Homework 5 - Random Forests

Objectives: In your own words comment ***Every Line*** of the below code to understand how Random Forests are implemented in R.

Your comments for the libraries and similar lines of code can be fairly simple but lines more directly related to the Random Forest should contain more detail: explain what each term means and why it is used not only in the code but through the theory of Random Forests.

I have also left questions commented in the code. Make sure to answer each.

```
require(randomForest)
require(MASS)
library(ggplot2)
attach(Boston)
dim(Boston)

set.seed(101)
train=sample(1:nrow(Boston),300)

#Why was the training set created differently than how we created training and
testing data sets when writing code for Bayesian Classification and Logistic
Regression?

Boston.rf=randomForest::randomForest(medv ~ . , data = Boston ,
                                     subset = train)

Boston.rf
plot(Boston.rf)
oob.err=double(13)
test.err=double(13)
for(mtry in 1:13){
  rf=randomForest::randomForest(medv ~ . ,
                               data = Boston ,
                               subset = train,
                               mtry=mtry,
                               ntree=400)
  oob.err[mtry] = rf$mse[400]

  pred<-predict(rf,Boston[-train,])
  test.err[mtry]=with(Boston[-train,],mean((medv - pred)^2))
  cat(mtry," ")
}

#Why is there a for loop?

test.err
```

```
oob.err # what is OOB stand for and what does it have to do with Random
Forests? How does Bootstrap come into this?
matplot(1:mtry, cbind(oob.err,test.err), pch=19 ,
col=c("red","blue"),type="b",ylab="Mean Squared Error",xlab="Number of
Predictors Considered at each Split")
#What is a split?
legend("topright",legend=c("Out of Bag (Validation Set) Error","Test (Training
Set) Error"),pch=19, col=c("red","blue"))

#Is the above code doing a classification or using regression to predict a
value? What is the difference between those two things? How is each calculated?
(use words to explain, not code)

ImpData <- as.data.frame(importance(Boston.rf))
ImpData$Var.Names <- row.names(ImpData)

ggplot(ImpData, aes(x=Var.Names, y=`IncNodePurity`)) +
  geom_segment( aes(x=Var.Names, xend=Var.Names, y=0, yend=`IncNodePurity`),
color="skyblue") +
  geom_point(aes(size = IncNodePurity), color="blue", alpha=0.6) +
  theme_light() +
  coord_flip() +
  theme(
      legend.position="bottom",
      panel.grid.major.y = element_blank(),
      panel.border = element_blank(),
      axis.ticks.y = element_blank()
  )

#Explain how to read and understand each visualization the code produces.
```