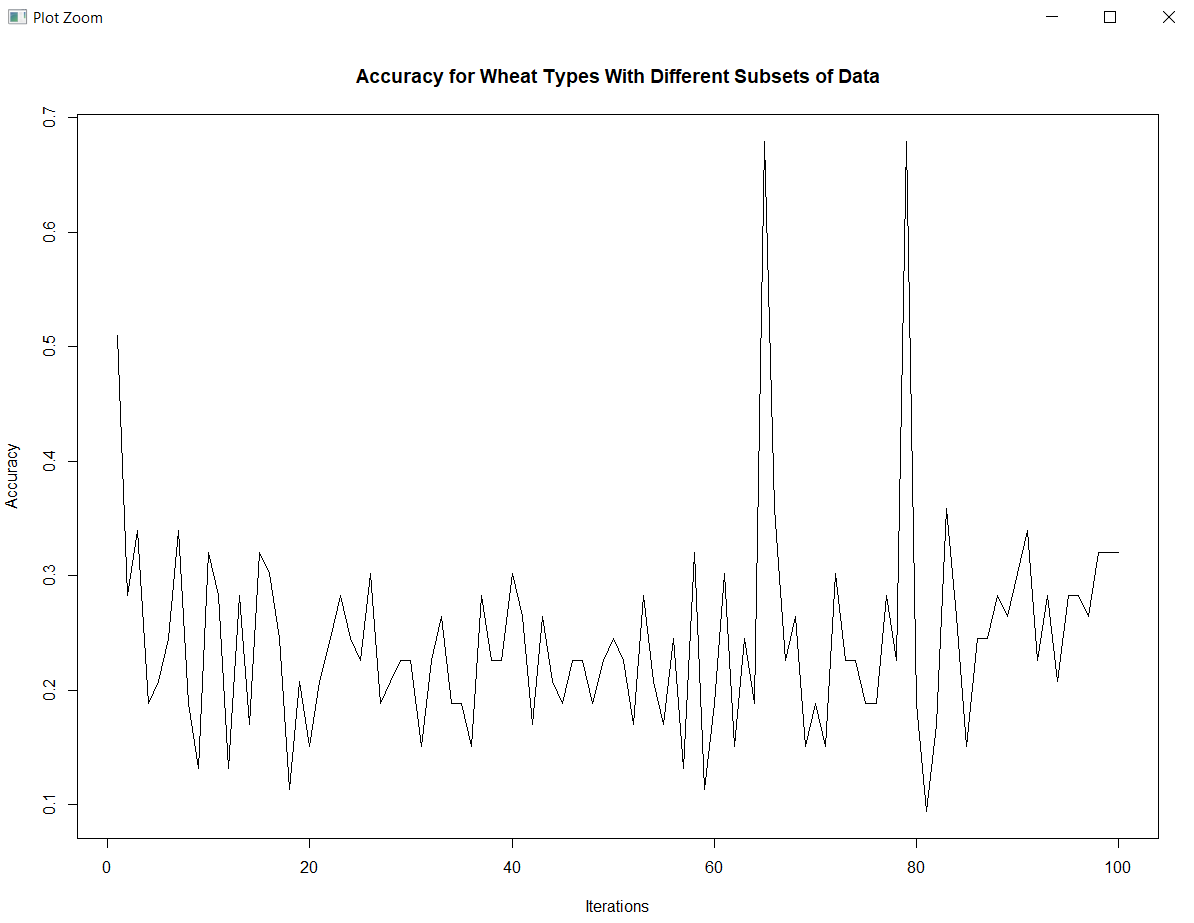
**max\_acc <- sum(diag(test\_predict)) / sum(test\_predict)**

**}**

**}**

* Draw a plot to show how accuracy changes over 100 iterations

plot(dt\_acc, type="l", ylab="Accuracy", xlab="Iterations", main="Accuracy for Wheat Types With Different Subsets of Data")

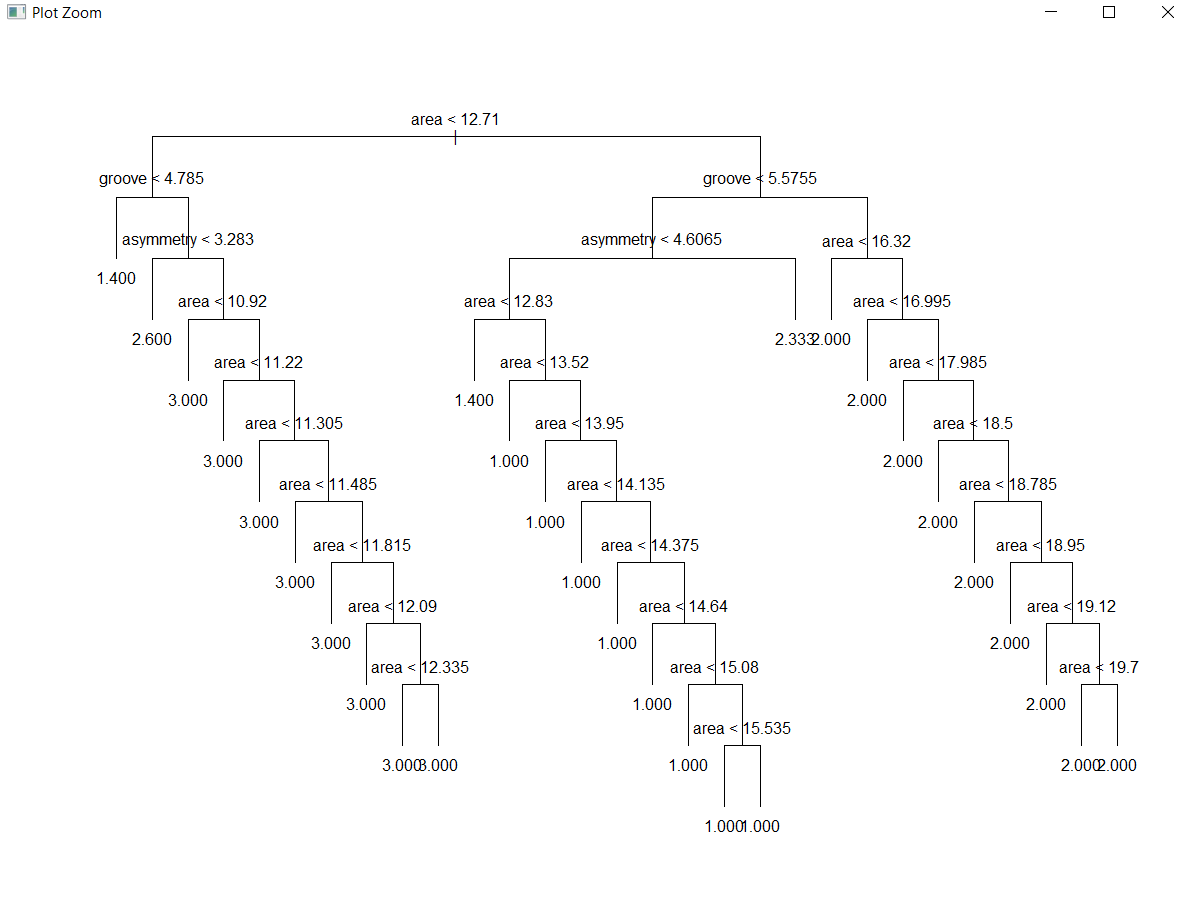
****

* What is the average accuracy of DT after 100 iterations?

mean(dt\_acc)

[1] 0.2443396

* Show the DT (with parent and internal node’s decision criteria) which has the highest accuracy over 100 iterations.



1. You will code a **k-nearest neighbor** (kNN) algorithm, which will be similar with the one given in the lab examples. You will use built-in data set “Smarket” that is provided in the “ISLR” library. The target feature is the “Direction” column. You will use two combinations of descriptive features and different ***k*** values. Apply the following steps and reply to the following questions in your report.

* You need to test two sets of descriptive features to improve the prediction accuracy of the target feature. The combinations:
  + Lag1, Lag2, Lag3, Volume
  + Lag4, Lag5, Volume, Today
* Use 60% of samples for training and 40% of them for test purposes (for each combination).

**library(class) # Contains the "knn" function**

**library(ISLR)**

**set.seed(1234113) # Set the seed for reproducibility**

**sm\_sample <- sample(1:nrow(Smarket), size=nrow(Smarket)\*0.6)**

**sm\_train <- Smarket[sm\_sample, ] #Select the 60% of rows**

**sm\_test <- Smarket[-sm\_sample, ] #Select the 40% of rows**

* Test different ***k*** values (1 to 11) for each combination.

**sm\_acc <- numeric() #holding variable**

**## Desc Feature Lag1, Lag2, Lag3, Volume**

**for(i in 1:11){**

**#Apply knn with k = i**

**predict <- knn(train=sm\_train[,c(2,3,4,7)], test=sm\_test[,c(2,3,4,7)], cl=sm\_train$Direction, k=i)**

**sm\_acc <- c(sm\_acc, mean(predict==sm\_test$Direction))**

**}**

**plot(sm\_acc, type="l", ylab="Accuracy", xlab="K", main="Accuracy for Smarket data set with varying K")**

**## Desc Feature Lag4, Lag5, Volume, Today**

**for(i in 1:11){**

**#Apply knn with k = i**

**predict <- knn(train=sm\_train[,c(5,6,7,8)], test=sm\_test[,c(5,6,7,8)], cl=sm\_train$Direction, k=i)**

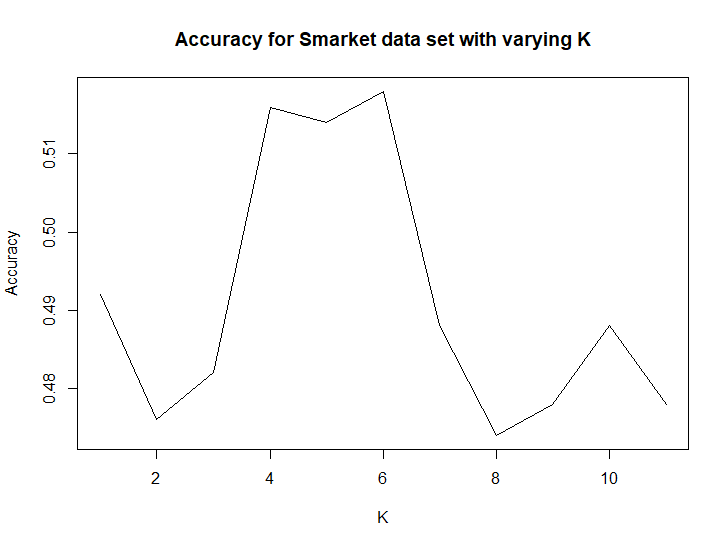
**sm\_acc <- c(sm\_acc, mean(predict==sm\_test$Direction))**

**}**

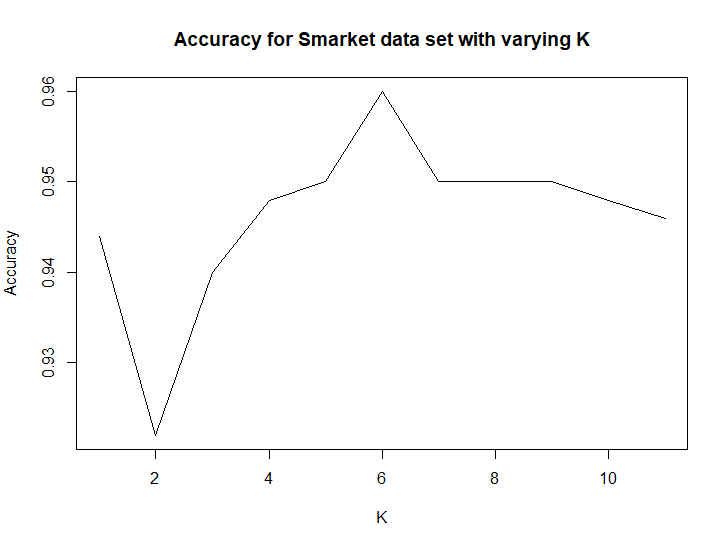
**plot(sm\_acc, type="l", ylab="Accuracy", xlab="K", main="Accuracy for Smarket data set with varying K")**

* Make a plot to show how accuracy changes for different ***k*** values (for each combination).

## Desc Feature Lag1, Lag2, Lag3, Volume



## Desc Feature Lag4, Lag5, Volume, Today



* Which ***k*** value did provide the highest accuracy for each combination? Report both ***k*** values and the highest accuracy value.

max(sm\_acc) :

## Desc Feature Lag1, Lag2, Lag3, Volume :

0.518 = %51,8 for k = 6

## Desc Feature Lag4, Lag5, Volume, Today

0.960 = %96 for k = 6

* What is the highest accuracy to predict the “Direction” column when you set a specific ***k*** value and a descriptive feature combination?

## Desc Feature Lag4, Lag5, Volume, Today

0.960 = %96 for k = 6