Vertebral column pathologies diagnosis using Multinomial Logistic Regression

Elena Lahoz Moreno, Claudia Lucio Sarsa, Antonio Martínez Payá

January 25th 2017

1 Abstract

Orthotics is the part of the medicine in charge of correcting or avoiding the traumas of the locomotor system of the human body. An essential part of the musculoskeletal system is the vertebral column. But, the column can suffer from several pathologies that cause backaches. We aim to conceive a model that can be used as an auxiliary system to aid on medical decision making.

We have choosen a dataset with 310 rows -which are the patients- and 6 variables, plus the response one -which classify if the patient has a column pathology or not in to 3 classes-.

We have designed a multinomial logistic regression model considering differents variables, and applying diverse techniques and measures such as k-fold cross-validation (measuring the sensitivity of the model), step-AIC and VIF. In the end, we obtained a model with 4 attributes. The reasons why we chose this model were that the sensitivity in the predictions obtained with k-fold cross-validation were considerably good, it has less attributes and a smaller VIF. The interpretation of the betas of the model could help medicine professionals to identify how the probability of having one of these two pathologies is been influenced by the biomechanical measures of the patients.

2 Introduction

Over the last few years, machine learning techniques have been applied to several medical fields such as cardiology, pulmonology or oncology. One reason for this practice is that the capacity of human diagnosis can be affected by unfavourable factors like fatigue, stress and insufficient technical knowledge, while a model diagnosis is not influenced by these conditions.

Even though the use of this type of techniques are already widely extended in Medicine Diagnosis, their application in Traumatic Orthopedics is uncommon in the literature due to the absence of numerical attributes that describe the pathologies, so it is unable to create an appropiate database in order to design classification models.

Orthotics is the medical field in charge of correcting or avoiding the deformities and traumas of the locomotor system of the human body. An essential part of the musculoskeletal system is the vertebral column. Its main functions are to serve as the body support element, to provide protection to the spinal cord and nervers roots, and to make movement possible. But, this system can suffer from several pathologies that cause backaches with different pain intensities and symptoms.

The purpose of this project is to design a multinomial logistic regression model that is capable of identifying and classifying vertebral column pathologies based on biomechanical features. We aim to conceive a model that can be used as an auxiliary system to aid on medical decision making. Since the objective of the model is to perform a diagnosis, we prefer a pessimistic model, this means

that the classification mistakes will be identifying normal spines as damaged spines, rather than the other way around. It is preferable to misclassify healthy patients than unhealthy ones because further medical tests would prove the good state of health of the patients. Moreover, we want to help medicine professionals to identify which are the main important variables for each pathology in order to avoid extra work in the task of taking measures of the biomechanical features.

The dataset was obtained from Kaggle website. It consists of 310 instances were each one has 6 biomechanical attributes -the predictors- and one response variable, which specifies if the patient has some kind of column pathology or not. There are 100 healthy patients, 60 patients with disc hernia and 150 patients with spondylolisthesis.

2.1 Response variable

Our response variable is a categorical variable that takes three different values:

Normal: A healthy spine without any pathology.

Hernia: Disc hernia appears when the soft core of the intervertebral disc migrates its place (from the center to the periphery of the disc). Once heading towards the medullary channel or to the spaces where the nervous roots lie, this leads inevitability to their compression which causes intense pain, see Fig. 1.

Spondylolisthesis: It occurs when one of the 33 vertebras of the vertebral column slips in relation to the others (Fig.2). This slipping occurs generally towards the base of the spine in the lumbar region, causing pain or symptomatology irritation of the nervous roots.

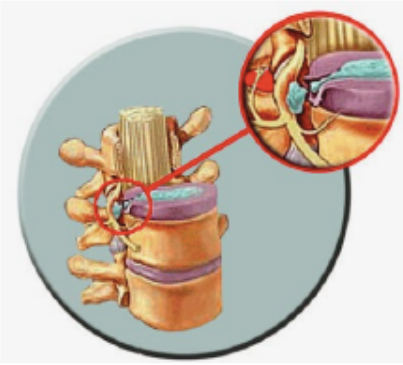

Fig.1: Disc hernia

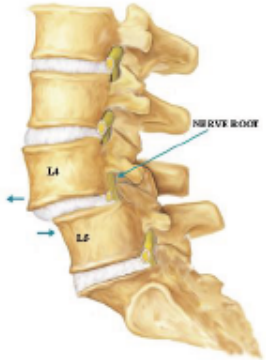

Fig.2: Spondylolisthesis

Below is shown the piechart representing the percentage of each category of the variable class.

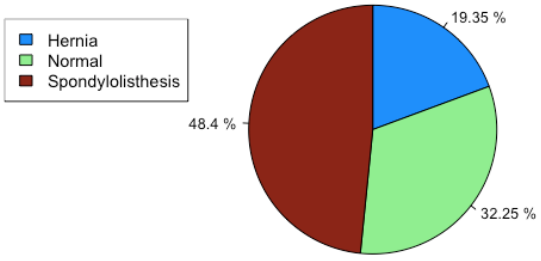

Fig.3: Percentage of variable class.

2.2 Predictors

Pelvic incidence: It is defined as an angle subtended by line $\overline{oa}$, which is drawn from the center of the femoral head to the midpoint of the sacral endplate (line segment $\overline{bc}$) and a line perpendicular to the center of the sacral endplate in Fig. 4.

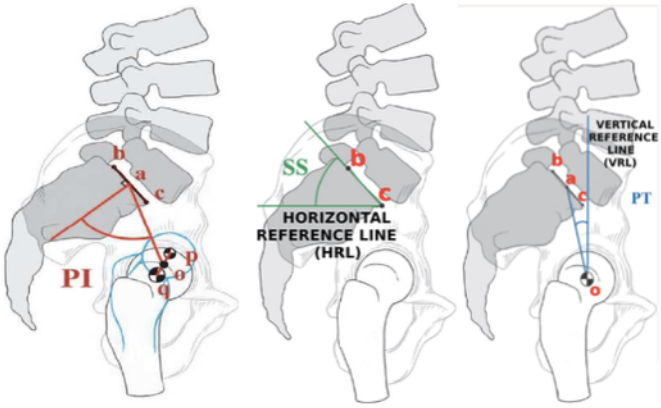

Fig.4: Spino-pelvic system

Pelvic tilt: It is defined as the angle between the vertical reference line (VRL) and the line joining the middle of the sacral endplate and the axis of the femoral heads in Fig. 4. It is positive when the hip axis lies in front of the middle of the sacral endplate.

Lumbar lordosis angle: It is the bigger sagittal angle between the sacrum superior plate and the lumbar vertebra superior plate or thoracic limit.

Sacral slope: It is defined as the angle between the sacral endplate ($\overline{bc}$) and the horizontal reference line (HRL), in Fig. 4.

Pelvic radius: It is the distance from the center of the bicoxofemoral axis to the center of the sacral endplate. It is represented as the segment line $\overline{ao}$ in Fig. 4. It is measured in millimeters.

Degree of spondylolisthesis: It is the percentage level of slipping between the inferior plate of the fifth lumbar vertebra and the sacrum.

In the next table, we can see a summary for all the predictors.

Table 1: Predictors' measures

Predictors	Min	1st Qu.	Median	Mean	3r Qu.	Max
pelvic_incidence	26.15	46.43	58.69	60.50	72.88	129.83
pelvic_tilt	-6.555	10.667	16.358	17.543	22.120	49.432
lumbar_lordosis_angle	14.00	37.00	49.56	51.93	63.00	125.74
sacral_slope	13.37	33.35	42.40	42.95	52.70	121.43
pelvic_radius	70.08	110.71	118.27	117.92	125.47	163.07
degree_spondylolisthesis	-11.058	1.604	11.768	26.297	41.287	418.543

2.3 Correlation between variables

If we want to define a generalize linear model, it is very important to analyse the linear correlation between the variables. The following corrplot -which represents the correlations with ellipses- gives us an idea of which are the linear relations between the predictors and the response variable.

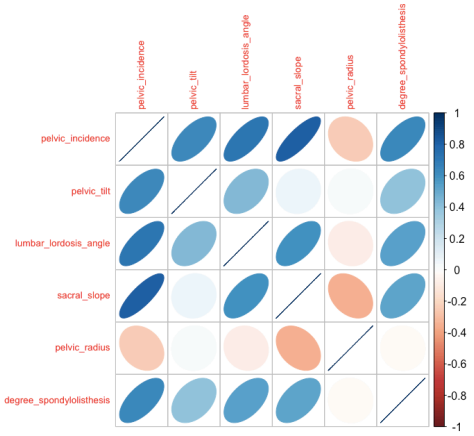

Fig. 4: Correlation between predictors

With the graphic we can see that some variables are highly correlated. This could gives us problems in terms of multicollinearity, due to those variables that share information. We will have to analyse the VIF for our model.

3 Methods

Our purpose is to perform a classification of n individuals attending to a qualitative variable Y with K classes by means of estimating the probabilities of belonging to each of these classes. Given $x_1, ..., x_p$ continuous variables meassured over the individuals, *Multinomial Logistic Regression* is a method that allows us to estimate these probabilities with $K-1$ linear predictors $\eta_k := \beta_{0k} + \beta_{1k}x_1 + \ldots + \beta_{pk}x_p$, for $k = 1, ..., K-1$, and a link function g^{-1}. Let's build the model starting from the linear regression. Consider that Y has only two classes and we want to predict the probability of belonging to the first class linearly:

$$\mathbb{E}[Y|X_1 = x_1, \ldots, X_p = x_p] = \mathbb{P}[Y = 1|X_1 = x_1, \ldots, X_p = x_p] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p =: \eta$$

We can not assure η is going to be between 0 and 1, and therefore if it is a probability. So we can define the following link function that maps each η into that range:

$$g^{-1}(\eta) := \frac{1}{1 + e^{-\eta}}$$

Then if we define odds(Y) such as the cocient between the probability of Y being 1 and being 0,

$$\text{odds}(Y|X_1 = x_1, \ldots, X_p = x_p) := \frac{\mathbb{P}[Y = 1|X_1 = x_1, \ldots, X_p = x_p]}{\mathbb{P}[Y = 0|X_1 = x_1, \ldots, X_p = x_p]} =$$

$$= \frac{\mathbb{P}[Y = 1|X_1 = x_1, \ldots, X_p = x_p]}{1 - \mathbb{P}[Y = 1|X_1 = x_1, \ldots, X_p = x_p]} = e^{\eta} = e^{\beta_0} e^{\beta_1 x_1} \ldots e^{\beta_p x_p},$$

we can easily interpret the effect of the increment in one unit in the variable X_l as the multiplicative increment of e^{β_l} in $\text{odds}(Y|X_1 = x_1, \ldots, X_p = x_p)$.

We can rebuild this idea for a categorical variable Y having K different classes. Now let's predict $K-1$ probabilities linearly. For $k = 1, ..., 1-K$, we have

$$\mathbb{P}[Y = k|X_1 = x_1, \ldots, X_p = x_p] = \beta_{0k} + \beta_{1k}x_1 + \ldots + \beta_{pk}x_p =: \eta_k,$$

but again we need to map these vales into $[0, 1]$ using g^{-1}, so for $k = 1, ..., 1-K$:

$$\mathbb{P}[Y = k|X_1 = x_1, \ldots, X_p = x_p] = g^{-1}(\eta_k)$$

The idea of performing Multinomial Logistic Regression is to do $K-1$ independent logistic regressions for the probability of $Y = k$ versus the probability of the reference $Y = K$. The natural definition of odds is to have odds_k for $k = 1, ..., 1-K$:

$$\text{odds}_\text{k}(Y = k|X_1 = x_1, \ldots, X_p = x_p) := \frac{\mathbb{P}[Y = k|X_1 = x_1, \ldots, X_p = x_p]}{\mathbb{P}[Y = K|X_1 = x_1, \ldots, X_p = x_p]} = e^{\eta_k} = e^{\beta_{0k}+\beta_{1k}x_1+\ldots+\beta_{pk}x_p}$$

Note that $\mathbb{P}[Y = K|X_1 = x_1, \ldots, X_p = x_p] = 1 - \sum_{k=1}^{K-1} \mathbb{P}[Y = k|X_1 = x_1, \ldots, X_p = x_p]$. For $k = 1, ..., 1-K$ we can obtain:

$$\mathbb{P}[Y = k|X_1 = x_1, \ldots, X_p = x_p] = (\mathbb{P}[Y = K|X_1 = x_1, \ldots, X_p = x_p])e^{\beta_{0k}+\beta_{1k}x_1+\ldots+\beta_{pk}x_p}$$

Summing these $K-1$ probabilities:

$$\sum_{l=1}^{K-1} \mathbb{P}[Y = l|X_1 = x_1, \ldots, X_p = x_p] = (\mathbb{P}[Y = K|X_1 = x_1, \ldots, X_p = x_p]) \sum_{l=1}^{K-1} e^{\beta_{0l}+\beta_{1l}x_1+\ldots+\beta_{pl}x_p}$$

Then

$$\mathbb{P}[Y = K|X_1 = x_1, \ldots, X_p = x_p] = 1 - (\mathbb{P}[Y = K|X_1 = x_1, \ldots, X_p = x_p]) \sum_{l=1}^{K-1} e^{\beta_{0l}+\beta_{1l}x_1+\ldots+\beta_{pl}x_p},$$

and:

$$p_K(\mathbf{x}) := \mathbb{P}[Y = K|X_1 = x_1, \ldots, X_p = x_p] = \frac{1}{1 + \sum_{l=1}^{K-1} e^{\beta_{0l}+\beta_{1l}x_1+\ldots+\beta_{pl}x_p}}$$

Substituting in the $K-1$ first probabilitites:

$$p_k(\mathbf{x}) := \mathbb{P}[Y = k|X_1 = x_1, \ldots, X_p = x_p] = \frac{e^{\beta_{0k}+\beta_{1k}x_1+\ldots+\beta_{pk}x_p}}{1 + \sum_{l=1}^{K-1} e^{\beta_{0l}+\beta_{1l}x_1+\ldots+\beta_{pl}x_p}}$$

The natural question now is how to obtain the betas. The multinomial distribution is appropriate for this purpose. We are going to discuss the two-class case, since the algorithms simplify considerably. Let's consider the log-likelihood function:

$$\ell(\beta) = \sum_{i=1}^{n} [Y_i \log p(x_i; \beta) + (1 - Y_i) \log(1 - p(x_i; \beta))] = \sum_{i=1}^{n} \left[Y_i \beta^T x_i + \log(1 + e^{\beta^T x_i})\right]$$

where $p(x_i;\beta) = P(Y = 1|X = x_i;\beta)$. Calculating the derivative and matching it to cero

$$\frac{\delta\ell(\beta)}{\delta\beta} = \sum_{i=1}^{n} x_i(Y_i - p(x_i;\beta)) = 0,$$

which means that we have $p+1$ nonlinear equations. We could use Newton-Rapshon alrogithm, which also requieres the Hessian:

$$\frac{\delta^2\ell(\beta)}{\delta\beta\delta\beta^T} = -\sum_{i=1}^{n} x_i x_i^T p(x_i;\beta)(1 - p(x_i;\beta))$$

Let X the $n \times (p+1)$ matrix of x_i values, Y the vector containing Y_i, p the vector of fitted probabilities and W a $n \times n$ diagonal matrix of weights such that the ith diagonal element is $p(x_i;\beta)(1 - p(x_i;\beta))$. So we can write

$$\frac{\delta\ell(\beta)}{\delta\beta} = X^T(Y - p),$$

$$\frac{\delta^2\ell(\beta)}{\delta\beta\delta\beta^T} = -X^T W X,$$

and then applying the algorithm

$$\beta^{new} = \beta^{old} + (X^T W X)^{-1} XT(y-p) = (X^T W X)^{-1} X^T W(X\beta^{old} + W^{-1}(y-p)) = (X^T W X)^1 X^T W z,$$

being $z = X\beta^{old} + W^{-1}(y - p)$. This z plays a fundamental role, it is known as the adjusted response (see reference 8). $\beta = 0$ is a good starting value for the iterative procedure, nevertheless convergence is never guaranteed.

When building a model is very important to take into account the complexity of the model. We may be focused on understanding a particular problem and how the predictors affect the response variable. There are some criteria that help us to understand if adding more variables to our model worths or not. One of them is **Bayesian information criterion (BIC)**. This criterion considers the number of individuals n, the number of parameters estimated by the model, k, and $\hat{L}$, the maximum value of the likelihood function of the model. So if we take a look at the formula, when the BIC takes low values it means that the complexity is low and the fitness is high, so the model is better.

$$BIC = \underbrace{ln(n)k}_{Complexity} - \underbrace{2ln(\hat{L})}_{Fitness}$$

Now, we can check from a practical approach if a model is good or not using k-fold cross-validation. It performs a random division of the dataset in k folds, and for each of the k folds it builds the desired model with the remaining $k-1$ folds, using the chosen fold for testing the model. K-fold cross-validation can be executed several times if wanted. In the case of multinomial logistic

regression, we can use this technique for testing the sensitivity for each class in each fold and decide how good is our model.

$$Sensitivity = \frac{TruePositives}{TruePositives + FalseNegatives}$$

4 Statistical analysis

In this part, we are going to design a proper model for our dataset using the theory explained in the previous section. This model will be a multinomial logistic regression one, setting as class reference $K = Normal$.

Firstly, we have built a model using all the predictors of the dataset. In order to prove its sensitivity, we performed cross-validation with five folds and five repetitions, the results obtained were the following:

Table 2: Mean sensitivity per class with six predictors.

Class	Normal	Hernia	Spondylolisthesis
Sensitivity	0.836	0.660	0.957

As we can see, both Normal and Spondylolisthesis classes have an elevated sensitivity but for the Hernia class, this rate is a bit lower. This value could be explained by the fact that Hernia class is the less common in the dataset. Since it has a medical aim and we preferred a pessimistic model, we analysed further the results of the classification and computed the mean number of Hernia and Spondylolisthesis patients that were misclassified as healthy patients. We have a mean of 33% of Hernia patients classified as Normal ones, while 3% of Spondylolisthesis patients are classified as Normal patients.

We have examined the correlations for the predictors and we obtained that some of them are highly correlated. Therefore, we have analysed the VIF for this model. These are the results.

Table 3: VIF of all predictors.

pelvic_inc.	pelvic_tilt	lordosis_angle	sacral_slope	pelvic_radius	degree_spond.
1.532+14	1.068+16	1.545e+01	-1.585e+16	4.945e+01	1.464

Because of the fact that we have a very high VIF for the majority of the variables, we decided to build new models reducing the number of predictors, until we found a set of attributes where their VIFs were less than 10. The model with the pelvic tilt and degree of spondylolisthesis variables obtained the smallest VIFs, 7.3 and 1.40 respectively.

Moreover, we are going to test this model with the cross-validation method, with the same parameters as before. The results are shown below.

Table 4: Mean sensitivity per class with two predictors.

Class	Normal	Hernia	Spondylolisthesis
Sensitivity	0.838	0.303	0.967

The table shows that the sensitivity for the Hernia class has dropped drastically. Again, we computed the mean number of Hernia and Spondylolisthesis patients that were misclassified as healthy patients. We have a mean of 67% of Hernia patients classified as Normal ones, while 2% of Spondylolisthesis patients are classified as Normal patients.

Even though this model has not multicollinearity, the sensitivity is worse. This is something curious regarding to the theory. The prediction results are not addecuate for a model whose purpose is to identify and classify medical pathologies, it is inadmissible that the model classifies most of the patients with Hernia as healthy ones. Therefore, we applied a step-AIC to the model with all predictors in order to find the relevant attributes that can perform a desirable prediction.

The search of the suitable predictors have been computed choosing the BIC option in the step-AIC function. It returned the variables pelvic incidence, pelvic tilt, pelvic radius and degree of spondylolisthesis as the optimal attributes.

Table 5: VIF of step-AIC predictors.

pelvic_incidence	pelvic_tilt	pelvic_radius	degree_spondylolisthesis
63.18	15.29	261.93	1.79

Although the VIF of the majority of the variables is quite large, we performed cross-validation for this model with the same parameters as before.

Table 6: Mean sensitivity per class with step-AIC predictors.

Class	Normal	Hernia	Spondylolisthesis
Sensitivity	0.846	0.650	0.961

This model obtained better sensitivity for Hernia class than the model with only two predictors. Once again, we computed the mean number of Hernia and Spondylolisthesis patients that were misclassified as healthy patients. We have a mean of 34% of Hernia patients classified as Normal ones, while 3% of Spondylolisthesis patients are classified as Normal patients.

After all the analysis, we decided that we are going to use this model for these main reasons:

1. The prediction sensitivity is overall high. It classifies pretty well both pathologies, specially Spondylolisthesis, while also obtaining proper results for the Normal class.

2. It has less attributes than the first model, this can be a critical fact in terms of computational time if the size of the dataset grows significantly. Moreover, as we explained in the

introduction, one of the problems for the use of computational models in this medical field is the absence of numerical attributes that describe the pathologies. Thus, a fewer number of predictors would ease the task of obtaining quantitative variables.

3. The variables have smaller VIF than the first model (even though it is still elevated).

Additionally, we present an interpretation of the coefficients of this model. We are going to analyse the values of e to the coefficients for the attributes individually, keeping the others constant. These are the odds expressions:

$$\frac{p_{k=1}(x)}{p_{K=3}(x)} = e^{20.85}e^{-0.19X_1}e^{0.27X_2}e^{-0.14X_3}e^{0.0002X_4}$$

$$\frac{p_{k=2}(x)}{p_{K=3}(x)} = e^{-0.52}e^{0.04X_1}e^{0.05X_2}e^{-0.06X_3}e^{0.31X_4}$$

Notice that k takes values 1 and 2 which represents Hernia and Spondylolisthesis classes respectively; and K takes value 3, representing the Normal class. The variables X_1, X_2, X_3 and X_4 represent the predictors pelvic incidence, pelvic tilt, pelvic radius and degree of spondylolisthesis.

We have obtained the following conclusions for the odds expression of the **Hernia** pathology:

- If we increase in one unit the **pelvic incidence** (X_1), the probability of having Hernia **decreases 17%** regarding to having a normal column.
- If we increase in one unit the **pelvic tilt** (X_2), the probability of having Hernia **increases 31%** regarding to having a normal column.
- If we increase in one unit the **pelvic radius** (X_3), the probability of having Hernia **decreases 13%** regarding to having a normal column.
- If we increase in one unit the **degree of spondylolisthesis** (X_4), the probability of having Hernia **increases 0.02%** regarding to having a normal column.

We have obtained the following conclusions for the odds expression of the **Spondylolisthesis** pathology:

- If we increase in one unit the **pelvic incidence** (X_1), the probability of having Spondylolisthesis **increases 4%** regarding to having a normal column.
- If we increase in one unit the **pelvic tilt** (X_2), the probability of having Spondylolisthesis **increases 5%** regarding to having a normal column.
- If we increase in one unit the **pelvic radius** (X_3), the probability of having Spondylolisthesis **decreases 6%** regarding to having a normal column.
- If we increase in one unit the **degree of spondylolisthesis** (X_4), the probability of having Spondylolisthesis **increases 36%** regarding to having a normal column.

If we wanted to know the variable that has the greatest impact we should standardize first, but as the results would be shown to medicine professionals and also two of the predictors are angles with a similar range, we decided to maintain the original values due to the fact that they may want to know the increments of the odds in terms of the original variables.

To sum up, without taking into account the standardized values, Hernia is mostly influenced by the pelvic incidence and the pelvic tilt (with an inverse and positive relation respectively), while Spondylolisthesis has mostly a positive relation with the degree of Spondylolisthesis.

5 Conclusions

- The final model is a multinomial logistic regression considering the Normal class as reference. The attributtes selection was obtained with the function step-AIC. The chosen predictors were pelvic incidence, pelvic tilt, pelvic radius and degree of spondylolisthesis.
- The reasons why we chose this model were that the precision in the predictions obtained with k-fold cross-validation were pretty reasonable, it has less attributes and a smaller VIF than the original model with all the predictors.
- The interpretation of the betas of the model could help medicine professionals to identify how the probability of having one of these two pathologies is been influenced by the biomechanical measures of the patients.

6 References

1. UCI Machine Learning. Biomechanical features of orthopedic patients. Recovered from: https://www.kaggle.com/uciml/biomechanical-features-of-orthopedic-patients
2. Ajalmar R. da Rocha Neto, Ricardo Sousa, Guilherme de A. Barreto and Jaime S. Cardoso. Diagnostic of Pathology on the Vertebral Column with Embedded Reject Option. Recovered from: https://ai2-s2-pdfs.s3.amazonaws.com/a972/7a403fc0b6c9324be101295fd6a42577f3c2.pdf
3. MedlinePlus. Trusted Health Information for You. Recovered from: https://medlineplus.gov
4. A. R. Rocha Neto and G. A. Barreto.(2009).On the Application of Ensembles of Classifiers to the Diagnosis of Pathologies of the Vertebral Column: A Comparative Analysis. Recovered from: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5349049&tag=1
5. Wikipedia. Orthotics. Recovered from: https://en.wikipedia.org/wiki/Orthotics
6. Wikipedia. Human musculoskeletal system. Recovered from: https://en.wikipedia.org/wiki/Human_musculoskeletal_system
7. Wikipedia. Vertebral column. Recovered from: https://en.wikipedia.org/wiki/Vertebral_column
8. Trevor Hastie, Robert Tibshirani, Jerome Friedman. The elements of statistical learning: Data mining, inference, and prediction. Springer Series in Statistics (2009).