STA6714 DATA PREPARATION

Project Report

Name: Sejal Wadekar

UCF ID: 4736729

NID: se843064

1. Purpose

As part of the course STA6714 Data Preparation, I have learned about the very important aspect in data analysis which is 'Data Preprocessing'. Handling the data well, before applying a model to it plays a very crucial part in analysis and is more important than the modelling itself. If data is not processed well before handing it over to any model, the results produced will not be good or won't be efficient.

My purpose in this project was to use all the knowledge taught through this course and apply it to a real data set. I believe that without applying the knowledge practically, it's not really complete knowledge. This project helped me work with a real data set and analyze it using it all the concepts taught. It made me understand better on how to actually deal with data step-by-step. The difficulties faced during the project, raised new questions which did not come into picture during classroom lectures. Answers to these questions have made me more confident about handling and pre-processing data sets.

2. Problem Statement

Pelvic pain is common amongst many people as they age and especially in women after delivery. The pain could be temporary or due to some serious problem. The pelvic x-ray is the very first thing examined by doctors, to understand the cause of pain. Various measurements of the pelvic bone and the lower end of the spine provide an insight for understanding the issue and the cause.

The problem statement of my project is to use different measurements of the pelvic bone and predict if the bone being examined is 'Normal' or 'Abnormal'. These measurements basically measure the shape and orientation of the pelvic bone and the intersection of the pelvic bone with the spine. Since, the outcome is to classify the bone as belonging to one of the two categories, this is a classification problem.

The business outcome of this project is that it can help doctors to quickly classify the examined pelvic bone as Normal or Abnormal and begin with the further investigation immediately. As future scope, the bone measurements could be fed into the model in real time and the model would immediately return an output. Hence, the doctors and patients won't have to wait until the reports came.

3. Data Set

3.1. Source

The dataset was selected from Kaggle (https://www.kaggle.com/uciml/biomechanical-features-of-orthopedic-patients). It classifies the patients as Normal or Abnormal where Abnormal accounts for both Disk Hernia and Spondylolisthesis patients.

3.2. Data Description

The data set consists of seven attributes and 310 observations. Six attributes are various measurements of the pelvic bone and the intersection the bone with the spine. These will be used as inputs for training

the model. The last variable is the output categorical variable which classifies the bone as Normal or Abnormal. The seven variables are Pelvic Incidences, Pelvic Tilt, Lumbar Lordosis Angle (LLA), Sacral Slope, Pelvic Radius, Spondylolisthesis Degree and Class.

4. Preprocessing

Various properties of the data were analyzed and processed before handing the data to a model. This is done to help our model perform better without redundant data.

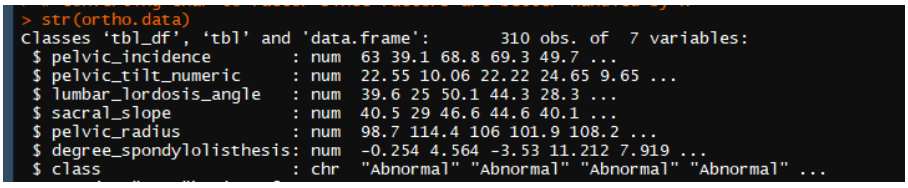
4.1. Missing Values

In most of the real data sets, there will be missing values. These values are to be treated before we move ahead with any analysis. There are two ways of dealing with missing values. First, you can directly delete the observations consisting of missing values or replace the missing values by the respective column's mean or median.

For the data set used in this project, the number of missing values was calculated using the *is.na()* function in R. The results showed that there were no missing values in the data set.

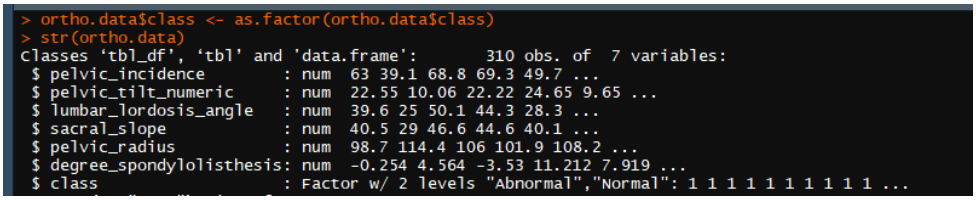
4.2. Variable Data Type

We use *str()* function to know the data-type of each variable. The output variable 'Class' is a categorical variable which is of type Character, as shown below.



Data-type of all variables

We convert it to type Factor because R can handle factors better. We use *as.factor()* function to convert it. As shown below, the data-type is updated to Factor.



Data-type of Class updated to Factor

4.3. Data Summary

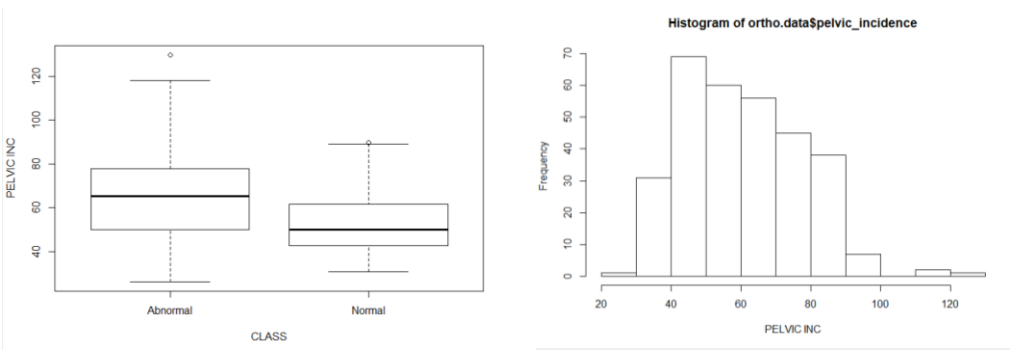
We observe the summary statistics of all the variables using *summary()* function. The output is as shown below. If we observe the range of min and max for all numerical variables, we see these ranges differ a lot, across variables. Pelvic_tilt_numeric has a maximum value of 49.432 and degree_spondylolisthesis has a maximum value of 418.543. The higher values may dominate the result and hence we see a need for normalizing the data. Normalization will be covered later in section 4.7.

```
> summary(ortho.data)
 pelvic_incidence pelvic_tilt_numeric lumbar_lordosis_angle  sacral_slope    pelvic_radius    degree_spondylolisthesis      class
 Min.   : 26.15   Min.   :-6.555      Min.   : 14.00        Min.   : 13.37   Min.   : 70.08   Min.   :-11.058          Abnormal:210
 1st Qu.: 46.43   1st Qu.:10.667      1st Qu.: 37.00        1st Qu.: 33.35   1st Qu.:110.71   1st Qu.:  1.604          Normal  :100
 Median : 58.69   Median :16.358      Median : 49.56        Median : 42.40   Median :118.27   Median : 11.768
 Mean   : 60.50   Mean   :17.543      Mean   : 51.93        Mean   : 42.95   Mean   :117.92   Mean   : 26.297
 3rd Qu.: 72.88   3rd Qu.:22.120      3rd Qu.: 63.00        3rd Qu.: 52.70   3rd Qu.:125.47   3rd Qu.: 41.287
 Max.   :129.83   Max.   :49.432      Max.   :125.74        Max.   :121.43   Max.   :163.07   Max.   :418.543
>
```

Variable summary

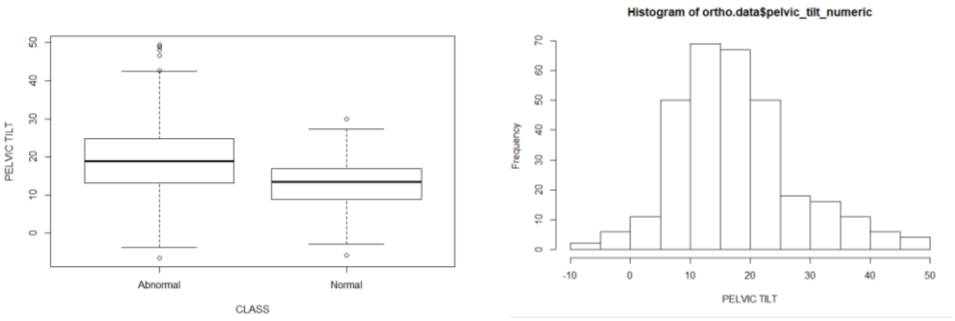
4.4. Data Visualization

4.4.1. Pelvic Incidence



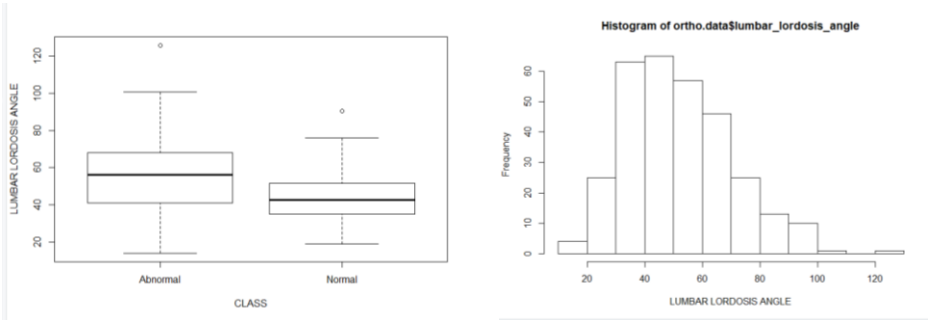
As shown above, the boxplot and histogram for 'Pelvic Incidence' is plotted. We observe the presence of outliers in boxplot. The histogram looks acceptable.

4.4.2. Pelvic Tilt



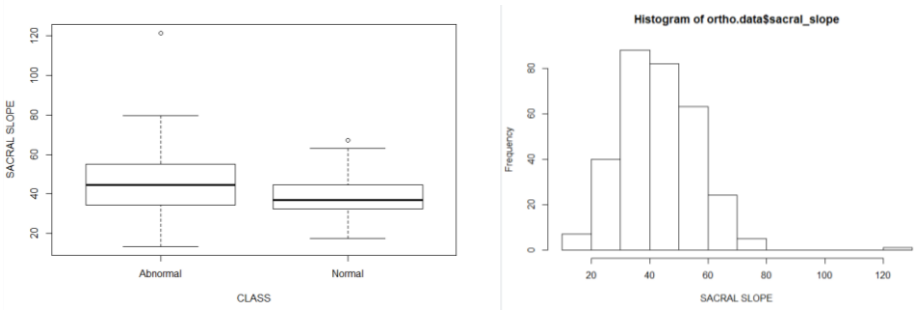
As shown above, the boxplot and histogram for ‘Pelvic Tilt’ is plotted. A lot of outliers are seen in the boxplot. Hence, these outliers need to be treated later in our analysis. The histogram looks acceptable.

4.4.3. Lumbar Lordosis Angle

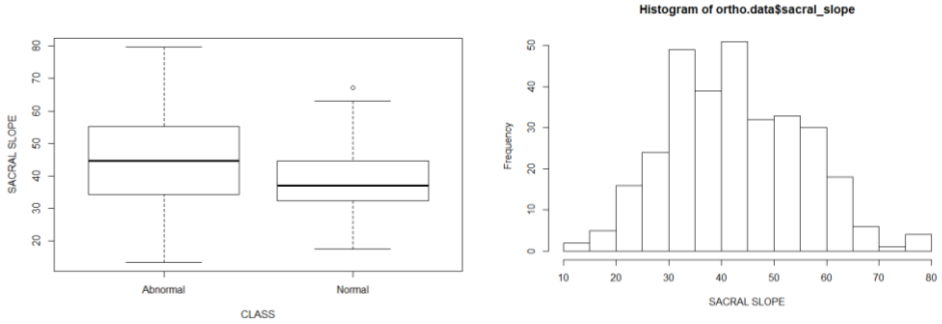


As shown above, the boxplot and histogram for ‘Lumbar Lordosis Angle’ is plotted. Outliers are observed and the histogram looks acceptable.

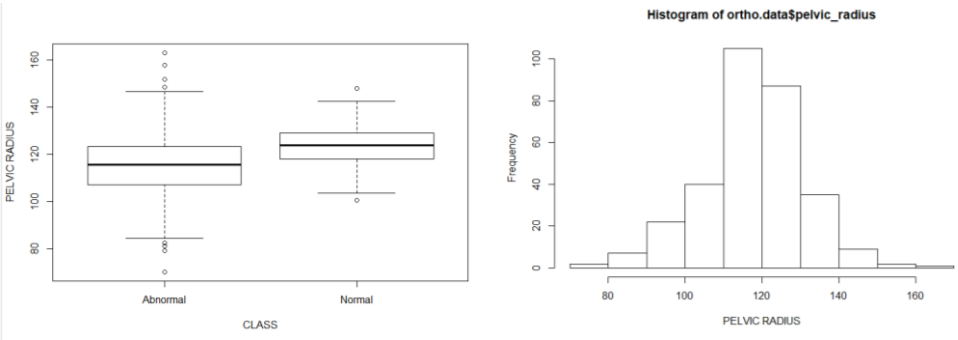
4.4.4. Sacral Slope



As shown above, the boxplot and histogram for ‘Sacral Slope’ is plotted. Few outliers are present in the boxplot. The histogram appears very much right skewed. One observation at the right end is causing this skewness. Hence, I replaced it with the column mean and the updated plots are as shown below. One outlier is removed and the histogram has no skewness now.

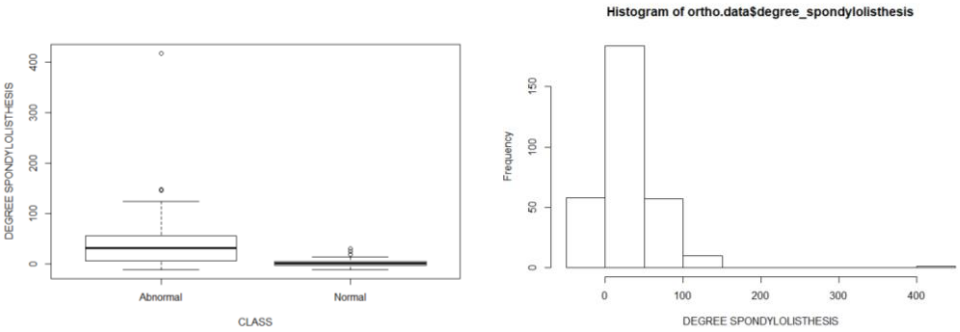


4.4.5. Pelvic Radius

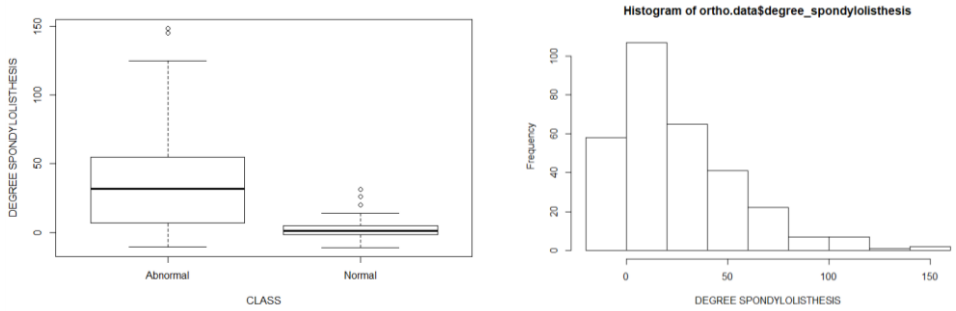


As shown above, the boxplot and histogram for 'Pelvic Radius' is plotted. Many outliers are observed, which will be treated later. Histogram looks acceptable.

4.4.6. Degree Spondylolisthesis



As shown above, the boxplot and histogram for 'Degree Spondylolisthesis' is plotted. Many outliers are observed, and the histogram is highly right skewed due to the single observation to the extreme right. Hence, we replace this observation by column mean. The updated plots are shown below. The data is still a bit skewed, but it is acceptable and much better than before.



4.5. Frequency Distribution

I checked the frequency distribution of the number of observations in the original data set that were classified as Normal and Abnormal. As shown below, 210 were classified as Abnormal and 100 were

classified as Normal. Hence, the data set is biased with a greater number of Abnormal observations and this bias should be removed before we train our model. If not, our model learn will learn to classify the Abnormal observations more accurately and hence the misclassification for the Normal observations will increase.

	freq	percentage
Abnormal	210	67.74194
Normal	100	32.25806

4.6. Removing Outliers

As observed in the bar plots, a lot of outliers are present, and they need to be treated. Directly deleting all the outliers will reduce the number of observations to a great extent. Instead, we selected the observations lying in a particular range. Quantile function was used for this task with which we selected the observations that were in the range of 1% - 99% for each variable. From these sets, we then selected the observations common between all the variables.

4.7. Normalization

As observed in the data summary section, the dataset should be normalized so that the large range variables don't dominate the calculation and all variables are on same scale. Also, normalized data will help deduce better results in the neural network model used in section 5.1.

4.8. Sampling

We sample the data in 70:30 ratio for training and validation respectively. From the distribution of the training set shown below, we can say that the sampled data also is biased with a greater number of Abnormal observations. Hence, we will oversample the training data and keep the validation data as it is so that it helps us examine the true performance of the model.

	freq	percentage
Abnormal	126	66.31579
Normal	64	33.68421

4.9. Oversampling

We oversample Normal observations by creating their duplicates. The final distribution is as shown below.

	freq	percentage
Abnormal	126	50
Normal	126	50

5. Methods Used

5.1. Neural Network

5.1.1. Dummy Variables

Dummy variables are created because neural network needs the categorical variables divided into their dummy variables.

5.1.2. Variable Selection

For variable selection, the model was first executed with only one input variable and the validation error was error was noted. Then we add another variable and check the error again. If the error has reduced, we keep the new variable; else we discard it and add another variable and perform the same steps. The summary of this process on all six input variables is as shown in the table below. The variable Sacral Slope was discarded because its addition did not bring any change in the validation error.

Variables	No. of Hidden Nodes	Train Error	Validation Error	Status
Pelvic Incidence	2	0.323	0.362	--
Pelvic Incidence + Pelvic Tilt	3	0.246	0.337	Accept
Pelvic Incidence + Pelvic Tilt + Lumbar Lordosis Angle	4	0.237	0.301	Accept
Pelvic Incidence + Pelvic Tilt + Lumbar Lordosis Angle + Sacral Slope	5	0.219	0.301	Rejected
Pelvic Incidence + Pelvic Tilt + Lumbar Lordosis Angle + Pelvic Radius	5	0.107	0.277	Accept
Pelvic Incidence + Pelvic Tilt + Lumbar Lordosis Angle + Pelvic Radius + Degree Spondylolisthesis	6	0.025	0.193	Accept

5.1.3. Selecting Number of Hidden Nodes

Five variables were selected for the neural network method and only one hidden layer was used. The number of hidden nodes was decided on the basis of the validation error produced by the model when trained and tested with different number of hidden nodes. It is ideal to start with the number of nodes equal to the number of variables and then increase the number. The summary of this analysis is as shown

below. We began with five nodes and analyzed up to eight hidden nodes. The validation error is minimum for six hidden nodes and therefore, we select the number of hidden nodes as six.

No. of Hidden Nodes	Validation Error	Status
5	0.21	Rejected
6	0.13	Accept
7	0.19	Rejected
8	0.16	Rejected

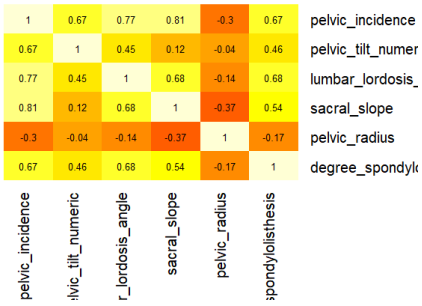
5.1.4. Results

The final model was executed with 5 input variables and 6 hidden nodes. Training error (misclassification error) obtained was 0.0630 and the validation error (misclassification error) was obtained as 0.2289. Validation error is greater than training error as expected. The accuracy obtained was 77.1 %.

5.2. Logistic Regression

5.2.1. Eliminating Correlated Variables

Logistic regression performs better if the input variables are not or less correlated to each other. Hence, we observe the heat map shown below, to identify the highly correlated variables.



Sacral Slope and Pelvic Incidence are correlated by 81%, hence we can drop one of them. Accuracy of the model by deleting either one of them was the same i.e. 0.8313. Finally, Sacral Slope was dropped.

5.2.2. Results

The final model was executed with 5 variables which gave an accuracy of 83.13 %.

6. Conclusion

The accuracy obtained by using neural networks and logistics regression was 77.1% and 83.13% respectively. Hence, we finally select logistic regression for modelling our data because it provides higher accuracy. It will be a better choice in predicting the pelvic bone status correctly as compared to neural networks.

7. R Code and Data Set