```
# DATA230 Decision Trees and CART (in-class demo code)

# Don't forget to install the following R packages first!
#install.packages("naniar")
#install.packages("rpart")
#install.packages("rpart.plot")
#install.packages("ggformula")

library(naniar) #naniar is a package to make it easier to summarise and handle missing values
data("iris")
#iris <- subset(iris,Species!="setosa") #remove "setosa" species
any_na(iris) #Dtree algorithms cannot take N/A value, we need to check the N/A value first.
n <- nrow(iris)
set.seed(1117) #specify seeds
#new <- iris[sample(n),]
t_idx <- sample(seq_len(n), size = round(0.7 * n)) #random sample 70% data as training data,
the rest 30% is the test data
traindata <- iris[t_idx,]
testdata <- iris[ - t_idx,]
library(rpart) #"rpart" package is used to implement CART, the split default to Gini, it can
be replace to Information
library(rpart.plot)
tree <- rpart(Species ~ ., data = traindata,
               method = "class") #change to anova for numerical

#pruning process, find the best stopping point.
printcp(tree) #find the best/optimal stopping point (CP, complexity paramater)
tree.pruned <- prune(tree,cp = tree$cptable[which.min(tree$cptable[,"xerror"]),"CP"])

rpart.plot(tree.pruned,digits=2)

future <- predict(tree.pruned, testdata, type="class")
future <- as.data.frame(future)
final <- cbind(future, testdata)
confusion <- table(final$Species,final$future, dnn = c("truth", "predicted"))
confusion
accuracy <- sum(diag(confusion)) / sum(confusion)
accuracy

#plot the iris dataset on "Sepal Length + Sepal Width" and "Petal Length + Petal Width" plane
library(ggformula)
scatterplot1=gf_point(Sepal.Length ~ Sepal.Width, data = iris, color = ~ Species) %>%
  gf_labs(title = "Figure 1: Scatterplot of Iris Data")
scatterplot2=gf_point(Petal.Length ~ Petal.Width, data = iris, color = ~ Species) %>%
  gf_labs(title = "Figure 1: Scatterplot of Iris Data")
scatterplot1
scatterplot2
```