Midterm Project for ELE: 581

Name: Muhammad Enayetur Rahman
Student ID: 100635221
Data set: [Vertibral Column Data Set]

Part A:

Total 7 attributes are:

1. Pelvic Incidence
2. Pelvic Tilt
3. Lumber Lordosis Angle
4. Sacral Slope
5. Pelvic Radius
6. Grade of spondylolisthesis (SL)
7. Decision

Here is the basic statistical summary for each attribute. Attribute 'Decision' would be my dependent attribute and it has the Binary labels: Abnormal (AB) and Normal (NO).
In the summary, for each attribute minimum, 1st quartile, Median, Mean, 3rd Quartile and Maximum values are given using R's command: *summary(dataframe$attribute)*. Also, the histogram of each attribute is found using R's command: *hist(dataframe$attribute)*

Pelvic Incidence:

Pelvic.Incidence
Min. : 26.15
1st Qu.: 46.43
Median : 58.69
Mean : 60.50
3rd Qu.: 72.88
Max. : 129.83

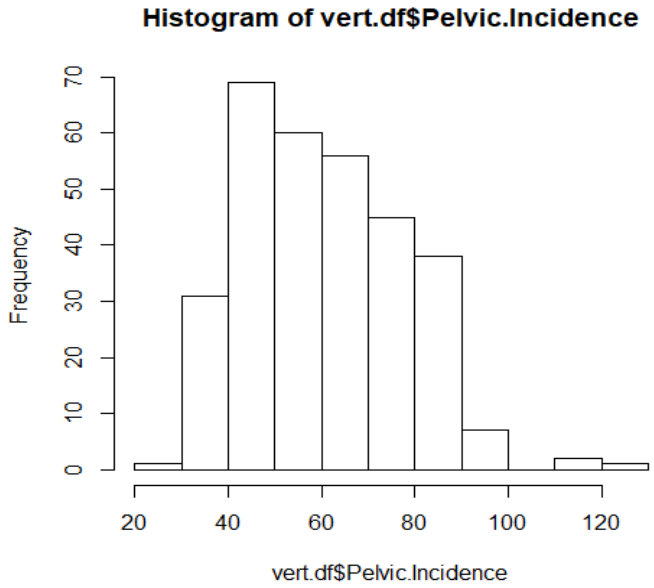


Figure 1: Histgram of Pelvic Incidence

Pelvic Tilt:

Pelvic.Tilt
Min. :-6.55
1st Qu.:10.67
Median :16.36
Mean :17.54
3rd Qu.:22.12
Max. :49.43

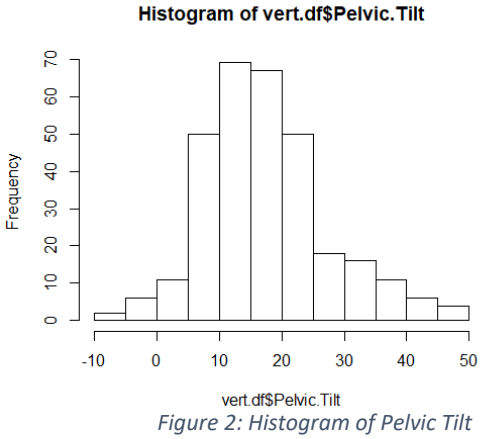


Figure 2: Histogram of Pelvic Tilt

Lumber Lordosis Angle:

Lumbar.Lordosis.Angle
Min. : 14.00
1st Qu.: 37.00
Median : 49.56
Mean : 51.93
3rd Qu.: 63.00
Max. : 125.74

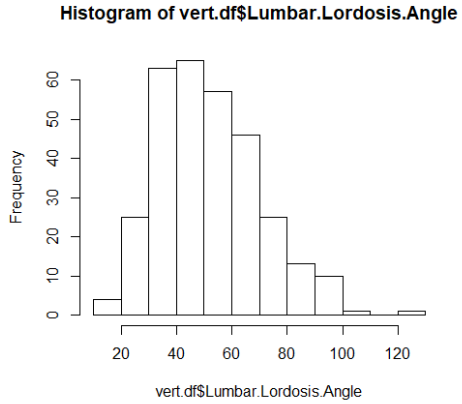


Figure 3: Histogram of Lumber Lordosis Angle

Sacral Slope

Sacral.Slope
Min. : 13.37
1st Qu.: 33.35
Median : 42.41
Mean : 42.95
3rd Qu.: 52.69
Max. : 121.43

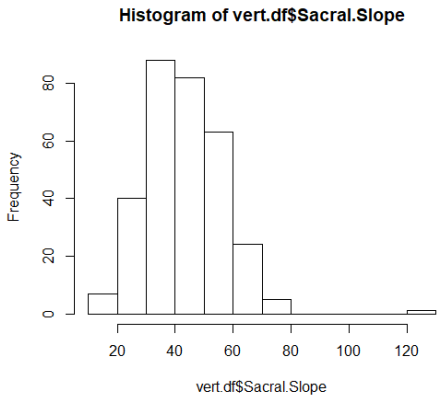


Figure 4: Histogram of Sacral Slope

Pelvic Radius

Pelvic.Radius
Min. : 70.08
1st Qu.:110.71
Median :118.27
Mean :117.92
3rd Qu.:125.47
Max. :163.07

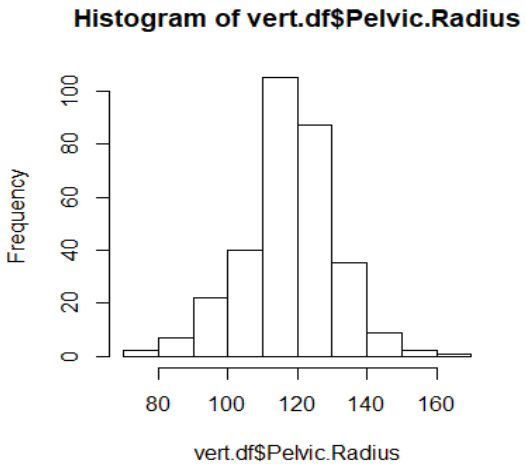


Figure 5: Histogram of Pelvic Radius

Grade of spondylolisthesis (SL)

Grade.of.SL
Min. :-11.06
1st Qu.:1.60
Median :11.77
Mean :26.30
3rd Qu.:41.28
Max. :418.54

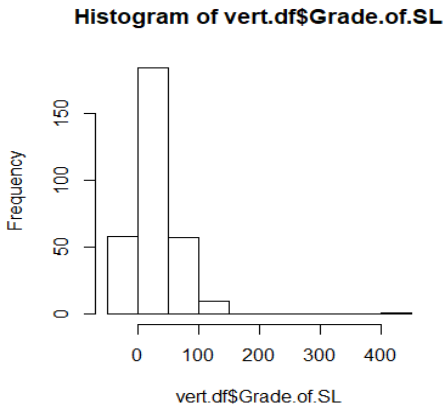


Figure 6: Histogram of Grade of Spondylolisthesis

Decision

Decision
AB:210
NO:100

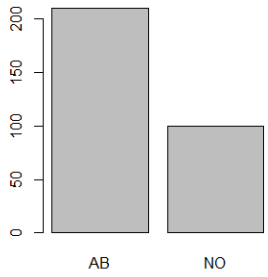
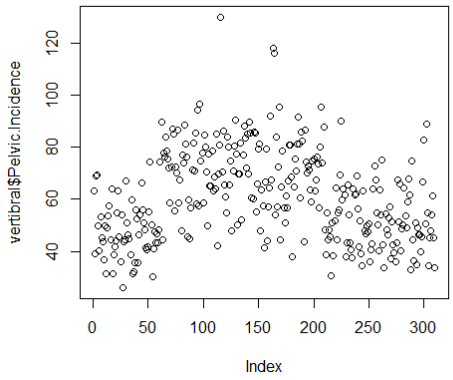


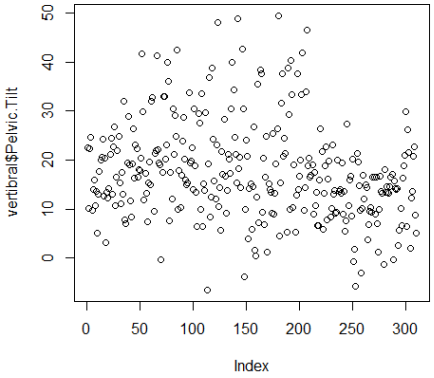
Figure 7: : Bar Plot of Dependent variable - Decision

Graphs of each independent variable:

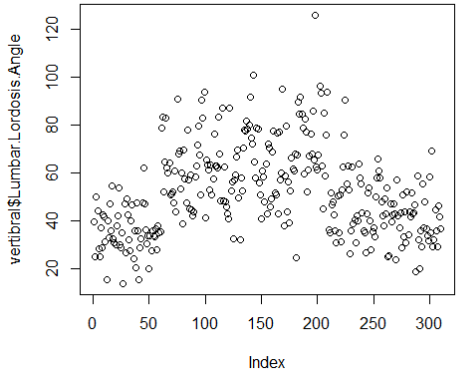
Attribute: Pelvic Incidence:



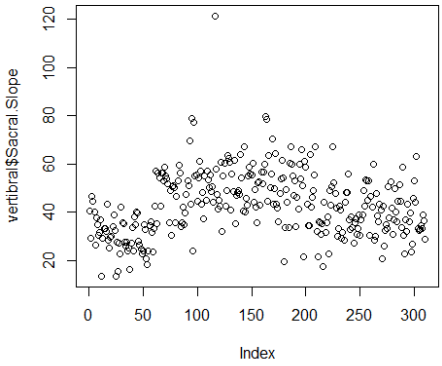
Pelvic Tilt



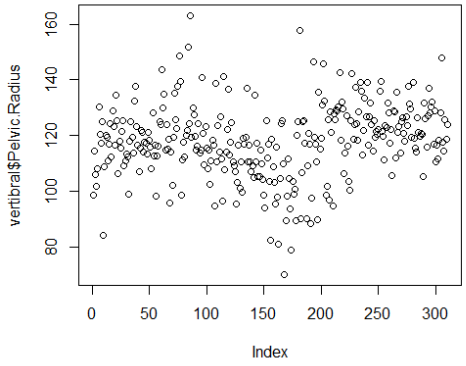
Lumbar Lordosis Angle



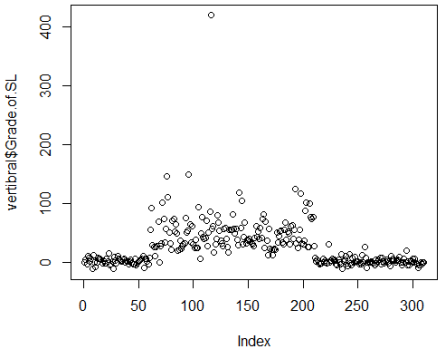
Sacral Slope



Pelvic Radius



'Grade of SL' Attribute:



Part: B

This table can be built using linear and polynomial kernels with the codes attached.
10-fold cross validation is used.

Different Models building using the codes in R [File: ***linearKernel.R*** & ***polynomialKernel.R***].
while Cross validated Error calculation, 10 Fold Cross Validation is used in the Codes.

ID	Kernel	Cost Constant	Training Error (%)	10 fold Cross-validated Error
1	Linear	0.001	32.25806	32.25806
2	Linear	0.01	28.3871	29.35484
3	Linear	0.1	13.22581	13.87097
4	Linear	1	15.16129	15.48387
5	Linear	10	13.87097	15.48387
6	Linear	100	13.54839	14.19355
7	Linear	1000	13.54839	14.83871
8	Linear	10000	13.54839	15.16129
9	Polynomial, degree = 2	10	24.83871	30.64513
10	Polynomial, degree = 2	100	23.22581	26.77419
11	Polynomial, degree = 2	1000	25.16129	29.03226
12	Polynomial, degree = 3	10	15.48387	19.03226
13	Polynomial, degree = 3	100	13.22581	19.35484
14	Polynomial, degree = 3	1000	10.96774	17.74194

It can be seen from the above table with linear kernel, increasing the complexity (cost, C) training error and cross validated errors are decreasing. Also, with polynomial kernel increasing the degree and cost increases the complexity; hence, reduces cross validated error.

Two best models are chosen considering the *complexity and cross validated errors* are:

ID	Kernel	Cost Constant	Training Error (%)	10 fold Cross-validated Error
3	Linear	0.1	13.22581	13.87097
6	Linear	100	13.54839	14.19355

Part C:

AB = Abnormal and NO = Normal. If we consider AB = +1, and NO = -1 then,
The confusion matrix is for *ID: 3*

		Predicted value	
Observed value		AB	NO
AB		195	15
NO		26	74

The confusion matrix is for *ID: 6*

		Predicted value	
Observed value		AB	NO
AB		189	21
NO		21	79

1. It can be seen from the confusion matrix that, for model: 3, among 41 (26+15 = 41) errors 15 are the false negative errors, which is somehow balanced.
   On the other hand, for model 6, among 42 (21+21 = 42) errors both false negative and false positive are same. So, it is neither balanced nor unbalanced.
2. Yes, model 3: kernel: linear, cost = 0.1 is preferable over model 6 because, in model 3 among 41 errors 15 s are false negative. Which is much less severe than model 6. In model 6 among 42 errors there are 21 s are false negative.

Part D:

Using the Bootstrap, hold-out method with my top two models and a 95% confidence interval, the lower and upper bound errors are given as follows:

	Lower Bound Error	**Upper Bound Error**
Model ID: 3 (Kernel = Linear, C = 0.1)	0.0753	0.2473
Model ID: 6 (Kernel = Linear, C = 100)	0.0753	0.2151

Code attached (File: **bootstrap.R**)

1. For a total of 1000 Bootstrap samples with the 95% error confidence interval, the lower bound is 2.5^{th} percentile which is: 25^{th} value of the corresponding error array and the upper bound is 97.5^{th} percentile which is: 975^{th} value of the corresponding error array.
   For Model ID: 3, the 95% confidence interval is: **[0.0753, 0.2473].**
   For Model ID: 6, the 95% confidence interval is: **[0.0753, 0.2151]**
2. Two models are: f_{D1} = [kernel = Linear, C = 0.1] with 95% confidence interval **[0.0753, 0.2473]**, Cross Validation Error (CVE_{D1}) = 0.1387 and f_{D2} = [kernel = Linear, C = 100] with the 95% confidence interval **[0.0753, 0.2151]**, Cross Validation Error (CVE_{D1}) = 0.1419.

we found: model 6's confidence interval is completely overlapped with the model 3's confidence interval. As a result, the performance of these two models are **not significantly** different.

To select the best models among these two, it is needed to see the other parameters besides Cross Validation Errors (CVE). So, **complexity** is considered here. As a result, model ID: 3 is which has Cost = 0.1 is much less complex than model ID: 6, which is Cost = 100. So, model ID: 3 ie. $\mathbf{f_{D1}}$ **= [kernel = Linear, C = 0.1]** is selected.

Moreover, model ID: 3 has the better performance of CVE than model ID: 6.