Predicting Spinal Disorder Based on Biomechanical Features of Vertebral Column

University of British Columbia

Daniel Lee

I. Introduction

Slipped disc, also known as disk herniation, is a disk rupture symptom when the fibrous connecting agent in between the intervertebral disc escapes out of its location, which consequently arouses a constant and severe degree of pain from irritating the nerves around the area ("Slipped Disk: Overview."). As well, Spondylolisthesis, is a commonly diagnosed spinal disorder that is often misinterpreted as the same symptom as disk herniation. In fact, Spondylolisthesis is a vertebral deformation where one of the columns forming the vertebrae (lower back) slips out of its past the other (above or below) ("Spondylolisthesis Treatment, Surgery &amp; Symptoms."). In the analysis, we take three classifications - Normal, Disk Hernia and Spondylolisthesis, and construct a predictive model with the training set for the diagnosis of a respective spinal disorder based on the predicting variables that we find the most appropriate examining through the exploratory data analysis. The question we are trying to answer is: Can we observe the distinction between the Disk Hernia and Spondylolisthesis based on two of the geometric vertebral column variables?

II. Dataset

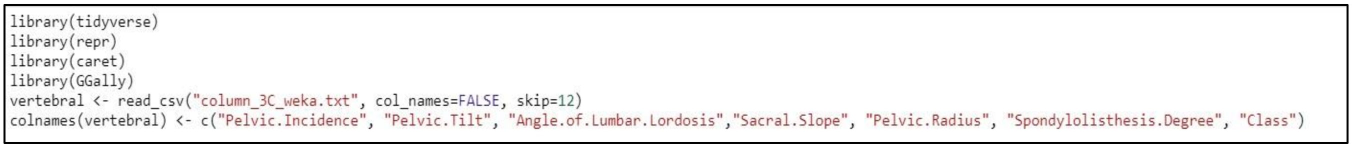
II.I. Data – Vertebral Columns

For this project, we will be examining the *Vertebral Columns Data* from UCI Machine Learning Laboratory that consists of 6 predictor variables (biomechanical features of vertebral column) and 1 class variable (classification of vertebral disorder). The dataset is consisted of 310 observations of patients: 60 diagnosed with Hernia, 150 diagnosed with Spondylolisthesis and 100 diagnosed normal. Additionally, the attributes are as following:

- Angle of Lumbar Lordosis
- Sacral Slope
- Spondylolisthesis Degree
- Pelvic Incidence
- Pelvic Radius
- Pelvic Tilt

II.II. Features and Preprocessing

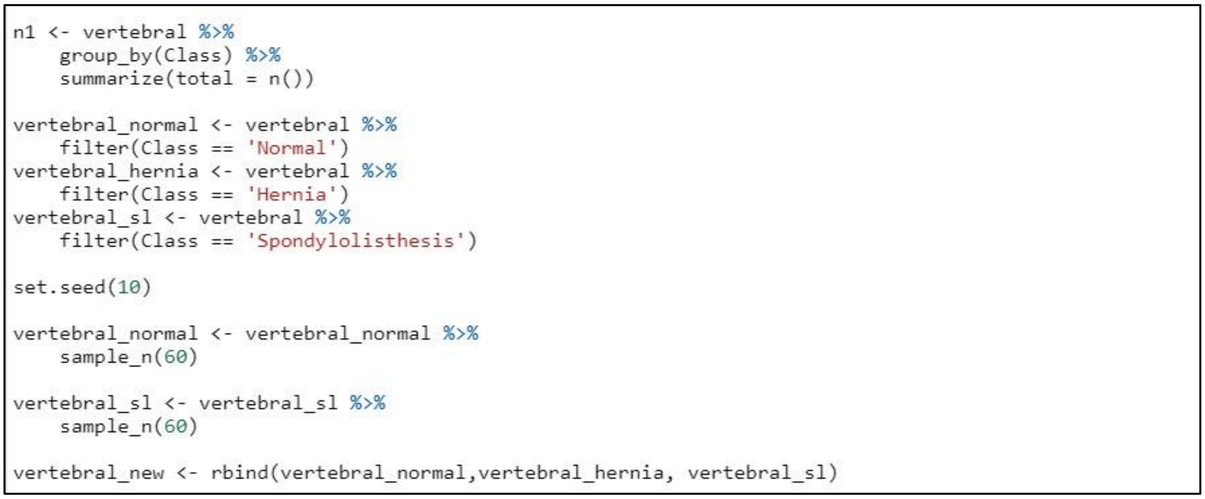
Since the original dataset comes in the DAT and ARFF File, the format must be changed to txt file, which can be done by simply changing the extension of the file name from .dat(or .arff) to .txt. From there we choose the Column_3C_WEKA, since we are concerned about the 3 levels of classification with proper attribute labels. In the original text file, we see that it's a comma separated file, with 12 missing rows on top. We can wrangle by using read_csv function with col_names set to FALSE and Skip = 12 (Fig 1.1). There is no white space or NA included in the table, so we do not have to worry about cleaning the dataset to a further extent.



```
library(tidyverse)
library(repr)
library(caret)
library(GGally)
vertebral <- read_csv("column_3C_weka.txt", col_names=FALSE, skip=12)
colnames(vertebral) <- c("Pelvic.Incidence", "Pelvic.Tilt", "Angle.of.Lumbar.Lordosis","Sacral.Slope", "Pelvic.Radius", "Spondylolisthesis.Degree", "Class")
```

Figure 1.1

Since there are 150 Spondylolisthesis diagnosed patients as opposed to the 60 Hernia diagnosed patients, we were concerned with the class imbalance; the method we use in the prediction of classification is sensitive to the balance of the class size, density and the noise level of the data. For this reason, we have decided to randomly subset the 60 observations for each class (Fig 1.2), and with this we performed the exploratory data analysis on the dataset with a total of 180 observations. It is important to note that we have set the seed to make the code reproducible for the random sampling.



```
n1 <- vertebral %>%
    group_by(Class) %>%
    summarize(total = n())

vertebral_normal <- vertebral %>%
    filter(Class == 'Normal')
vertebral_hernia <- vertebral %>%
    filter(Class == 'Hernia')
vertebral_sl <- vertebral %>%
    filter(Class == 'Spondylolisthesis')

set.seed(10)

vertebral_normal <- vertebral_normal %>%
    sample_n(60)

vertebral_sl <- vertebral_sl %>%
    sample_n(60)

vertebral_new <- rbind(vertebral_normal,vertebral_hernia, vertebral_sl)
```

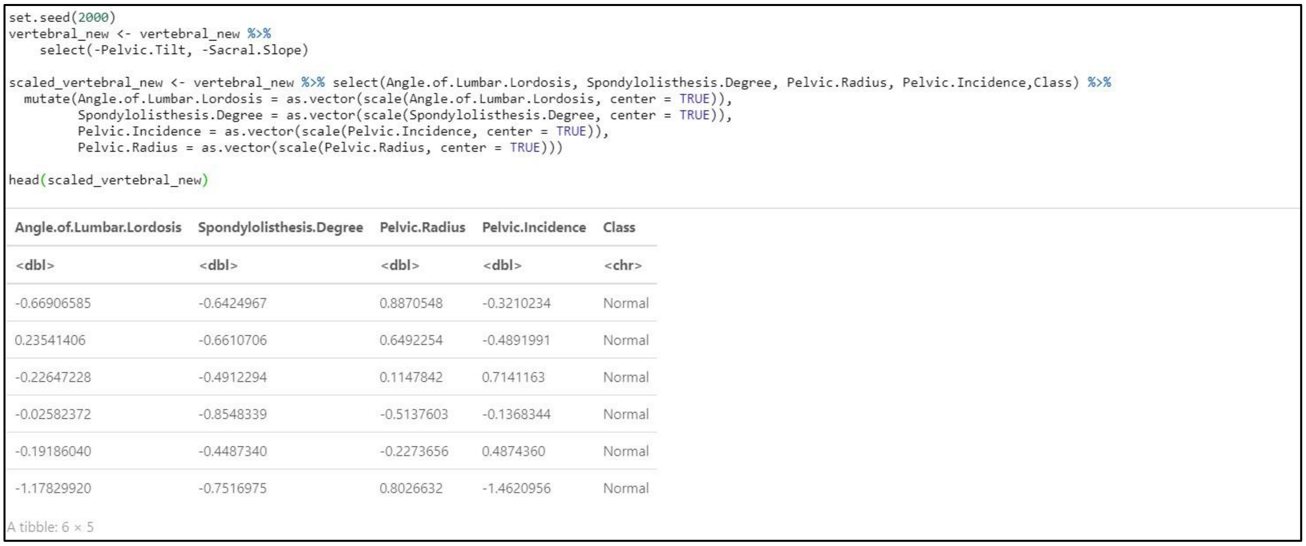
Figure 1.2

III. Methods

III.I. Exploratory Data Analysis

We chose to perform the exploratory data analysis using ggpairs visualization to get preliminary insights on the distribution and correlation of the predictor variables and classes. Prior to the performance of the analysis, we excluded the sacral slope and pelvic tilt, since the two values combined to form a value for pelvic incidence. It is also integral to scale the values before visualization takes place, since there are some discrepancies in between the scale of the variables.

Fig 2.1 shows the filtering and scaling of the variables.



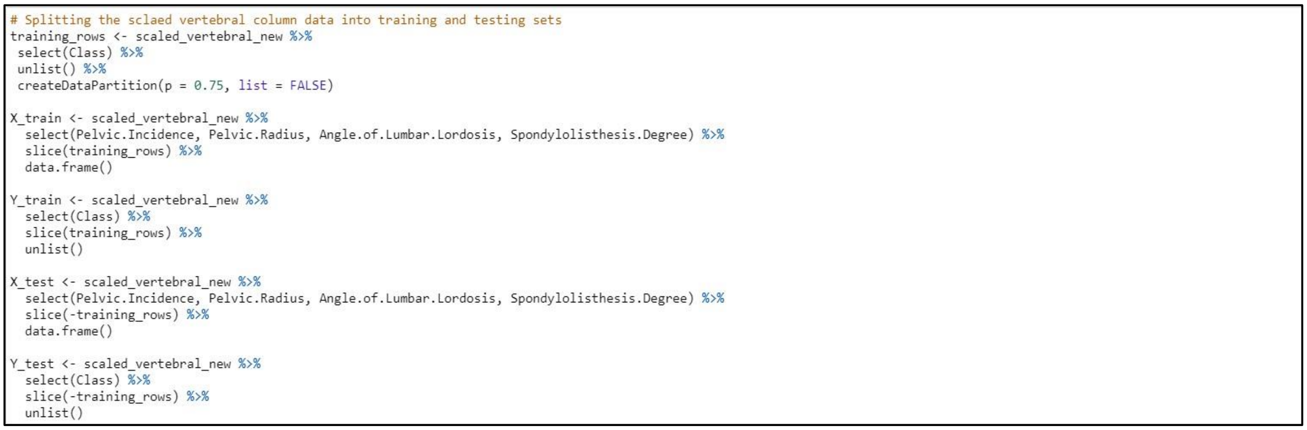
```
set.seed(2000)
vertebral_new <- vertebral_new %>%
    select(-Pelvic.Tilt, -Sacral.Slope)

scaled_vertebral_new <- vertebral_new %>% select(Angle.of.Lumbar.Lordosis, Spondylolisthesis.Degree, Pelvic.Radius, Pelvic.Incidence,Class) %>%
  mutate(Angle.of.Lumbar.Lordosis = as.vector(scale(Angle.of.Lumbar.Lordosis, center = TRUE)),
         Spondylolisthesis.Degree = as.vector(scale(Spondylolisthesis.Degree, center = TRUE)),
         Pelvic.Incidence = as.vector(scale(Pelvic.Incidence, center = TRUE)),
         Pelvic.Radius = as.vector(scale(Pelvic.Radius, center = TRUE)))

head(scaled_vertebral_new)
```

Angle.of.Lumbar.Lordosis	Spondylolisthesis.Degree	Pelvic.Radius	Pelvic.Incidence	Class
<dbl>	<dbl>	<dbl>	<dbl>	<chr>
-0.66906585	-0.6424967	0.8870548	-0.3210234	Normal
0.23541406	-0.6610706	0.6492254	-0.4891991	Normal
-0.22647228	-0.4912294	0.1147842	0.7141163	Normal
-0.02582372	-0.8548339	-0.5137603	-0.1368344	Normal
-0.19186040	-0.4487340	-0.2273656	0.4874360	Normal
-1.17829920	-0.7516975	0.8026632	-1.4620956	Normal

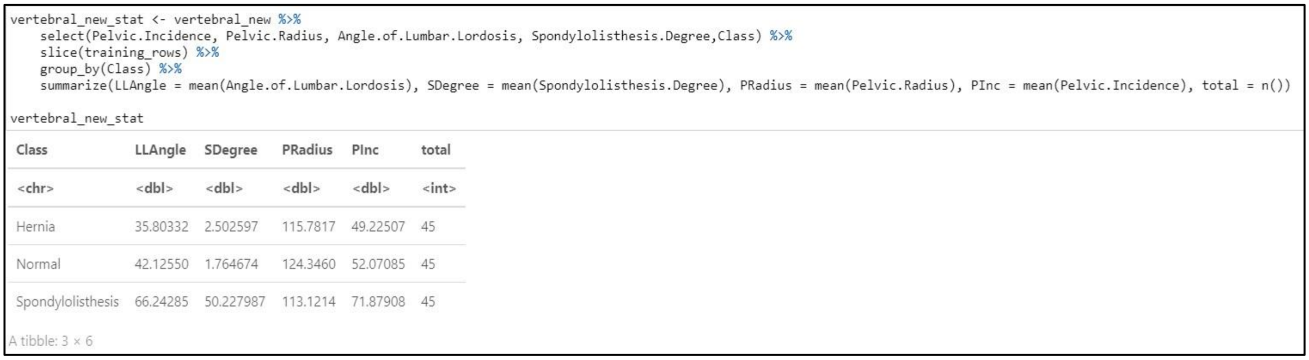
A tibble: 6 × 5

Figure 2.1

We then split the scaled vertebral column data into training dataset and testing dataset, which leads us to examine the statistical summary of the distribution for the potential predictor variables (Fig 2.2).



```
# Splitting the sclaed vertebral column data into training and testing sets
training_rows <- scaled_vertebral_new %>%
 select(Class) %>%
 unlist() %>%
 createDataPartition(p = 0.75, list = FALSE)

X_train <- scaled_vertebral_new %>%
  select(Pelvic.Incidence, Pelvic.Radius, Angle.of.Lumbar.Lordosis, Spondylolisthesis.Degree) %>%
  slice(training_rows) %>%
  data.frame()

Y_train <- scaled_vertebral_new %>%
  select(Class) %>%
  slice(training_rows) %>%
  unlist()

X_test <- scaled_vertebral_new %>%
  select(Pelvic.Incidence, Pelvic.Radius, Angle.of.Lumbar.Lordosis, Spondylolisthesis.Degree) %>%
  slice(-training_rows) %>%
  data.frame()

Y_test <- scaled_vertebral_new %>%
  select(Class) %>%
  slice(-training_rows) %>%
  unlist()
```



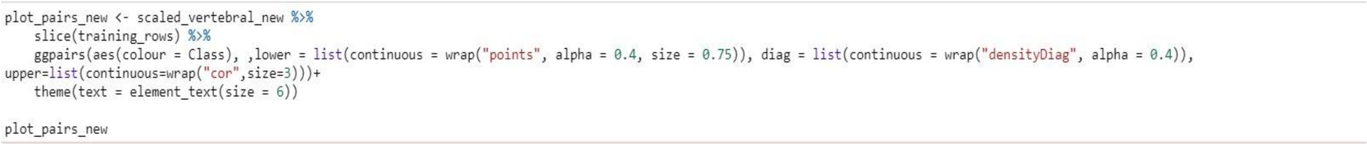
```
vertebral_new_stat <- vertebral_new %>%
    select(Pelvic.Incidence, Pelvic.Radius, Angle.of.Lumbar.Lordosis, Spondylolisthesis.Degree,Class) %>%
    slice(training_rows) %>%
    group_by(Class) %>%
    summarize(LLAngle = mean(Angle.of.Lumbar.Lordosis), SDegree = mean(Spondylolisthesis.Degree), PRadius = mean(Pelvic.Radius), PInc = mean(Pelvic.Incidence), total = n())

vertebral_new_stat
```

Class	LLAngle	SDegree	PRadius	PInc	total
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<int>
Hernia	35.80332	2.502597	115.7817	49.22507	45
Normal	42.12550	1.764674	124.3460	52.07085	45
Spondylolisthesis	66.24285	50.227987	113.1214	71.87908	45

A tibble: 3 × 6

Figure 2.2

Fig 2.4 shows the visualization of the cross-comparison of the explanatory variables and the class distribution.



```
plot_pairs_new <- scaled_vertebral_new %>%
    slice(training_rows) %>%
    ggpairs(aes(colour = Class), ,lower = list(continuous = wrap("points", alpha = 0.4, size = 0.75)), diag = list(continuous = wrap("densityDiag", alpha = 0.4)),
upper=list(continuous=wrap("cor",size=3)))+
    theme(text = element_text(size = 6))

plot_pairs_new
```

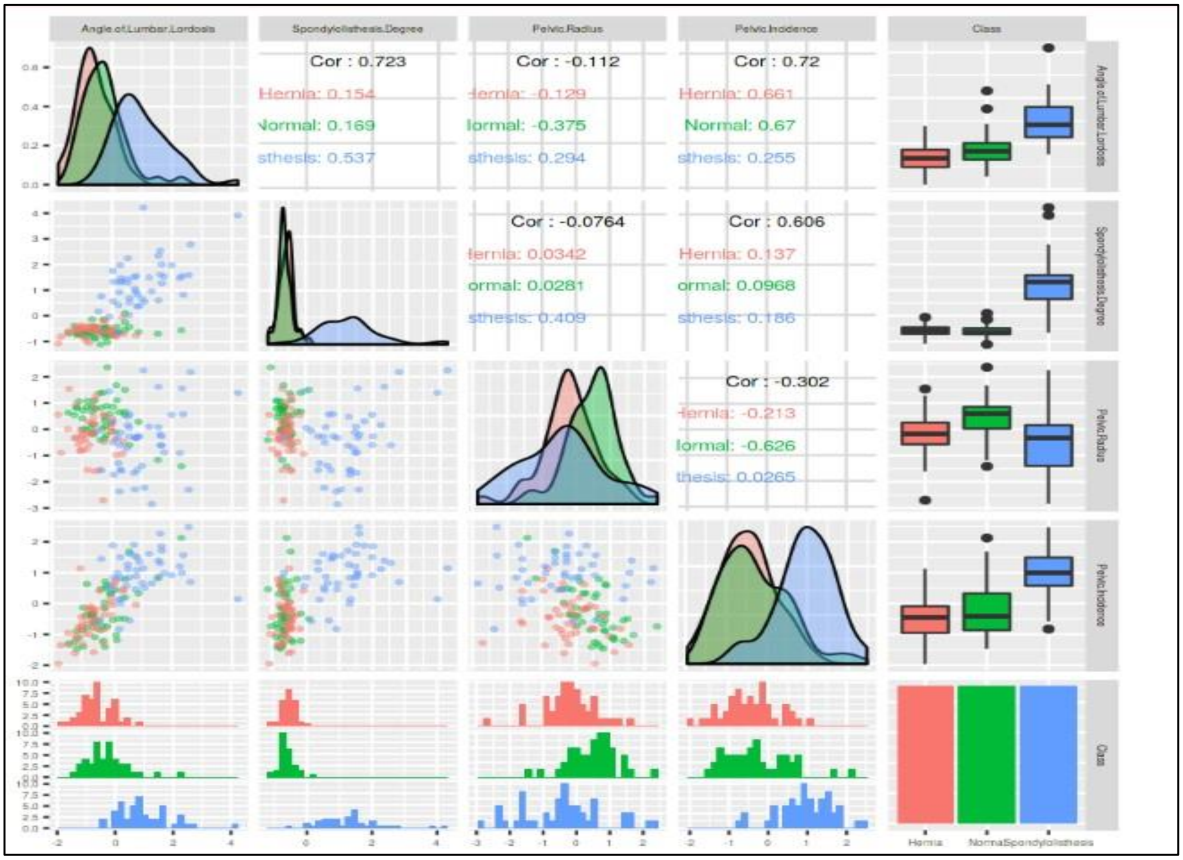


Figure 2.4

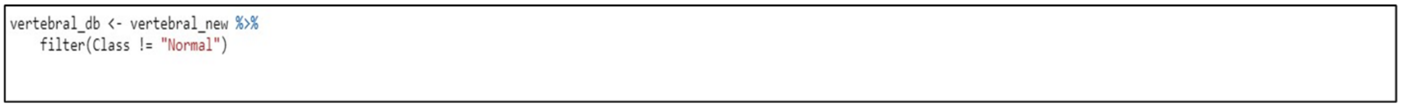
It is clear, at a first glance, that there are potential positive common trends between Pelvic Incidence against Classification and Angle of Lumbar Lordosis against Classification. In addition, we can learn from the above visualization that both the angle of Lumbar Lordosis and pelvic incidence are the more skewed-to-right attribute by classification - also can be seen in the vertebral_new_stat data frame above - and that both data has sufficient amount of variation in terms of its distribution by classes. It may be reasonable and worthwhile to take a look at the two predictors - Angle of Lumbar Lordosis and Pelvic Incidence - and how they perform on the k-nearest-neighbours classification. The K-NN classification mainly depends on the surrounding continuous samples, and more suitable for the crossover or overlapping of the sample groups to be divided.

III.II. K-NN Classification

We used K-nearest-neighbours classification in training the predictive model for the test dataset of the spinal disorder diagnosis. It is appropriate to choose K-NN over regression, since we are concerned about the classification of the labels than a precise number.

In doing so, we have constructed two different cases of predicting three different class labels with Angle of Lumbar Lordosis and Spondylolisthesis Degree (Case A) versus two abnormal class labels with Angle of Lumbar Lordosis and Pelvic Incidence (Case B), because we believe that Spondylolisthesis degree is an effective indicator to separate the normal patients from the patients with any of two disc symptoms, but not a good indicator for differentiating which disc symptom.

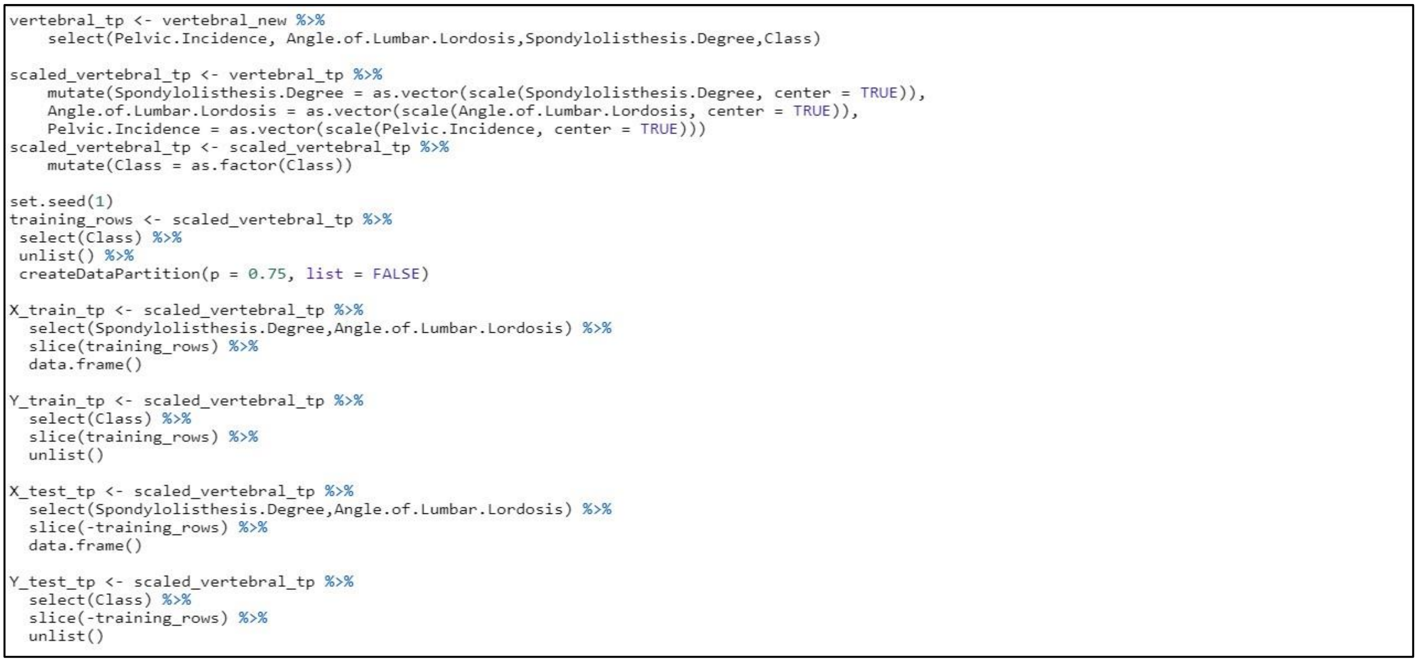
Fig 3.1 shows the process of filtering out the normal class from the 180 observations.



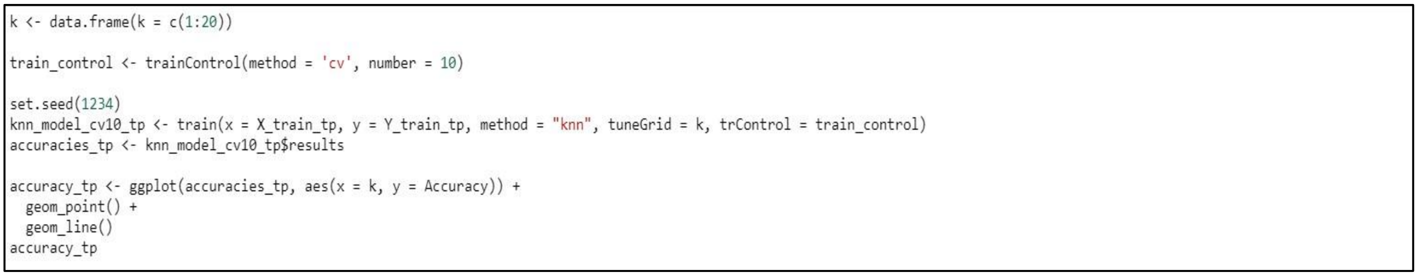
```r
vertebral_db <- vertebral_new %>%
    filter(Class != "Normal")
```

Figure 3.1

Fig 3.2 shows the process of training and tuning the model for Case A.



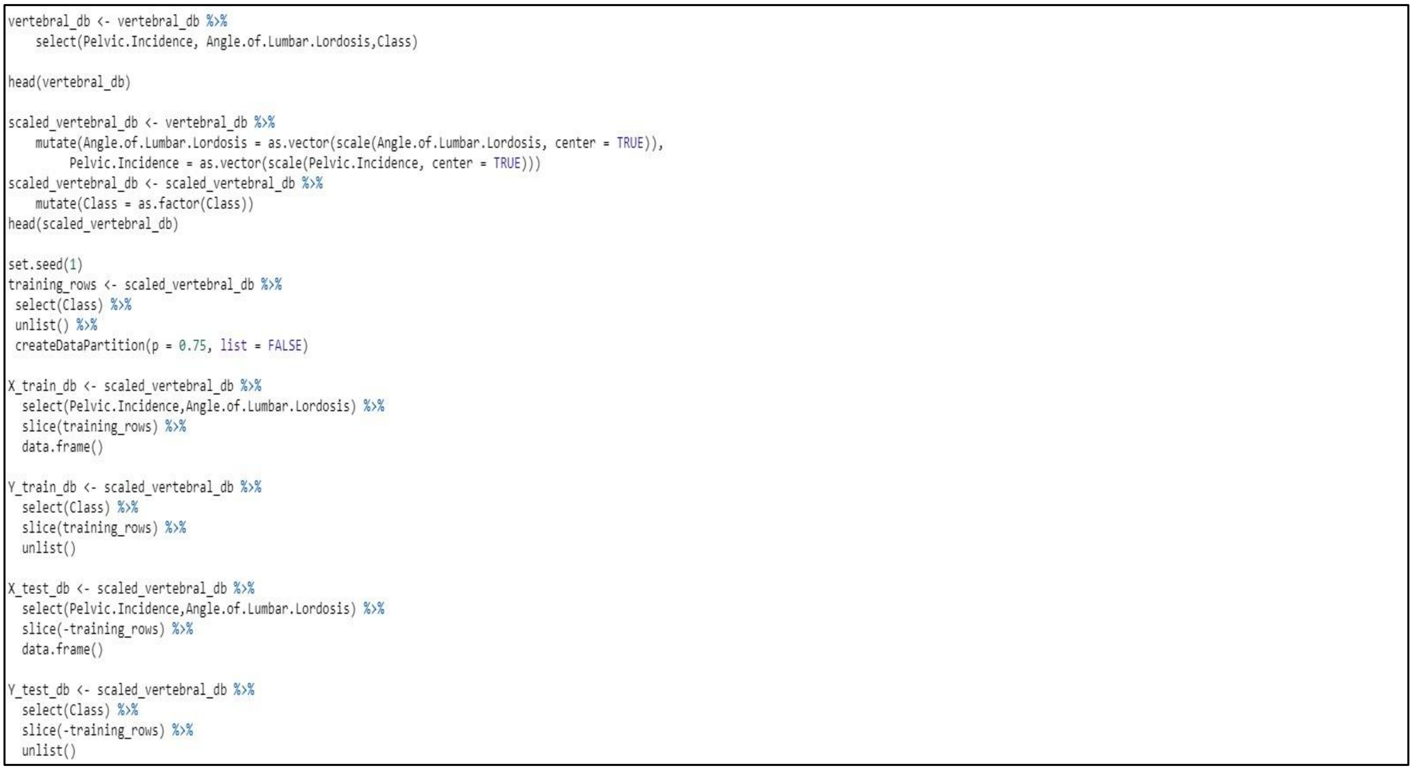
```r
vertebral_tp <- vertebral_new %>%
    select(Pelvic.Incidence, Angle.of.Lumbar.Lordosis,Spondylolisthesis.Degree,Class)

scaled_vertebral_tp <- vertebral_tp %>%
    mutate(Spondylolisthesis.Degree = as.vector(scale(Spondylolisthesis.Degree, center = TRUE)),
    Angle.of.Lumbar.Lordosis = as.vector(scale(Angle.of.Lumbar.Lordosis, center = TRUE)),
    Pelvic.Incidence = as.vector(scale(Pelvic.Incidence, center = TRUE)))
scaled_vertebral_tp <- scaled_vertebral_tp %>%
    mutate(Class = as.factor(Class))

set.seed(1)
training_rows <- scaled_vertebral_tp %>%
 select(Class) %>%
 unlist() %>%
 createDataPartition(p = 0.75, list = FALSE)

X_train_tp <- scaled_vertebral_tp %>%
  select(Spondylolisthesis.Degree,Angle.of.Lumbar.Lordosis) %>%
  slice(training_rows) %>%
  data.frame()

Y_train_tp <- scaled_vertebral_tp %>%
  select(Class) %>%
  slice(training_rows) %>%
  unlist()

X_test_tp <- scaled_vertebral_tp %>%
  select(Spondylolisthesis.Degree,Angle.of.Lumbar.Lordosis) %>%
  slice(-training_rows) %>%
  data.frame()

Y_test_tp <- scaled_vertebral_tp %>%
  select(Class) %>%
  slice(-training_rows) %>%
  unlist()
```



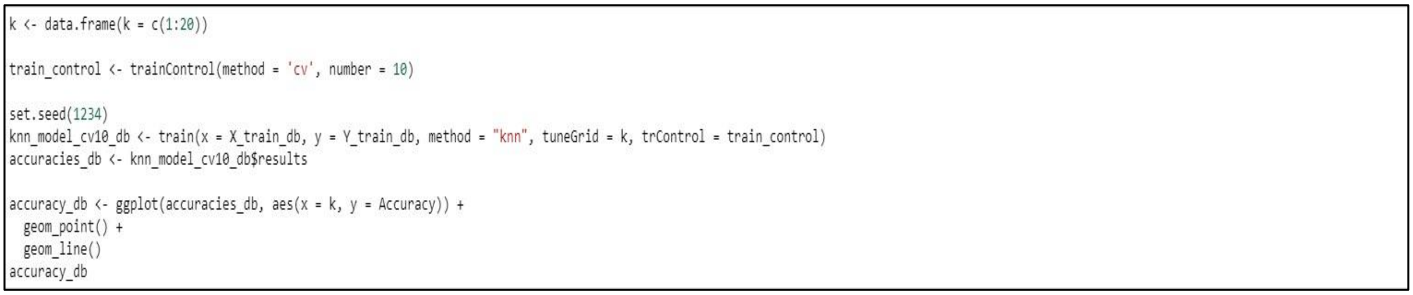
```r
k <- data.frame(k = c(1:20))

train_control <- trainControl(method = 'cv', number = 10)

set.seed(1234)
knn_model_cv10_tp <- train(x = X_train_tp, y = Y_train_tp, method = "knn", tuneGrid = k, trControl = train_control)
accuracies_tp <- knn_model_cv10_tp$results

accuracy_tp <- ggplot(accuracies_tp, aes(x = k, y = Accuracy)) +
  geom_point() +
  geom_line()
accuracy_tp
```

Figure 3.2

Fig 3.3 shows the process of training and tuning the model for Case B.



```
vertebral_db <- vertebral_db %>%
    select(Pelvic.Incidence, Angle.of.Lumbar.Lordosis,Class)

head(vertebral_db)

scaled_vertebral_db <- vertebral_db %>%
    mutate(Angle.of.Lumbar.Lordosis = as.vector(scale(Angle.of.Lumbar.Lordosis, center = TRUE)),
         Pelvic.Incidence = as.vector(scale(Pelvic.Incidence, center = TRUE)))
scaled_vertebral_db <- scaled_vertebral_db %>%
    mutate(Class = as.factor(Class))
head(scaled_vertebral_db)

set.seed(1)
training_rows <- scaled_vertebral_db %>%
 select(Class) %>%
 unlist() %>%
 createDataPartition(p = 0.75, list = FALSE)

X_train_db <- scaled_vertebral_db %>%
  select(Pelvic.Incidence,Angle.of.Lumbar.Lordosis) %>%
  slice(training_rows) %>%
  data.frame()

Y_train_db <- scaled_vertebral_db %>%
  select(Class) %>%
  slice(training_rows) %>%
  unlist()

X_test_db <- scaled_vertebral_db %>%
  select(Pelvic.Incidence,Angle.of.Lumbar.Lordosis) %>%
  slice(-training_rows) %>%
  data.frame()

Y_test_db <- scaled_vertebral_db %>%
  select(Class) %>%
  slice(-training_rows) %>%
  unlist()
```



```
k <- data.frame(k = c(1:20))

train_control <- trainControl(method = 'cv', number = 10)

set.seed(1234)
knn_model_cv10_db <- train(x = X_train_db, y = Y_train_db, method = "knn", tuneGrid = k, trControl = train_control)
accuracies_db <- knn_model_cv10_db$results

accuracy_db <- ggplot(accuracies_db, aes(x = k, y = Accuracy)) +
  geom_point() +
  geom_line()
accuracy_db
```

Figure 2.3

IV. Results

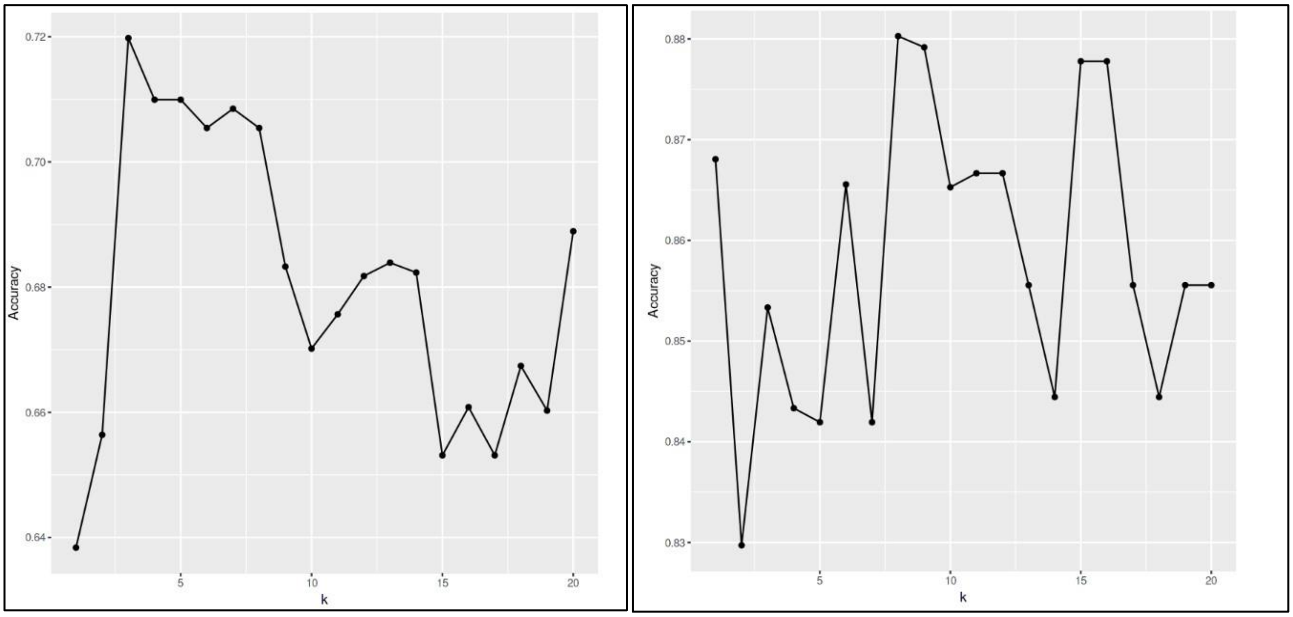


Figure 4.1 Case A vs Case B

Figure 4.1 suggests the comparison of the plots for the optimal K for K-NN classification. We can see that the optimal K for Case A (Three class levels vs Angle of Lumbar Lordosis and Spondylolisthesis Degree) is 5 (chose 5 instead of 3 considering the uncertainties around the number) and despite the high uncertainty and noise level, 11 may be a good choice for Case B (Two abnormal classes vs Angle of Lumbar Lordosis and Pelvic Incidence).

V. Discussion

The discussion can be made by retrieving back to the question that was initially proposed: Can we observe the distinction between the Disk Hernia and Spondylolisthesis based on two of the geometric vertebral column variables?

We were able to train and tune the predictive model for distinguishing the Disk hernia and Spondylolisthesis based on the two predictor variables we chose: Pelvic Incidence and Angle of Lumbar Lordosis. We were able to observe that the model, despite the high uncertainties for each of the K selection, attained the accuracy of ≈ 88% (Case B) In comparison, we were also able to observe that the model distinguishing the three conditions with Angle of Lumbar Lordosis and Spondylolisthesis Degree attained the highest accuracy of ≈ 72% with relatively less significant uncertainties of K (Case A).

For Case A, we were aware that the spondylolisthesis degree is a good indicator in assessing the normal/abnormal condition of the vertebral column. The near 72% accuracy is not sufficiently convincing enough to realistically apply the model in the real medical settings. This also implies that the model can further be developed to higher accuracy with the introduction of a new, additional predictor variable.

For Case B, we are confident in the potential of the model to be an accurate predictor building block for the future. The near 88% accuracy is sufficiently convincing for us to utilize the model in the future for the diagnosis of incoming patients, with further tuning and testing of the model. With this being said, the model suggests the value of pelvic incidence and angle of lumbar lordosis as the predictors of spinal disorders and presents potential worth in further analyzing to find out which, between the sacral slope and pelvic tilt, has the negative or positive correlation to each case of the symptom.

A few things to keep in mind are that the sample size was not significantly large for us to fully develop the model to a more intricate level of accuracy and bias-free, and that the model was randomly seeded during the process of balancing and cross validation. Method such as bootstrapping or randomly repeated sampling maybe helpful in the case of tuning the model to a finer accuracy.

VI. Reference

"Slipped Disk: Overview." InformedHealth.org [Internet]., U.S. National Library of Medicine, 1 June 2017, www.ncbi.nlm.nih.gov/books/NBK279472/.

"Spondylolisthesis Treatment, Surgery & Symptoms." Cleveland Clinic, my.clevelandclinic.org/health/diseases/10302-spondylolisthesis.