# Biomechanical Features of Orthopedic Patients

*Pamela Gipe*

*6/6/2019*

## Introduction

The dataset Biomechanical Features of Orthopedic Patients was downloaded from Kaggle, and was selected based on a a clear expectation that the variables included could provide reasonable predictions for certain outcomes. There were two datasets that included 6 independent variables, specificially physical charateristics (the shape and orientation) of the spine and pelvis regions. There was also a column for the diagnosis of the spine, which included either three outcomes or two outcomes depending on the dataset. The relationships between the outcomes and the independent variables was evaluated for this project.

## Summary

The data was previously cleaned and curated by Kaggle, so there was not much more to be done. The data was analyzed for missing values but none were found. There were two .csv files in the dataset, and both were used in this project. One dataset (3C_weka) had three outcomes, and the other (2C_weka), had two outcomes, identified as 'Normal' and 'Abnormal'. These outcomes were converted to factors in order to anaylze them. Initially exploratory analysis and data visualization was performed separately on the dataset with three outcomes which showed an overview of the data, but then the use of an early stage Random Forest led to the observation that one variable(degree_spondylolisthesis) had more influence over the outcomes than the other variables.

Based on this observation simple linear regression was run (one variable), which showed less correlation than expected. Then exploratory analysis and data visualization was performed on the second dataset(2c_weka), This second dataset was then split into two groups, one with 80% of the data for training purposes, and the other with the remaining 20% of the data for validation and testing. This allowed for the machine learning processes, specifically KNN, Random Forest, and a classification tree to see if there were more robust algorythms that could predict outcomes better. The second dataset was used for this more complex set of analyses because two of the outcomes could be classified as abnormal, and it made more sense to evaluate them together rather than trying to predict two different abnormal outcomes and one normal outcome. Another approach would have been to merge the two datasets, and predict either the three outcomes or the two outcomes. It seemed cleaner to keep the datasets separate however.

## Methods

Data exploration was done through the use of ggplot generating a bar graph illustrating the different numbers of the three outcomes, Hernia, Spondylothisthesis and Normal. A summary of the first dataset was also provided. A correlation of all the variables (except the outcomes) was also performed in order to see what variables were influencing other variables and to what extent.

An initial statistical analysis was performed on the first dataset. A Random Forest evaluation provided MSE (mean squared error) values for each variable. A tree visualization supported the results of the random forest evaluation.

At this point exploratory data analysis was performed on the second dataset, in which the outcomes had been combined into just two, with hernia and Spondylothisthesis combined into one outcome defined as

abnormal. Here the same data exploration was done to see if there were major differences. The correlation was not done again because the data in those columns was the same.

The next step was to examine the data usin machine learning techniques. The data from dataset two were split into a training set (80% of the data), and a validation/test set (20%) of the data. Two machine learning techniques were evaluated, K nearest neighbors (KNN), which is a simple algorithm that calculates distances between variables and outcomes, and rpart, which evaluates a series of factors to determine the one most effective at generating a classification tree. Rpart was also used earlier in the analysis, but without separating into training and validation sets.

# Results

```
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: tidyverse
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.1
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: caret
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```r
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(ggplot2)
library(randomForest)
```

```
## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following object is masked from 'package:dplyr':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin
```

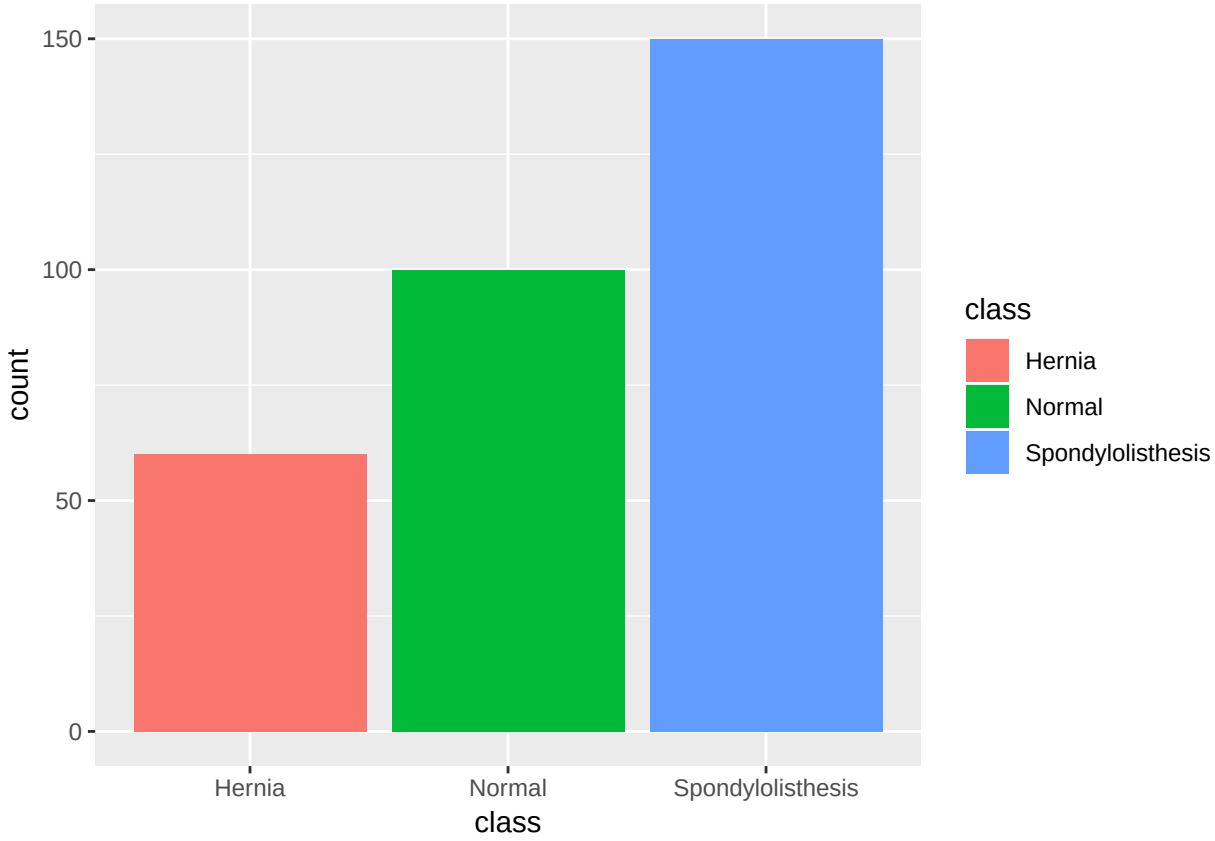
```r
library(rpart)
library(dplyr)
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```r
library(e1071)
#Data collection
dl <- tempfile()
download.file("https://github.com/pjg2016/chooseyourown/files/3267565/CYO.zip", dl)
dataset <- read.csv(unzip(dl, "3C_weka.csv"))
dataset_two <- read.csv(unzip(dl, "2C_weka.csv"))
dataset$class<-as.factor(dataset$class)
rm(dl)

#Random Forest to identify variables with most influence
set.seed(1)
rf_spine<-randomForest(class~.,data = dataset)
```

```
RF_importance <- importance(rf_spine)
RF_dataframe <- data.frame((RF_importance), MSE = RF_importance[,1])

#Tree visualization ~ supporting Random Forest dataframe
fit<- rpart(dataset$class~., data = dataset)

#Exploratory analysis, Dataset 2
dataset_two$class<-as.factor(dataset_two$class)

#Random Forest to identify variables with most influence, Dataset 2
set.seed(1)
rf_spine_two<-randomForest(class~.,data = dataset_two)
RF_importance_two <- importance(rf_spine_two)
RF_dataframe_two <- data.frame((RF_importance_two), MSE = RF_importance_two[,1])

#Tree visualization ~ supporting Random Forest dataframe
fit_two<- rpart(dataset_two$class~., data = dataset_two)

#create training and test sets
stats_index <- createDataPartition(dataset_two$class, p=0.80, list=FALSE)
 #select 20% of the data for testing the models
test <- dataset_two[-stats_index,]
 #use the remaining 80% of data for training the models
datasetTR <- dataset_two[stats_index,]

#knn
set.seed(1)
knnfit <- train(class ~ ., method = "knn",
             tuneGrid = data.frame(k = seq(1, 15, 2)),
             data = datasetTR)

#classification tree
train_rpart <-train(class ~ ., method = "rpart",
tuneGrid = data.frame(cp = seq(0,0.1, len = 25)),
data = datasetTR)
```

**Exploratory Data Analysis Dataset 1 ~ Comparing counts of Hernia, Spondylothisthesis and Normal Outcomes**



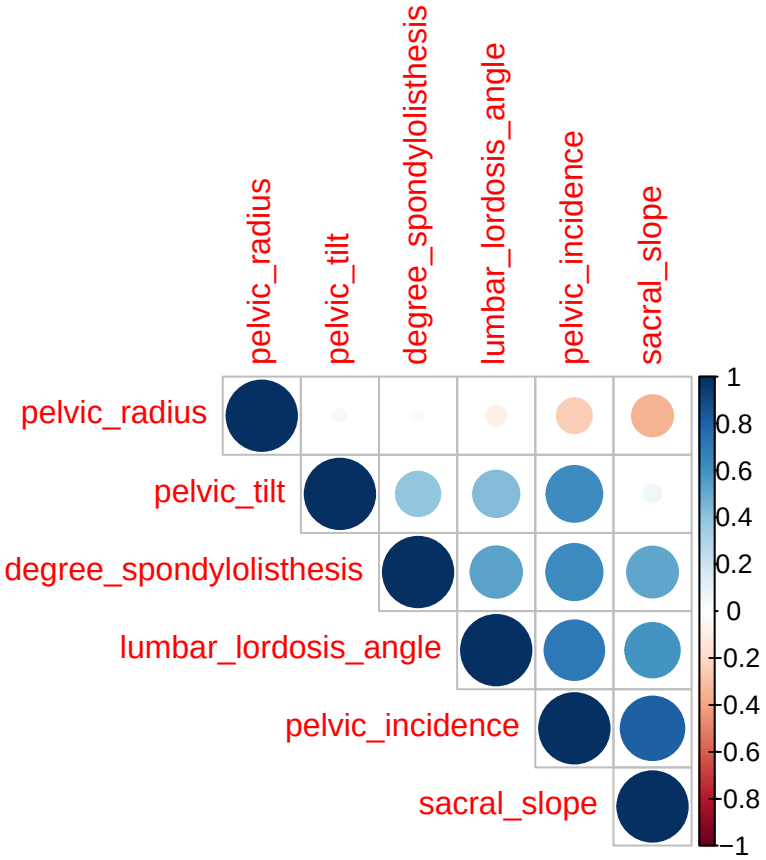

**Exploratory Data Analysis Dataset 1 ~ Five Number Summary including Mean.**

```
##  pelvic_incidence  pelvic_tilt     lumbar_lordosis_angle  sacral_slope
##  Min.   : 26.15   Min.   :-6.555   Min.   : 14.00        Min.   : 13.37
##  1st Qu.: 46.43   1st Qu.:10.667   1st Qu.: 37.00        1st Qu.: 33.35
##  Median : 58.69   Median :16.358   Median : 49.56        Median : 42.40
##  Mean   : 60.50   Mean   :17.543   Mean   : 51.93        Mean   : 42.95
##  3rd Qu.: 72.88   3rd Qu.:22.120   3rd Qu.: 63.00        3rd Qu.: 52.70
##  Max.   :129.83   Max.   :49.432   Max.   :125.74        Max.   :121.43
##  pelvic_radius    degree_spondylolisthesis               class
##  Min.   : 70.08   Min.   :-11.058          Hernia           : 60
##  1st Qu.:110.71   1st Qu.:  1.604          Normal           :100
##  Median :118.27   Median : 11.768          Spondylolisthesis:150
##  Mean   :117.92   Mean   : 26.297
##  3rd Qu.:125.47   3rd Qu.: 41.287
##  Max.   :163.07   Max.   :418.543
```

**Exploratory Data Analysis Dataset 1 ~ Correlations between independent variables.**

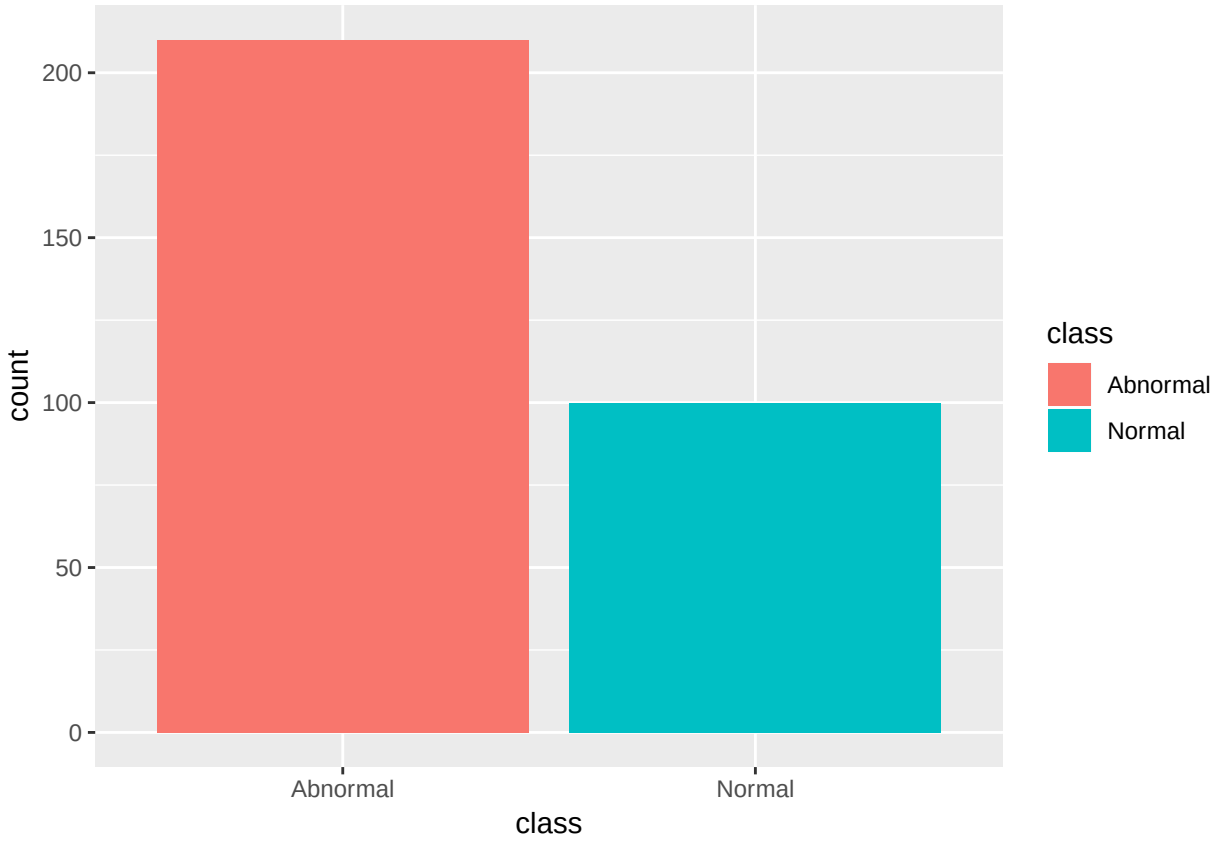
```
##                          pelvic_incidence pelvic_tilt
```

```
## pelvic_incidence                1.0000000  0.62919877
## pelvic_tilt                     0.6291988  1.00000000
## lumbar_lordosis_angle           0.7172824  0.43276386
## sacral_slope                    0.8149600  0.06234529
## pelvic_radius                  -0.2474672  0.03266781
## degree_spondylolisthesis        0.6387427  0.39786228
##                          lumbar_lordosis_angle sacral_slope pelvic_radius
## pelvic_incidence                    0.71728236   0.81495999   -0.24746721
## pelvic_tilt                         0.43276386   0.06234529    0.03266781
## lumbar_lordosis_angle               1.00000000   0.59838689   -0.08034361
## sacral_slope                        0.59838689   1.00000000   -0.34212835
## pelvic_radius                      -0.08034361  -0.34212835    1.00000000
## degree_spondylolisthesis            0.53366701   0.52355746   -0.02606501
##                          degree_spondylolisthesis
## pelvic_incidence                       0.63874275
## pelvic_tilt                            0.39786228
## lumbar_lordosis_angle                  0.53366701
## sacral_slope                           0.52355746
## pelvic_radius                         -0.02606501
## degree_spondylolisthesis               1.00000000
```




The sacral slope and the pelvic incidence show the highest correlations, but pelvic incidence has a strong correlation with all of the variables except pelvic radius. Pelvic radius does not appear to be correlated with any of the other variables however. It is interesting to note that sacral slope is very poorly correlated with pelvic tilt, not withstanding is high correlation with pelvic incidence (which has a high correlation with pelvic tilt).

**Exploratory Data Analysis Dataset 2 ~ Count of Normal vs. Abnormal outcomes.**

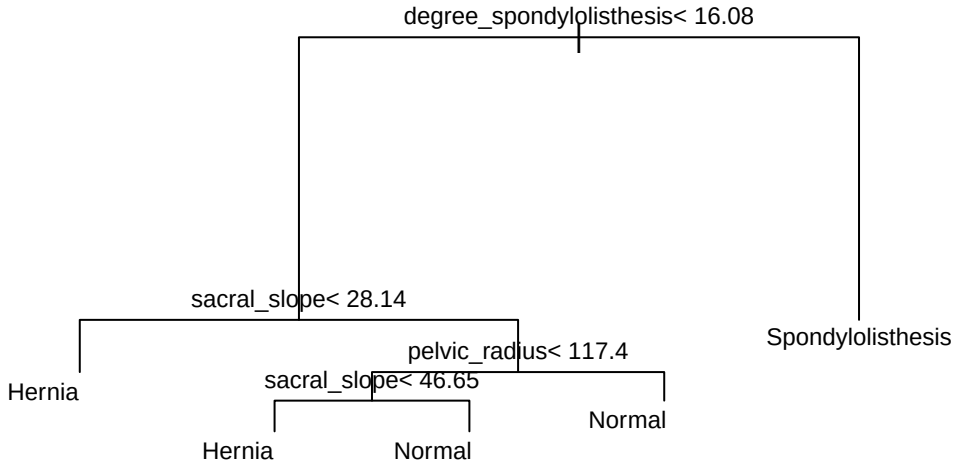



## Random Forest to identify variables with most influence, Dataset 1

```
##                           MeanDecreaseGini      MSE
## pelvic_incidence                  23.29097 23.29097
## pelvic_tilt                       16.62990 16.62990
## lumbar_lordosis_angle             27.10469 27.10469
## sacral_slope                      24.97032 24.97032
## pelvic_radius                     22.64306 22.64306
## degree_spondylolisthesis          78.02385 78.02385
```

The degree of Spondylolisthesis has a very high Mean Square Error (MSE), which suggests that it has the least influence over the outcome.

**Tree visualization Dataset 1 ~ supporting Random Forest dataframe**

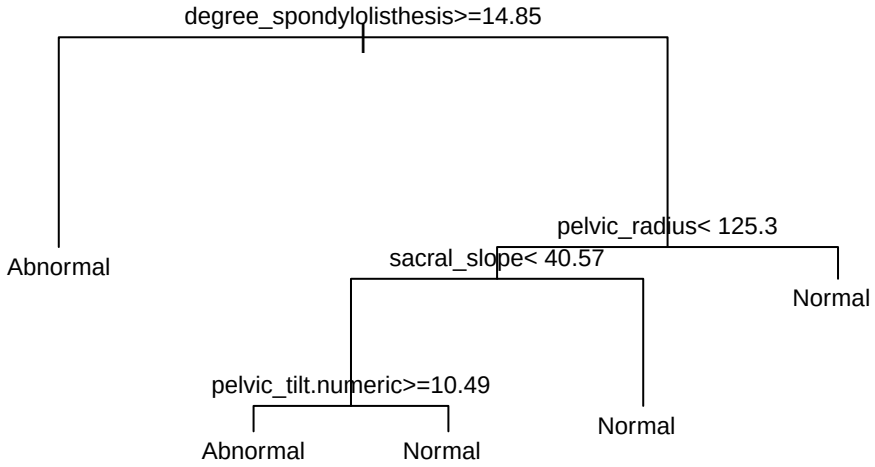



## Random Forest to identify variables with most influence, Dataset 2

```
##                           MeanDecreaseGini      MSE
## pelvic_incidence                   16.97846 16.97846
## pelvic_tilt.numeric                15.67576 15.67576
## lumbar_lordosis_angle              15.12642 15.12642
## sacral_slope                       16.90331 16.90331
## pelvic_radius                      24.04899 24.04899
## degree_spondylolisthesis           46.37651 46.37651
```

When the two abnormal outcomes are combined into one 'abnormal' outcome, the degree of Spondylolisthesis shows more influence over the outcome. This suggests that continuing to include this variable is important.

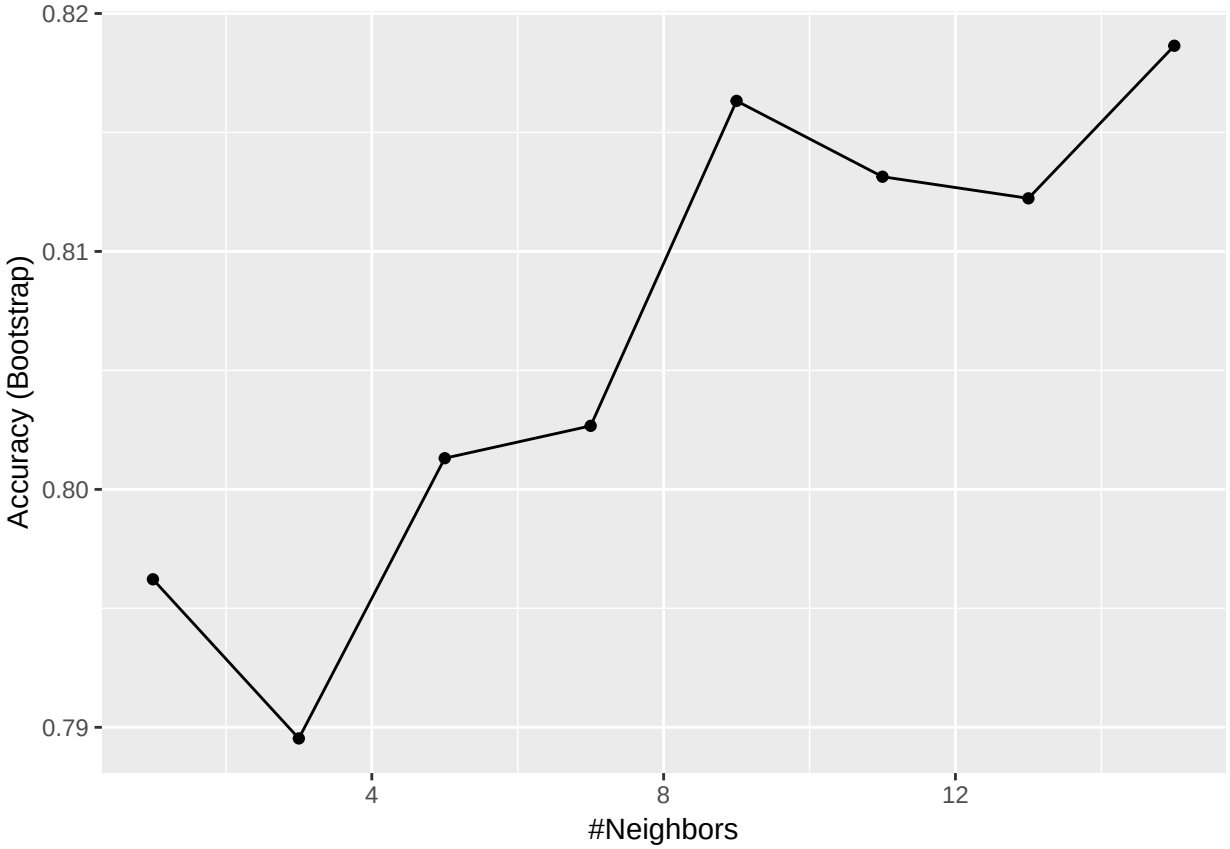
**Tree visualization Dataset 2 ~ supporting Random Forest dataframe**




**Summary of Dataset 2 Training Set ~ Five Number Summary including Mean.**

```
##  pelvic_incidence pelvic_tilt.numeric lumbar_lordosis_angle
##  Min.   : 26.15   Min.   :-6.555      Min.   : 14.00
##  1st Qu.: 46.39   1st Qu.:10.739      1st Qu.: 36.67
##  Median : 58.56   Median :15.917      Median : 49.56
##  Mean   : 60.19   Mean   :17.184      Mean   : 51.31
##  3rd Qu.: 72.72   3rd Qu.:21.998      3rd Qu.: 62.65
##  Max.   :118.14   Max.   :49.432      Max.   :100.74
##   sacral_slope   pelvic_radius    degree_spondylolisthesis      class
##  Min.   :13.37   Min.   : 70.08   Min.   :-11.058          Abnormal:168
##  1st Qu.:33.40   1st Qu.:110.70   1st Qu.:  1.622          Normal  : 80
##  Median :42.65   Median :117.98   Median : 12.388
##  Mean   :43.01   Mean   :117.34   Mean   : 24.086
##  3rd Qu.:53.04   3rd Qu.:125.02   3rd Qu.: 39.570
##  Max.   :79.70   Max.   :163.07   Max.   :148.754
```

### Knn using Dataset 2

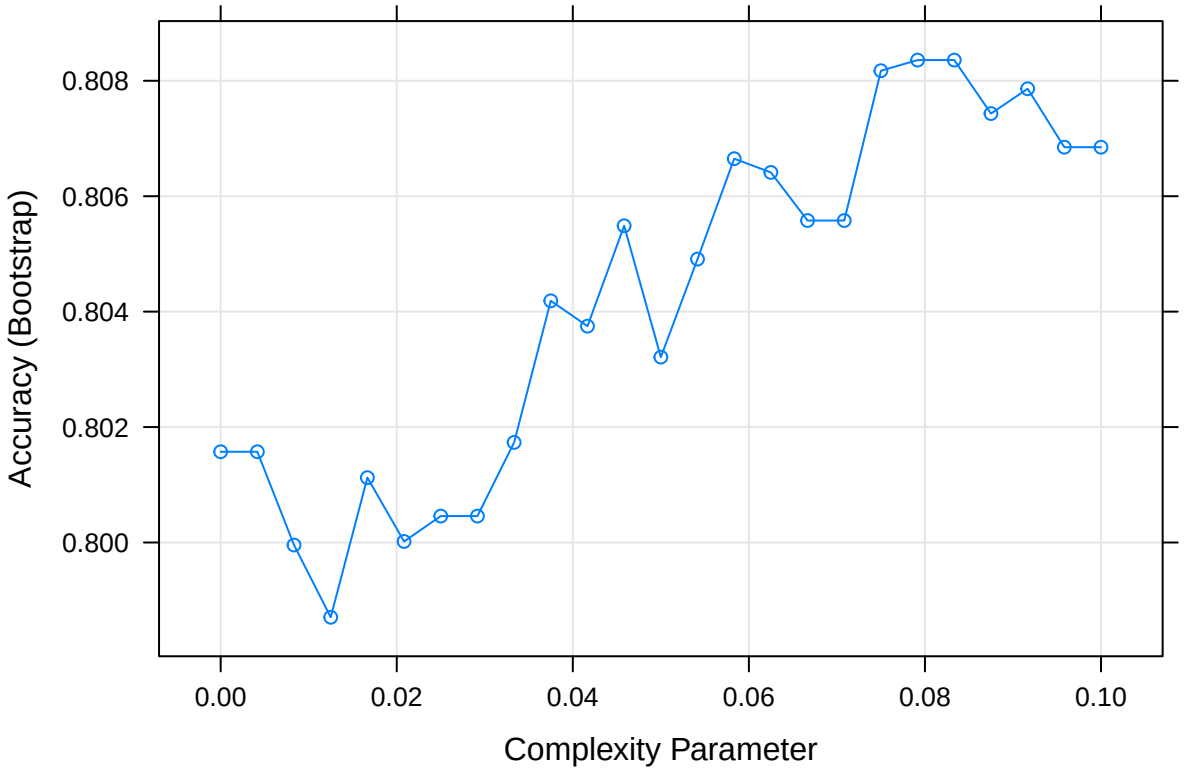



```
## k-Nearest Neighbors
##
## 248 samples
##   6 predictor
##   2 classes: 'Abnormal', 'Normal'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 248, 248, 248, 248, 248, 248, ...
## Resampling results across tuning parameters:
##
##   k   Accuracy   Kappa
##    1  0.7962206  0.5411719
##    3  0.7895325  0.5296484
##    5  0.8013092  0.5552899
##    7  0.8026686  0.5579265
##    9  0.8163251  0.5885780
##   11  0.8131392  0.5800043
##   13  0.8122310  0.5798297
##   15  0.8186429  0.5944798
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 15.
```

```
##  Accuracy
## 0.8870968
```

All of the values of k showed high accuracy, but $\mathbf{k = 15}$ was the highest. This showed a very high accuracy. A concern with this value of k is that higher values of k take longer to execute, and are susceptible to underfitting. However, this dataset is small enough that execution time is not a concern, and low values of k may need to overfitting.

### Classification Tree using Dataset 2




```
## CART
##
## 248 samples
##   6 predictor
##   2 classes: 'Abnormal', 'Normal'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 248, 248, 248, 248, 248, 248, ...
## Resampling results across tuning parameters:
##
##   cp           Accuracy   Kappa
##   0.000000000  0.8015723  0.5398317
##   0.004166667  0.8015723  0.5398317
```

```
##   0.008333333  0.7999562  0.5369133
##   0.012500000  0.7987050  0.5358631
##   0.016666667  0.8011240  0.5390943
##   0.020833333  0.8000182  0.5377785
##   0.025000000  0.8004578  0.5396844
##   0.029166667  0.8004578  0.5403850
##   0.033333333  0.8017344  0.5438099
##   0.037500000  0.8041878  0.5499399
##   0.041666667  0.8037483  0.5515384
##   0.045833333  0.8054874  0.5580861
##   0.050000000  0.8032118  0.5516421
##   0.054166667  0.8049103  0.5560380
##   0.058333333  0.8066496  0.5623298
##   0.062500000  0.8064117  0.5669344
##   0.066666667  0.8055783  0.5654006
##   0.070833333  0.8055783  0.5654006
##   0.075000000  0.8081738  0.5737731
##   0.079166667  0.8083582  0.5746894
##   0.083333333  0.8083582  0.5746894
##   0.087500000  0.8074317  0.5661778
##   0.091666667  0.8078618  0.5688895
##   0.095833333  0.8068491  0.5660728
##   0.100000000  0.8068491  0.5660728
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was cp = 0.08333333.
```

```
##  Accuracy
## 0.7903226
```

The classification tree shows a lower accuracy than the KNN evaluation. This is the same function run earlier as part of the exploratory data analysis, but in this case a tree was not plotted because the goal is to identify accuracy rather than visualize relationships.

## Conclusion

The K-Nearest Neighbors algorythm showed higher accuracy that the classification tree process. This makes sense because while one variable (the degree of Spinal Spondylothisthesis) had significantly less influence over the outcomes, the others did show strong influence over the outcome. Further data analysis (additional correlations) should look at mixes of the variables, and pull out how much the other variables influence each other and how that can obscure how the variables actually influence the outcome (Normal or Abnormal Spines). Additionally, more information about how Normal and Abnormal are determined would be useful. It is possible that there are many borderline cases that could be considered either Normal or Abnormal which can cloud the results. Separating the outcomes into three does not make sense, because two categories are clearly Abnormal, however, changing the classifications to a continuous numerical scale could be very informative. Then the designation of Normal could be given for a range of results, which is more informative than simply a binary result. This would also allow for clearly understandable ranges of abnormal (from slightly abnormal to severely abnormal for example).