Harvard_Task

Hadi Yolasigmaz

3/24/2019

INTRODUCTION

Dataset

Dataset selected is one of the machine learning ready dataset downloaded from 'Kaggle.com'. It was in curated list of datasets

Subject of selected Dataset is 'Biomechanical Features of Orthopedic Patients'. Each patient in the data set (line); six biomechanics derived from the shape and orientation of the pelvis and the lumbar spine (each one is a column) are particularly represented: Pelvic incidence, Pelvic tilt numeric, Lumbar lordosis angle, Sacral slope, Pelvic radius, Degree spondylolisthesi.

It is a clean data. After data visulization, Multiclass Classification will be applied as machine learning analyses to this data.

Column_3C_weka.csv will be imported as dataset and it has 7 colummns data. First six columns are six biomechanics and last column is used to classify patients. In file, 100 patients are as normal, 60 patients are as Disk Hernia and 150 patients are Spondilolistez. Total 310 patients.

The goal of the project

The main goal of the project is to be known that we are ready to datascience but the specific goal of the project is to find a prediction way by checking Biomechanical Features of Orthopedic Patients and predict whether they are normal, 'Disk Hernia' or 'Spondilolistez'

Structure of data

```
Data_3class <- read.csv(file="column_3C_weka.csv", header=TRUE, sep=",")
str(Data_3class)
```

```
## 'data.frame':    310 obs. of  7 variables:
##  $ pelvic_incidence        : num  63 39.1 68.8 69.3 49.7 ...
##  $ pelvic_tilt             : num  22.55 10.06 22.22 24.65 9.65 ...
##  $ lumbar_lordosis_angle   : num  39.6 25 50.1 44.3 28.3 ...
##  $ sacral_slope            : num  40.5 29 46.6 44.6 40.1 ...
##  $ pelvic_radius           : num  98.7 114.4 106 101.9 108.2 ...
##  $ degree_spondylolisthesis: num  -0.254 4.564 -3.53 11.212 7.919 ...
##  $ class                   : Factor w/ 3 levels "Hernia","Normal",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Summary of data

```
summary(Data_3class)
```

```
##  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
```

First 6 data lines

```
head(Data_3class)
```

```
##   pelvic_incidence pelvic_tilt lumbar_lordosis_angle sacral_slope
## 1         63.02782   22.552586              39.60912     40.47523
## 2         39.05695   10.060991              25.01538     28.99596
## 3         68.83202   22.218482              50.09219     46.61354
## 4         69.29701   24.652878              44.31124     44.64413
## 5         49.71286    9.652075              28.31741     40.06078
## 6         40.25020   13.921907              25.12495     26.32829
##   pelvic_radius degree_spondylolisthesis  class
## 1      98.67292                -0.254400 Hernia
## 2     114.40543                 4.564259 Hernia
## 3     105.98514                -3.530317 Hernia
## 4     101.86850                11.211523 Hernia
## 5     108.16872                 7.918501 Hernia
## 6     130.32787                 2.230652 Hernia
```

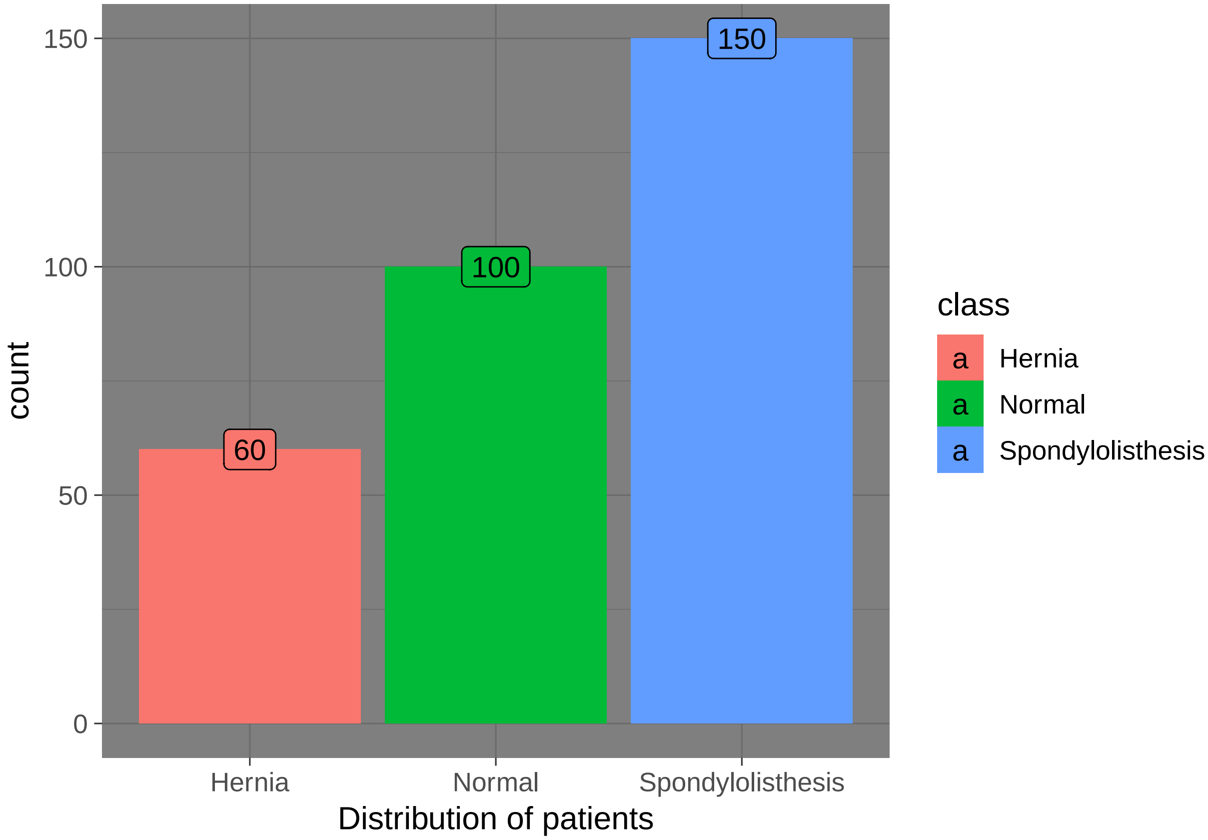
ANALYSES

Firstly, rapid data analyses will be done by visualization and lastly machine learning Multiclass Classification will be applied

Data Visulization

Distribution of Patients in 3 class items dataset

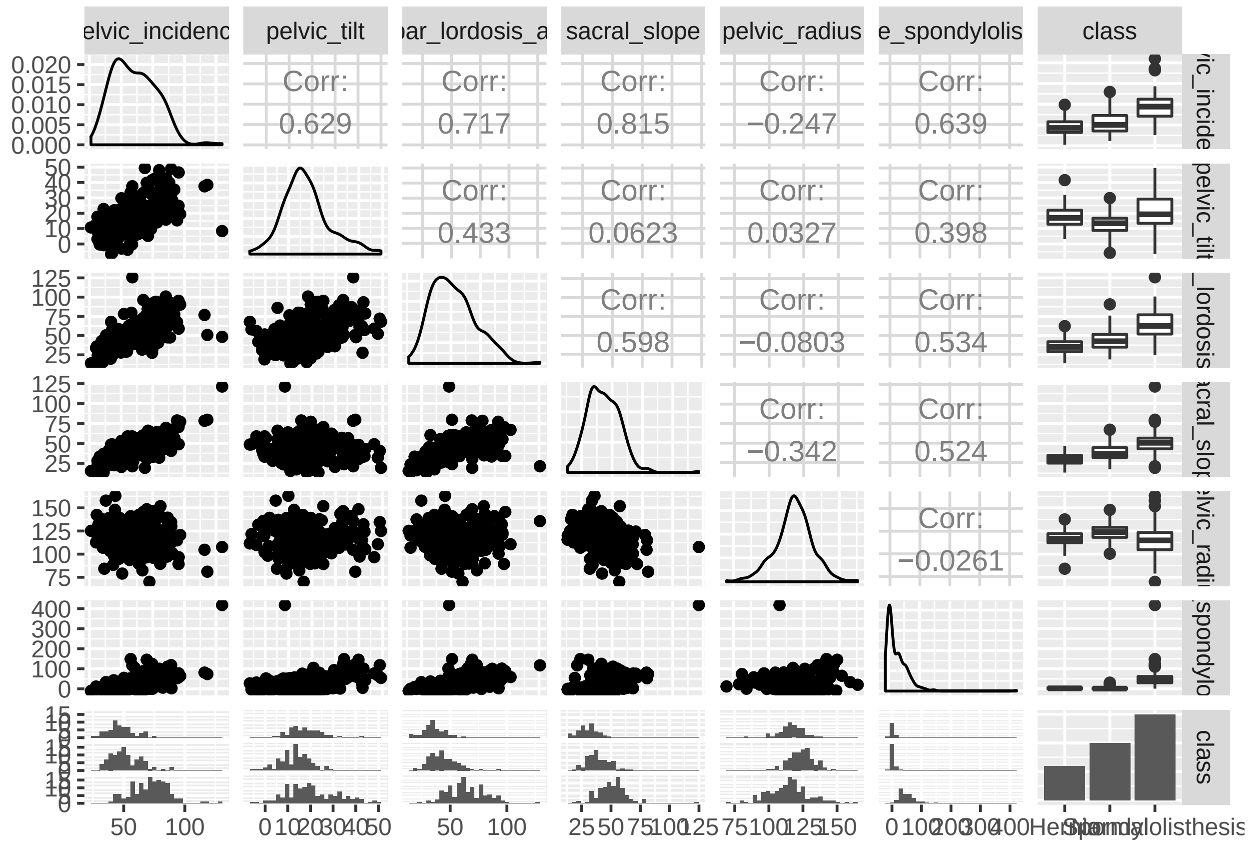
```
ggplot(Data_3class,aes(x=class,fill=class))+geom_bar(stat = 'count')+labs(x = 'Distribution of patients
  geom_label(stat='count',aes(label=..count..), size=4) +theme_dark(base_size = 12)
```



Pair graphs of data.

```
ggpairs(data=Data_3class, columns=c(1:7))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



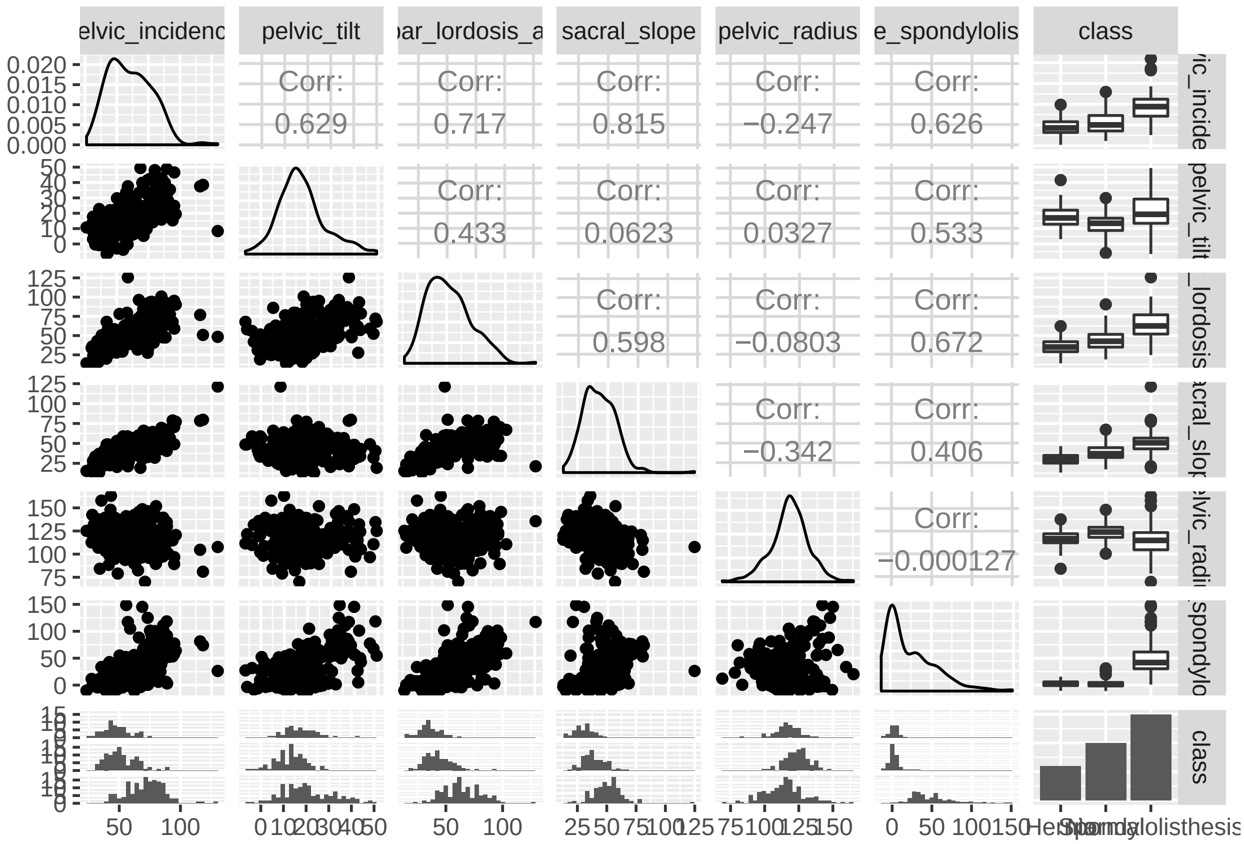
It is seen that from diagonally looking there is an outlier problem for degree_spondylolisthesis. it should be corrected first by redefine it as mean of column, mean of degree_spondylolisthesis.

```
outlier_3class <- which.max(Data_3class$degree_spondylolisthesis)
Data_3class$degree_spondylolisthesis[outlier_3class] <- mean(Data_3class$degree_spondylolisthesis)
```

Refreshing pair graph it is viewed that outlier problem is Solved.

```
ggpairs(data=Data_3class, columns=c(1:7))
```

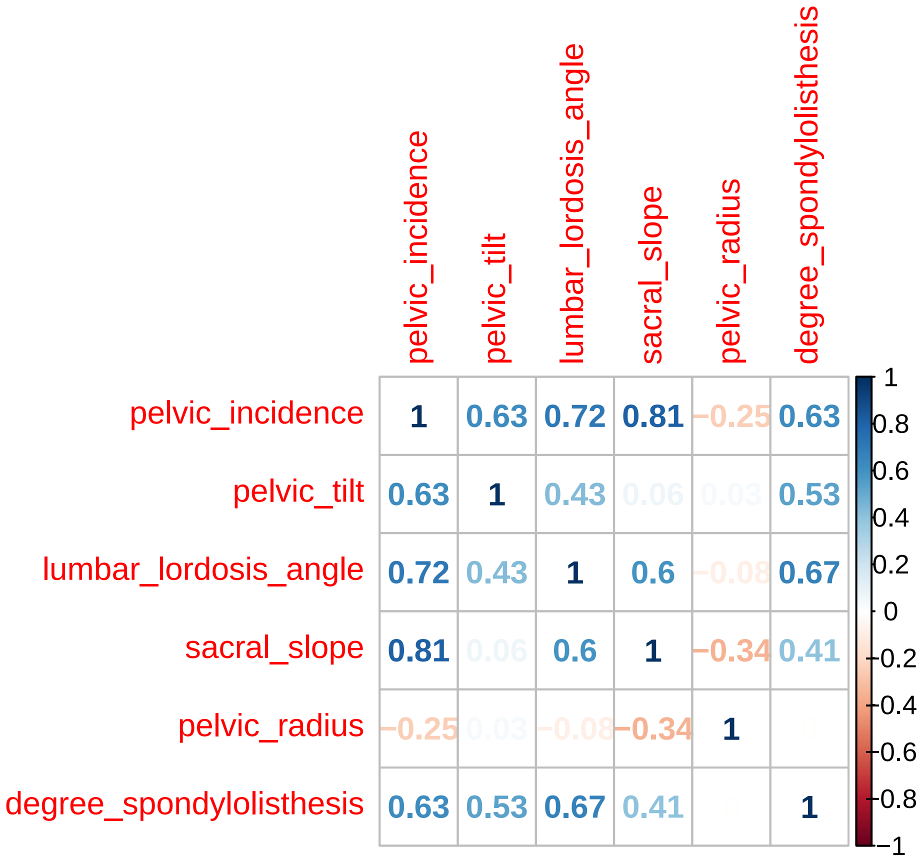
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Correlation between six biometrics

It is better to see this relations in number format not only graph format.

```
suppressMessages(library(corrplot))
corr_mat <- cor(Data_3class[,1:6])
corrplot(corr_mat, method = "number")
```



There aren't good correlations and relations between 6 biometrics that easily formalize.

Machine learning techniques

As seen, it is not easily formalize, use mathematic relation we will try to do well defined, well known machine learning techniques for this dataset to classify. Classification with More than Two Classes machine learning techniques will be applied. They are Decision trees and random forest.

Prepare Train set and test set

Firstly, train dataset and test dataset will be prepared from analyses.

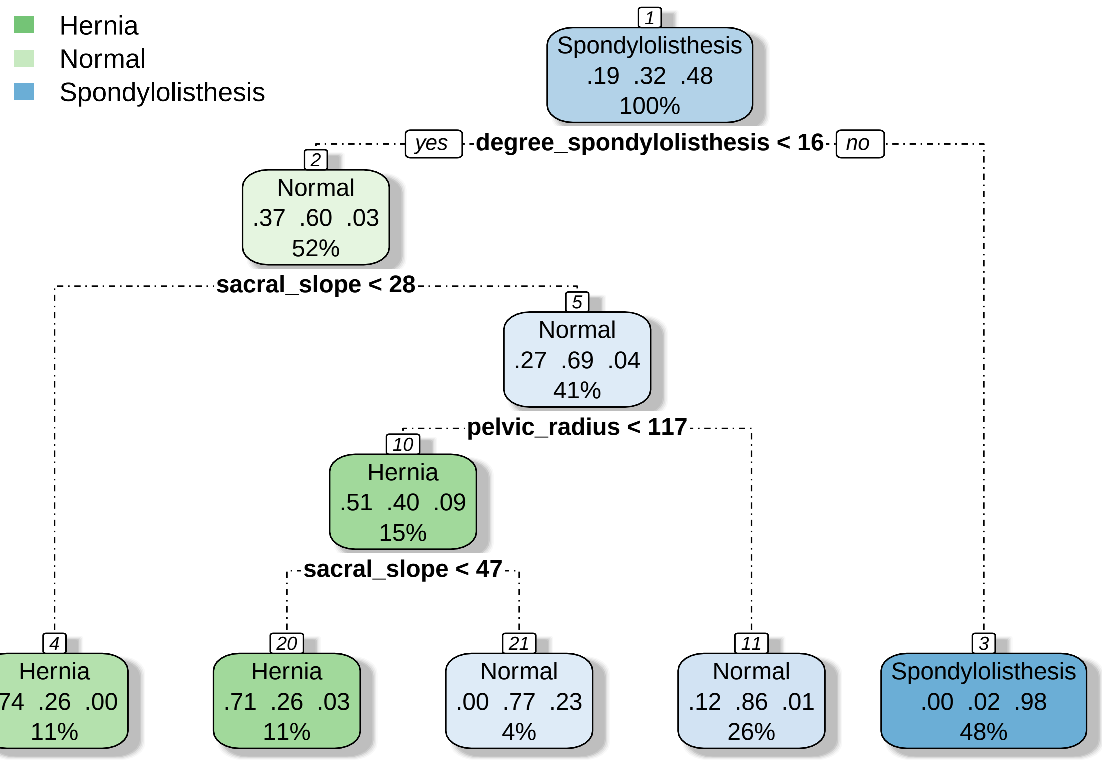
```
y <- Data_3class$class
set.seed(1)
test_index <- createDataPartition(y, times = 1, p = 0.5, list = FALSE)
train_set <- Data_3class %>% slice(-test_index)
test_set <- Data_3class %>% slice(test_index)
```

Classification (decision) trees.

This is a technique that it can be controlled and has interpretability. we can follow all branches of decision tree. For this rpart.plot will be used visualization of decision tree.

```
class.tree <- rpart(Data_3class$class~.,data = Data_3class,control = rpart.control(cp = 0.01))
rpart.plot(class.tree,
```

```
          box.palette="GnBu",
          branch.lty=10, shadow.col="gray", nn=TRUE)
```

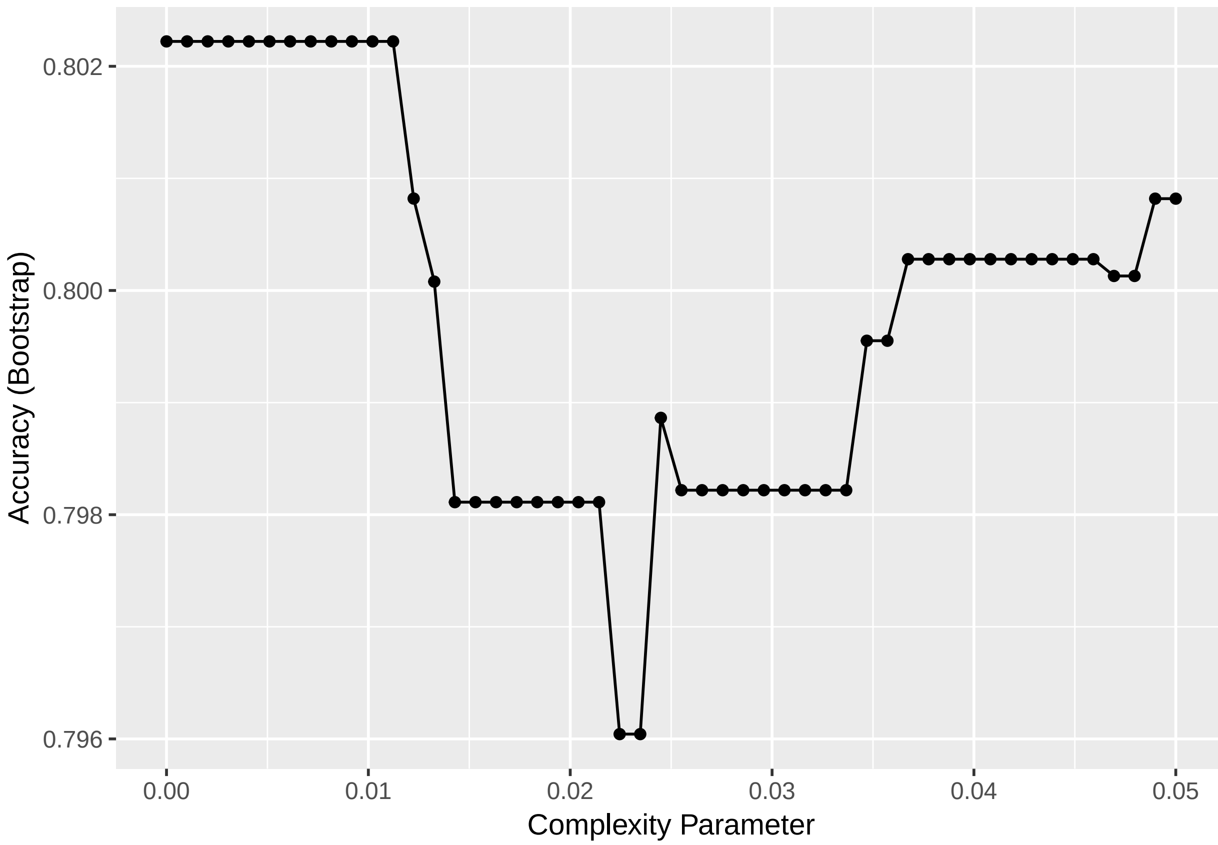


First branch of decision tree is degree_spondylolisthesis whether degree_spondylolisthesis is bigger than 16, If it is yes then patient class is spondylolisthesis. Numbers in the box. 0.19 - 0.193548387 - Number of Hernia in the dataset - (60 / 310) 0.32 - 0.322580645 - Number of Normal in the dataset - (100 / 310) 0.48 - 0.483870968 - Number of spondylolisthesis in the dataset - (150 / 310)

Right side of the first branche is spondylolisthesis patients 48% of all data Numbers in the box. 0.00 - 0 - Number of Hernia in the dataset - (0 / 150) 0.02 - 0.2 - Number of Normal in the dataset - (3 / 150) 0.98 - 0.98 - Number of spondylolisthesis in the dataset - (147 / 150) and so on.

cp is selected as 0.01 by controlling accuracy with respect to below approach.

```
train_rpart <- train(class ~ .,  method = "rpart",
                     tuneGrid = data.frame(cp = seq(0, 0.05, len = 50)),
                     data = train_set)
ggplot(train_rpart)
```



And Accuracy is

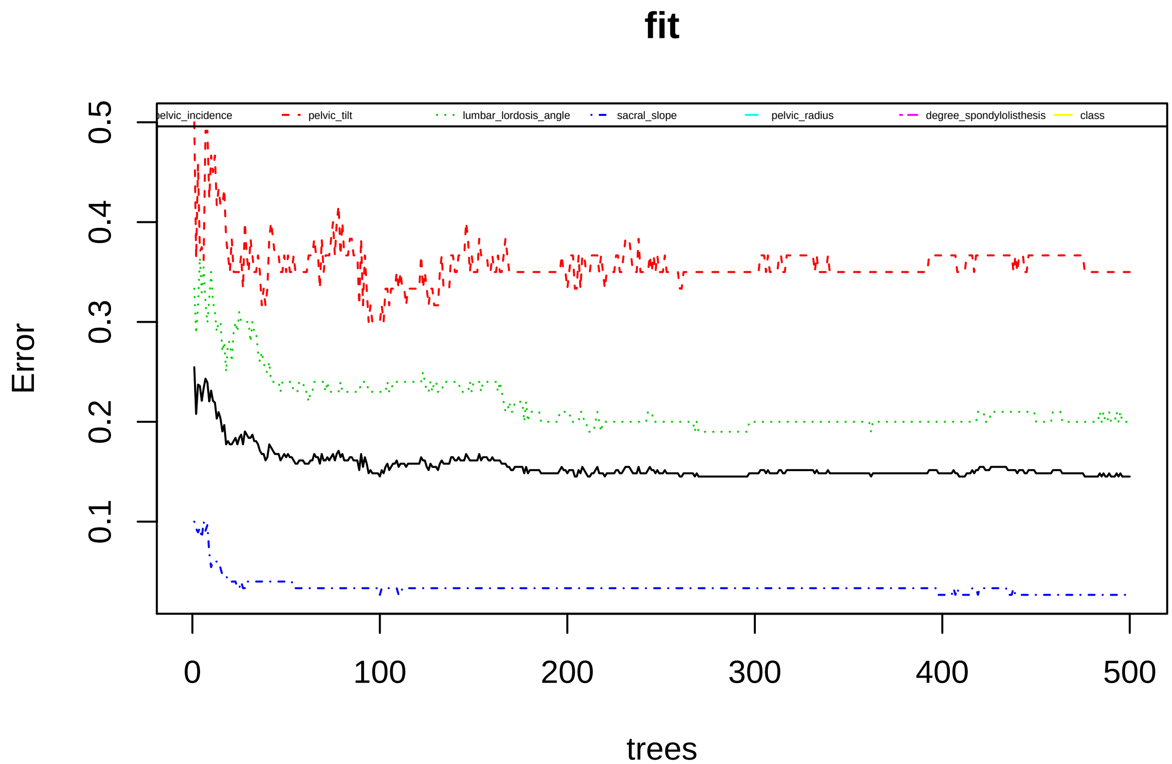
```
confusionMatrix(predict(train_rpart, test_set),test_set$class)$overall["Accuracy"]
```

```
## Accuracy
##      0.8
```

random forest

This is a technique that it can't be easily interpretability. Excepted give better accuracy in prediction.

```
fit <- randomForest(class~., data = Data_3class, ntree=500, proximity=T)
plot(fit)
fit.legend <- colnames(Data_3class)
legend("top", cex =0.3, legend=fit.legend, lty=c(1,2,3,4,5,6,7), col=c(1,2,3,4,5,6,7), horiz=T)
```



Around 100 is the value can be used. It is better to find lowest err.rate. It is 100.

```
which.min(fit$err.rate[,1])
```

```
## [1] 100
```

```
suppressMessages(library(randomForest))
rf.train_set <-randomForest(class ~., data=train_set)
rf.train_set
```

```
##
## Call:
##  randomForest(formula = class ~ ., data = train_set)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 14.84%
## Confusion matrix:
##                     Hernia Normal Spondylolisthesis class.error
## Hernia                  18     12                 0  0.40000000
## Normal                   8     40                 2  0.20000000
## Spondylolisthesis        0      1                74  0.01333333
```

Accuracy is better as expected.

```
train_rf <- randomForest(class ~ ., data=train_set)
confusionMatrix(predict(train_rf, test_set), test_set$class)$overall["Accuracy"]
```

```
##  Accuracy
## 0.8387097
```

Results

When Decision Tree is used. Accuracy was

```
confusionMatrix(predict(train_rpart, test_set),test_set$class)$overall["Accuracy"]
```

```
## Accuracy
##      0.8
```

When Random Forest is used. Accuracy was

```
confusionMatrix(predict(train_rf, test_set), test_set$class)$overall["Accuracy"]
```

```
##  Accuracy
## 0.8387097
```

CONCLUSION

Of course, to have a predictionwith accuracies above 0.8, it can be seen as good. But if we talk about patients, 2 things should be done. One side is Biometric data collection side. It should be continue to understand whether there is any misadded variable that affect patients conditions. Amount or way of data collection should be changed. Another side is Datascience part, to apply methods to data, may be sometimes new grouping from owned data, may be sometimes only concentrate one part of problem but at the end come to a point that increase accuracy of owned data. Prediction should be say more. This is time consuming operations that sometimes, I believe this, work together is better to predict more close.