CS771A Assignment 1: Decision Trees

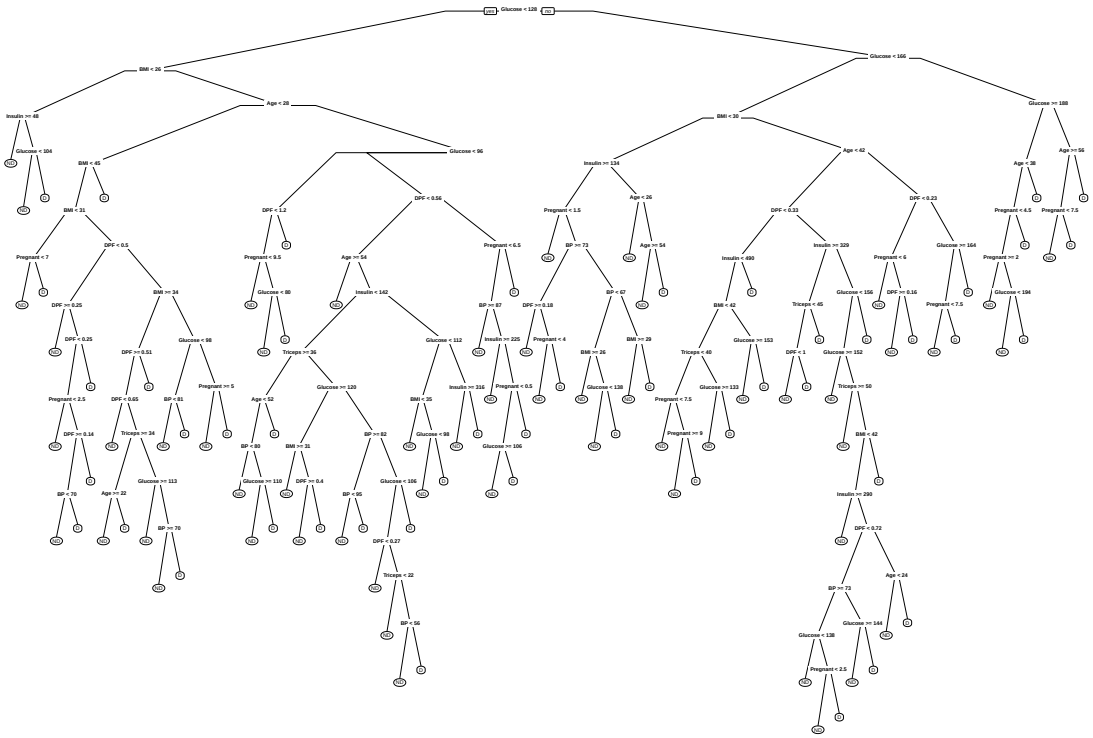
Saurav Kumar (12641)

January 19, 2014

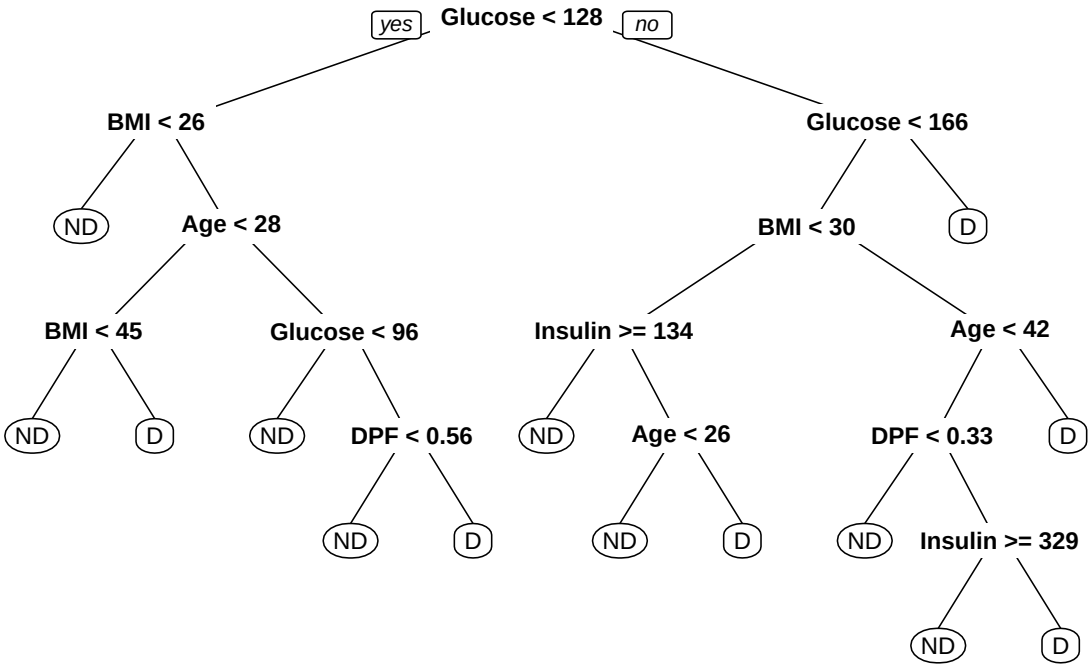
```
library(rpart)
library(rpart.plot)
set.seed(10)
rawData = read.csv(file="data", header=F, sep=",")
originalData = rawData[sample(nrow(rawData)),]
colnames(originalData) = c("PregnantCount","Glucose","BP","Triceps",
                           "Insulin","BMI","DPF","Age","Class")
N = nrow(originalData)
K = 5
foldWidth = floor(N/K)
# Missing Values can be in Glucose, BP, Triceps, Insulin, BMI
Accuracy = 0
for (i in (1:K))
{
    data = originalData
    data$Glucose[data$Glucose==0] = NA          # Missing Data # c
    data$BP[data$BP==0] = NA
    data$Triceps[data$Triceps==0] = NA
    data$Insulin[data$Insulin==0] = NA
    data$BMI[data$BMI==0] = NA

    start = as.integer((i-1)*foldWidth)+1
    end = as.integer(i*foldWidth)
    if(i==K)
    {
        end = N
    }
    testData = data[c(start:end),]
    learnData = data[c(-start:-end),]
    diabStat = factor(learnData$Class, levels=0:1, labels=c('ND','D'))
    cfit = rpart(
                    diabStat ~ PregnantCount+Glucose+BP+Triceps+Insulin+BMI+DPF+Age,
                    data = learnData,
                    na.action = na.rpart,
          method ='class',
                    parms = list(split = "information"),
                    control = rpart.control(
                                                cp = 0.0,    # Threshold complexity parameter
                                                minsplit = 1,    # Min no. of obs. for which the routine
                                                minbucket = 1,  # Min no. of obs in leaf. Default = min
                                            )
                )
    opt = cfit$cptable[which.min(cfit$cptable[,"xerror"]),"CP"]                          # Pruning
    prunedTree = prune(cfit, cp = opt)
    predictedFactor = predict(prunedTree, testData, type="class")
    predictedFrame = as.data.frame.factor(predictedFactor)
    predicted = c(predictedFrame[ ,1]) - 1
```

```
    actual = testData$Class
    TP = sum(predicted & actual)
    TN = nrow(testData) - sum(predicted | actual)
    # Accuracy
    print((TP+TN)/nrow(testData))
    Accuracy = Accuracy + (TP+TN)/nrow(testData)
    print("Unpruned Tree")
    rpart.plot(cfit)
    print("Pruned Tree")
    rpart.plot(prunedTree)
}
```

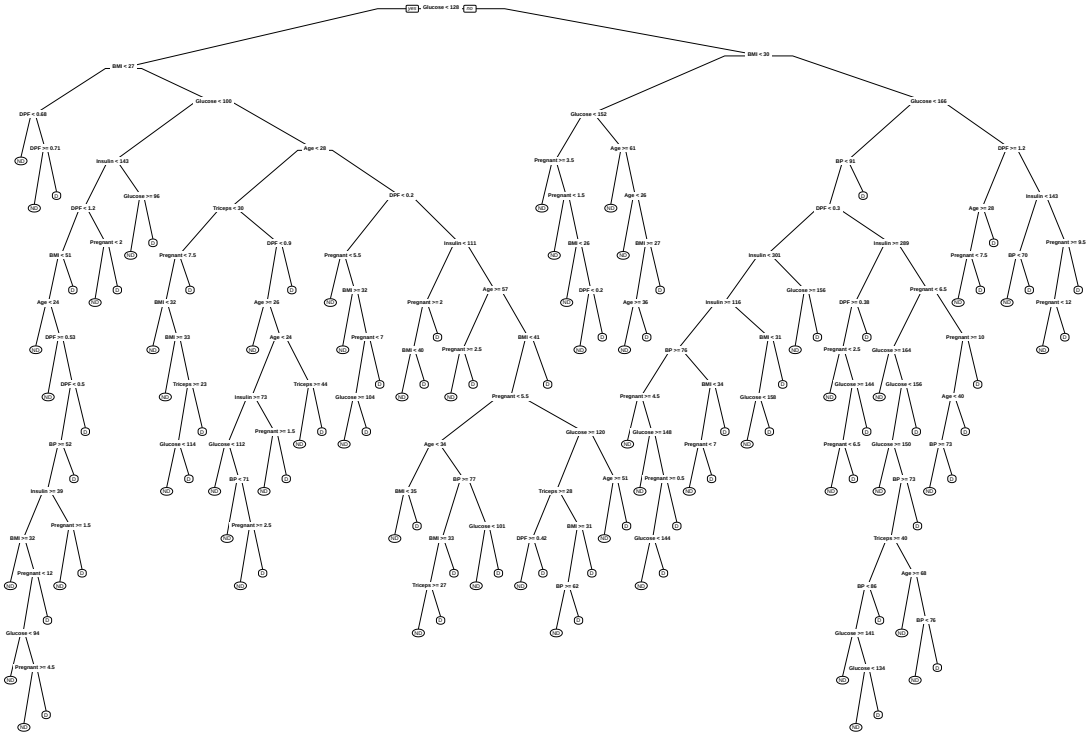
```
## [1] 0.7254902
## [1] "Unpruned Tree"
```



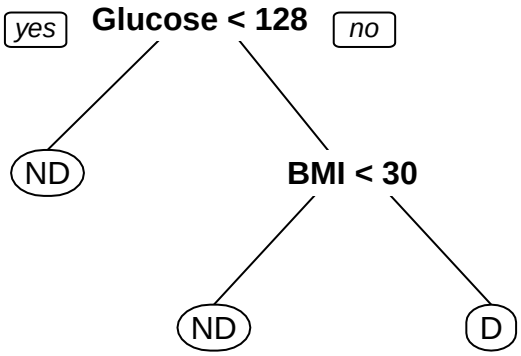

```
## [1] "Pruned Tree"
```



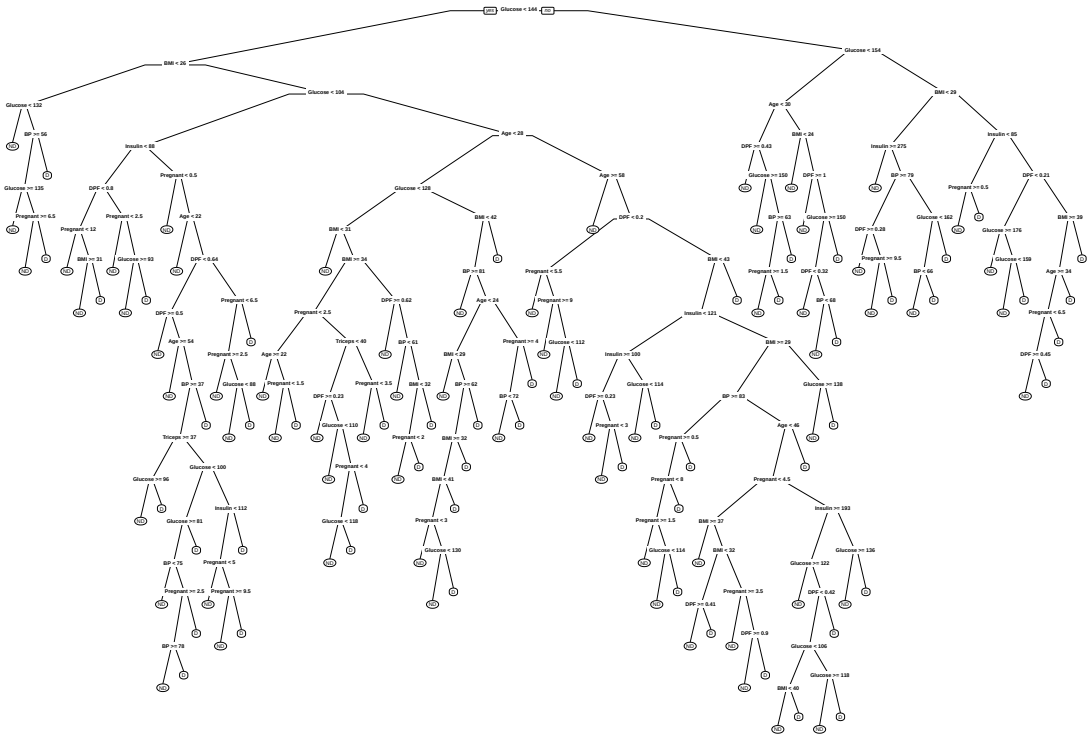

```
## [1] 0.7777778
## [1] "Unpruned Tree"
```



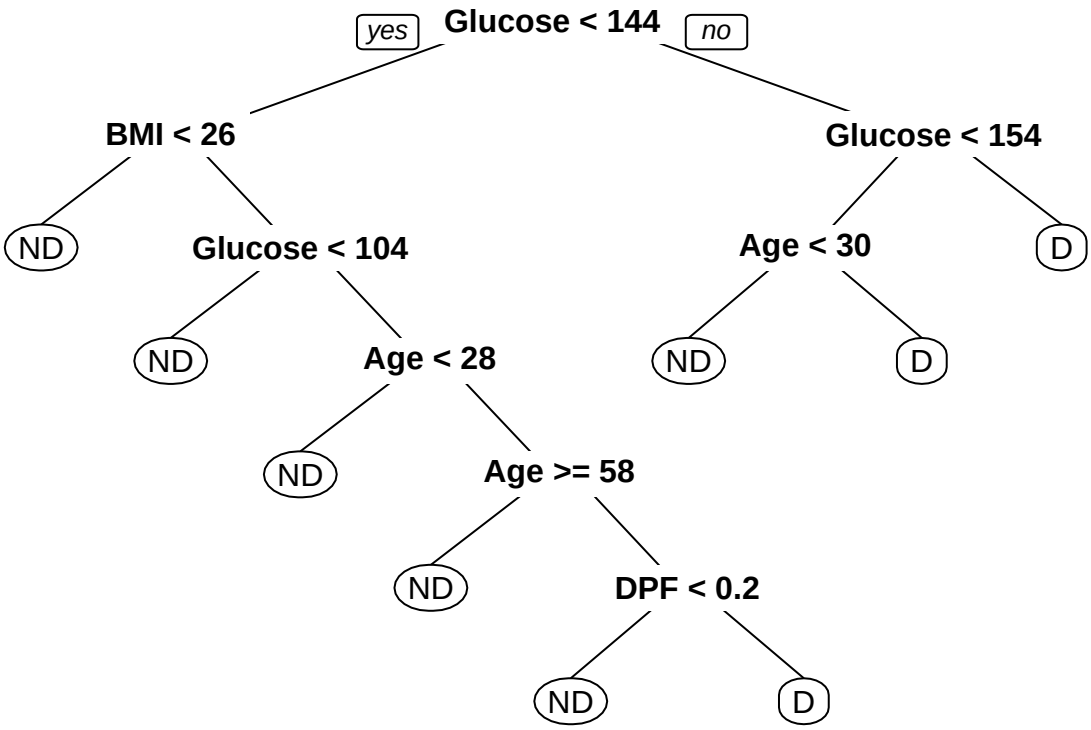
```
## [1] "Pruned Tree"
```



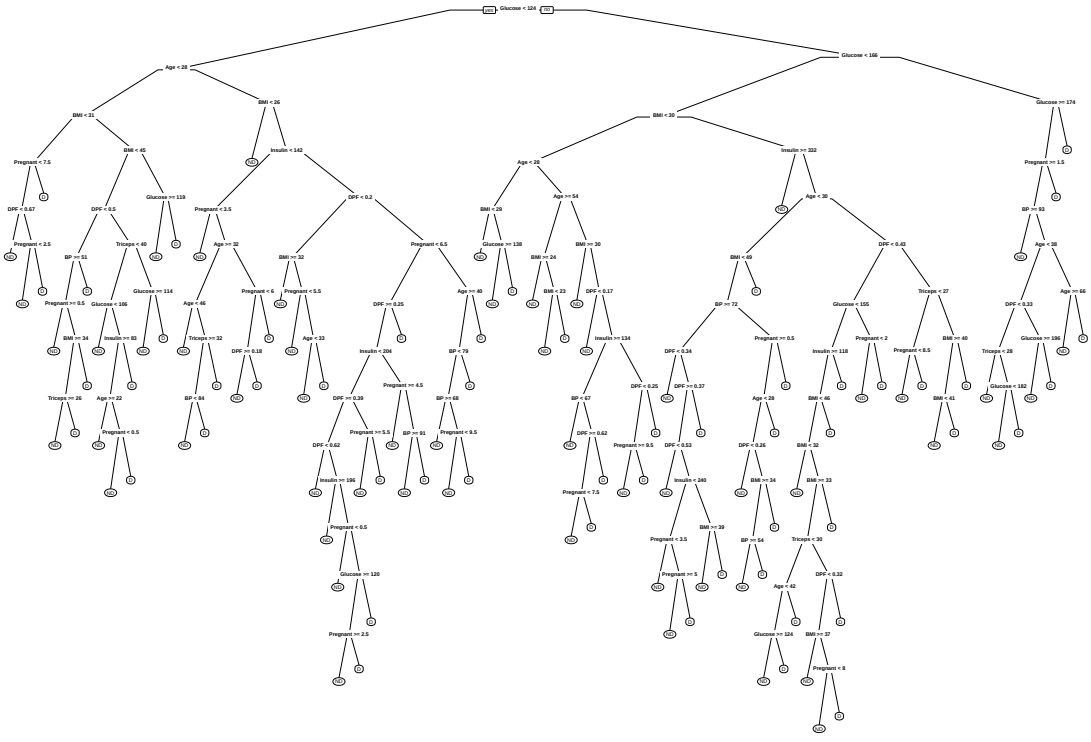

```
## [1] 0.745098
## [1] "Unpruned Tree"
```



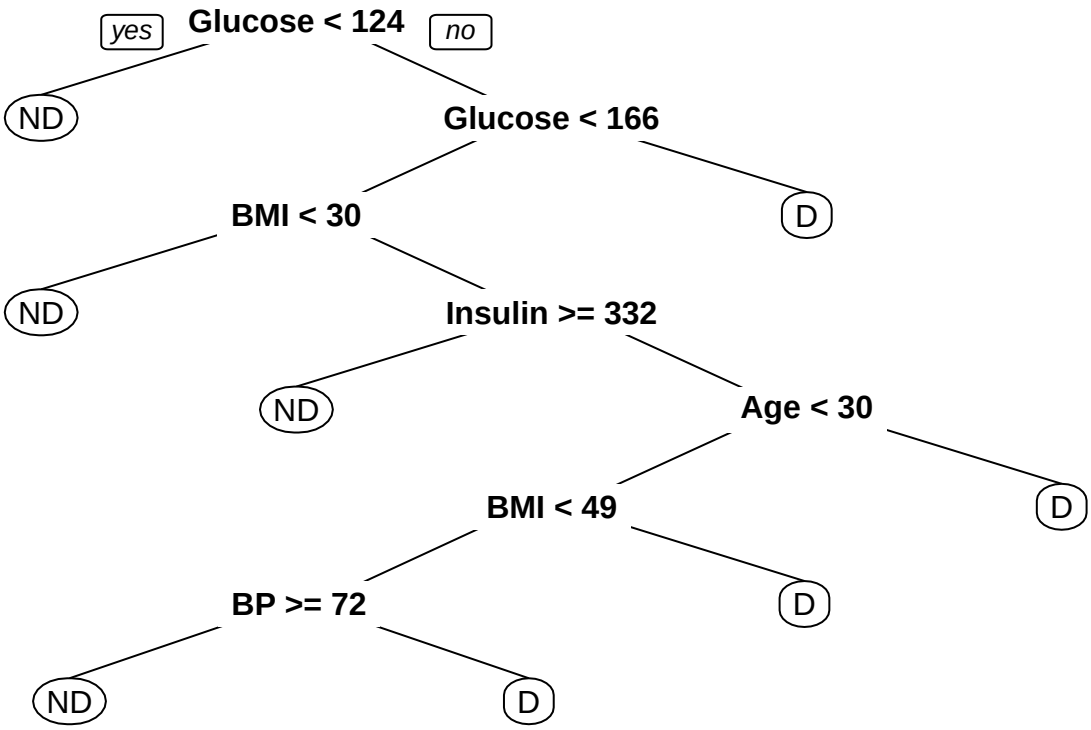
```
## [1] "Pruned Tree"
```




```
## [1] 0.7254902
## [1] "Unpruned Tree"
```

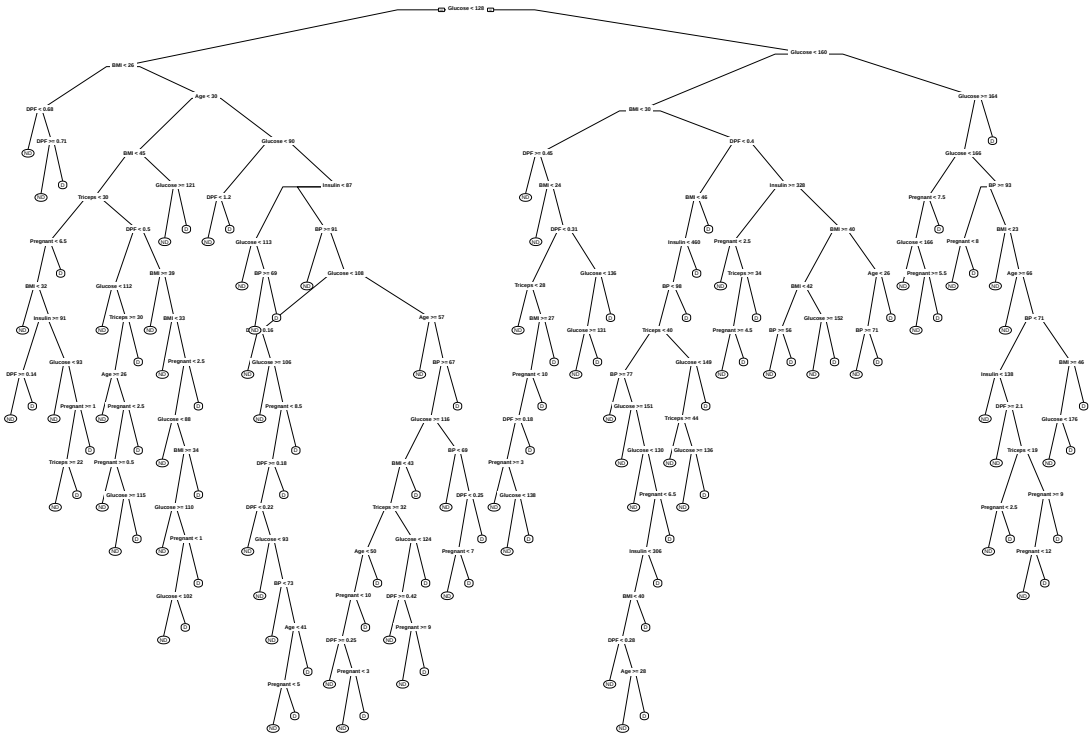


```
## [1] "Pruned Tree"
```

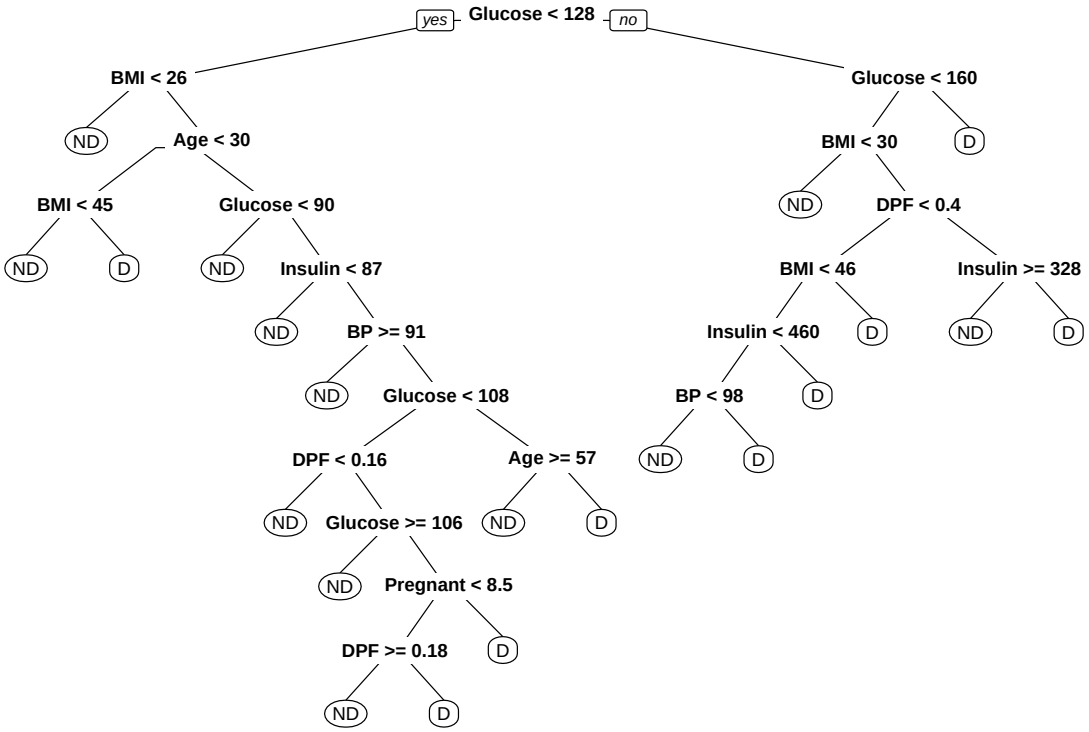



```
## [1] 0.7628205
## [1] "Unpruned Tree"
```

```
## Warning: labs do not fit even at cex 0.15, there may be some overplotting
```



```
## [1] "Pruned Tree"
```




```
#Mean Accuracy
print(Accuracy/K)
```

```
## [1] 0.7473353
```