HW05_#5

Philip Lee

11/4/2019

#Part A

```
set.seed(2013)
sample.data<-sample.int(nrow(students), floor(.50*nrow(students)), replace = F)
train<-students[sample.data, ]    ##splits the data in halves
test<-students[-sample.data, ]
```

#Part B

```
OLSresult<-lm(GPA~Gender+Smoke+Marijuan+DrivDrnk+PartyNum+DaysBeer+StudyHrs,data=train)
##Fits OLS in training data
preds<-predict(OLSresult,newdata=test)     ##calculates predicted values of test data
mean((preds-test$GPA)^2)  ##calculates the test MSE
```

```
## [1] 0.1962592
```

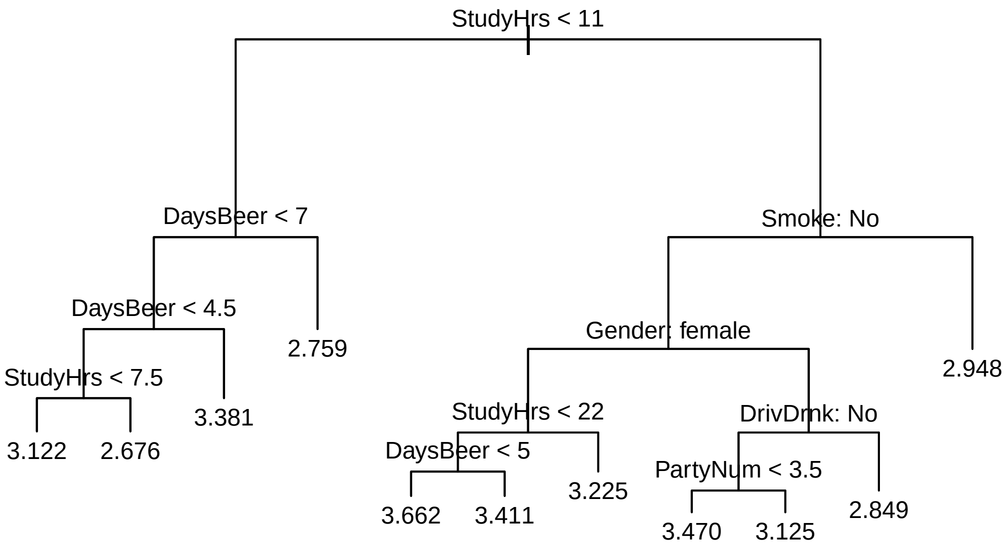
0.1963 for test MSE

#Part C

```
tree.class.train<-tree(GPA~., data=train) ##Fit a regression tree
summary(tree.class.train)  ##Finds the # of terminal nodes
```

```
##
## Regression tree:
## tree(formula = GPA ~ ., data = train)
## Variables actually used in tree construction:
## [1] "StudyHrs" "DaysBeer" "Smoke"    "Gender"   "DrivDrnk" "PartyNum"
## Number of terminal nodes:  11
## Residual mean deviance:  0.1833 = 19.61 / 107
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -1.05500 -0.26630  0.05151  0.00000  0.26920  0.87110
```

```
plot(tree.class.train) ##plots the tree
text(tree.class.train, cex=0.75, pretty=0)
```



11 terminal nodes

#Part D

```
yhat<-predict(tree.class.train, newdata=test) ##Gets the predicted y values based on tree
students.test<-test[,"GPA"]  ##Stores the GPA from test data
mse.tree<-mean((students.test-yhat)^2) ##calculates the MSE
mse.tree
```

```
## [1] 0.3057565
```

0.3058 for test MSE

#Part E

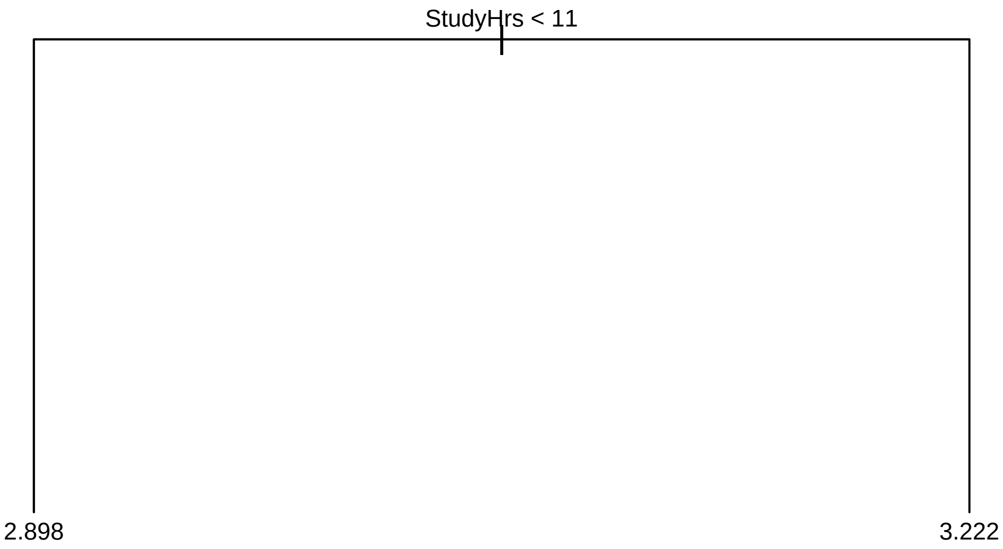
```
set.seed(1)
cv.class<-cv.tree(tree.class.train, K=10) ##Performs 10-fold CV
trees.num.class<-cv.class$size[which.min(cv.class$dev)] ##gets the size leading to smallest
##deviance
trees.num.class
```

```
## [1] 2
```

2 terminal nodes

#Part F

```
prune.full<-prune.tree(tree.class.train, best=trees.num.class) ##Prunes the tree
plot(prune.full)  ##plots the pruned tree
text(prune.full, cex=0.75)
```



Students who studied less than 11 hours are predicted as having 2.898 for GPA and students who studied more than 11 hours are predicted as having 3.222 GPA.

#Part G

```
prune.yhat<-predict(prune.full, newdata=test) ##Gets predicted y values for test data
##based on pruned tree
mse.prune<-mean((students.test-prune.yhat)^2)  ##calculates the test MSE
mse.prune
```

```
## [1] 0.2170533
```

0.2171 for test MSE for pruning

#Part H

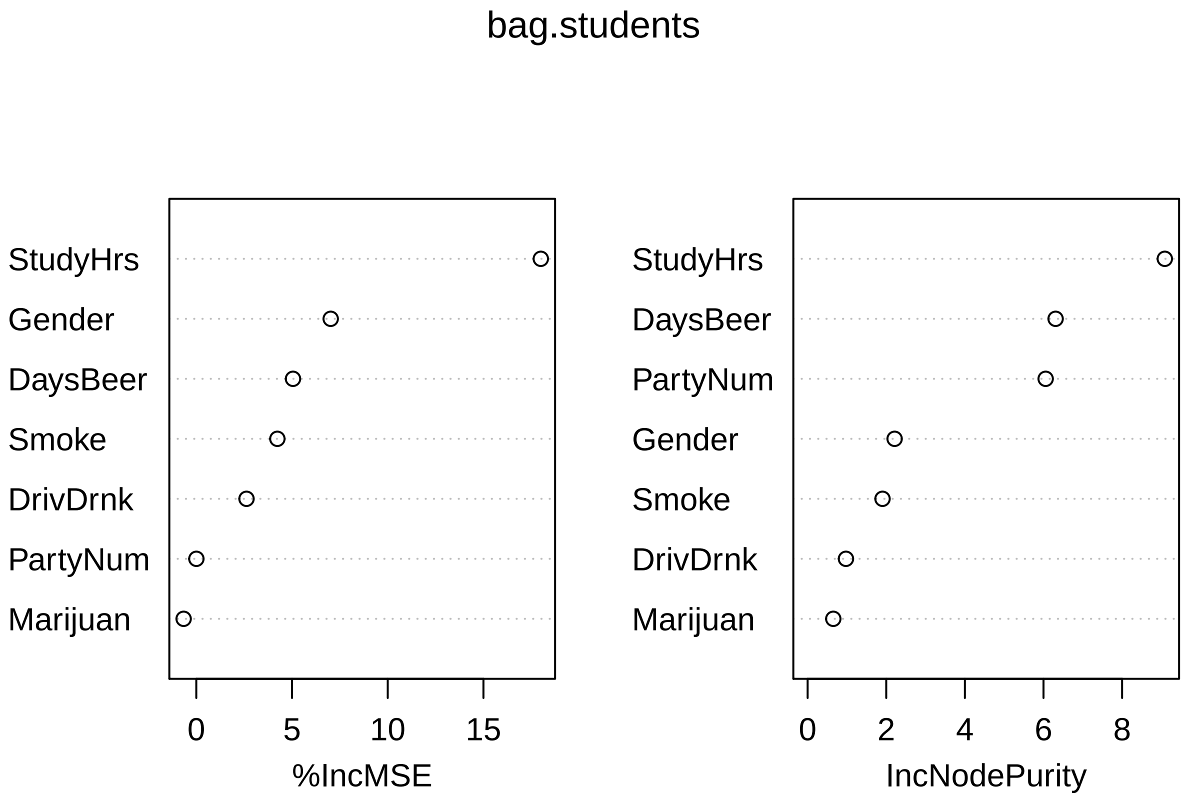
```
set.seed(2)
bag.students<-randomForest(GPA~., data=train, mtry=7, importance=TRUE) ##Uses bagging to
##fit regression tree on training data
yhat.bag<-predict(bag.students, newdata=test) #Gets the predicted values for test
mse.bag<-mean((students.test-yhat.bag)^2)  ##calculates test MSE for Bagging
mse.bag
```

```
## [1] 0.2671621
```

```
round(importance(bag.students),2)
```

```
##          %IncMSE IncNodePurity
## Gender      7.02          2.21
## Smoke       4.24          1.91
## Marijuan   -0.66          0.65
## DrivDrnk    2.62          0.97
## PartyNum    0.01          6.05
## DaysBeer    5.05          6.31
## StudyHrs   18.00          9.09
```

```
varImpPlot(bag.students)
```



0.2672 for test MSE for bagging.

For the importance based on mean decrease of accuracy (graph on left side), variable StudyHrs is the most important variable. For the importance based on the decrease in training RSS (graph on right side), StudyHrs, DaysBeer, and PartyNum are the important variables.

#Part I

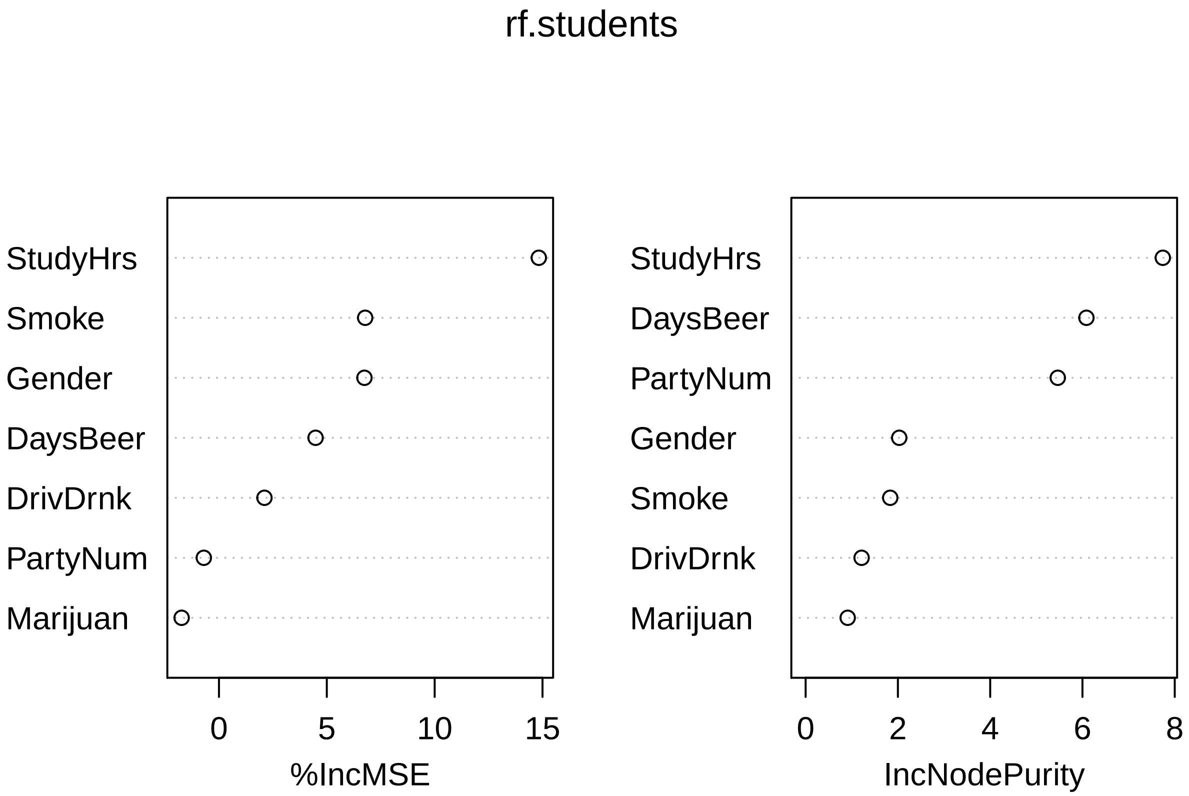
```
set.seed(2)  ##same as part H but with random forest
rf.students<-randomForest(GPA~., data=train, mtry=3, importance=TRUE) ##Uses random forests
yhat.rf<-predict(rf.students, newdata=test)
mse.rf<-mean((students.test-yhat.rf)^2)
mse.rf
```

```
## [1] 0.2389143
```

```
round(importance(rf.students),2)
```

```
##          %IncMSE IncNodePurity
## Gender      6.74          2.03
## Smoke       6.78          1.83
## Marijuan   -1.73          0.91
## DrivDrnk     2.11           1.21
## PartyNum    -0.70           5.47
## DaysBeer     4.48           6.09
## StudyHrs    14.83           7.74
```

```
varImpPlot(rf.students)
```



Test MSE for random forest = 0.2389

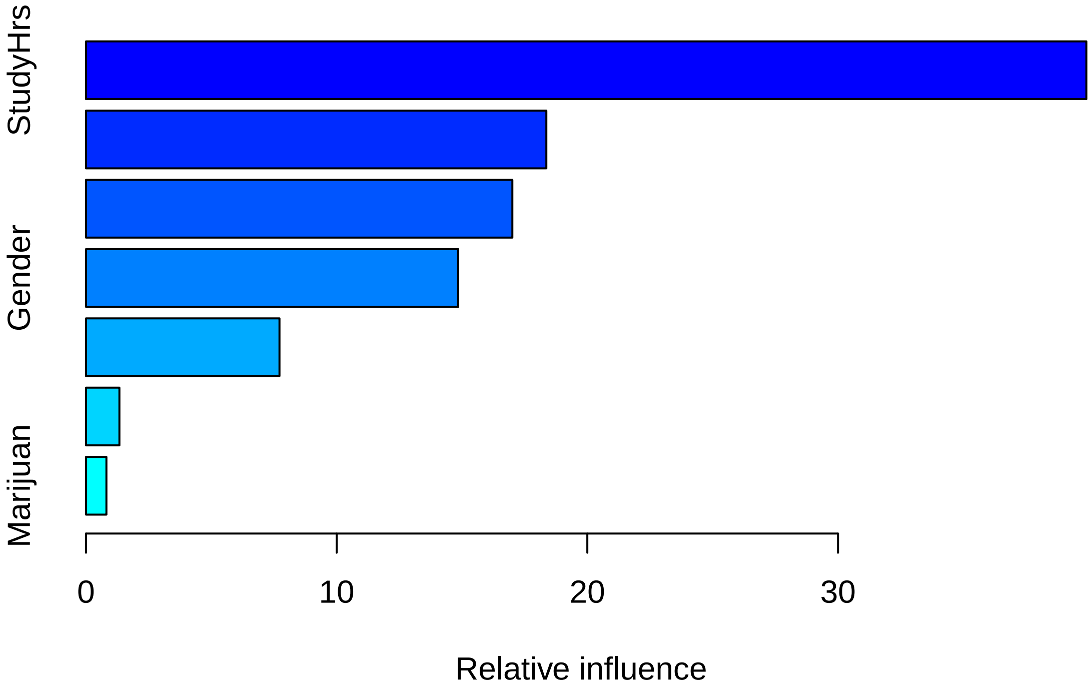
For the importance based on mean decrease of accuracy (graph on left side), variable StudyHrs is the most important variable. For the importance based on the decrease in training RSS (graph on right side), StudyHrs, DaysBeer, and PartyNum are the important variables. Similar to the result for part H

#Part J

```
set.seed(2) ##same as Parth H but with Boosting
boost.students<-gbm(GPA~., data=train, distribution="gaussian", n.trees=5000, shrinkage = 0.0001, intera
yhat.boost<-predict(boost.students, newdata=test,n.trees=5000)
mse.boost<-mean((students.test-yhat.boost)^2)
mse.boost
```

```
## [1] 0.202228
```

```
summary(boost.students)
```



```
##               var   rel.inf
## StudyHrs StudyHrs 39.907643
## DaysBeer DaysBeer 18.365373
## Smoke       Smoke 17.009407
## Gender     Gender 14.849067
## PartyNum PartyNum  7.718825
## DrivDrnk DrivDrnk  1.333953
## Marijuan Marijuan  0.815733
```

Boost test MSE = 0.2022

Can't use importance() for boosting but summary() shows that StudyHrs is the most important variable.