Git Url : <https://github.com/ShilpaAChavan/Datascience-With-R/blob/master/Assignment/Randomforest/RandomforestIrisEg.R>

Code :

##################################################################################

#Problem Statement : build a Random forest for iris dataset.

##################################################################################

install.packages("caret",dependencies = TRUE)

install.packages("randomForest")

library(randomForest)

#First Case : Random forests classifier in simple way

# all the data is used for training as well as testing.

model<- randomForest(iris$Species~.,data=iris,ntree=500)

#randomForest(formula = iris$Species ~ ., data = iris, ntree = 500)

#Type of random forest: classification

#Number of trees: 500

#No. of variables tried at each split: 2

#OOB estimate of error rate: 4.67%

#Confusion matrix:

# setosa versicolor virginica class.error

# setosa 50 0 0 0.00

# versicolor 0 47 3 0.06

# virginica 0 4 46 0.08

# View the forest results.

print(model)

# OOB estimate of error rate: 4.67%

# wrongly classified data/total records (3+4)/150 \*100

# the smallest value for this error rate ie preferred

#Importance of the variable - Lower Gini

print(importance(model))

# Petal.Width is the important feature with highest MeanDecreaseGini value.

#MeanDecreaseGini

#Sepal.Length 9.937739

# Sepal.Width 2.284887

# Petal.Length 40.191219

# Petal.Width 46.791835

#prediction

pred <- predict(model,iris[,-5])

pred

#pred setosa versicolor virginica

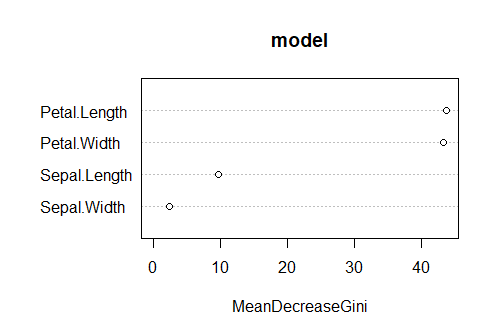
#setosa 50 0 0

#versicolor 0 50 0

#virginica 0 0 50

table(pred,iris$Species)

varImpPlot(model)



#Second Case : with training data-70% & testing data-30%

index\_row <- sample(2,nrow(iris),replace = T,prob = c(0.7, 0.3))

train\_data <- iris[index\_row == 1,]

test\_data <- iris[index\_row == 2,]

#build model on training dataset

iris\_model <- randomForest(Species~.,data=train\_data,importance=T)

# randomForest(formula = Species ~ ., data = train\_data, importance = T)

#Type of random forest: classification

#Number of trees: 500

#No. of variables tried at each split: 2

#OOB estimate of error rate: 5.05%

#Confusion matrix:

# setosa versicolor virginica class.error

#setosa 33 0 0 0.00000000

#versicolor 0 30 2 0.06250000

#virginica 0 3 31 0.08823529

#Out of bag estimates error rate is 5.05% in this random forest model.

pred1 <- predict(iris\_model,train\_data)

confusionMatrix(pred1,train\_data$Species)

# Accurancy is 100% with traindata and model built.

importance(iris\_model)

# setosa versicolor virginica MeanDecreaseAccuracy MeanDecreaseGini

#Sepal.Length 5.607518 6.950099 6.680433 10.056898 6.022583

#Sepal.Width 4.630033 1.221292 4.875114 5.338872 2.519420

#Petal.Length 21.332776 28.611753 28.242696 31.887375 30.558005

#Petal.Width 22.127443 29.115110 32.887261 33.115285 32.068452

varImpPlot(iris\_model)

qplot(Petal.Width,Petal.Length,data=iris,color= Species)

#predicting test data on built model

pred\_table <- predict(iris\_model,test\_data[,-5])

table(observed=test\_data[,5],predicted=pred\_table)

predicted

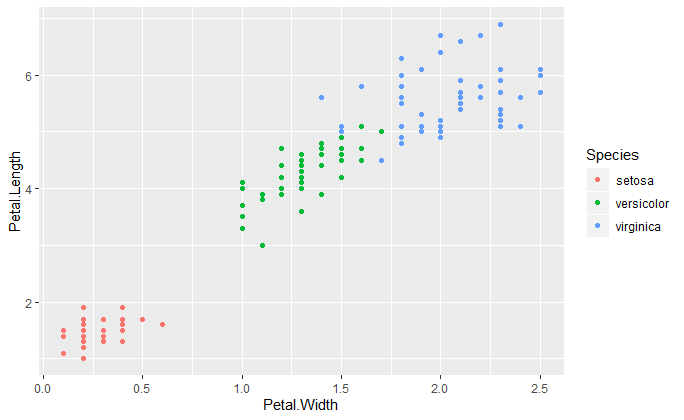
#observed setosa versicolor virginica

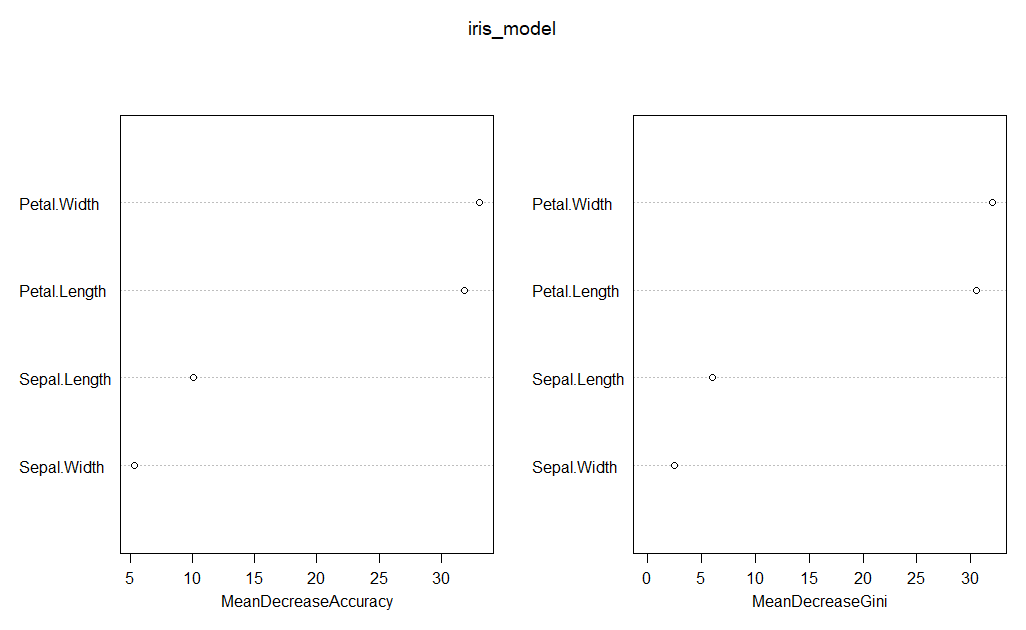
#setosa 13 0 0

#versicolor 0 13 0

#virginica 0 0 16

Qplot:





**Interpretation**:

With qplot,MeanDecreaseAccuracy and MeanDecreaseGini petal variable are important

feature for predicting species.