

Using Logistic Regression to Classify Spine Disorders via Normal/ Abnormal Cases

Instructor: Dr. Todd Fernandez

Student: Fereshteh Shahmiri

Course: Statistics in Biomedical Engineering

Date: 07. 19. 2020

INTRODUCTION

The common diagnosis and monitoring methods for spine deformations and poor body postures are clinical methods like X-Ray, MRI, and CT Scan. Although their necessity is indispensable, it is usually difficult and not feasible for patients to frequently have clinical visits to check their health situations over a long period of time. That is why wearable sensing which allows self-assessment and self-monitoring by patients as well as remote reporting to care givers/ technicians, could be a valuable way of tracking body postures, not only for patients with serious posture problems, but also, for most of us who can benefit from wearable sensing, capable of notifying us to correct our posture, anytime we are working at desk, sitting, walking or standing, etc. generally while performing our daily routine activities.

As part of my PhD research, I have already implemented the ShArc sensor, a multi-bend sensor which can detect dynamically varying shapes and complex bends in real time with high accuracy. Though the main question has remained **if its high-accuracy bend measurement is high enough to track spine vertebrae flexure and secondly, where on the spine is the best location for posture tracking and a variety of posture health monitoring systems.** Thus, I have planned for:

- **Exploratory Data Analysis through Descriptive Statistics** - I explore a publicly available dataset to learn better about the spine anatomy and key measurement parameters for both normal and abnormal cases with chronic back pain.
- **Factor Analysis to determine which predictors to some extent impact classifying the normal and abnormal spine cases.**
- **Binary classification of normal and abnormal spine cases** using a wide variety of predictors which are deflection radii relevant to three major parts in spinal cord.

Hopefully, such a study will help me to evaluate if my proposed bend sensor with its performance characteristics, could be a practical solution as a wearable sensor for monitoring the range of motions in either or all pelvic, lumbar, and thoracic areas in spine.¹ Therefore, as the first step I restrict my research question(s) to first, detecting what predictors and how impact on binary status of spine normality. And second, proposing an appropriate classification method for categorizing such abnormalities. To do so, I will use Factor Analysis (FA) and Principal Component Analysis (PCA), as well as logistic regression.

DATA

Description of Dataset:

The “Vertebral Column” dataset ² is a multivariate biomedical data set built by Dr. Henrique da Mota during a medical residence period in the Group of Applied Research in Orthopedics (GARO) in Lyon, France. All measures are collected through X-Rays. Dataset has 310 sample size and two different but related classification tasks. The first task consists in classifying patients as belonging to one out of three categories: Normal (100 patients), Disk Hernia (60 patients) or Spondylolisthesis (150 patients). For the second task, the categories Disk Hernia and Spondylolisthesis were merged into a single category labelled as 'abnormal'. Thus, the second task consists in classifying patients as belonging to one out of two categories: Normal (100 patients) or Abnormal (210 patients). The attributes are continuing real numbers. For comprehensive overview of data, please check the references. ^{1 2 3 4}

Descriptive Statistics

This dataset includes twelve independent variables in numeric form. I only focus on one dependent variable in binary form for patients with binary status of normal or abnormal spines. The independent variables are related to the deflection radii collected from three core sections of the spine. Such deflection degrees collectively, determine if the patient’s spine is in normal or abnormal status.

Predictors are 1)pelvic incidence (PI), 2)pelvic tilt (PT), 3) Lumbar Lordosis Angle(LLA), 4) sacral slope (SS), 5)pelvic radius (PR), 6) degree spondylolisthesis (DS), 7) pelvic slope (PS), 8) direct tilt(DT), 9)thoracic slope (TS), 10)cervical tilt(CT), 11)sacrum angle(SA), and 12)scoliosis slope (ScolS). Table 1 demonstrates some basic information about data.

	PI	PT	LLA	SS	PR	DS	PS	DT	TS	CT	SA	ScolS	Class_att
Mean	60.27	17.57	51.94	42.70	117.95	25.03	0.47	21.33	13.07	11.94	-13.99	25.64	0.32
Std dev	16.80	10.01	18.58	12.68	13.33	30.23	0.29	8.65	3.40	2.90	12.19	10.47	0.47
Min	26.15	-6.55	14.00	13.37	70.08	7.03	7.04	7.03	-35.29	7.01	26.15	-6.55	14.00

¹ The proposal for this project was submitted earlier in the semester, for project 1. After discussing with instructor, I postponed this option for the second project.

² <https://www.kaggle.com/caesarlupum/vertebralcolumndataset>

Max	129.83	49.43	125.74	121.43	163.07	36.74	19.32	16.82	6.97	44.34	129.83	49.43	125.74
Count	310	310	310	310	310	310	310	310	310	310	310	310	310
Skewness	0.37	0.67	0.60	0.23	-0.18	1.30	0.02	0.01	0.02	0.01	-0.02	0.07	0.76
Kurtosis	-0.36	0.67	0.15	-0.26	0.94	1.59	-1.15	-1.25	-1.09	-1.15	-1.20	-1.15	-1.44

Table 1: Descriptive Statistics

Having a pre-assumption of normal distribution of data, it was important to me to explore the skewness and kurtosis of the data.

Skewness – By default, my understanding was that since logistic regression and many other classification methods are sensitive to skewed data and outliers, it would be a necessary preprocessing task to remove all outliers from dataset. The skewness values for PI, SS, PR, PS, DT, TS, CT, SA, and ScolS, are between -0.5 and +0.5. Thus, we can consider their distribution as **fairly symmetrical**. The skewness values for PT and LLA, are between 0.5 and 1.0. Thus, we count these two variables as **moderately skewed**. The last parameter DS with skewness equal to 1.3 is considered as **highly skewed** since it is larger than 1. ⁵

Kurtosis - Since kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution, its calculation is very helpful to gain insight of the number of outliers in our data. We know that a data set with high kurtosis tends to have heavy tails, or outliers and a dataset with low kurtosis tends to have light tails, or lack of outliers. Figure 1 and 2, a histogram and a box and whisker plots illustrate both the skewness and kurtosis of data set. ⁶ By removing the outliers from data set and going through the whole process I noticed that such outlier reduction does not improve the accuracy of classification. (I eventually got the score =84.05 with including all predictors which is about 2% less than the calculation including all outliers.) Thus, I redo the whole process, including outliers in dataset.

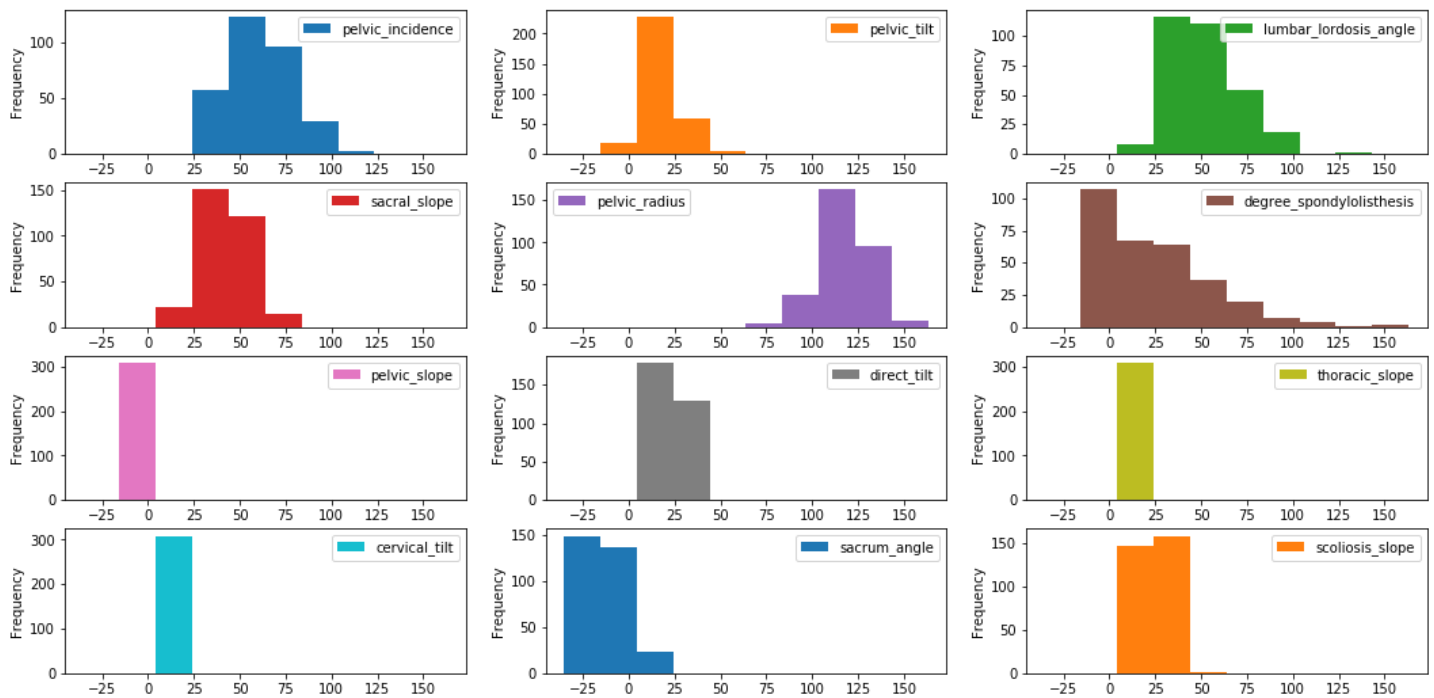


Figure 1: Histogram – it clearly shows the distribution of data, its skewness and kurtosis.

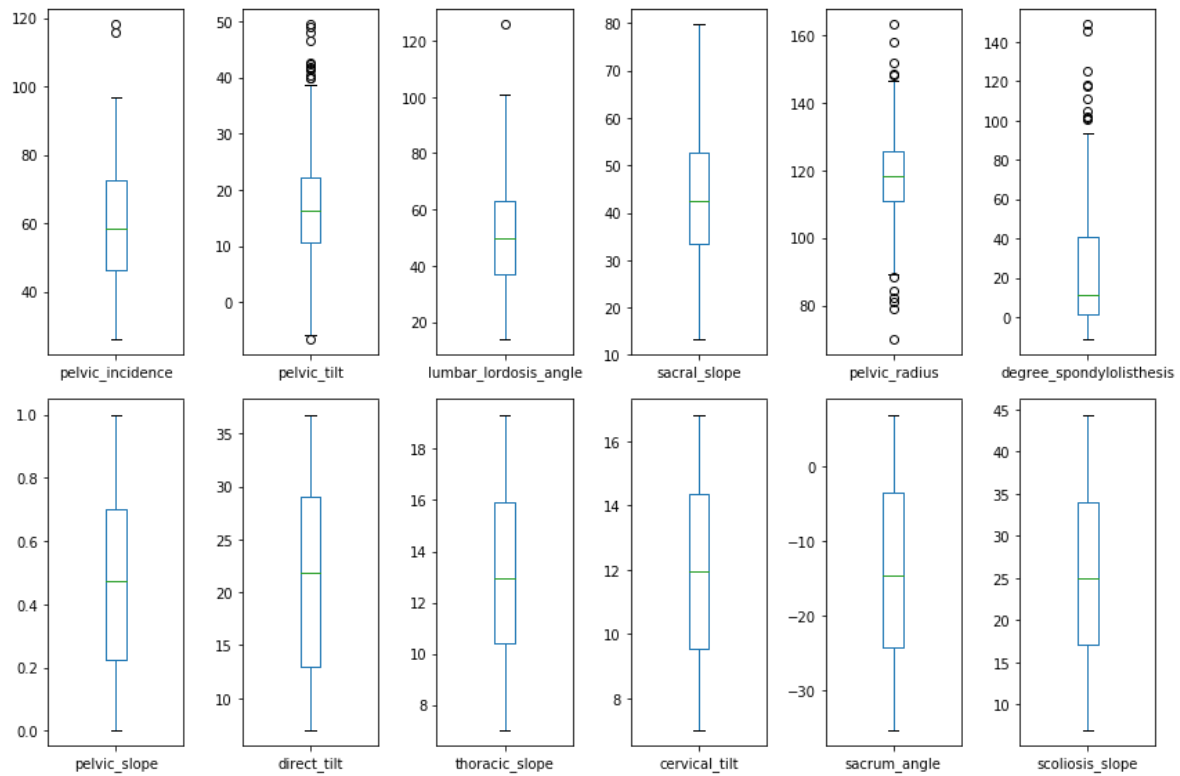


Figure 2: box and whisker plot – Pelvic Tilt, Lumbar Lordosis Angle, Pelvic Radius, and degree spondylolisthesis are impacted by outliers. All outliers in belongs to abnormal spine class. As a note, I will add the point that one time, I went through the calculation with all predictors, although got the significant p-value, the accuracy did not improve comparing the calculation including all outliers. Thus, in all next steps, I continue my analysis including outliers. y axes are not in the same scale in all plots. Thus, we cannot compare the variance among variables in this plot.

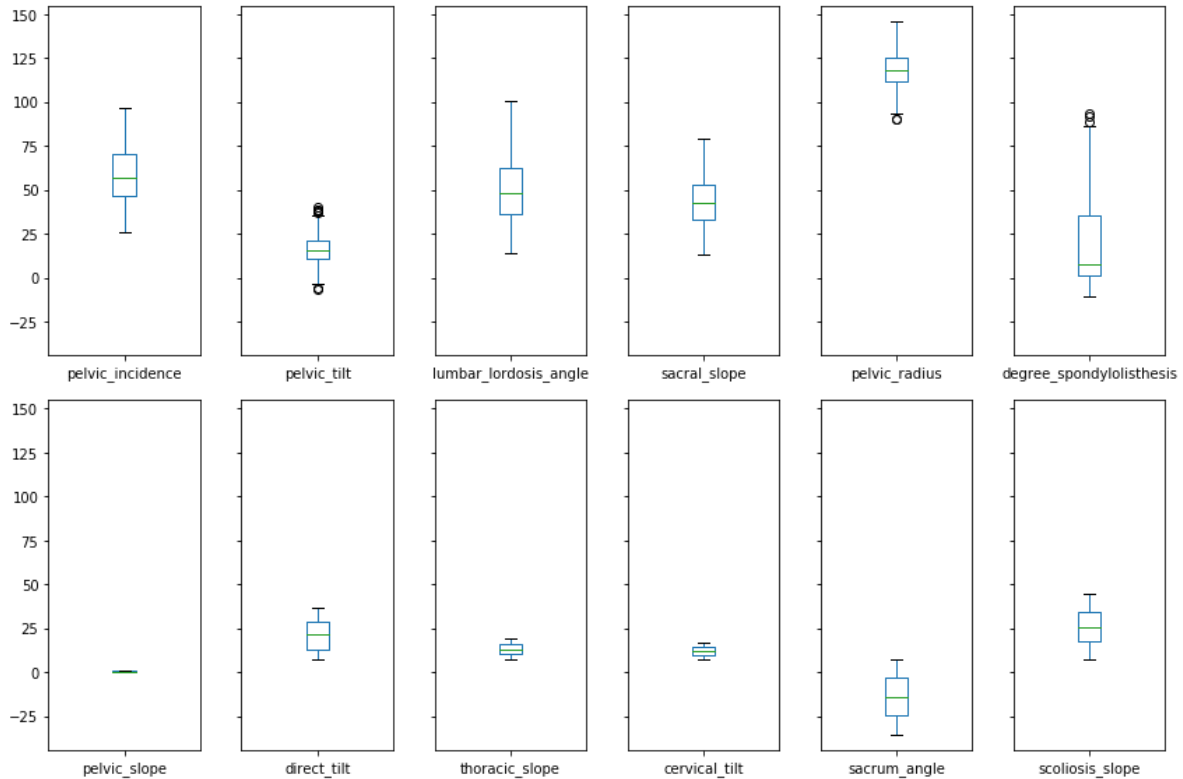


Figure 3: box and whisker plot – y axes are in the same scale in all plots. Thus, we can compare the variance among variables in this plot in which we have removed most of the outliers.

Correlation among Independent Variables - As table 2 and figure 4 show, there are moderate correlations between first three independent variables. And for the rest of the predictors such a pairwise correlation is too low and almost close to zero.

Correlation Matrix													
	PI	PT	LLA	SS	PR	DS	PS	DT	TS	CT	SA	ScolS	Class_att
PI	1.00	0.66	0.74	0.80	-0.24	0.64	0.03	-0.08	-0.08	0.02	0.04	-0.01	-0.35
PT	0.66	1.00	0.43	0.08	0.03	0.53	0.01	-0.07	-0.07	0.03	0.03	-0.06	-0.33
LLA	0.74	0.43	1.00	0.64	-0.08	0.67	0.03	-0.11	-0.06	0.06	0.06	-0.05	-0.31
SS	0.80	0.08	0.64	1.00	-0.35	0.43	0.02	-0.04	-0.06	0.00	0.03	0.03	-0.21
PR	-0.24	0.03	-0.08	-0.35	1.00	0.00	0.02	0.06	0.06	-0.04	0.03	-0.03	0.31
DS	0.64	0.53	0.67	0.43	0.00	1.00	0.05	-0.07	-0.05	0.08	0.10	-0.06	-0.52
PS	0.03	0.01	0.03	0.02	0.02	0.05	1.00	0.01	-0.01	0.09	0.07	-0.07	-0.05
DT	-0.08	-0.07	-0.11	-0.04	0.06	-0.07	0.01	1.00	0.01	0.07	-0.04	-0.02	0.04
TS	-0.08	-0.07	-0.06	-0.06	0.06	-0.05	-0.01	0.01	1.00	0.05	0.01	0.01	0.05
CT	0.02	0.03	0.06	0.00	-0.04	0.08	0.09	0.07	0.05	1.00	0.06	0.02	-0.10
SA	0.04	0.03	0.06	0.03	0.03	0.10	0.07	-0.04	0.01	0.06	1.00	0.02	-0.03
ScolS	-0.01	-0.06	-0.05	0.03	-0.03	-0.06	-0.07	-0.02	0.01	0.02	0.02	1.00	0.07
Class_att	-0.35	-0.33	-0.31	-0.21	0.31	-0.52	-0.05	0.04	0.05	-0.10	-0.03	0.07	1.00
	2.24	1.04	1.73	1.42	0.29	1.64	0.03	0.04	0.03	0.04	0.03	0.02	0.77

Table 2: Correlation Matrix - Pairwise correlation between all twelve Independent Variables

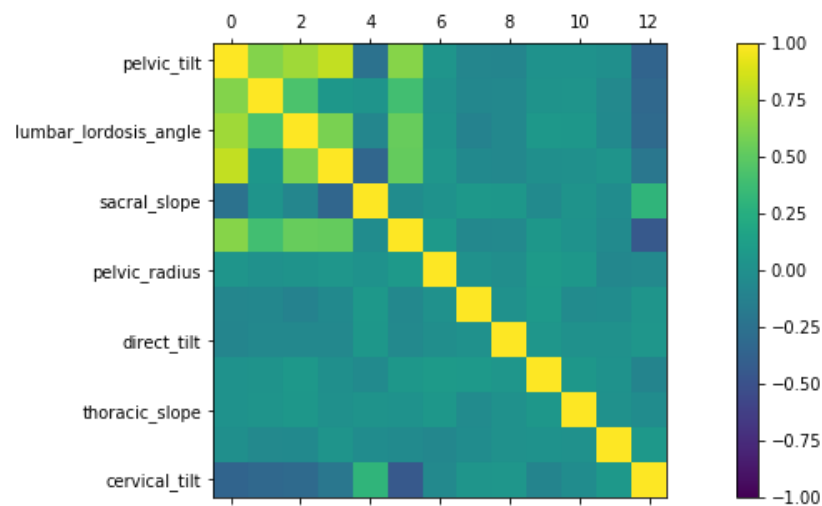


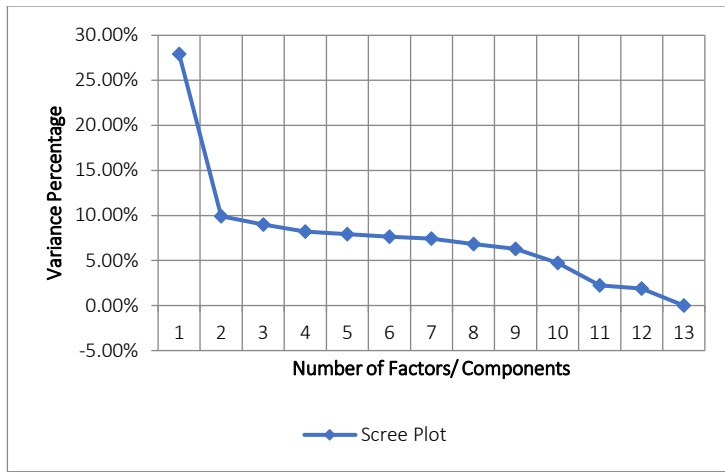
Figure 4: Correlation Matrix – Pairwise correlation between all twelve Independent Variables

Factor Analysis (FA) & Principal Component Analysis (PCA)

One of the main questions in this study is determining which predictors impact the predicted value, and how is the magnitude of such impact. Further, having answer to this question allows more effective performance for the classification task.

Determine the number of factors

Initially we need to apply PCA to determine how many factors explain the variability of data and an acceptable level of variance. By referring to the literature, 80% -90% of the variance explained by factors, determine the number of factors that we need. Using Kaiser criterion ⁷, we may use the factors with eigenvalues that are greater than 1. According to Kaiser criterion, the variance explained in dataset is about 63% (Figure 5 and Table 3)



Scree Plot		
Variance (eigenvalue)	Variance (%)	Cum %
3.632402	27.94%	27.94%
1.289234	9.92%	37.86%
1.168034	8.98%	46.84%
1.068472	8.22%	55.06%
1.028111	7.91%	62.97%
0.993878	7.65%	70.62%
0.966898	7.44%	78.05%
0.889118	6.84%	84.89%
0.816374	6.28%	91.17%
0.612999	4.72%	95.89%
0.291358	2.24%	98.13%
0.243122	1.87%	100.00%
-3.9E-16	0.00%	100.00%

Figure 5: Scree plot to determine the number of factors to retain in an exploratory factor analysis or principal components to keep in principal component analysis (PCA). It shows the eigenvalues in downward curve, ordering the eigenvalues from largest to smallest. The elbow of the graph where the eigenvalues seem to level off is found and factors or components to the left of this point should be retained as significant. ⁸

Table 3: first column is Variance. Since I have used the PCA to extract factors, the variance equals the eigenvalue. Second column is the percentage of variance to determine the amount of variance each factor explain. Third column is cumulative variance.

Factor Matrix (unrotated)												
Predictors	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10	Factor 11	Factor 12
PI	0.93	-0.09	-0.10	-0.06	0.09	-0.14	-0.03	0.03	-0.01	0.23	-0.18	-0.03
PT	0.64	0.43	-0.23	0.07	0.05	0.24	0.15	0.07	0.13	0.49	-0.02	0.03
LLA	0.84	0.04	-0.10	-0.12	0.07	-0.17	-0.08	0.04	-0.15	-0.15	0.29	0.31
SS	0.73	-0.46	0.05	-0.13	0.09	-0.37	-0.16	-0.02	-0.11	-0.09	-0.22	-0.06
PR	-0.27	0.71	-0.38	-0.21	0.16	-0.22	0.05	0.12	-0.10	-0.29	-0.23	0.05
DS	0.82	0.29	-0.04	-0.02	0.04	0.09	0.04	-0.02	0.06	-0.28	0.19	-0.34
PS	0.06	0.28	0.49	0.08	-0.34	-0.43	-0.12	0.50	0.32	0.04	0.02	0.00
DT	-0.13	0.11	0.30	0.23	0.67	-0.36	0.24	-0.33	0.29	0.04	0.04	0.02
TS	-0.11	0.13	0.16	-0.40	0.33	0.26	-0.74	-0.04	0.21	0.05	0.00	0.01
CT	0.08	0.19	0.67	-0.16	0.22	0.22	0.19	0.18	-0.56	0.06	-0.03	-0.01
SA	0.09	0.22	0.24	-0.53	-0.46	-0.14	0.15	-0.59	0.06	0.08	-0.01	0.02
ScolS	-0.05	-0.33	-0.06	-0.62	0.18	0.15	0.47	0.34	0.35	-0.06	0.02	0.01
Class_att	-0.57	-0.04	-0.29	-0.29	0.11	-0.48	-0.08	0.11	-0.29	0.33	0.20	-0.15
	3.63	1.29	1.17	1.07	1.03	0.99	0.97	0.89	0.82	0.61	0.29	0.24

Table 4: Unrotated Factor Matrix - These results show the unrotated factor loadings for all the factors using the principal components method of extraction. The first five factors have variances (eigenvalues) that are greater than 1. From second till tenth factor the variance only changes about 5%. The percentage of variability explained by the first factor is 27.94%, and the next eight factors each explains about 10% - 6.5% (in descending order) of the variance, respectively. Therefore, as the scree plot and this table shows, **the first 9 factors account for most of the total variability in data equal to 91.17%**. The remaining factors account for a small proportion of the variability and are likely unimportant.

Factor Optimization – Maximum Likelihood Function

As discussed for table 4, by determining the number of factors equal to 9, I repeated the factor analysis and optimized the weights using the maximum likelihood function. Then examined the loading pattern to determine those factors that has the most influence on each variable ordinaly. Loadings close to -1 or 1 indicate that the factor strongly influences the variable. Loadings close to 0 indicate that the

factor has a weak influence on the variable. And some variables may have high loadings on multiple factors. factor matrix table in table 5, allows an easily assessment of the load(s) on each predictor. ⁹

Factor Matrix (rotated Varimax)									
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9
PI	0.83	-0.12	0.00	-0.02	-0.03	0.46	0.05	0.01	0.02
PT	0.23	0.17	-0.02	0.00	-0.06	0.82	0.07	0.04	0.02
LLA	0.80	0.03	0.01	0.05	-0.09	0.38	0.02	-0.03	-0.07
SS	0.92	-0.30	0.02	-0.03	0.01	-0.04	0.02	-0.02	0.02
PR	-0.16	0.91	0.02	0.03	0.05	0.06	-0.04	-0.03	0.01
DS	0.50	0.01	0.03	0.05	-0.02	0.71	0.00	-0.10	-0.07
PS	0.02	0.01	1.00	0.04	0.01	0.01	0.01	-0.03	-0.04
DT	-0.03	0.04	0.01	0.01	1.00	-0.03	0.00	0.02	-0.04
TS	-0.03	0.04	-0.01	-0.01	0.00	-0.03	-1.00	0.00	-0.03
CT	0.01	-0.02	0.04	-0.01	0.04	0.03	-0.03	-0.03	-0.99
SA	0.02	0.02	0.03	-0.01	-0.02	0.02	0.00	-0.99	-0.03
ScolS	0.00	-0.01	-0.04	-1.00	-0.01	-0.03	-0.01	-0.01	-0.01
Class_att	-0.01	0.55	-0.06	-0.07	-0.02	-0.71	0.01	0.04	0.08
	2.50	1.27	1.00	1.01	1.01	2.04	1.00	1.01	1.01

Table 5: Rotated Factor Matrix - In these results, a varimax rotation is performed on the data. Using the rotated factor loadings, I interpret the factors as follows: Pelvic Incidence (PI) (0.83), Lumbar Lordosis Angle (LLA) (0.80), and Sacral Slope (SS) (0.92) have large positive loadings on factor 1, so this factor describes PI, LLA and SS in data very well. Pelvic Radius (PR) (0.91) has large positive loadings on factor 2. Continuing the same way of analysis, we observe that all predictors have a large loading either positive or negative loadings on at least one of the nine factors. Thus, I conclude it is important to keep all twelve independent variables for the classification task. Together, all nine factors explain 91.17% of the variation in the data.

Binary Classification: Logistic Regression

Using logistic regression, we can classify the predicted variables which are in binary or binomial formats. Here, with two status of normal or abnormal spine cases, I implement a classification model. The key questions regarding the classifier would be:

- Can the categories be correctly predicted given a set of predictors? (What would be the highest accuracy, sensitivity, and specificity?)
- How good is the model at classifying cases for which the outcome is known?

The main steps are first, **determining whether the association between predictor(s) and predicted value is statistically significant.** Second, **understanding the effect of predictors and determining how well the model fits data.**

P-Value

To determine whether the association between the predictors (IVs) and predicted value (DV) in the model is statistically significant, I compared the calculated p-value (Shown in Table 6) to significance level of 0.05 to assess the null hypothesis. Since the p value is much smaller than the significance level, I reject the null hypothesis that states the predictors' coefficient is equal to zero, indicating that there is no association between the predictors and the predicted value. In table 7, none of the twelve coefficients are zero. Those closer to ± 1 have more impact on DV, and those closer to 0, less impact. For instance, SS (0.297) and PS (0.219) have the highest positive weights, PI (-0.202) and DS (-0.171) have the highest negative weights.

LLO	LL1	Chi-Square	df	p-value	alpha	significant	R_L^2	R_{CS}^2	R_N^2
-194.93	-88.07	213.708	12	4.78E-39	0.05	yes	0.55	0.50	0.70

Table 6: P-Value, Optimization Parameters, test for goodness of the fit

Log-Likelihood Statistic -

The increase in number of coefficients(b), causes the inflation in the standard error and the Wald statistic become inflated. Such inflation increases the probability that b (combination of all coefficients) is viewed as not making a significant contribution to the model even

when it does (i.e. a type II error). To overcome this problem it is better to test on the basis of the log-likelihood statistic. Maximum Likelihood Estimation (MLE) is a proper method for probabilistic problems. Using log-likelihood statistic we can substitute the Wald test with Chi Square and avoid discussed inflation problem. We can apply such substitution because:

$$2 (LL1 - LL0) \sim X^2 (df)$$

LL1 refers to the full log-likelihood model and LL0 refers to a model with fewer coefficients (especially the model with only the intercept b_0 and no other coefficients). This is equivalent to:

$$-2 \ln(L^0/L^1) \sim X^2 (df)$$

Where the L^1 is the maximum likelihood for the full model with coefficient b and L^0 is the maximum likelihood for the reduced model without coefficient b and only includes the intercept b_0 .

Model Optimization – Error Minimization

We use MLE to minimize the classification error. Two models are required. First, the logistic regression model with coefficient b (The model with all predictors' coefficients), and second, the model without coefficient b . We need to calculate the likelihood ratio test statistic to gain the log likelihood for $LL0$ and $LL1$. Hence, having the formula, $2 (LL1 - LL0) \sim X^2 (df)$, we calculate the p -value which is equal to $CHIDIST (X^2 (df))$. As table 6 shows $p = 4.78E-39 < 0.05 = \alpha$.

Goodness of the Fit

For the goodness of the fit test, R^2 should be minimized. There are many ways to calculate an R^2 for logistic regression, and no consensus on which one is best.¹⁰ (Discussing this topic is beyond my project). As discussed, logistic regression is estimated by maximizing the likelihood function. Assuming L_0 is the value of the likelihood function for a model with no predictors, and L_1 is the likelihood for the model being estimated with all predictors, an R^2 is calculated by some adjustment in relation of these two log likelihood values. The most optimum fit would be achieved by the most minimum value for such R^2 . In the table 6, you see three different values for R^2 . **Cox and Snell's R^2** is R_{CS}^2 , **Nagelkerke's R^2** is R_N^2 and **McFadden's R^2** is R_L^2 . The pseudo- R^2 statistics are considering all these three possibilities. Either formula could be practical in different cases.

As results in table 7 show, the predictors PR and DS are statistically significant (less than the significance level of 0.05). although in combination of all predictors we got the significant difference (Table 6), we can see that most predictors do not have p -value less than 0.05.

We can conclude that predictors' changes are associated with changes in the probability that the event occurs. To assess the coefficient to determine whether a change in a predictor variable makes the event more likely or less likely, we need to consider the magnitude and sign of the coefficient. The relationship between the coefficient and the probability depends on several aspects of the analysis, including the link function. Generally, positive coefficients like SS, PS or PR coefficients indicate that the event becomes more likely as the predictor increases. Negative coefficients like PI, DS or DT coefficients indicate that the event becomes less likely as the predictor increases.

	<i>Coefficient b</i>	<i>Standard error</i>	<i>Wald</i>	<i>p-value</i>	<i>exp(b)</i>
Intercept	-15.17	3.55	18.218235124946300	0.00001969838033	0.00
Pelvic incidence	-0.20	206138.95	0.0000000000000960	0.99999921825248	0.82
Pelvic tilt	0.12	206138.95	0.0000000000000343	0.99999953290786	1.13
Lumbar lordosis angle	0.02	0.02	0.898323675090759	0.34323159344827	1.02
Sacral slope	0.30	206138.95	0.0000000000002071	0.99999885172410	1.35
Pelvic radius	0.11	0.02	20.081590423353800	0.00000742075557	1.11
Degree spondylolisthesis	-0.17	0.02	51.779200895388500	0.000000000000062	0.84
Pelvic slope	0.22	0.69	0.100991410136835	0.75064313137854	1.24
Direct tilt	-0.01	0.02	0.343710287185613	0.55769536322167	0.99
Thoracic slope	0.05	0.06	0.813620545507619	0.36705199572960	1.05
Cervical tilt	-0.05	0.07	0.559178249961515	0.45459152772124	0.95
Sacrum angle	0.00	0.02	0.097834319462646	0.75444417836692	1.00
Scoliosis slope	0.01	0.02	0.279500649937074	0.59702869532261	1.01

Table 7

In the results, the model uses the deflection radii in twelve predictors to predict the presence or absence of spine abnormality in patients. The odd ratio relevant to predictors can show that for what amount of deflection radii increase or decrease, the likelihood that the patient's spine is in normal status increases by one unit.

Accuracy

To gauge the fit of the model to the observed data we can use the % accuracy. **In this analysis, the accuracy is 85.16%.** This statistic says that 85.16% of the observed cases are predicted accurately by the model. For any observed value of the independent variables, when the predicted value of p is greater or equal to 0.5, it is viewed as predicting success and then the % correct is equal to the value of the observed number of successes divided by the total number of observations. When the $p < 0.5$, it is viewed as predicting failure then the % correct is equal to the value of the observed number of successes divided by the total number of observations. These values are weighted by the number of observations of that type and then summed to provide the % correct statistic for all the data.

Sensitivity, Specificity, FPR and Threshold

As we defined earlier, sensitivity or recall is the true positive rate. It measures the proportion of actual positives that are correctly identified as such. Thus, we can see a high sensitivity value in the table 7. There is sensitivity of 89% for normal detection and 81% sensitivity for abnormal detection. However, the table lacks presenting the specificity or true negative rate which measures the proportion of actual negatives that are correctly identified as such. We see there are 23 false negative and 19 false positive in the confusion metrics, table 8.

Metrics		Precision (Positive Predictive Value)	Recall (Sensitivity)	F1-Score	Support
Logistic Regression with Twelve Predictors	0	0.91	0.89	0.90	209
	1	0.78	0.81	0.79	100
	Avg/Total	0.87	0.86	0.86	309

Table 7: statistical measures of the performance of a binary classification test

Confusion Matrix		
	Predicted 0s	Predicted 1s
Actual 0s	186	23
Actual 1s	19	81

Table 8: Confusion Metrics

AUC - ROC Curve

AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represent degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing between true positive and true negative. The ROC curve in figure 6 is plotted with true positive rate against the false positive where TPR is on y-axis and FPR is on the x-axis. Since sensitivity and specificity are inversely proportional to each other. when we increase Sensitivity, Specificity decreases and vice versa. Here, as table 7, 8 and figure 6 we are trading an almost high sensitivity against a low specificity. That means our classifier, is susceptible to more false positive.

To generalize my understanding, I can conclude that the threshold of 0.5 in our logistic regression model is potent to false positive, thus, such a low classification threshold, though increases the sensitivity, decrease the specificity. In figure 6 we observe that since FPR is 1 - specificity. So, when we increase TPR, FPR also increases and vice versa.

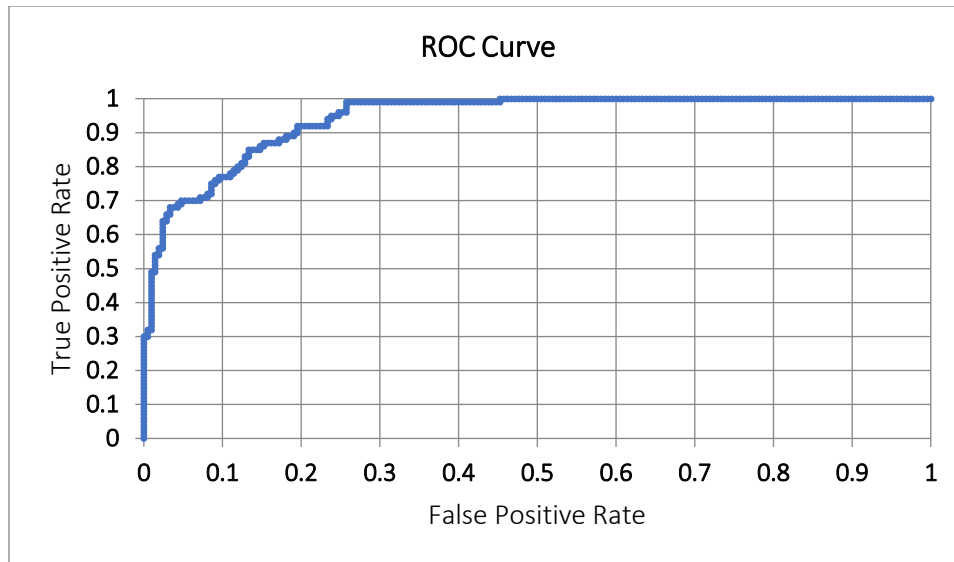


Figure 6 – ROC Curve & AUC

DISCUSSION

For my own curiosity, I decided to run the classification model several times with only one predictor (results are shown in Table 7, 8 and Appendix section). I wondered those predictors with no significant impact on the multi-predictor model, would be significant or not if there is only one predictor in the model. As shown in table 7, 8, and Appendix,

All statistical results for this part are added to the Appendix part. We can see the p-values for "pelvic slope", "direct tilt", "thoracic slope", "cervical tilt", "sacrum angle", and "scoliosis slope" are bigger than significance level of 0.05, thus those factors do not cause significant difference in our classification. In fact, such result is predictable even with observing the visualization part in table 10. We can see for those predictors the sigmoid probability curve (illustrated in green dots) never reach the threshold of 0.5. That means inherently it is not classifying our binary classes.

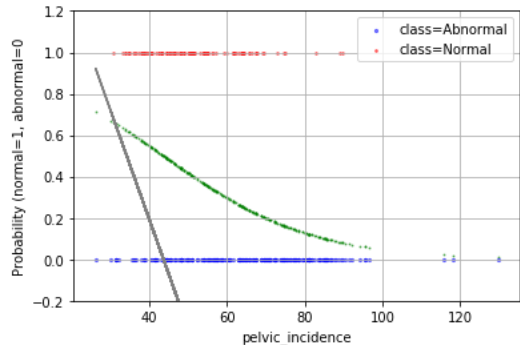
For the other predictors like "pelvic incidence", "pelvic tilt", "lumbar lordosis angle", "sacral slope", "pelvic radius", and "degree spondylolisthesis", the p-value is < 0.05 . Thus, those factors significantly impact on the performance of the classification if the logistic regression uses them respectively as a single predictor in the classification model. The key point that we must discuss is that significance by itself is not enough. We always need to consider the effect size as well. Checking the coefficient for these predictors, we observe the very small values for coefficients demonstrate the small effects here.

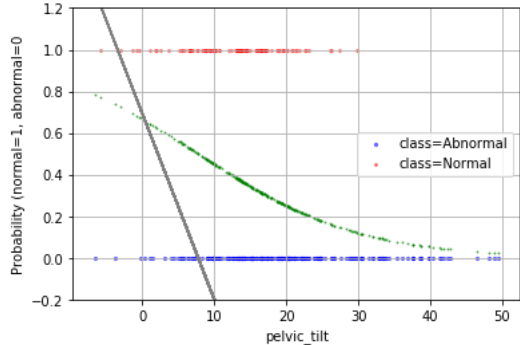
CONCLUSION

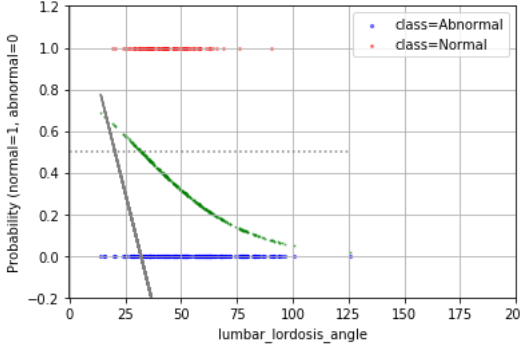
To recap, I conclude the implemented logistic regression classifier has accuracy about 85.16%. It has 86 % sensitivity and 87% precision. Intuitively, and experimentally, I conclude that logistic regression, in which we are dealing with the probability of occurring failure and success events here, has high a false positive rate. Moreover, we can observe the last 6 predictors do not have much impact on our binary classification. However, such impact was not low enough in our factor analysis and PCA part that allows us to get rid of these predictors for our predictive model of binary classification.

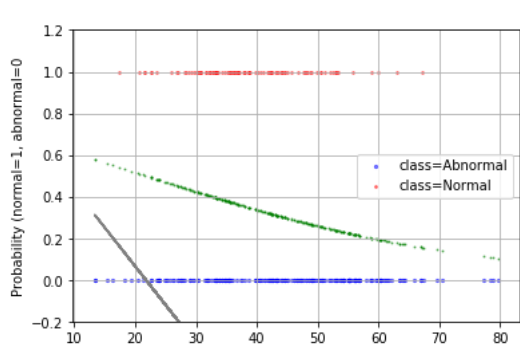
Pelvic Incidence		Precision	Recall	F1-Score	Support
	0	0.72	0.88	0.79	209
	1	0.54	0.29	0.38	100
	Avg/Total	0.66	0.69	0.66	309
Pelvic Tilt	0	0.71	0.90	0.79	209
	1	0.51	0.22	0.31	100
	Avg/Total	0.64	0.68	0.63	309
Lumbar Lordosis Angle	0	0.69	0.89	0.78	209
	1	0.44	0.17	0.24	100
	Avg/Total	0.61	0.66	0.61	309
Sacral Slope	0	0.67	0.96	0.79	209
	1	0.27	0.03	0.05	100
	Avg/Total	0.54	0.66	0.55	309
Pelvic Radius	0	0.69	0.90	0.78	209
	1	0.39	0.13	0.20	100
	Avg/Total	0.59	0.65	0.59	309
Degree spondylolisthesis	0	0.88	0.81	0.84	209
	1	0.66	0.78	0.72	100
	Avg/Total	0.81	0.80	0.80	309
Pelvic Slope	0	0.68	1.00	0.81	209
	1	0.00	0.00	0.00	100
	Avg/Total	0.46	0.68	0.55	309
Direct Tilt	0	0.68	1.00	0.81	209
	1	0.00	0.00	0.00	100
	Avg/Total	0.46	0.68	0.55	309
Thoracic Slope	0	0.68	1.00	0.81	209
	1	0.00	0.00	0.00	100
	Avg/Total	0.46	0.68	0.55	309
Cervical Tilt	0	0.68	1.00	0.81	209
	1	0.00	0.00	0.00	100
	Avg/Total	0.46	0.68	0.55	309
Sacrum Angle	0	0.68	1.00	0.81	209
	1	0.00	0.00	0.00	100
	Avg/Total	0.46	0.68	0.55	309
Scoliosis Slope	0	0.68	1.00	0.81	209
	1	0.00	0.00	0.00	100
	Avg/Total	0.46	0.68	0.55	309

Table 9: Metrics Table - statistical measures of the performance of a binary classification test

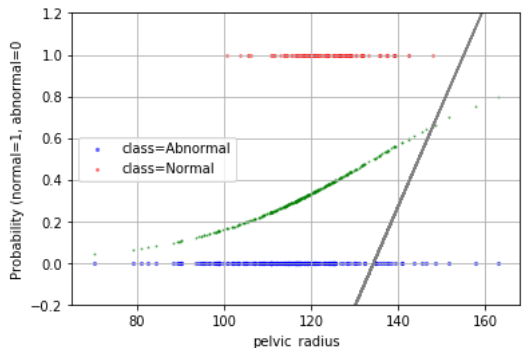
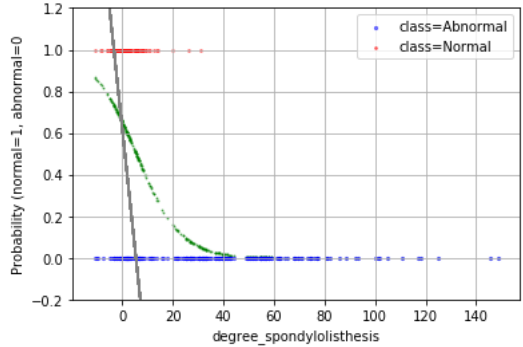
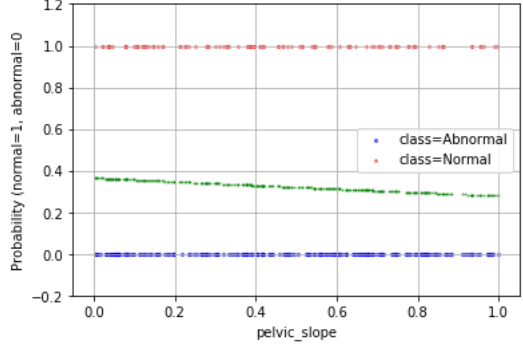
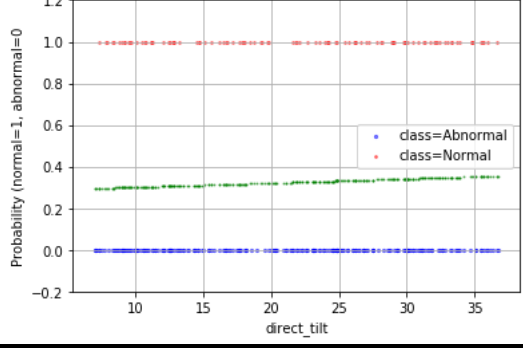
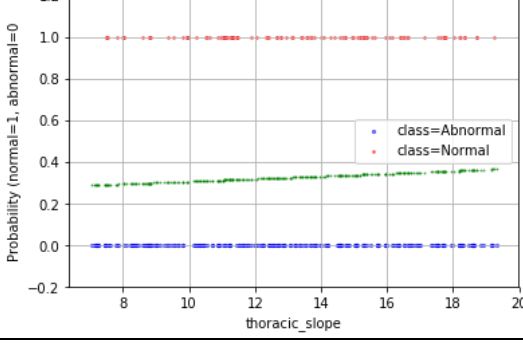
Pelvic Incidence	Intercept	Coefficient	Accuracy	
	2.30	-0.05	0.69	
	Confusion Matrix			
		Predicted 0s	Predicted 1s	
	Actual 0s	185	25	
Actual 1s	71	29		

Pelvic Tilt	Intercept	Coefficient	Accuracy	
	0.70	-0.09	0.70	
	Confusion Matrix			
		Predicted 0s	Predicted 1s	
	Actual 0s	188	21	
Actual 1s	78	22		

Lumbar Lordosis Angle	Intercept	Coefficient	Accuracy	
	1.37	-0.04	0.66	
	Confusion Matrix			
		Predicted 0s	Predicted 1s	
	Actual 0s	187	22	
Actua 1s	83	17		

Sacral Slope	Intercept	Coefficient	Accuracy	
	0.80	-0.037	0.66	
	Confusion Matrix			
		Predicted 0s	Predicted 1s	
	Actual 0s	201	8	
Actual 1s	97	3		

Pelvic Radius	Intercept	Coefficient	Accuracy	
	-6.47	0.05	0.65	
	Confusion Matrix			
		Predicted 0s	Predicted 1s	
	Actual 0s	190	20	

	Actual 1s	87	13	
Degree spondylolisthesis	Intercept	Coefficient	Accuracy	
	0.61	-0.11	0.80	
	Confusion Matrix			
		Predicted 0s	Predicted 1s	
	Actual 0s	169	40	
Actual 1s	22	78		
Pelvic Slope	Intercept	Coefficient	Accuracy	
	-0.55	-0.40	0.68	
	Confusion Matrix			
		Predicted 0s	Predicted 1s	
	Actual 0s	209	0	
Actual 1s	109	0		
Direct Tilt	Intercept	Coefficient	Accuracy	
	-0.93	0.009	0.68	
	Confusion Matrix			
		Predicted 0s	Predicted 1s	
	Actual 0s	209	0	
Actual 1s	109	0		
Thoracic Slope	Intercept	Coefficient	Accuracy	
	-1.11	0.028	0.68	
	Confusion Matrix			
		Predicted 0s	Predicted 1s	
	Actual 0s	209	0	
Actual 1s	100	0		
Cervical Tilt	Intercept	Coefficient	Accuracy	
	0.14	-0.07	0.68	

	Confusion Matrix			<p>Probability (normal=1, abnormal=0)</p> <p>cervical_tilt</p>
		Predicted 0s	Predicted 1s	
	Actual 0s	209	0	
	Actual 1s	100	0	
Sacrum Angle	Intercept	Coefficient	Accuracy	<p>Probability (normal=1, abnormal=0)</p> <p>sacrum_angle</p>
	-0.81	-0.005	0.67	
	Confusion Matrix			
		Predicted 0s	Predicted 1s	
	Actual 0s	209	0	
	Actual 1s	100	0	
Scoliosis Slope	Intercept	Coefficient	Accuracy	<p>Probability (normal=1, abnormal=0)</p> <p>scoliosis_slope</p>
	-1.09	0.014	0.68	
	Confusion Matrix			
		Predicted 0s	Predicted 1s	
	Actual 0s	209	0	
	Actual 1s	100	0	

Table 10

APPENDIX

Logistic Regression - All Twelve Predictors are Applied

```

Current function value: 0.285028
Iterations: 35
st_model = <statsmodels.discrete.discrete_model.Logit object at 0x000002629A156F28>
st_coef = [ -1.51816034e+01 -4.83096917e+00  4.74978916e+00  2.20989036e-02
  4.92568564e+00  1.05369329e-01 -1.70853147e-01  2.18791611e-01
 -1.36885995e-02  5.21952322e-02 -4.95838438e-02 -4.94493321e-03
  9.78456422e-03]
Predict_table [[ 186.  23.]
 [ 23.  77.]]
summary1 =
Logit Regression Results
=====
Dep. Variable:                y      No. Observations:                309
Model:                    Logit      Df Residuals:                  296
Method:                    MLE        Df Model:                      12
Date:                Mon, 20 Jul 2020      Pseudo R-squ.:                0.5473
Time:                22:58:23      Log-Likelihood:               -88.074
converged:                False      LL-Null:                     -194.54
                                   LLR p-value:                6.932e-39
=====
               coef      std err          z      P>|z|      [0.025      0.975]
-----
const         -15.1816      3.554      -4.271      0.000      -22.148      -8.215
x1             -4.8310       nan       nan       nan       nan       nan
x2              4.7498       nan       nan       nan       nan       nan
x3              0.0221      0.023       0.948      0.343      -0.024      0.068
x4              4.9257       nan       nan       nan       nan       nan
x5              0.1054      0.023       4.484      0.000       0.059      0.151
x6             -0.1709      0.024      -7.196      0.000      -0.217     -0.124
x7              0.2188      0.688       0.318      0.750      -1.130      1.567
x8             -0.0137      0.023      -0.587      0.557      -0.059      0.032
x9              0.0522      0.058       0.902      0.367      -0.061      0.166
x10            -0.0496      0.066      -0.747      0.455      -0.180      0.081
x11            -0.0049      0.016      -0.311      0.756      -0.036      0.026
x12             0.0098      0.019       0.527      0.598      -0.027      0.046
=====

```

Possibly complete quasi-separation: A fraction 0.12 of observations can be perfectly predicted. This might indicate that there is complete quasi-separation. In this case some parameters will not be identified.

```

summary2 =
Results: Logit
=====
Model:                Logit      Pseudo R-squared: 0.547
Dependent Variable: y      AIC:                202.1476
Date:                2020-07-20 22:58      BIC:                250.6810
No. Observations:    309      Log-Likelihood:    -88.074
Df Model:            12      LL-Null:          -194.54
Df Residuals:        296      LLR p-value:       6.9323e-39
Converged:            0.0000      Scale:            1.0000
No. Iterations:      35.0000
=====
               Coef.      Std.Err.          z      P>|z|      [0.025      0.975]
-----
const         -15.1816      3.5545      -4.2711      0.0000      -22.1483      -8.2149
x1             -4.8310       nan       nan       nan       nan       nan
x2              4.7498       nan       nan       nan       nan       nan
x3              0.0221      0.0233       0.9482      0.3430      -0.0236      0.0678

```

```

x4      4.9257      nan      nan      nan      nan      nan
x5      0.1054      0.0235      4.4841      0.0000      0.0593      0.1514
x6     -0.1709      0.0237     -7.1958      0.0000     -0.2174     -0.1243
x7      0.2188      0.6880      0.3180      0.7505     -1.1296      1.5672
x8     -0.0137      0.0233     -0.5870      0.5572     -0.0594      0.0320
x9      0.0522      0.0579      0.9022      0.3670     -0.0612      0.1656
x10     -0.0496      0.0664     -0.7465      0.4553     -0.1798      0.0806
x11     -0.0049      0.0159     -0.3105      0.7562     -0.0362      0.0263
x12      0.0098      0.0186      0.5267      0.5984     -0.0266      0.0462
=====

```

Pelvic Incidence

Optimization terminated successfully.

Current function value: 0.560668

Iterations 6

st_model = <statsmodels.discrete.discrete_model.Logit object at 0x0000020796F25EF0>

st_coef = [2.35971108 -0.05371534]

Predict_table [[181. 28.]

[71. 29.]]

summary1 =

Logit Regression Results

```

=====
Dep. Variable:          y      No. Observations:          309
Model:                Logit      Df Residuals:          307
Method:                MLE      Df Model:              1
Date:                Mon, 20 Jul 2020      Pseudo R-squ.:          0.1094
Time:                22:40:01      Log-Likelihood:        -173.25
converged:              True      LL-Null:              -194.54
                                LLR p-value:          6.777e-11
=====

```

```

=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
const          2.3597      0.521      4.533      0.000      1.339      3.380
x1           -0.0537      0.009     -5.893      0.000     -0.072     -0.036
=====

```

summary2 =

Results: Logit

```

=====
Model:                Logit      Pseudo R-squared:    0.109
Dependent Variable:  y      AIC:                350.4927
Date:                2020-07-20 22:40      BIC:                357.9594
No. Observations:    309      Log-Likelihood:     -173.25
Df Model:            1      LL-Null:           -194.54
Df Residuals:        307      LLR p-value:       6.7766e-11
Converged:           1.0000      Scale:             1.0000
No. Iterations:      6.0000
=====

```

```

-----
              Coef.      Std.Err.          z      P>|z|      [0.025      0.975]
-----
const          2.3597      0.5206      4.5327      0.0000      1.3394      3.3801
x1           -0.0537      0.0091     -5.8929      0.0000     -0.0716     -0.0358
=====

```


Pelvic Tilt

Optimization terminated successfully.

Current function value: 0.567781

Iterations 6

st_model = <statsmodels.discrete.discrete_model.Logit object at 0x00000250E60E15F8>

st_coef = [0.70228365 -0.08970192]

Predict_table [[187. 22.]

[78. 22.]]

summary1 =

Logit Regression Results

```
=====
Dep. Variable:          y      No. Observations:          309
Model:                Logit      Df Residuals:            307
Method:                MLE       Df Model:                1
Date:                 Mon, 20 Jul 2020      Pseudo R-squ.:        0.09815
Time:                 22:37:15      Log-Likelihood:       -175.44
converged:             True      LL-Null:            -194.54
                                LLR p-value:        6.429e-10
=====
```

```
=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
const          0.7023      0.277      2.534      0.011      0.159      1.245
x1          -0.0897      0.017     -5.414      0.000     -0.122     -0.057
=====
```

summary2 =

Results: Logit

```
=====
Model:                Logit      Pseudo R-squared:    0.098
Dependent Variable:    y      AIC:                354.8885
Date:                 2020-07-20 22:37 BIC:                362.3552
No. Observations:     309      Log-Likelihood:    -175.44
Df Model:              1      LL-Null:          -194.54
Df Residuals:          307      LLR p-value:      6.4292e-10
Converged:             1.0000      Scale:            1.0000
No. Iterations:        6.0000
=====
```

```
-----
              Coef.      Std.Err.          z      P>|z|      [0.025      0.975]
-----
const          0.7023      0.2771      2.5345      0.0113      0.1592      1.2454
x1          -0.0897      0.0166     -5.4141      0.0000     -0.1222     -0.0572
=====
```

Lumbar Lordosis Angle

Optimization terminated successfully.

Current function value: 0.575137

Iterations 6

st_model = <statsmodels.discrete.discrete_model.Logit object at 0x000001D8523C19B0>

st_coef = [1.40537097 -0.04346933]

Predict_table [[185. 24.]

[82. 18.]]

summary1 =

Logit Regression Results

```
=====
Dep. Variable:          y      No. Observations:          309
Model:                Logit      Df Residuals:            307
Method:                MLE       Df Model:                1
Date:                 Mon, 20 Jul 2020      Pseudo R-squ.:      0.08646
Time:                 22:33:27      Log-Likelihood:     -177.72
converged:             True      LL-Null:            -194.54
                                LLR p-value:      6.630e-09
=====
```

```
=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
const          1.4054      0.409      3.436      0.001      0.604      2.207
x1          -0.0435      0.008     -5.248      0.000     -0.060     -0.027
=====
```

summary2 =

Results: Logit

```
=====
Model:                Logit      Pseudo R-squared: 0.086
Dependent Variable: y      AIC:          359.4347
Date:                 2020-07-20 22:33 BIC:          366.9014
No. Observations:    309      Log-Likelihood: -177.72
Df Model:            1      LL-Null:        -194.54
Df Residuals:        307      LLR p-value:  6.6299e-09
Converged:           1.0000      Scale:          1.0000
No. Iterations:      6.0000
=====
```

```
-----
              Coef.      Std.Err.          z      P>|z|      [0.025      0.975]
-----
const          1.4054      0.4090      3.4365      0.0006      0.6038      2.2069
x1          -0.0435      0.0083     -5.2475      0.0000     -0.0597     -0.0272
=====
```

Sacral Slope

Optimization terminated successfully.

Current function value: 0.606853

Iterations 5

st_model = <statsmodels.discrete.discrete_model.Logit object at 0x000001CDF11DCDD8>

st_coef = [0.82929849 -0.03761463]

Predict_table [[201. 8.]

[97. 3.]]

summary1 =

Logit Regression Results

```
=====
Dep. Variable:          y      No. Observations:          309
Model:                Logit      Df Residuals:            307
Method:                MLE       Df Model:                1
Date:                 Mon, 20 Jul 2020      Pseudo R-squ.:      0.03609
Time:                 22:28:34      Log-Likelihood:      -187.52
converged:            True      LL-Null:            -194.54
                               LLR p-value:      0.0001789
=====
```

```
=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
const          0.8293      0.441      1.880      0.060      -0.035      1.694
x1          -0.0376      0.010     -3.613      0.000      -0.058     -0.017
=====
```

summary2 =

Results: Logit

```
=====
Model:                Logit      Pseudo R-squared: 0.036
Dependent Variable: y      AIC:          379.0349
Date:                 2020-07-20 22:28 BIC:          386.5016
No. Observations:    309      Log-Likelihood: -187.52
Df Model:            1      LL-Null:      -194.54
Df Residuals:        307      LLR p-value: 0.00017894
Converged:           1.0000      Scale:          1.0000
No. Iterations:      5.0000
=====
```

```
-----
              Coef.      Std.Err.          z      P>|z|      [0.025      0.975]
-----
const          0.8293      0.4410      1.8805      0.0600      -0.0351      1.6936
x1          -0.0376      0.0104     -3.6126      0.0003      -0.0580     -0.0172
=====
```

Pelvic Radius

Optimization terminated successfully.

Current function value: 0.576912

Iterations 6

st_model = <statsmodels.discrete.discrete_model.Logit object at 0x000001734AABDC88>

st_coef = [-7.67410735 0.05799199]

Predict_table [[190. 20.]

[86. 14.]]

summary1 =

Logit Regression Results

```
=====
Dep. Variable:          y      No. Observations:          310
Model:                Logit      Df Residuals:            308
Method:                MLE       Df Model:              1
Date:                Sun, 19 Jul 2020      Pseudo R-squ.:      0.08252
Time:                21:31:31      Log-Likelihood:     -178.84
converged:            True       LL-Null:            -194.93
                                LLR p-value:      1.413e-08
=====
```

```
=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
const         -7.6741      1.368      -5.609      0.000     -10.356     -4.992
x1              0.0580      0.011       5.143      0.000       0.036       0.080
=====
```

summary2 =

Results: Logit

```
=====
Model:                Logit      Pseudo R-squared: 0.083
Dependent Variable: y      AIC:          361.6856
Date:                2020-07-19 21:31      BIC:          369.1587
No. Observations:    310      Log-Likelihood: -178.84
Df Model:            1      LL-Null:        -194.93
Df Residuals:        308      LLR p-value:   1.4125e-08
Converged:            1.0000      Scale:        1.0000
No. Iterations:      6.0000
=====
```

```
-----
              Coef.      Std.Err.          z      P>|z|      [0.025      0.975]
-----
const         -7.6741      1.3683      -5.6086      0.0000     -10.3559     -4.9923
x1              0.0580      0.0113       5.1429      0.0000       0.0359       0.0801
=====
```

Degree Spondylolisthesis

Optimization terminated successfully.

Current function value: 0.392158

Iterations 8

st_model = <statsmodels.discrete.discrete_model.Logit object at 0x0000017D65405E80>

st_coef = [0.61556026 -0.11462397]

Predict_table [[169. 40.]

[21. 79.]]

summary1 =

Logit Regression Results

```
=====
Dep. Variable:          y      No. Observations:          309
Model:                Logit      Df Residuals:          307
Method:                MLE       Df Model:              1
Date:                 Sun, 19 Jul 2020      Pseudo R-squ.:      0.3771
Time:                 21:34:34      Log-Likelihood:     -121.18
converged:            True       LL-Null:           -194.54
                               LLR p-value:      9.029e-34
=====
```

```
=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
const          0.6156      0.179      3.440      0.001      0.265      0.966
x1          -0.1146      0.016     -6.952      0.000     -0.147     -0.082
=====
```

summary2 =

Results: Logit

```
=====
Model:                Logit      Pseudo R-squared: 0.377
Dependent Variable: y      AIC:          246.3537
Date:                 2020-07-19 21:34      BIC:          253.8204
No. Observations:    309      Log-Likelihood:  -121.18
Df Model:            1      LL-Null:       -194.54
Df Residuals:        307      LLR p-value:   9.0287e-34
Converged:           1.0000      Scale:        1.0000
No. Iterations:      8.0000
=====
```

```
-----
              Coef.      Std.Err.          z      P>|z|      [0.025      0.975]
-----
const          0.6156      0.1789      3.4399      0.0006      0.2648      0.9663
x1          -0.1146      0.0165     -6.9523      0.0000     -0.1469     -0.0823
=====
```

Pelvic Slope

Optimization terminated successfully.

Current function value: 0.628163

Iterations 5

st_model = <statsmodels.discrete.discrete_model.Logit object at 0x000002A3AC855550>

st_coef = [-0.5513133 -0.39880779]

Predict_table [[209. 0.]

[100. 0.]]

summary1 =

Logit Regression Results

```
=====
Dep. Variable:          y      No. Observations:          309
Model:                Logit      Df Residuals:            307
Method:                MLE       Df Model:                1
Date:                  Mon, 20 Jul 2020      Pseudo R-squ.:      0.002237
Time:                  21:25:48      Log-Likelihood:      -194.10
converged:              True      LL-Null:            -194.54
                                LLR p-value:      0.3509
=====
```

```
=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
const          -0.5513      0.232      -2.379      0.017      -1.005      -0.097
x1             -0.3988      0.428      -0.931      0.352      -1.238      0.441
=====
```

summary2 =

Results: Logit

```
=====
Model:                Logit      Pseudo R-squared: 0.002
Dependent Variable: y      AIC:          392.2049
Date:                  2020-07-20 21:25      BIC:          399.6716
No. Observations:      309      Log-Likelihood:  -194.10
Df Model:              1      LL-Null:      -194.54
Df Residuals:          307      LLR p-value:   0.35090
Converged:              1.0000      Scale:        1.0000
No. Iterations:        5.0000
=====
```

```
-----
              Coef.      Std.Err.          z      P>|z|      [0.025      0.975]
-----
```

Direct Tilt

Optimization terminated successfully.

Current function value: 0.628814

Iterations 5

st_model = <statsmodels.discrete.discrete_model.Logit object at 0x0000021BEA501518>

st_coef = [-0.94400763 0.00963984]

Predict_table [[209. 0.]

[100. 0.]]

summary1 =

Logit Regression Results

```
=====
Dep. Variable:          y      No. Observations:          309
Model:                Logit      Df Residuals:            307
Method:                MLE       Df Model:                1
Date:                  Mon, 20 Jul 2020      Pseudo R-squ.:      0.001204
Time:                  21:29:11      Log-Likelihood:      -194.30
converged:              True      LL-Null:            -194.54
                                   LLR p-value:            0.4938
=====
```

```
=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
const          -0.9440      0.328      -2.881      0.004      -1.586      -0.302
x1              0.0096      0.014       0.684      0.494      -0.018      0.037
=====
```

summary2 =

Results: Logit

```
=====
Model:                Logit      Pseudo R-squared: 0.001
Dependent Variable: y      AIC:          392.6068
Date:                  2020-07-20 21:29      BIC:          400.0735
No. Observations:      309      Log-Likelihood: -194.30
Df Model:              1      LL-Null:        -194.54
Df Residuals:          307      LLR p-value:   0.49377
Converged:              1.0000      Scale:         1.0000
No. Iterations:        5.0000
=====
```

```
-----
              Coef.      Std.Err.          z      P>|z|      [0.025      0.975]
-----
const          -0.9440      0.3277      -2.8807      0.0040      -1.5863      -0.3017
x1              0.0096      0.0141       0.6838      0.4941      -0.0180      0.0373
=====
```

Thoracic Slope

Optimization terminated successfully.

Current function value: 0.628411

Iterations 5

st_model = <statsmodels.discrete.discrete_model.Logit object at 0x000001BFB04FC278>

st_coef = [-1.13552156 0.03033395]

Predict_table [[209. 0.]

[100. 0.]]

summary1 =

Logit Regression Results

```
=====
Dep. Variable:          y      No. Observations:          309
Model:                Logit      Df Residuals:            307
Method:                MLE       Df Model:                1
Date:                 Mon, 20 Jul 2020      Pseudo R-squ.:      0.001843
Time:                 22:07:18      Log-Likelihood:      -194.18
converged:             True      LL-Null:            -194.54
                                LLR p-value:      0.3971
=====
```

```
=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
const          -1.1355      0.489      -2.324      0.020      -2.093      -0.178
x1              0.0303      0.036       0.846      0.398      -0.040      0.101
=====
```

summary2 =

Results: Logit

```
=====
Model:                Logit      Pseudo R-squared: 0.002
Dependent Variable: y      AIC:          392.3581
Date:                 2020-07-20 22:07      BIC:          399.8248
No. Observations:    309      Log-Likelihood: -194.18
Df Model:            1      LL-Null:        -194.54
Df Residuals:        307      LLR p-value:   0.39711
Converged:           1.0000      Scale:         1.0000
No. Iterations:      5.0000
=====
```

```
-----
              Coef.      Std.Err.          z      P>|z|      [0.025      0.975]
-----
const      -1.1355      0.4886      -2.3240      0.0201      -2.0932      -0.1779
x1          0.0303      0.0359       0.8458      0.3977      -0.0400      0.1006
=====
```


Cervical Tilt

Optimization terminated successfully.

Current function value: 0.624494

Iterations 5

st_model = <statsmodels.discrete.discrete_model.Logit object at 0x000002AC80283940>

st_coef = [0.14837265 -0.07488787]

Predict_table [[209. 0.]

[100. 0.]]

summary1 =

Logit Regression Results

```
=====
Dep. Variable:          y      No. Observations:          309
Model:                Logit      Df Residuals:            307
Method:                MLE       Df Model:                1
Date:                  Mon, 20 Jul 2020      Pseudo R-squ.:      0.008065
Time:                  22:10:10      Log-Likelihood:      -192.97
converged:              True      LL-Null:              -194.54
                                LLR p-value:      0.07650
=====
```

```
=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
const          0.1484      0.513      0.289      0.772      -0.857      1.154
x1          -0.0749      0.043     -1.761      0.078      -0.158      0.008
=====
```

summary2 =

Results: Logit

```
=====
Model:                Logit      Pseudo R-squared: 0.008
Dependent Variable: y      AIC:          389.9374
Date:                  2020-07-20 22:10      BIC:          397.4041
No. Observations:      309      Log-Likelihood:  -192.97
Df Model:              1      LL-Null:        -194.54
Df Residuals:          307      LLR p-value:   0.076499
Converged:             1.0000      Scale:        1.0000
No. Iterations:        5.0000
=====
```

```
-----
              Coef.      Std.Err.          z      P>|z|      [0.025      0.975]
-----
const          0.1484      0.5129      0.2893      0.7724      -0.8569      1.1536
x1          -0.0749      0.0425     -1.7613      0.0782      -0.1582      0.0084
=====
```

Sacrum Angle

Optimization terminated successfully.

Current function value: 0.629099

Iterations 5

st_model = <statsmodels.discrete.discrete_model.Logit object at 0x000001CA174E85F8>

st_coef = [-0.81345234 -0.00539943]

Predict_table [[209. 0.]

[100. 0.]]

summary1 = Logit Regression Results

```
=====
Dep. Variable:          y      No. Observations:          309
Model:                Logit      Df Residuals:            307
Method:                MLE      Df Model:                1
Date:                 Mon, 20 Jul 2020      Pseudo R-squ.:      0.0007503
Time:                 22:23:52      Log-Likelihood:      -194.39
converged:             True      LL-Null:            -194.54
                                LLR p-value:            0.5890
=====
```

```
=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
const         -0.8135      0.187      -4.339      0.000      -1.181      -0.446
x1            -0.0054      0.010      -0.540      0.589      -0.025      0.014
=====
```

summary2 = Results: Logit

```
=====
Model:                Logit      Pseudo R-squared: 0.001
Dependent Variable: y      AIC:          392.7832
Date:                 2020-07-20 22:23 BIC:          400.2499
No. Observations:    309      Log-Likelihood: -194.39
Df Model:            1      LL-Null:      -194.54
Df Residuals:        307      LLR p-value: 0.58900
Converged:           1.0000      Scale:          1.0000
No. Iterations:      5.0000
=====
```

```
-----
              Coef.      Std.Err.          z      P>|z|      [0.025      0.975]
-----
const         -0.8135      0.1875      -4.3386      0.0000      -1.1809      -0.4460
x1            -0.0054      0.0100      -0.5400      0.5892      -0.0250      0.0142
=====
```

Scoliosis Slope

Optimization terminated successfully.

Current function value: 0.627136

Iterations 5

st_model = <statsmodels.discrete.discrete_model.Logit object at 0x000001F121CC0160>

st_coef = [-1.10765584 0.01429691]

Predict_table [[209. 0.]

[100. 0.]]

summary1 =

Logit Regression Results

```
=====
Dep. Variable:          y      No. Observations:          309
Model:                Logit      Df Residuals:            307
Method:                MLE       Df Model:                1
Date:                Mon, 20 Jul 2020      Pseudo R-squ.:      0.003869
Time:                22:26:01      Log-Likelihood:      -193.78
converged:            True      LL-Null:            -194.54
                                LLR p-value:      0.2199
=====
```

```
=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
const          -1.1077      0.329      -3.365      0.001      -1.753      -0.462
x1              0.0143      0.012       1.224      0.221      -0.009      0.037
=====
```

summary2 =

Results: Logit

```
=====
Model:                Logit      Pseudo R-squared: 0.004
Dependent Variable: y      AIC:          391.5699
Date:                2020-07-20 22:26      BIC:          399.0366
No. Observations:    309      Log-Likelihood: -193.78
Df Model:            1      LL-Null:      -194.54
Df Residuals:        307      LLR p-value:  0.21987
Converged:           1.0000      Scale:        1.0000
No. Iterations:      5.0000
=====
```

```
-----
              Coef.      Std.Err.          z      P>|z|      [0.025      0.975]
-----
const          -1.1077      0.3292      -3.3648      0.0008      -1.7529      -0.4625
x1              0.0143      0.0117       1.2243      0.2208      -0.0086      0.0372
=====
```

REFERENCES

¹ <http://archive.ics.uci.edu/ml/machine-learning-databases/00212/>

² Berthonnaud, E., Dimnet, J., Roussouly, P. & Labelle, H. (2005). 'Analysis of the sagittal balance of the spine and pelvis using shape and orientation parameters', Journal of Spinal Disorders & Techniques.

³ Rocha Neto, A. R. & Barreto, G. A. (2009). 'On the Application of Ensembles of Classifiers to the Diagnosis of Pathologies of the Vertebral Column: A Comparative Analysis', IEEE Latin America Transactions, 7(4):487-496.

⁴ Rocha Neto, A. R., Sousa, R., Barreto, G. A. & Cardoso, J. S. (2011). 'Diagnostic of Pathology on the Vertebral Column with Embedded Reject', Proceedings of the 5th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2011), Gran Canaria, Spain, Lecture Notes on Computer Science, vol. 6669, p. 588-595.

⁵ <https://www.spcforexcel.com/knowledge/basic-statistics/are-skewness-and-kurtosis-useful-statistics>

⁶ <https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm>

⁷ <https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/factor-analysis/interpret-the-results/key-results/>

⁸ https://en.wikipedia.org/wiki/Scree_plot#:~:text=The%20scree%20plot%20is%20used,known%20as%20a%20scree%20test.

⁹ Factor analysis and PCA

¹⁰ <https://statisticalhorizons.com/r2logistic#:~:text=The%20Cox%20and%20Snell%20R%20is,predictors%20by%20precisely%20this%20formula>