Machine Learning Model to predict allocation of patients with chronic back pain for integrated practice units in a system of value based health care.

Authors:

Vitor Pereira Barbosa[1]
João Lucas Maehara Said dos Reis [2]
Natália Neto Pereira Cerize [3]
Vinicius Monteiro de Paula Guirado [4]

[1] Institute for Technological Research (IPT) – Laboratory of Chemical Processes and Particle Technology, Group for Bionanomanufacturing (BIONANO). Address: Av. Prof. Almeida Prado, 532 - Butantã, São Paulo - SP, Brazil – zip code 05508-901. Phone +55 11 3767 4211. E-mail: vitorpbarbosa@ipt.br

[2] Institute for Technological Research (IPT) – Laboratory of Chemical Processes and Particle Technology, Group for Bionanomanufacturing (BIONANO). Address: Av. Prof. Almeida Prado, 532 - Butantã, São Paulo - SP, Brazil – zip code 05508-901. Phone +55 11 3767 4581. E-mail: joaomaehara@ipt.br

[3] Institute for Technological Research (IPT) – Laboratory of Chemical Processes and Particle Technology, Group for Bionanomanufacturing (BIONANO). Address: Av. Prof. Almeida Prado, 532 - Butantã, São Paulo - SP, Brazil – zip code 05508-901. Phone +55 11 3767 4682. E-mail: ncerize@ipt.br

[4] University of São Paulo Faculty of Medicine Clinics Hospital – Division of Neurosurgery. Address: Av. Dr. Enéas de Carvalho Aguiar, 255 Cerqueira César, São Paulo – SP, Brazil – zip code: 05403-001.
Phone +55 11 3069 6000 (7152). E-mail: PREENCHER

Abstract

Chronic back pain is responsible for a great part of the global costs of healthcare, due to the necessity of long term specialized care . The immense variability of patients in its therapeutical itineraries between hospitals and healthcare providers turns the value based healthcare into a potentially less viable option. To more effectively advance in the management and the efforts of secondary, tertiary and quaternary prevention, the adoption of data science tools, methods and techniques , such as supervised and unsupervised machine learning algorithms is necessary, since these models have been proved trustworthy in the forecast of specific results in some neurological illnesses. However, the development of management strategies for the integrated practice units has not been tested with the support of those techniques. In this study, data from 6 different questionnaires were used for the evaluation of pain conditions in patients. Correlation techniques between ordinal and nominal variables were applied for determination of which questions present greater correlation with the low back pain, back pain and leg pain. Furthermore, 5 different machine learning algorithms were applied to predict the absence or presence of pain in the low back region in the patients. As a result, 7 variables were selected to use as input for the prediction models, of which the one that presented better accuracy was *XBoost Classifier*, with accuracy, precision and recall of 0.8 and F1-Score of 0.78. This work allows the reduction in the number of questions currently applied and assists in the screening process based on the supplied answers of each patient.This machine learning model could help the decision maker to predict allocation of patients with chronic back pain for integrated practice units in a system of value based health care.

Keywords: Chronic Low Back Pain, Value Based Healthcare, Machine Learning

1 Introduction

One of the primordial questions proposed by Michael E. Porter in the book "*Redefining Health Care*" was the reason why the competitive model of management to be failing in the healthcare system. The author argues, for example, that throughout history in the economy, the competition in the private market is one of the biggest forces for the improvement of quality and costs reduction in goods and services. However, this was not observed in the case of the health sector, in which the competition was only acting on the costs, which exclusively grew and the quality of the given services did not necessarily improve [1].

The necessity of a value based healthcare system arose from the extreme and unsustainable costs of the current practiced system [2]. This analysis was initially carried through the American healthcare system, which presented a fundamental paradox related to the increase in the biomedicine knowledge, which was protagonist in innovations in therapies and surgical procedures and the treatment of conditions which were previously fatal, however, this system started to present problems in basic questions related to quality, results achieved for the patients and the costs [3].

In 2006, the *National Academy of Medicine* established the basis for an evidence based medicine with the intention to provide a reliable base for the national leaders in healthcare, in order to allow the generation of a system that can generate real value for the patients and the society. The purpose to advance up to one "*Learning Health System*" quickly emerged and was defined as a system where science, computer science, incentives and culture are lined up for improvement and continuous innovation [4].

Considering global scenario, the problems related to the increase of costs is also cited in a report elaborated by Deloitte in 2019, which pointed that the global expenditure in healthcare was expected to grow to an annual tax of 5,4% between 2018-2022, compared with an increase of 2,9% between 2013-2017 [5]. This estimate was based mainly on the strengthening of the dollar when compared to the euro and other currencies; to the expansion of the coverage of medical assistance in the developing countries; to the aging population; to the sprouting of new treatments and technologies in healthcare and to the increase of labor costs in the health sector [6].

In the specific field of neurological surgery, diagnostic and therapeutical options are available in a high complex scenario at the beginning of the 21st century. Likewise, the range of possible results is varied because many dimensions of interpretation exist. Still influenced by the social context, patients and doctors are overwhelmed with information from the digital age and, thus, the decision making process today is cardinal and critical. The contemporary resource for this challenging demand is the application of information management technologies, such as artificial intelligence.

Recent advances in artificial intelligence (AI) are creating new opportunities to personalize technology-based health interventions for patients with chronic pain. Tools available in the AI field - intelligent learning environments, interaction narrative generation, user modeling and adaptative training - can be used to model the learning and the involvement of patients with chronic pain and provide personalized support in adaptative health technologies. Many of these technologies have emerged from applications centered on human activities for education, training and entertainment. However, its application in health improvement, so far, has been comparatively limited.

An example of a study that makes use of statistical techniques and tools was developed by Depintor et al. (2016), in which the prevalence of chronic spinal pain was estimated in individuals with 15 or more years old and tried to identify associated factors. This work made use of Cox Regression (or Proportional Risks model). For bivariate analysis, statistical associations were determined through the Log-Rank test. For ordinal variables the chi-square test was used to find trends and the analyses were performed using the STATA 13.0 software.

This research interviewed 826 participants and the result indicated that the prevalence of chronical vertebral pain was estimated at 22 % with a confidence interval, at the significance level of 5 %, 19.3 % - 25 %. The factors associated with chronic vertebral pain were: female, 30 years old or more, four years or less of schooling, symptoms compatible with anxiety and intense physical effort during the main occupation [7].

Chronic spinal algias are part of the category of pain classified as chronic pain, which affect approximately 20 % of the world population. Primary chronic pain is defined as a pain that persists for more than 3 months and has a significant impact in the emotional welfare, being a strong cause of distress, demoralization and functional disability in the patients, which makes it one of the main sources of suffering [8].

Currently, there are two well-established pain classification systems, named *STarT Back* and *McKenzie*. The *McKenzie* method makes use of the patient`s symptom history and pain presented after conducting certain movements and classifies them in 3 different groups according to their syndrome. The *STarT Back* classifies the patients as high, medium and low risk of developing persistent symptoms that disable them, based on physical and psychosocial factors. These two methods are examples of approaches that do not consider only the anatomical basis to perform the diagnosis [9].

A literature review of the application of Machine Learning (ML) algorithms for back and lower back pain was carried out by Tagliaferri et al. (2020), in which 48 articles were selected to evaluate the approaches and compared them with the methods already established as the STarT Back and McKenzie. From the 48 selected articles, 45 used samples smaller than 1000, 19 used less than 5 parameters in the final model, 13 applied multiples models and had achieved high accuracy and 25 evaluated low back pain through binary classification (patient presents or does not presents the pain) [10].

From the presented studies and the presented scenario, simplifying the complexity is not a solution and, therefore, science applied to data, known as human knowledge added to digital technologies, is the best alternative available for the decision-making process in the field of neurological surgery. Discussing data-based options is a more important decision that an incision and offering the opportunity to use digital technologies to facilitate and support the clinical routine and patient management.

We then here illustrate the opportunities provided by AI-driven adaptive technologies for preventive healthcare for patients with chronic pain, describing a vision of how future preventive health interventions for this large group of patients can be carried out inside and outside of the specialized clinic.

2. Objective

The main purpose within this study is to, through the use of exploratory data analysis, pre-processing and machine learning models on the available database, select the main characteristics which are most correlated with the presence of chronic low back pain, back pain and leg pain.

3 Materials and Methods

Data collection

The available data are the results of clinical evaluation in a multidisciplinary integrated care unit specialized in chronic pain from February to December 2019. The initial available sample included 240 patients undergoing clinical evaluation. The data from the questionnaires are of a socio-demographic and clinical nature, which are listed below:

1 - Basic clinical assessment of spine symptoms (Gothenburg Protocol)
2 - Brief Pain Inventory (BPI)
3 - Oswestry 2.0 Disability Index
4 - Roland Morris Disability Questionnaire (RMDQ)
5 - Questionnaire to assess quality of life 12-Item Short Form Health Survey (SF-12)
6 - Questionnaire for Diagnosis of Neuropathic Pain 4 (DN-4)

Data pre-processing

The questions available in the questionnaires were divided into two categories, nominal (all converted to binary format) and ordinal. This approach allows the observation of the correlation between the binary variables to be performed through a contingency table and the correlation between binary variables and ordinal variables to be measure with the Point-Biserial Correlation Coefficient.

The initial database is a result of a six questionnaires junction and composed of 118 variables (also mentioned as characteristics or column vectors). Regarding the variables which are presented as nominal, the dummy encoding technique was used to generate new columns.
As a consequence of this first pre-processing measure, 146 column vectors were obtained, among which, 101 are binary and 45 are ordinal.

For the application of the machine learning model for patient classification regarding the presence or absence of low back pain, patients who did not fill this question, corresponding to pain of number 30 in the Brief Pain Inventory questionnaire, were removed from the dataset. As a result, there were 138 patients in the final database. This dataset can, therefore, be submitted to correlation analysis, graph construction and used as input for machine learning algorithms.

It is known that with this amount of column vectors (146), and only 138 patients, there is a case of sparse data, in which the phenomenon of *Curse of Dimensionality*, introduced by Bellman, (1957). The sparse data becomes a problem to obtain results with statistical significance in a machine learning model, because the number of observations should grow exponentially with the dimension (number of variables or factors) [12].

To reduce the number of predictor variables, we used the results derived from the analysis of correlation between binary variables, and also the correlation between binary variables and ordinal variables, in order to select those that have the highest absolute values.

Additionally, as a tool which allows to see the groups division, the PCA analysis with 2 components was used to visualize the data belonging to the group of people who present low back pain and those who answered that do not present this pain.

Correlations and statistics model validation

In this section, the methods for correlating binary variables, and also for correlating binary and ordinal variables are described, which were used for variable selection (also named as feature selection).

Correlation between binary variables

To evaluate the correlation between binary variables, a contingency table was used, with its structure presented in Table 1 :

Table 1 - Example of cross table

		Variable B	
		0 (not)	1 (yes)
Variable A	0 (not)	*a*	*b*
	1 (yes)	*c*	*d*

In Table 1 there is an example of a contingency table, in which a, b, c and d are integers. This table structure makes it possible to assess whether the change in proportion between no / yes answers of variable A is correlated with variable B

To assess the statistical relevance of the results, the chi-square hypothesis test was applied:

H0 Independence - A does not depend on B

H1 Dependence - A depends on B

The chi-square calculation is done using the following expression:

$$\chi^2 = \frac{(ad - bc) * N}{n_1 * n_2 * n_3 * n_4}$$

In which:

a, b, c and d, are the observation counts presented in Table 1

$n_1 : a + b$

$n_2 : c + d$

$n_3 : a + c$

$n_4 : b + d$

$N : n_1 + n_2 + n_3 + n_4$

This hypothesis test can be interpreted as the difference in the frequency distribution of variable A due to the presence of variable B. In such manner, when the calculated chi-square value is greater than or equal to the tabulated chi-square, when adopting level of significance of 5%, one can reject the null hypothesis of independence and consider the alternative hypothesis of dependence on variable A in relation to variable B [13].

Correlation between binary and ordinal questions

The correlation between questions that appear to be dichotomous and questions that present on an ordinal scale, was performed through the Point-Biserial Correlation Coefficient.

The division of a group into (0 - no) and (1 - yes) is considered and, inasmuch as , the coefficient can be calculated according to the following expression:

$$r_{pb} = \frac{(M_1 - M_0)}{s_n} * \sqrt{\frac{n_1 n_0}{n^2}}$$

In which:

s_n : Standard deviation considering all population data

M_0 and M_1: Average of the ordinal variables, for patients who answered the question as negative (0 - no) and positive (1 - yes), respectively

n_0 and n_1: Number of people belonging to each group (0 - no) and (1 - yes)

The evaluation of the statistical significance of this coefficient is performed by the hypothesis test of Pearson's correlation, on the assumption that the biserial point correlation is a specific case of it for one of the variables being dichotomous. The hypothesis are stated as follows:

H0: $r_{pb} = 0$

H1:$r_{pb} \neq 0$

The test makes use the t of *student* distribution, so that the calculated *t value* is given by:

$$t = \frac{r_{pb} * \sqrt{n-2}}{\sqrt{1 - r_{pb}^2}}$$

In which:

r_{pb}: Bisserial Point Correlation Coefficient

n: Number of observations

According to the adopted statistical criteria, if the value of calculate t is higher than the value of t tabulated when adopting a significance level of 5 %, the null hypothesis can be rejected [14].

Thus, in this paper, only the correlations that passed this test were presented.

Sampling

The sampling technique used was that of K stratified folds (*StratifiedKFold*), with 10 subdivisions (K = 10), in order to maintain the percentage of the original sample base for each class.

Supervised machine learning model

The performances of the following supervised machine learning models for classification were evaluated:

• Logistic regression with l1 and l2 regularization (*elasticnet*), with regularization factor l1 of 0.8. The optimization algorithm used to obtain the coefficients was lbfgs (*Limited-memory Broyden – Fletcher – Goldfarb – Shanno*).

• Neural network with 1 hidden layer composed of 5 neurons with the *ReLu* (rectified linear unit) activation function and on the output layer with the sigmoid activation function. The network was trained through discrete sample sizes of 5 observations for 200 epochs.

• *Random Forest:* 50 decision trees were used, with an average depth of 8 levels. The criterion chosen to perform the split was *gini impurity*.

• SVM (*Suppor Vector Machine*): The Radial Base Function kernel, through the scalar product tools and the expansion of the Taylor series, was used to obtain a relationship between the observations in an infinite dimension. The probabilities generated by this model were calibrated using the Platt scaling calibration method, which applies the logistic regression function over the original generated probabilities, as detailed in [15].

• XBGClassifier: Implementation of the *Gradient Boosting* method with greater speed and design. It makes use of the so-called *weak learners*, decision trees with only one node and two leaves, so that through the method of *ensemble* (joining of several models) be able to obtain at the end an optimized and robust model to be used in classification in this study [16].

Model evaluation metrics

In order to evaluate the models performance on the correct classification of patients regarding the presence of chronic low back pain, the metrics Accuracy, *Precision*, *Recall* and *F1-Score* were employed within the process of cross-validation.

All of these metrics can be calculated from the confusion matrix, with its structure presented in Table 2 .

Table 2 - Example of confusion matrix

		Real	
		0 (not)	1 (yes)
Prediction	0 (not)	TP	FN
	1 (yes)	FP	TP

Based on what was exposed in Table 2 the following metrics are defined:

$$Acurácia = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

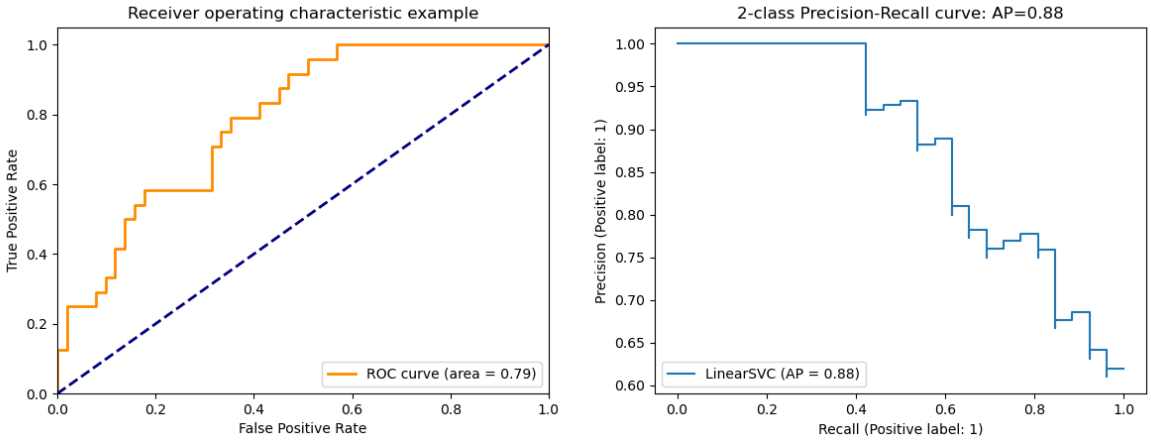
$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

The metric *Precision* calculates the ratio between the number of correctly predicted positives and all predicted positives. However, the Recall metric is particularly more important considering this analysis's focus, given that the calculation consists of the ratio between the total value of True Positives and the sum of True Positives and False Negatives [17].

The reason for using it is that a machine learning model for predicting the occurrence or not of chronic low back pain in patients must have a low amount of False Negatives and consequently a high *Recall* value. In addition, the F1-Score metric allows to obtain a harmonic average between *Precision* and *Recall* [18].

The evaluation of the model was also carried out using curves ROC curve and AUC (*Area Under Curve*) Figure 1 (a) and *Precision-Recall curve* Figure 1 (b).

Figure 1 - Example of ROC curves and *Precision-Recall*



Reference: [19]

The ROC Curve allows us to observe how the distribution between the True Positive Ratio (TPR) and the False Positive Ratio (FPR) occurs, for different *thresholds*, which is defined between 0 and 1.

The calculation of both is defined by:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

The same principle is present in the curve *Precision* x *Recall* curve.

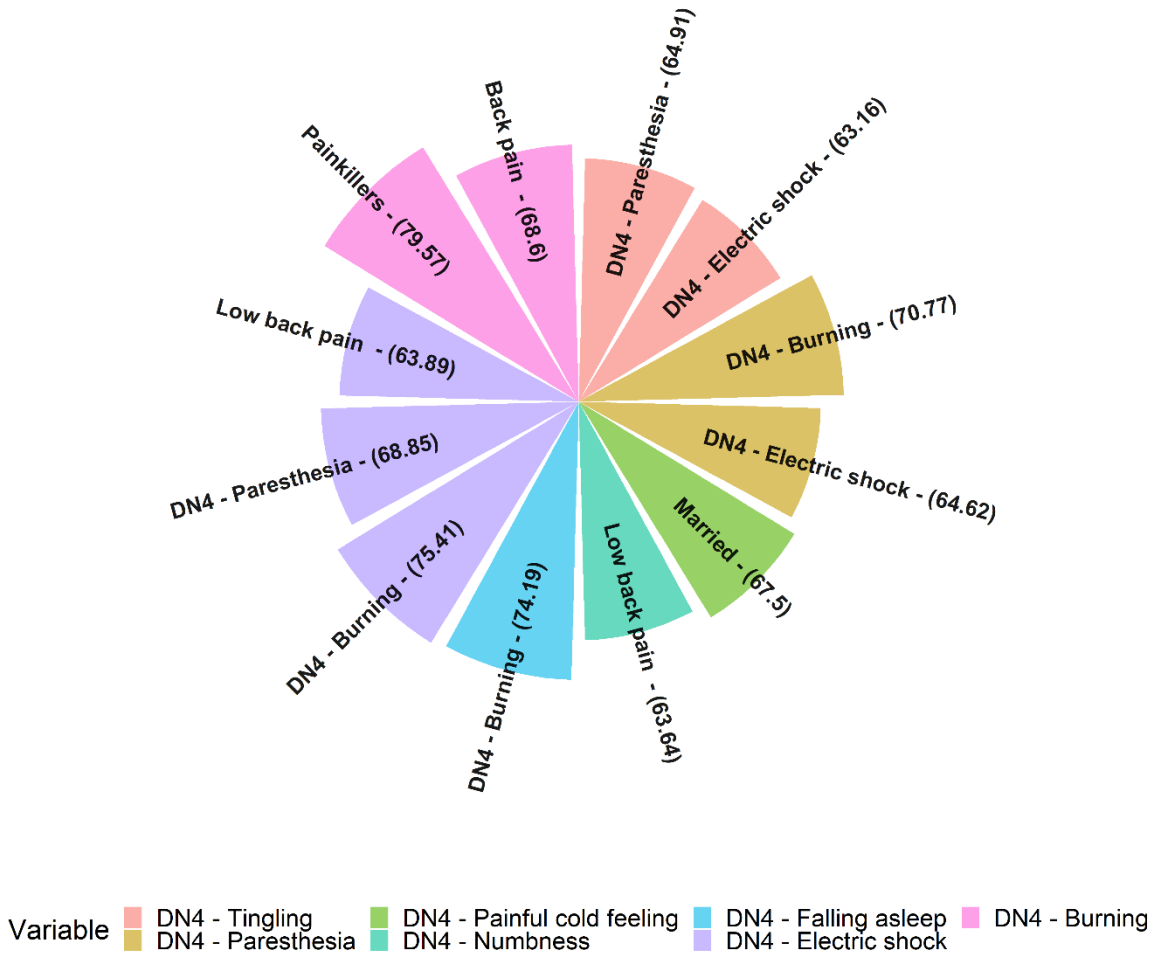
4 Results and discussion

Correlations

Main correlations of the variables in the Questionnaire for Diagnosis of Neuropathic Pain 4 (DN4)

Figure 2 presents the main correlations with the variables available in the Questionnaire for Diagnosis of Neuropathic Pain 4.

Figure 2 - Main correlations of the variables in the Questionnaire for Diagnosis of Neuropathic Pain 4 (DN4)



It can be seen from Figure 2 that 79.19% of the patients who indicated to feel the symptom of numbness in the region where they feel pain, also noted that the pain has the characteristic of burning. It is also noteworthy that 68.6% of people who said they had the burning symptom in the region where they feel pain, marked the option of feeling back pain.

Figure 3 - Main correlations between the questionnaire DN4, about whether or not the patient presents parenthesia - feeling “pins and needles” - at the pain’s region

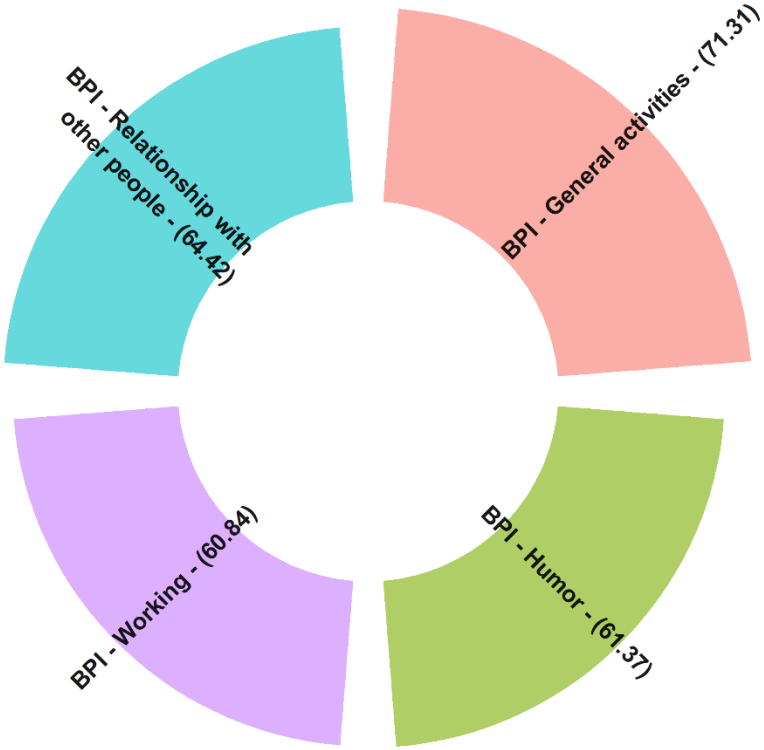


Figure 3 presents which were the questions with ordinal variables which have the highest correlation with the binary question variables, whether or not the patient presents paresthesia, that is, feeling “pins and needles” at the pain’s region. Through Figure 4 there is an example of why the correlation has the value of 64.42 for the question BPI - Relationship with other people.

Figure 4 - *Boxplot* graph to visualize the distribution of data in question 2-b of the DN4 questionnaire with the interference of pain in the relationship with other people

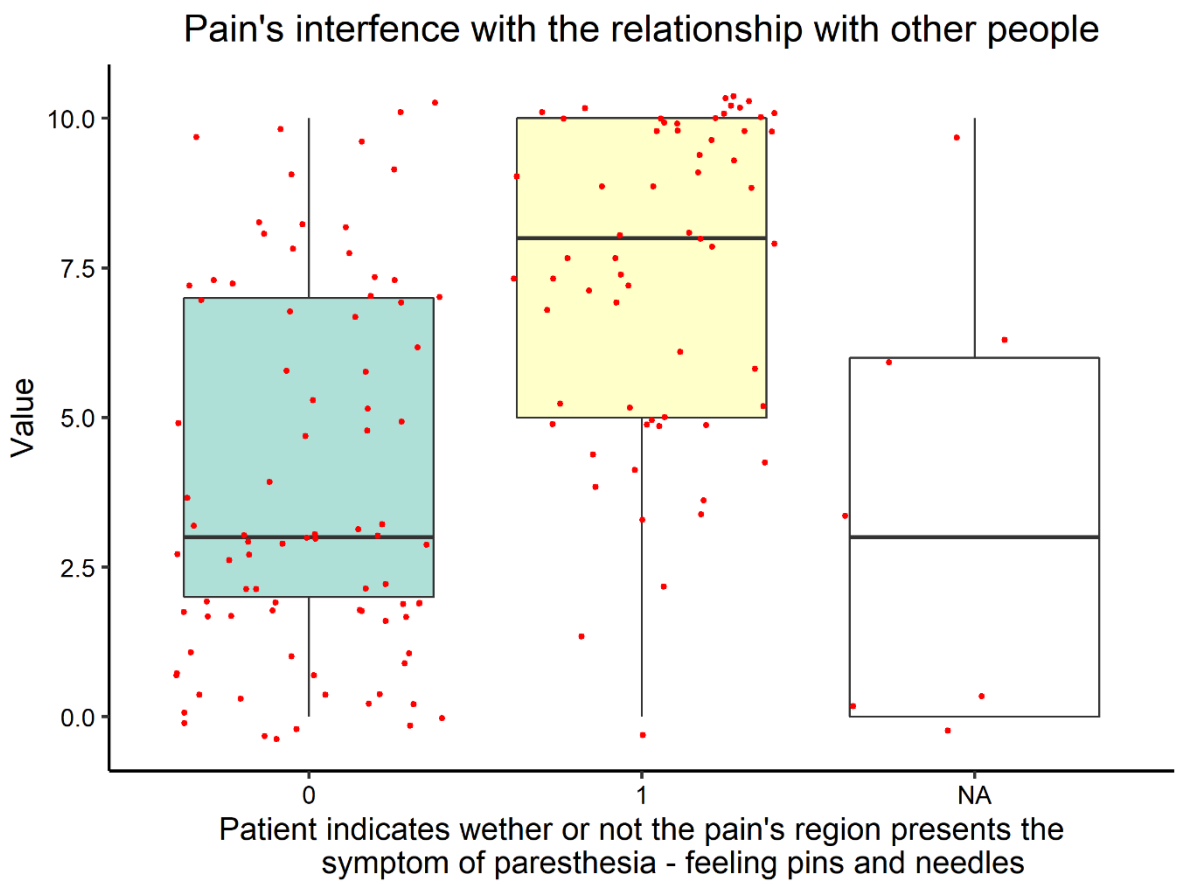


Figure 4 shows that the patients who reported having paresthesia at the pain's region also answered higher values for the interference of this pain in the relationship with other people, as observed in the higher density of points, and in yellow boxplot, corresponding to option 1 of the binary question.

It is prominent that only this question (2 - b - pins and needles) from the Questionnaire for the diagnosis of Neuropathic Pain presented a number greater than 30 people who answered yes or no to the question. The other questions in the DN4 questionnaire did not present statistical significance for the Point Biserial Correlation Coefficient or presented unbalanced data between yes or no, that is, a much larger proportion of people answered the yes or no option when compared to the other alternative.

Main correlations between the variables low back pain, leg pain and back pain with the questions presented in the questionnaires Oswestry 2.0 Disability and Quality of Life Index SF-12

Variables of ordinal character

Table 3 contains the main ordinal variables correlated with low back pain from the Oswestry 2.0 disability and quality of life questionnaire SF-12.

Table 3 - Ordinal variables of the Oswestry 2.0 disability and quality of life questionnaires SF-12 most strongly correlated with low back pain

Ordinal	Correlation coefficient	P-value
SF12_M1_1_v1	0.43	6·63E-05
Oswestry1_v1	0·38	5·46E-04
Oswestry2_v1	0·37	6·07E-04
SF12_M1_8_v1	0·35	1·25E-03
Oswestry9_v1	0·32	3·33E-03
Oswestry4_v1	0·31	4·13E-03
Sum_Oswestry	0·31	4·78E-03
Oswestry10_v1	0·31	5·08E-03
SF12_M1_10_v1	-0·30	5·92E-03
SF12_M1_12_v1	-0·36	1·07E-03
SF12_M1_11_v1	-0·37	7·40E-04
SF12_M1_9_v1	-0·38	5·77E-04
SF12_M1_2_v1	-0·40	1·97E-04
SF12_M1_5_v1	-0·46	1·53E-05
SF12_M1_6_v1	-0·47	1·07E-05
SF12_M1_3_v1	-0·48	9·29E-06
SF12_M1_7_v1	-0·49	4·68E-06
SF12_M1_4_v1	-0·51	1·59E-06

It should be noted that the first question in the SF-12 Quality of Life questionnaire, coded as SF12_M1_1_, presented the highest Bisserial Point Correlation Coefficient, with it's value of 0·43. The questions is written in order to enable the patient to inform in a general way, what is his life quality, on a scale that goes from: "weak", "reasonable", "good", "very good" and "excellent", listed such as 5,4,3,2 and 1, respectively. Thus, it indicates that patients who reported having the condition of low back pain are also more likely to answer this question with the option of life quality closer to "reasonable" or "weak".

The variable that presented the highest value in module was the one correlated to the question coded as SF12-M1_4_v1 in the value of -0.51, which asks the patient, in the last 4 weeks, they ended up doing less than they wanted as a result of their physical state in daily activities at work. The alternatives consists of a scale that goes from "always", "most of the time", "some time", "little time" and "never", listed as 1,2,3,4 and 5 respectively. This result indicates that patients with low back pain are more likely to indicate that "always" or "most of the time" performed less than they wanted at work or other regular daily activities, as a result of their physical condition.

Furthermore, the correlation regarding the low back pain condition with the question SF12_M1_3_v1, presented a value of -0·48, which indicates that the patient's current health limits him in the activity of climbing several stairways and also with the question SF12_M1_6_v1 (correlation value of -0·47), regarding how much the patient had emotional problems that led him to perform less than he wanted in regular daily activities.

It is also complemented that the question coded as Oswestry1_v1 measures the intensity of the patient's pain at the moment, on a scale ranging from 1 to 6. This variable had a correlation coefficient of 0.38, which indicates that patients with low back pain are more likely to have more severe pain.

Table 4 brings the two issues most correlated with the issue of leg pain.

Table 4 - Ordinal variables from the RMDQ, Oswestry and SF-12 questionnaires most strongly correlated with leg pain

Ordinal	Correlation coefficient	P-value
Oswestry1_v1	0·30	8·79E-04
SF12_M1_3_v1	-0·30	8·77E-04

The question coded as Oswestry1_v1 asks how much pain the patient currently feels. Thus, patients with leg pain are also more likely to have a positive result for this issue. In addition to that, the question available in the SF-12 Quality of Life questionnaire, coded as SF12-M1-3-v1, related to how much the patient's health limits him in climbing several flights of stairs, is correlated with the presence of leg pain.

Table 5 shows which issues are most correlated with the general presence of back pain in patients.

Table 5 - Ordinal variables from the SF-12 Quality of life questionnaire most strongly correlated with back pain

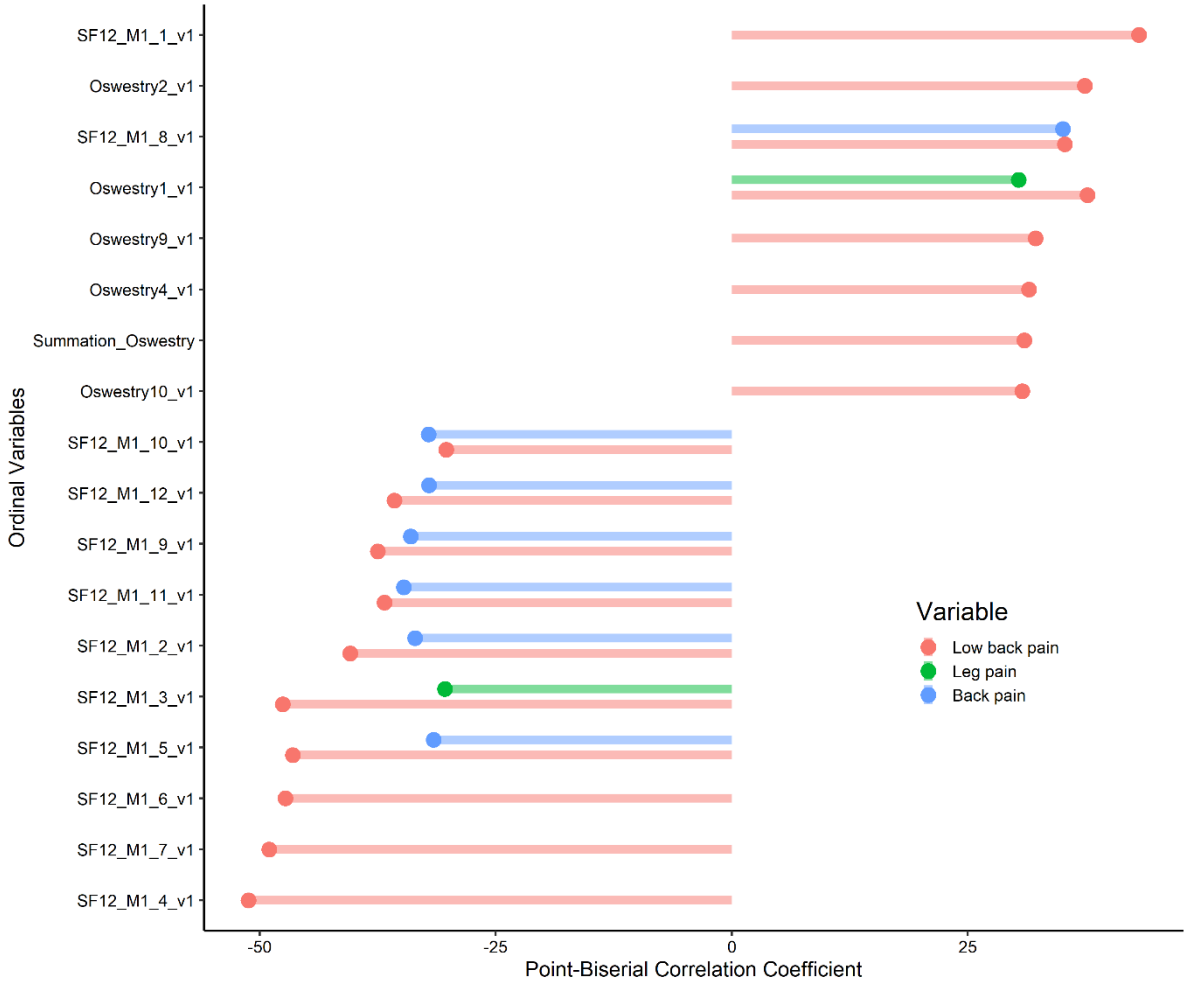
Ordinal	Correlation coefficient	P-value
SF12_M1_2_v1	-0·34	1·55E-04
SF12_M1_5_v1	-0·32	3·84E-04
SF12_M1_8_v1	0·35	7·57E-05
SF12_M1_9_v1	-0·34	1·25E-04
SF12_M1_10_v1	-0·32	3·03E-04
SF12_M1_11_v1	-0·35	8·81E-05
SF12_M1_12_v1	-0·32	3·05E-04

It appears that the question from the Quality of life SF-12 questionnaire coded as SF12_M1_8_V1, which asks how the pain interfered with the patient's normal work in the last 4 weeks, on a scale of: “absolutely nothing”, “a little”, “moderately”,“a lot ”and“ immensely ”, listed as 1,2,3,4 and 5 respectively, are correlated with the presence or absence of back pain in patients. This way, those who answered that they have back pain are more likely to indicate that the pain interfered “a lot” or “immensely” in their normal work in the last 4 weeks.

The question coded as SF12_M1_11_v1 raises the question of how long, during the last 4 weeks, the patient felt sad or depressed with the following answers: “always”, “most of the time”, “some time”, “little time” and “never ”, listed as 1,2,3,4 and 5. Hence, given the negative correlation coefficient -0.35, it is concluded that patients who have back pain are more likely to answer this question with the options "always" and "most of the time".

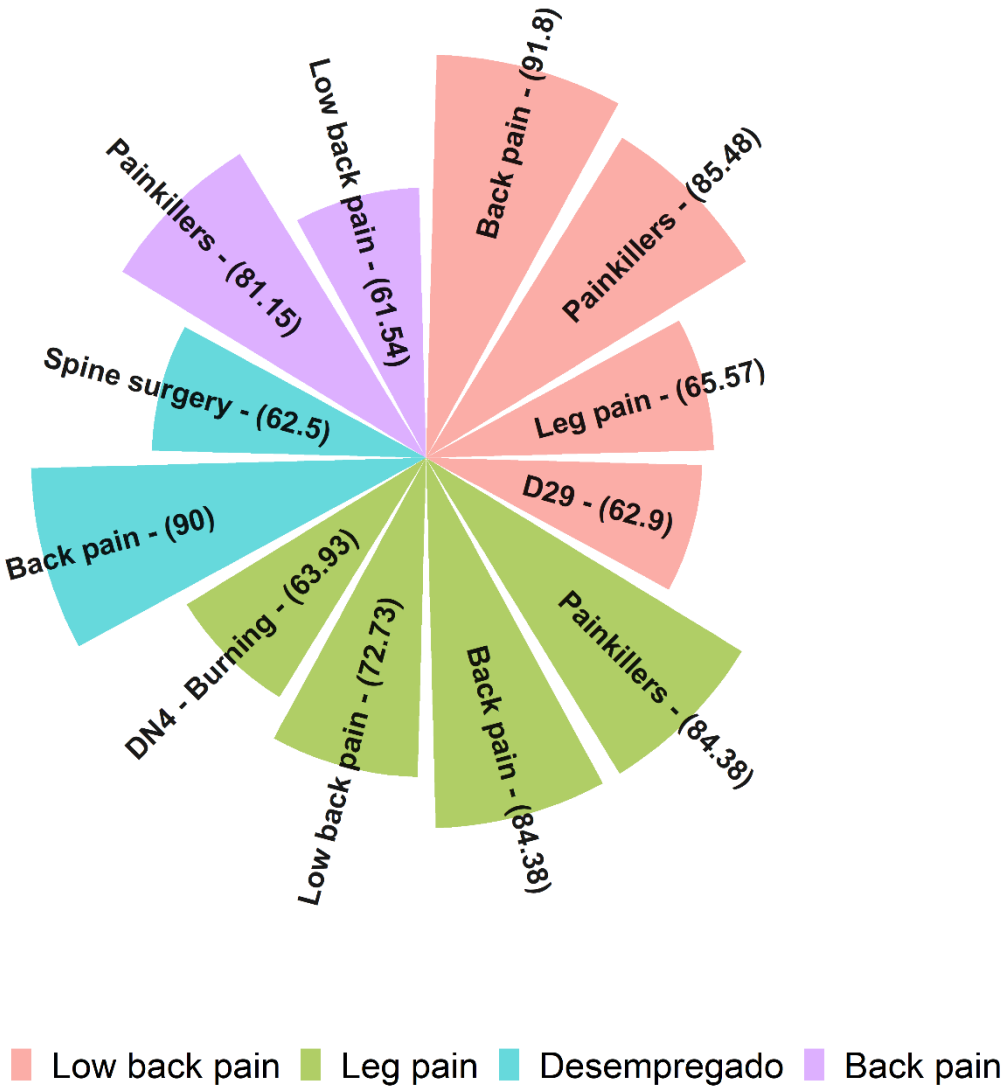
Figure 5 provides a visual summary of the correlations found in Table 3 , Table 4 and Table 5 and previously commented in text form.

Figure 5 - Ordinal questions from the RMDQ, Oswestry and SF-12 questionnaires most strongly correlated with Low Back Pain, Leg Pain and Back Pain



Main correlations between the binary variables back pain, low back pain, leg pain and being unemployed or not.

Figure 6 - Main correlations between the binary variables back pain, low back pain, leg pain and being unemployed or not.



In Figure 6, it is possible to visualize which issues are most correlated with the variables low back pain, leg pain, back pain and whether or not you are unemployed. It is noted that 65·57 % of people who checked the option of presenting pain in the lower back (low back pain) also indicated that they have pain in their legs.

It is also important to mention that 63·93 % of patients with leg pain also checked the option in Questionnaire DN4 that their pain has a burning characteristic (DN - burning). It is also noted the high percentage of patients who use painkillers, 84.38 % for those with leg pain and 90 % for those with back pain.

Besides that, there was no correlation greater than 60 % and with a significance level of 5 % for the variable that considers whether the patient has already undergone any spinal surgery

Feature Selection

Based on the correlations presented, the following criteria were defined for the selection of predictor variables that are used in the machine learning model:

- According to the contingency table, those variables that present a percentage higher than 60% correlation value.

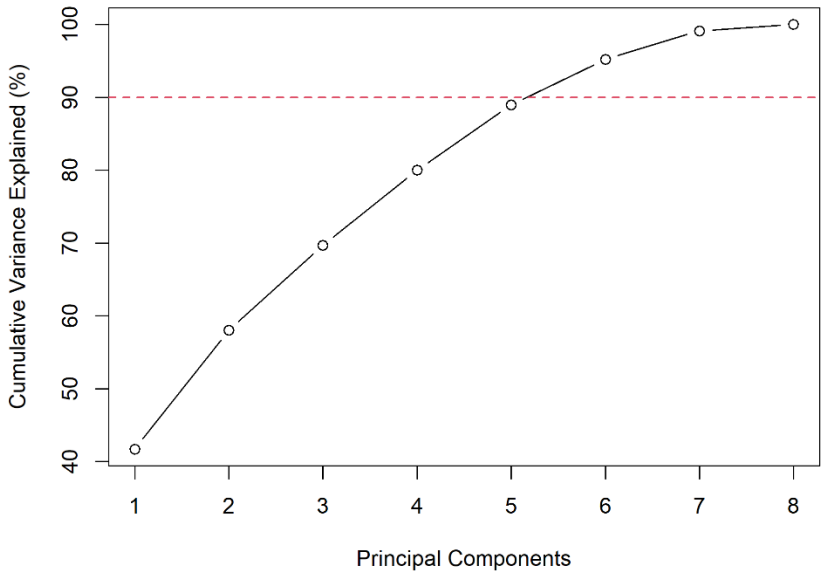
- From the correlation between the low back pain binary variable (answering yes or no for that question), were selected only those with a correlation greater than 0·5 or less than -0·5 (values presented in the figures as multiplied by 100).

Thus, the selected variables were:
- Gothenburg - Whether or not you use painkillers
- Gothenburg - Whether or not you have back pain
- Gothenburg - Whether or not you have leg pain
- Gothenburg - Which gender (Male or Female)
- BPI - If you have pain in the region 29 of question 2 from the Brief Pain Inventory Questionnaire
- Gothenburg - How long have you had pain in your legs
- BPI - Pain intensity at the moment
- Short Form Health Survey - Performed less than he wanted in daily activities due to pain issues

Principal component analysis

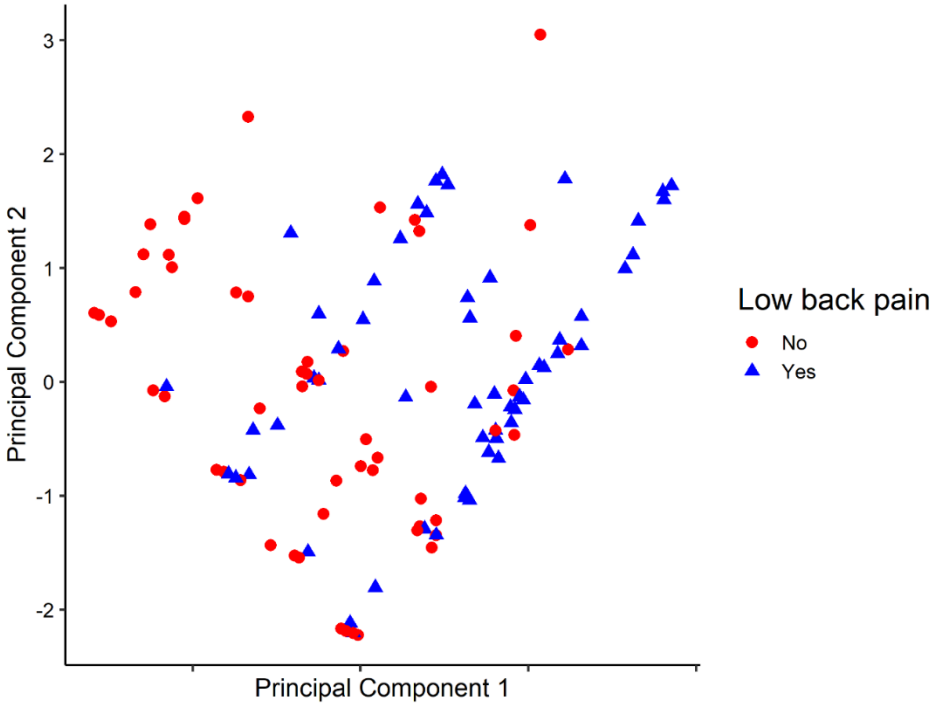
From the principal component analysis, it is possible to see the Scree Plot graph in Figure 7 , which allows us to infer the percentage of variance explained according to the number of main components used. Thus, it is observed that 2 components can explain up to approximately 59 % of the variability of the data and with 5 components, it explains up to 94 %.

Figure 7 -*Scree plot* from the principal component analysis



Since 2 main components are sufficient to explain up to approximately 59 % of the data variability, it can be seen in Figure 8 how the data are distributed in two dimensions.

Figure 8 - Two-axis view of the distribution of low back pain cases using the first 2 main components



It is possible to observe from Figure 8 that with only 2 dimensions there is a great presence of overlap between positive and negative cases.

Machine learning model to predict low back pain

From the application of the five machine learning models previously presented in the Methodological procedure section, Figure 9 shows the result of the ROC curve obtained and in Figure 10 there is the Precision-Recall curve, considering then the variation of the *threshold* for prediction between the values 0 and 1.

Figure 9 - *ROC curve for the 5 different machine learning models*

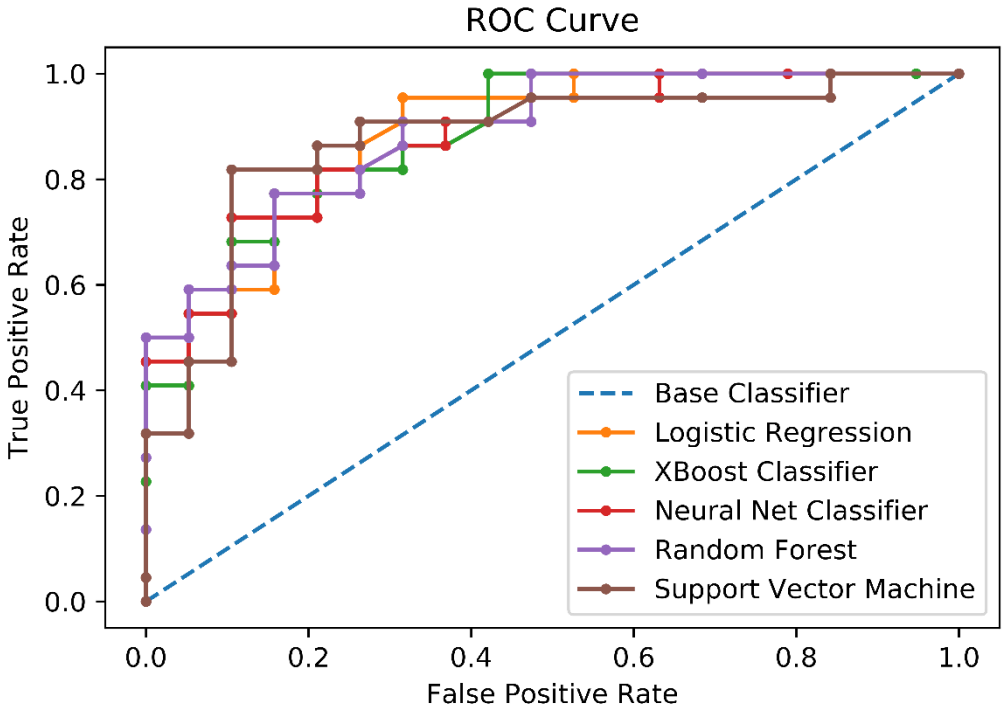
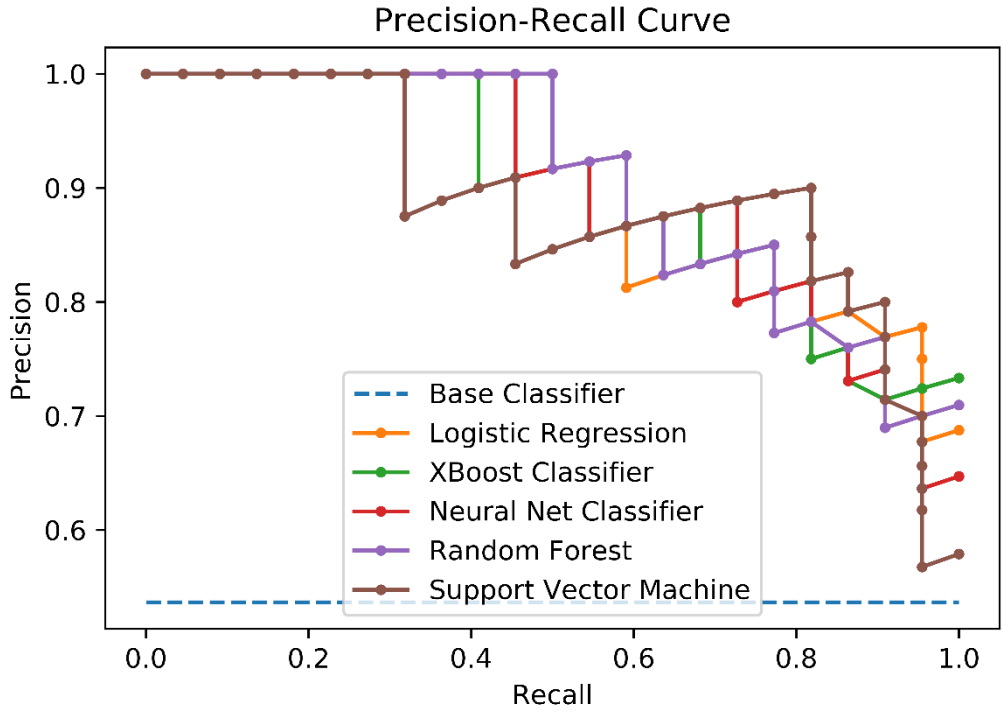


Figure 10 - ***Precision-Recall curve for the 5 different machine learning models***



For the *threshold* default value of 0·5, we present in Table 6 the final metrics obtained with the application of 5 different machine learning models performances for the classification of patients regarding the presence of low back pain.

Table 6 - Results of the performance metrics of the tested machine learning models

	Accuracy	*Precision*	*Recall*	*F1-Score*
Logistic regression	0.78	0.82	0.73	0.73
Neural network	0.71	0.83	0.81	0.82
Random Forest	0.78	0.77	0.74	0.74
Support Vector Machine	0.79	0.88	0.69	0.72
XBoost Classifier	0.81	0.81	0.80	0.79

Therefore, it is noted that with the use of the *threshold* standard of 0·5, the algorithm neural network presented the highest *Recall* and *F1-Score* value and the model built from the *XBoost* algorithm presented the highest accuracy value, as well as the second highest value of *Recall* and *F1-Score*. It should be mentioned that the performance of these models can be improved through the collection of larger databases and with a smaller proportion of missing values.

It is noteworthy that the task performed and exposed, the discovery of which are the main variables that have the highest correlation values with low back pain, allows the reduction of the number of questions in the applied questionnaires. Thus, one of the main benefits of adopting this measure is that patients are more likely to answer questionnaires with fewer questions than the current ones, thus contributing positively to a better quality database generation and data integrity, consequently assisting the patient screening process.

5 Conclusions

In summary, the presented method allowed to determine which questions are more correlated with the condition of back pain, leg pain and chronic low back pain. Moreover, the machine learning algorithms made use of only seven variables as input to predict the occurrence of chronic low back pain. The neural network and XBoost algorithm resulted in the best metrics performance with Recall and Precision near the value of 0·8. These results are important to assist in the patient screening and allocation process, contributing to reduce costs, as stated by the value based health care management system approach. Additionally, the reduction of number of questions, increases the probability of more patients to answer every item in a questionnaire and contribute to build larger datasets that can be used to build models with more even better metrics.

7 References

1. Porter ME, Teisberg EO. *Redefining Health Care - Creating Value-Bases Competition on Results*.; 2006.
2. Curfman GD, Morrissey S, Drazen JM. High-Value Health Care — A Sustainable Proposition. *N Engl J Med*. 2013;369(12):1163-1164. doi:10.1056/nejme1310884
3. Smith M, Saunders R, Stuckhardt L, Mcginnis JM. *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America*. Vol 51.; 2014. doi:10.5860/choice.51-3277
4. National Academy of Medicine. Artifical Intellingence in Health Care. Published online 2018.
5. Stephanie Allen P, Hammett DR, Vettori E de, et al. 2019 Global health care outlook Shaping the future. *Des Issues*. Published online 2019:41. https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Life-Sciences-Health-Care/gx-lshc-hc-outlook-2019.pdf%0Ahttp://www.ncbi.nlm.nih.gov/books/NBK2665/%0Ahttp://dx.doi.org/10.1016/j.bios.2016.09.038%0Ahttps://pdfs.semanticscholar.org/956b/6ee61
6. Medici AC, Monitor UH, Market L. Por André C . Medici. 2019;(December). doi:10.13140/RG.2.2.27521.40807
7. Depintor JDP, Bracher ESB, Cabral DMC, Eluf-Neto J. Prevalence of chronic spinal pain and identification of associated factors in a sample of the population of São Paulo, Brazil: cross-sectional study. *Sao Paulo Med J*. 2016;134(5):375-384. doi:10.1590/1516-3180.2016.0091310516
8. Kennedy R, Abd-Elsayed A. The International Association for the Study of Pain (IASP) Classification of Chronic Pain Syndromes. *Pain*. Published online 2019:1101-1103. doi:10.1007/978-3-319-99124-5_234
9. Werneke MW, Edmond S, Young M, Grigsby D, McClenahan B, McGill T. Association between changes in function among patients with lumbar impairments classified according to the STarT Back Screening Tool and managed by McKenzie credentialed physiotherapists. *Physiother Theory Pract*. 2020;36(5):589-597. doi:10.1080/09593985.2018.1490839
10. Tagliaferri SD, Angelova M, Zhao X, et al. Artificial intelligence to improve back pain outcomes and lessons learnt from clinical classification approaches: three systematic reviews. *npj Digit Med*. 2020;3(1). doi:10.1038/s41746-020-0303-x
11. Bellman RE. *Dynamic Programming*. Princeton University Press; 1957.
12. Venkat N. The Curse of Dimensionality. Published online 2010:169-181. doi:10.1201/ebk0824740993-10
13. NCSS. Contingency Tables Square Test. In: *NCSS - Statistical Software*. ; 2019:1-39. https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Contingency_Tables-Crosstabs-Chi-Square_Test.pdf
14. Bedrick EJ. Biserial Correlation. *Encycl Biostat*. Published online 2005. doi:10.1002/0470011815.b2a10007
15. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. *ICML 2005 - Proc 22nd Int Conf Mach Learn*. 2005;(1999):625-632. doi:10.1145/1102351.1102430
16. Browlee J. A Gentle Introduction to XGBoost for Applied Machine Learning. Published 2016. Accessed September 28, 2020. https://machinelearningmastery.com/gentle-introduction-xgboost-applied-

machine-learning/

17. Powers DMW. Evaluation : From Precision , Recall and F-Factor to ROC , Informedness , Markedness & Correlation. 2007;(December). Commonly used evaluation measures including Recall, Precision, F-Factor and Rand Accuracy are biased and should not be used without clear understanding of the biases, and corresponding identification of chance or base case levels of the statistic. Using t
18. Sasaki Y. The truth of the F-measure The truth of the F-measure. 2015;(January 2007):1-6.
19. Scikit-Learn Developers. Metrics and scoring: quantifying the quality of predictions. Published 2020. Accessed October 28, 2020. https://scikit-learn.org/stable/modules/model_evaluation.html