###Index

##sampling data

#Build a ctree

#sampling iris data

str(iris)

## 'data.frame': 150 obs. of 5 variables:

## $ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...

## $ Sepal.Width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1...

## $ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1...

## $ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0...

## $ Species : Factor w/ 3 levels "setosa","versicolor",....

# split data into two subsets: training (70%) and test (30%); set

# a fixed random seed to make results reproducible

set.seed(1234)

ind <- sample(2, nrow(iris), replace = TRUE, prob = c(0.7, 0.3))

train.data <- iris[ind == 1, ]

test.data <- iris[ind == 2, ]

#Build a ctree

library(party)

myFormula <- Species ~ Sepal.Length + Sepal.Width + Petal.Length +

Petal.Width

iris\_ctree <- ctree(myFormula, data = train.data)

# check the prediction

table(predict(iris\_ctree), train.data$Species)

print(iris\_ctree)

plot(iris\_ctree)

plot(iris\_ctree, type = "simple")

#predection

#predict on test data

testPred <- predict(iris\_ctree, newdata = test.data)

table(testPred, test.data$Species)

#The bodyfat Dataset

data("bodyfat", package = "TH.data")

dim(bodyfat)

## [1] 71 10

# str(bodyfat)

head(bodyfat, 5)

#Train a Decision Tree with Package rpart

data("bodyfat", package = "TH.data")

dim(bodyfat)

## [1] 71 10

# str(bodyfat)

head(bodyfat, 5)

#Train a Decision Tree with Package rpart

# split into training and test subsets

set.seed(1234)

ind <- sample(2, nrow(bodyfat), replace=TRUE, prob=c(0.7, 0.3))

bodyfat.train <- bodyfat[ind==1,]

bodyfat.test <- bodyfat[ind==2,]

# train a decision tree

library(rpart)

myFormula <- DEXfat ~ age + waistcirc + hipcirc + elbowbreadth +

kneebreadth

bodyfat\_rpart <- rpart(myFormula, data = bodyfat.train,

control = rpart.control(minsplit = 10))

# print(bodyfat\_rpart$cptable)

print(bodyfat\_rpart)

plot(bodyfat\_rpart)

text(bodyfat\_rpart, use.n=T)

# select the tree with the minimum prediction error

opt <- which.min(bodyfat\_rpart$cptable[, "xerror"])

cp <- bodyfat\_rpart$cptable[opt, "CP"]

# prune tree

bodyfat\_prune <- prune(bodyfat\_rpart, cp = cp)

# plot tree

plot(bodyfat\_prune)

text(bodyfat\_prune, use.n = T)

#Model Evaluation

DEXfat\_pred <- predict(bodyfat\_prune, newdata = bodyfat.test)

xlim <- range(bodyfat$DEXfat)

plot(DEXfat\_pred ~ DEXfat, data = bodyfat.test, xlab = "Observed",

ylab = "Prediction", ylim = xlim, xlim = xlim)

abline(a = 0, b = 1)

#Train a Random Forest

#split into two subsets: training (70%) and test (30%)

ind <- sample(2, nrow(iris), replace=TRUE, prob=c(0.7, 0.3))

train.data <- iris[ind==1,]

test.data <- iris[ind==2,]

# use all other variables to predict Species

library(randomForest)

rf <- randomForest(Species ~ ., data=train.data, ntree=100,proximity=T)

table(predict(rf), train.data$Species)

print(rf)

#Error Rate of Random Forest

plot(rf, main = "")

#Variable Importance

importance(rf)

#Variable Importance

varImpPlot(rf)

#Margin of Predictions

#The margin of a data point is as the proportion of votes for the

#correct class minus maximum proportion of votes for other classes.

#Positive margin means correct classication.

irisPred <- predict(rf, newdata = test.data)

table(irisPred, test.data$Species)

plot(margin(rf, test.data$Species))