Task 1

Read the dataset

Code:

setwd('/Users/duwonha/Desktop/University/Second year/T2/Big data/Assignment2')

drugdata <- read.csv('drug200.csv', stringsAsFactors = TRUE)

Task 2

Dataset split into training and test sets

Code:

training\_size <- floor(0.8 \* nrow(drugdata))

set.seed(101)

train\_int <- sample(seq\_len(nrow(drugdata)), size = training\_size)

trainingSet <- drugdata[train\_int, ]

testSet <- drugdata[-train\_int, ]

Task 3

Importing library

Code:

library(ISLR)

data(package = "ISLR")

require(tree)

Tree Construnction

Code:

tree\_accuracy <- tree(formula = Drug ~ Age+Sex+BP+Cholesterol+Na\_to\_K, data = trainingSet)

Tree Plot

Code:

plot(tree\_accuracy)

text(tree\_accuracy, pretty = 0)

Accuracy Calculation

Code:

tree\_pred = predict(tree\_accuracy, drugdata[-train\_int,], type = 'class')

with(drugdata[-train\_int,], table(tree\_pred, Drug))

Task 4

Cross-validation and Tree Construction

Code:

drug.cv = cv.tree(tree\_accuracy, FUN = prune.misclass)

drug.cv > plot(drug.cv)

drug.cv

#Once you run the drug.cv, you can see deviation of size 6 is 1 which means size 6 is the most accurate one. Size 6 will be used for making prune. (best = 6)

Text

Description automatically generated

prune.drug = prune.misclass(tree\_accuracy, best = 6)

Tree plot

Code:

plot(prune.drug)

text(prune.drug, pretty=0)

Accuracy Calculation

Code:

tree\_pred = predict(prune.drug, drugdata[-train\_int,], type = "class")

with(drugdata[-train\_int,], table(tree\_pred, Drug))

Screenshot

Graphical user interface, application

Description automatically generated

Graphical user interface

Description automatically generated

Graphical user interface, application

Description automatically generated

Graphical user interface

Description automatically generated