Diagnosis of neuropathies in diabetic patients by applying machine learning

Diagnóstico de neuropatías en pacientes diabéticos mediante la aplicación de machine learning

Claudio Meneses Villegas^{1*} David Coa Aqueveque²

Recibido 6 de Enero de 2021, aceptado 17 de Mayo de 2021

Received: January 6, 2021 Accepted: May 17, 2021

ABSTRACT

This article describes the construction of classification models by applying machine learning algorithms, in the kinesiological domain of neuropathy diagnosis, in patients with diabetes. The process is developed using the Cross Industry Standard Process for Data Mining guide (CRISP-DM), according to the general process of Knowledge Discovery in Databases (KDD). The main objective of the research is to build a model for classifying the condition of diabetic neuropathy in patients, using the data recorded by the expert and, mainly, by using a Wii Balance Board to capture the postural variations of a subject undergoing a series of clinical evaluations. In this research, the results obtained in each phase of the process are analyzed and the degree of achievement of objectives and performance metrics in the evaluation of *machine learning* models are measured. The results obtained will be expanded with the analysis of time series and the extraction of features, under the hypothesis that the center of pressure (COP) is a sufficient and reliable predictor to discriminate against the existence of a neuropathy.

Keywords: center of pressure, classification, CRISP-DM, data mining, diabetes, diabetic neuropathy.

RESUMEN

En este artículo se describe la construcción de modelos de clasificación, mediante la aplicación de algoritmos de machine learning, en el dominio kinesiológico del diagnóstico de neuropatías, en pacientes que presentan diabetes. El proceso se desarrolla mediante la metodología CRISP-DM, acorde al proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD). El objetivo principal de la investigación es construir un modelo de clasificación de la condición de neuropatía diabética en pacientes, mediante los datos registrados por el experto y, principalmente, por el uso de una Wii Balance Board para capturar las variaciones posturales de un sujeto sometido a una serie de evaluaciones clínicas. En esta investigación, se analizan los resultados obtenidos en cada fase del proceso y se mide el grado de cumplimiento de los objetivos y las métricas de rendimiento, en la evaluación de los modelos de machine learning. Los resultados obtenidos serán ampliados con el análisis de series de tiempo y la extracción de características, bajo la hipótesis de que el centro de presión (COP) es un predictor suficiente para discriminar la existencia de una neuropatía.

Palabras clave: centro de presión, clasificación, CRISP-DM, diabetes, minería de datos, neuropatía diabética.

¹ Universidad Católica del Norte. Departamento de Ingeniería de Sistemas y Computación. Antofagasta, Chile.
E-mail: cmeneses@ucn.cl

² Universidad Católica del Norte. Departamento de Ingeniería de Sistemas y Computación. Antofagasta, Chile.
E-mail: david.coo@alumnos.ucn.cl

* Autor de correspondencia: cmeneses@ucn.cl

INTRODUCTION

The problem of patient classification with diabetic neuropathy condition (DN) is a research issue in the health area, particularly in the Faculty of Health Sciences of the *Universidad de Antofagasta*. One of the activities of the professionals in the Department of Kinesiology of this institution is to support adults and elderly patients in the rehabilitation of musculoskeletal pathologies, being the main study group diabetic individuals. In 50% of cases of type 1 and 2 diabetes mellitus, people develop DN. This affects the peripheral nervous system and may cause loss of sensitivity, muscle pain and weakness, since the nervous system does not correctly encode the signals sent by the body [1]. In extreme cases, due to sensory loss and poor circulation, patients are unaware of injuries or wounds to their limbs, leading to amputation of affected parts. Hence, a patient classification model is needed, which through the analysis of quantitative attributes it identifies the presence of the pathology.

From the methodological point of view, a variation of the Cross Industry Standard Process for Data Mining (CRISP-DM) guide was utilized in this work, for the development of machine learning models. The evaluation of the results and the development process was based, in addition to the standard performance metrics, on the judgment of experts in the domain (kinesiology area) for the correct interpretation and evaluation of the results obtained. The initial situation of how the physical therapist approached the problem of determining whether a patient presents a neuropathy, considers the analysis of symptoms using standardized clinical evaluations and the posture tests using a Wii Balance Board (WBB) device.

The following business (clinical) objectives were considered in this domain:

- Improve patient care and treatment focus.
- Provide new sources of research, encouraging the use of machine learning techniques.
- Classification of patients using the data sampled from the WBB, to determine DN condition.
- Determine whether the measurement of center of pressure (COP) is a better predictor than clinical evaluations.

In the other hand, technical objectives listed in relation to data analysis are:

- Development of a predictive model that classifies, according to its attributes, whether the patient has diabetic neuropathy or not.
- Find the correlation between the attributes of the medical surveys and the features of the time series.
- Identify the relevance of each attribute in the classification task, establishing a hierarchy based on the class attribute.

Later, the results obtained from this work will allow us to expand the knowledge, refine the investigative process and to consider new important features, to improve the classification model.

In what follows, this article considers firstly the analysis of the state of the art, specifically related works using computational methods for the classification of pathologies, that may be applicable. Secondly, the materials required (data, hardware and software resources) and the methods used to generate results for each posed objective. Third, the results obtained from the data preparation and modeling stages are presented. Finally, the discussion of the obtained results and final conclusions are presented.

RELATED WORKS

As related works, those that use the COP time series are considered to differentiate between subjects of different age or health conditions. On the other hand, studies that consider electromyography signals are analyzed as an alternative, despite the fact that the nature of both signals are different.

Time series analysis

In time series analysis, different transformations are applied to obtain descriptive variables. This is useful when a time series lacks a clear trend or a certain cycle, considering that the COP is sensitive to various external factors such as diseases that alter posture, distractors, involuntary movements of the human body, etc. However, several studies have found that it is possible to identify conditions of the subject. As in Yamagata's research [2], in which older adults, with no history of pathologies, a series of clinical tests are applied to study the stiffness that they adopt as a method to compensate for the body roll. This has made it possible to differentiate not only the age range of a subject or the condition of eyes open or closed, but also to find disturbances

associated with processes of other systems, such as breathing, the influence of vision, the vestibular and somatosensory system, proprioceptive information, among others.

In COP time series studies [2], general statistical variables are obtained through calculations that do not represent a specific attribute of the subject, their health or a clinical assessment, but rather interpret the series as a single value, which then it is entered into a machine learning model. Examples of statistical variables include the slope of a series, the standard deviation, the mean, the variance, the maximum, the minimum, the median, etc. Because the physiology of the COP signal is complex, laboratory tests usually include a control group, that is, healthy people, since there is no standardized value that indicates whether a COP is normal or abnormal. For this reason, when researchers need to study physiological signals, they include both groups of patients, to compare the different results.

Feature extraction

The features extracted from a time series can be in the time domain or in the frequency domain. Subasi in his research [3], obtained the coefficients of each sub-band after applying discrete Wavelet transform, to differentiate between normal, myopathic or neurogenic electromyogram signals. On the other hand, in Chamnongthai's work [4], the characteristics are based on graph theory, where the time series peaks are discretized and transformed into weighted graphs, then the attributes are obtained to differentiate between normal signals or with myopathic or neuropathic pathologies. Gregory King [5] explores a mixture of unique attributes of the COP, based on parameters of frequencies, speed, body roll, among others. These studies have in common that, after feature extraction, the attributes are entered into three machine learning models: Support Vector Machine (SVM), a Random Forest, a Linear Discriminant Analysis (LDA) classifier and a neural network. The results obtained show that the time series feature extraction is applicable to the current kinesiological domain. However, it should be considered that electromyography signals are different from COP signals, since the former maintain a disruptive structure than postural signals, when an evaluated muscle is activated. On the other hand, the COP signal does not present the same behavior, since it represents the excursion of

the posture, influenced by internal and external factors. Nevertheless, it seems feasible to adapt the methodology used in the related works to the current domain, in which the time series is transformed into new values (based on feature extraction), with its respective label: diabetic or neuropathic.

BACKGROUND

This section describes the background to the COP, which is a variable that represents the excursion of a subject's posture, used in posturology analysis. The next subsection describes its study.

Study of the center of pressure

COP is the most widely used variable in balance control research, since it is feasible to detect postural changes due to diseases such as Parkinson, multiple sclerosis, myopathies, etc. It is defined as the point of application of all vertical forces performed by an individual on a support surface [6]. In Figure 1, it is visualized that the center of mass (COM) is located close to the pelvic area, which projects a vertical force on the WBB. This accessory captures the data, in two dimensions: the anteroposterior (AP) and mid-lateral (ML) movement, which represent the Y and X axis respectively.

Through an analysis of the COP excursion, it is possible to determine if an individual generates compensation to maintain his posture. The tests that are applied to the patient are carried out in 30 seconds and in two states: with eyes open and eyes closed, since without vision, the brain tries to compensate for the sway, using the proprioceptive

Figure 1. COP representation in a subject.

or the somatosensory system [7]. Additionally, distractors can be included as dual tasks and an unstable surface. This causes that the brain sends out more signals to compensate the posture, which makes COP more natural.

Machine learning algorithms

Machine learning is one of the branches of artificial intelligence, which seeks to build systems that work automatically in complex and changing environments. It is a combination of computer science and statistics. Advances in this subject have made it possible to create an approach based on learning behaviors that are mathematically modeled and that have their origin in stored data from the past. In the medical domain, the aim is to improve the diagnostic processes or the evolution of diseases, so it is of interest to have a data source and a model that assigns them to a certain class. The algorithms described below are based on the book on artificial intelligence by Peter Norvig and Stuart Russel [8] and data mining by Han, Kamber and Pei [9]:

1. Naïve Bayes: this classifier is based on the Bayes' theorem, in which given a set of classes C_i , the model predicts the value X to which it belongs, evaluating the highest conditional probability P of each xk to n elements, which is described in equation (1).

$$P(X|C_i) = \prod_{k=1}^n P(X_k|C_i) \quad (1)$$

- 2) K-Nearest Neighbor: this model allows searching in the pattern space for the unknown k values that are closest to the training values. The degree of "closeness" between two tuples $X_1 = (X_{11} \dots X_{1n})$ and $X_2 = (X_{21} \dots X_{2n})$ is based on the Euclidean distance. The mathematical model is described in equation (2).

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^n (x_{1i} - x_{2i})^2} \quad (2)$$

3. Artificial Neural Network: It consists of a set of interconnected nodes, in which each attribute I_j enters separately, in parallel and crosses a hidden layer, in which, given a sigmoid activation function, it allows to adjust non-linear

regressions, by means of the summation of weights w_{ij} , where i is the value in the previous layer and j is the current value, plus a bias θ that are adjusted iteratively. The mathematical model is described in equation (3).

$$I_j = \sum_i w_{ij} 0_i + \theta_j \quad (3)$$

4. Decision Tree: it is based on a graph structure, where the root node refers to the main attribute, the branches reference the result of the evaluation between attributes and each terminal node (leaf) represents the class. By the gain information metric, the selection of attributes is carried out, such that p_i corresponds to the probability that a tuple belongs to the class C_i and, using a logarithm function of base 2, the number of bits necessary to encode the information. The mathematical model is found in equation (4).

$$Info(D) = -\sum_{i=1}^m p_i \log_2(p_i) \quad (4)$$

METHODOLOGY

This section mainly describes an overview of the CRISP-DM methodology, used for the development of the project.

CRISP-DM

The CRISP-DM methodology was developed at the end of 1999, where five European companies led the project until its first version was developed, with the aim of providing a non-proprietary standard process model and freely available for use [10]. On the other hand, as it is the medical area there are certain considerations that are not specified. In Krzysztof and Moore's research, they warn that there must be some care in "privacy sensitivity mining" (social, legal and ethical issues) and in the heterogeneity of the data [11]. Because some clinical data come from individuals, their discreet use is necessary to prevent ethical conflicts with patients. To complement the warning of the mentioned authors, a model adjusted to the medical area is applied, known as Medical Domain (CRISP-MED-DM), proposed by Niaksu [12]. This model adds additional activities and renames others according to the clinical environment, without changing the

main structure. Therefore, the original phases of the methodology are maintained.

The phases of the CRISP-DM methodology are:

- Business understanding.
- Data understanding.
- Data preparation.
- Modelling
- Evaluation

The development of the research followed the model proposed by the KDD process and more specifically by the CRISP-DM guide. Thus, in the business understanding stage, the business objectives, success criteria and technical requirements and objectives of the project were established. In the stages of understanding and data preparation, the characteristics of the data, the selection of attributes and the transformations were determined to obtain the datasets in the form required to perform data mining techniques. This stage included the use of RapidMiner 9.4 tool (Technical University of Dortmund, Dortmund, Germany), Python 3 (Python Software Foundation) and use of spreadsheets in Microsoft Excel 2016. In the modeling stage, predictive models were built using the application of machine learning algorithms: Naïve Bayes (NB), K-Nearest Neighbor (KNN), neural networks (NN) and decision trees (DT), whose performances were based on the calculation of standard metrics such as accuracy, precision, sensitivity or recall and the area under curve (AUC) of operational characteristics of the receiver (ROC).

General process diagram

In Figure 2, the entire development process is modeled, based on the CRISP-DM methodology

phases. In each one, the outputs become the inputs of the following. The outputs of the Business Understanding phase are the business and technical objectives, in the Data Understanding phase there is an exploration of the original datasets (Excel file and time series). In the following, it includes clean datasets and extracted features from time series, then the Modeling phase includes the design of classification models, correlation analysis and attribute hierarchy analysis. Finally, the last phase includes the tabulated results.

EXPERIMENTS

In this section, the data sets used are detailed, as well as a brief description of the hardware and software used.

Patient datasets

There are two main data sets, one corresponds to clinical evaluations and surveys, tabulated in an Excel file, while the second corresponds to comma separated values files (CSV) for each posture evaluation performed on all patients. The description of attributes is listed in Table 1.

In the case of CSV files, they contain eight numerical attributes, without missing data. For analysis purposes, the most relevant attributes selected are COP X and COP Y. A representation of the COP for a diabetic subject is in Figure 3 and another one with diagnosed DN in Figure 4, where it cannot be clearly identified if there is a trend or a seasonality.

Time series have been grouped by type of examination, as detailed in Table 2 according to

Figure 2. General process representation.

Table 1. List of attributes of clinical evaluations in Excel spreadsheet.

Attribute	Type	Values	Observation
Age	Numerical	65 ± 8.5 SD	Patient's age.
Sex	Nominal	{0,1}	Male or female identifier.
Time MD	Numerical	21 ± 11.5 SD	Number of years with diabetes diagnosed by a medic.
Glucose	Numerical	156 ± 78.3 SD	Glucose level during test.
EFAM	Ordinal	{0, 1}	A scale that measures the loss of function in older adults who require or not support from orthoses or a third party.
NSS	Ordinal	{0, 1, 2, 3}	Evaluation of 17 items on muscle weakness, sensory disorders and automatic symptoms. The values are normal, mild, moderate and severe.
DNS	Ordinal	{0, 1}	Evaluation of four questions on neurological abnormalities. The values are absent or present.
SVAI - SVAD	Ordinal	{0, 1, 2, 3, AM}	Evaluation of the left foot (I) and right foot (D), using a 128Hz pitch. The values are serious, moderate, mild, no neuropathy and amputated.
SVBI - SVBD	Ordinal	{0, 1, 2, 3, AM}	Evaluation of the left foot (I) and right foot (D), using a 256Hz pitch. The values are the same as for SVAI-SVAD.
Class	Label	{0, 1}	Class attribute to predict. The values are absent or present.

Figure 3. COP X (green) and Y (blue) from a diabetic patient.

Figure 4. COP X (orange) and Y (red) from a diabetic patient with neuropathy diagnosed.

Table 2. CSV files grouped by exam type.

Examination	Files	Missing	Observation
EO	47	0	Eyes open.
EOUS	42	0	Eyes open, unstable surface.
EODT	45	0	Eyes open, dual task.
EODTUS	41	0	Eyes open, dual task, unstable surface.
EC	45	0	Eyes closed.
ECUS	42	0	Eyes closed, unstable surface.
ECDT	45	0	Eyes closed, dual task.
ECDTUS	42	0	Eyes closed, dual task, unstable surface.
Total	349	0	–

the indications of the domain expert, which gives a total of 349 different files or time series. The reason for adding different distractors (closed eyes, unstable surface and dual task), is to obtain several samples of the COP, altering the senses of the subject, generating a more autonomous postural balance.

Hardware

Include all physical components required to get and analyze data. These are:

- **Wii Balance Board (WBB):** device that obtains the coordinates of the patient's posture, through the calculations of the four internal sensors. The data is stored as a time series. Figure 5 represents the used device.
- **Laboratory computer:** equipment used by the physical therapist where the clinical data in Excel spreadsheet is generated and time series files are manipulated.
- **Analysis computer:** equipment used by the data mining analyst, with the necessary applications to carry out his activities.

Figure 5. Wii Balance Board device. Source: Amazon.

Software

All computer components as applications installed on hardware equipment are:

- **BrainBlox:** application used by the physical therapist that allows interaction with the WBB and stored data from it.
- **RapidMiner 9.4:** application used to design machine learning models and data preparation.
- **Microsoft Excel 2016:** Microsoft application that allows interaction with data sheets.
- **Google Collab with Python 3.7:** Google tool for programming ad-hoc applications in Python.

Feature extraction and selection

The activities developed allowed the understanding of the data recorded by the physical therapist in both data sources: clinical evaluations in an Excel spreadsheet and a set of time series files. No modifications in Excel data spreadsheet were required, but the exclusion of irrelevant attributes. To the second dataset, transformations were applied with the aim of extract features to represent the time series adequately. Using the RapidMiner tool, the following activities were realized:

- **Define class attribute (label).**
- **Parse attributes Age, Glucose and Time MD to numerical.**
- **Conversion of the attributes Sex into nominal category type and the tests EFAM, NSS, DNS, SVAD, SVAI, SVBD and SVBI into ordinal category.**
- **Exclusion of the attributes Name and Subject.** However, the subject identificatory would be used for labelling between the Excel spreadsheet and the time series.

Time series files were grouped according to the type of examination to which corresponds. Additionally,

two files are added, General EO and General EC, which consolidate all time series with eyes open and eyes closed respectively. It results in a total of ten files, in which a Moving Average technique is applied to smooth the signal with a window size of 100 samples⁽¹³⁾. Then, feature extraction is applied to get the maximum, minimum and median values from COP on the X axis and the Y axis. The process also includes the crossing of information between the subject identificatory in Excel spreadsheet and the time series of each patient, to label each group of files. Later, the final dataset is created with the following time series features:

- Maximum value in COP X.
- Minimum value in COP X.
- Median value in COP X.
- Maximum value in COP Y.
- Minimum value in COP Y.
- Median value in COP Y.
- Class from the Excel spreadsheet.

Construction of data models

In this phase, algorithms were designed according to the technical objectives initially set. Five different models were made and applied to each data source, according to its own limitations. Table 3 describes strategy used for each dataset and technical objectives. The algorithms used are the NB, the NN, the KNN, a covariance matrix and a DT. Cross-validation method is used, in which the data is divided into k units (10 – folds) and crossed with the training data, for each of the k units. Such, the last fold is used for testing, thus, 10% is the stratified testing data.

Default parameters of each model were used, according to the default RapidMiner settings. Minors settings were considered in KNN and DT models, just adjusting the k value and quantity of trees respectively. Parameters of models are detailed as follows: Table 4 for NN model, Table 5 for KNN model and Table 6 for DT model.

Table 3. Modeling strategy for each objective.

Objective	Source	Algorithm	Training / Testing
Classification model	Excel spreadsheet	NB	90% / 10%
	Time Series	RN KNN	90% / 10%
Attribute correlation	Excel spreadsheet	Covariance	n/a
	Time Series	Covariance	n/a
Attribute hierarchy	Excel spreadsheet	DT	90% / 10%

Table 4. Parameters of Neural Networks model.

Parameter	Value	Description
Times	200	Training cycles.
Learning rate	0.01	Degree of adjustment of the weights.
Momentum	0.9	Adjustment to prevent local maximums.
Randomize	True	Data is randomized.
Normalized	True	Data is normalized.

Table 5. Parameters of KNN model.

Parameter	Value	Description
K	5	K-neighbors to evaluate
Weighted Vote	True	Weigh distances
Type of measure	Numerical	Data type
Numerical measurements	Euclidean distance	Metrics to be used

Table 6. Parameters of Decision Tree model.

Parameter	Value	Description
Trees	3	Quantity
Criterion	Gain ratio	Type of gain for each attribute
Depth	10	Maximum number of leaves
Apply pruning	True	Prune leaves depending on confidence level

RESULTS

The metrics of evaluation are used to measure the performance of the models and verify the degree of compliance with the technical objectives.

* Accuracy: rate of correctly classified tuples. Calculation is shown in equation (5).

$$\frac{\text{True Positives} + \text{True Negatives}}{\text{Positives} + \text{Negatives}} \quad (5)$$

* Error: rate of incorrectly classified tuples. Calculation is shown in equation (6).

$$\frac{\text{False Positives} + \text{False Negatives}}{\text{Positives} + \text{Negatives}} \quad (6)$$

* Precision: the number of positive hits divided by the total of positive correct and incorrect instances. Calculation is shown in equation (7).

$$\frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (7)$$

* Sensitivity: completeness measure that corresponds to the number of correctly classified positive instances. Calculation is shown in equation (8).

$$\frac{\text{True Positives}}{\text{True Positives} + \text{False Negative}} \quad (8)$$

* AUC: rate of the degree of separability between classes. This curve plot two parameters: true positive rate (TPR) and false positive rate (FPR).

Calculation is shown in equation (9) and equation (10).

$$\text{TPR} = \frac{\text{True Positive}}{\text{True Positives} + \text{False Negative}} \quad (9)$$

$$\text{FPR} = \frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}} \quad (10)$$

Classification model

This section lists the results obtained in the machine learning models, for each planted objective. In Table 7, it is inferred that NB achieves an average performance of 70% of accuracy, by the probabilistic model to the 12 attributes of the Excel spreadsheet, considering the low number of instances and the high dimensionality. This implies that it is feasible to design a model based only on clinical attributes, which is an alternative to the feature extraction from the time series. However, the value of the AUC considers a low differentiation between classes, with a value of 0.6, due to the low number of instances (21 patients) in the data set. Another point to consider is that the model is still influenced by the human bias present in each exam. So, the model should be limited only for comparative purposes between data sets and not as a definitive solution.

On the other hand, Table 8 and Table 9 show the metrics obtained, in the datasets of the time series features, grouped by type of exam. In the KNN model results, there is a balanced performance between eyes closed and eyes open exams. The set that achieves the best performance, with 77.5% accuracy and 0.83 AUC is the EO set. Results over 70% of accuracy are obtained in the sets of eyes

Table 7. Performance of Naïve Bayes classifier model.

Dataset	Accuracy	Error	Precision	Sensitivity	AUC
Excel spreadsheet	66.6%	33.3%	62.5%	55.5%	0.6

open, except the EOUSDT data set. This is explained by the fact that the signal becomes more complex as more distractors are incorporated. In eyes closed data sets, on average there is a lower performance of the model, compared to eyes open. Although a high performance of accuracy is achieved in ECDT, with 86.5%, extracting new features is considered an alternative to improve performance.

In Table 9, similar results are obtained in the NN model. The EO data set obtained 74.5% correct answers, with values over 80% of precision and specificity. The ECDT data set obtained 74.5% correct answers, higher than the rest of the eyes closed exams, which reaffirms that the models are influenced by the type of features entered.

Attributes correlation

Regarding the second technical objective, the correlation matrix is used to verify the influence between the attributes. Table 10 shows ten results obtained from the model: the first five with the

highest positive correlation and the rest with negative correlation. The results indicate that a clear correlation is not found between the set of attributes. This gives an indication that none of the evaluations supersedes another.

Figure 6 shows the correlation between attributes, excluding those non-binary categorical, such as the NSS and the vibration tests. The results indicate that the direct proportionality is low between attributes, with the highest value being 0.33 between age and time of medical diagnosis of diabetes.

In Table 11, five results with the highest correlation between attributes are listed. The eyes open tests, including unstable surface and dual task datasets are grouped in EO and the eyes closed tests in EC respectively, then the calculation is made for each one. The results indicate that the high correlation (above 0.8) implies that there is redundancy when considering the maximum, minimum and median at the same time.

Table 8. Performance of KNN classifier with time series features.

Dataset	Accuracy	Error	Precision	Sensitivity	AUC
EO	77.5%	22.5%	82.4%	87.5%	0.83
EOUS	78.5%	21.5%	78.8%	92.9%	0.75
EODT	74%	26%	75.7%	90.3%	0.78
EODTUS	51%	49%	61.3%	70.4%	0.59
General EO	67.9%	32.1%	72.5%	84.8%	0.66
EC	76%	24%	85.7%	77.4%	0.86
ECUS	56.5%	43.5%	70.8%	60.7%	0.65
ECDT	86.5%	13.5%	87.9%	93.6%	0.92
ECDTUS	56.5%	43.5%	65.5%	70.4%	0.73
General EC	75%	25%	80.3%	83.8%	0.80

Table 9. Performance of Neural Network classifier with time series features.

Dataset	Accuracy	Error	Precision	Sensitivity	AUC
EO	74.5%	25.5%	81.3%	81.3%	0.76
EOUS	71%	29%	76.7%	82.1%	0.75
EODT	75.5%	24%	76.3%	93.6%	0.64
EODTUS	61%	39%	64.9%	88.9%	0.69
General EO	69.6%	30.4%	71%	93.2%	0.64
EC	62.5%	37.5%	70.6%	77.4%	0.64
ECUS	60%	40%	64.1%	89.3%	0.54
ECDT	74.5%	25.5%	77.1%	87.1%	0.85
ECDTUS	60.5%	39.5%	64.9%	88.9%	0.66
General EC	73.9%	26.1%	76.5%	88.9%	0.76

Table 10. Correlation among attributes.

Attribute 1	Attribute 2	Correlation	Level
Age	MD Time	0.33	Low
Age	Glucose	0.26	Low
Sex	DNS	0.22	Low
Age	EFAM	0.07	Null
Glucose	EFAM	0.06	Null
EFAM	DNS	-0.23	Low
Sex	MD Time	-0.31	Low
Glucose	DNS	-0.32	Low
Time MD	Glucose	-0.38	Low
Age	Sex	-0.41	Medium

In Figure 7 and Figure 8, high correlation is obtained only between COP X features likewise COP Y features, for both data sets. No correlation exists between X features and Y features.

Attribute hierarchy

In the case of the third technical objective, the DT algorithm is applied to identify which of all attributes

gets the highest gain information ratio to classify. Next, the graph of the generated DT is in Figure 9.

Figure 9 shows that the NSS attribute is better to identify the pathology (0 for diabetic and 1 for neuropathic). This attribute is divided into: 0 for normal, 1 for mild, 2 for moderate and 3 for severe symptoms. If NSS score is mild, then the DT considers to evaluate the Age attribute as a second option. So, if Age is greater than 70.5 years old then it classifies as DN, else not DN. Nevertheless, there is a margin of error when classifying patients with DN and without DN, selecting age > 70.5 as the discriminator, which is reflected in the AUC value.

Compared with KNN and NN classifiers, the DT model has higher performance than previous ones, reaching an accuracy rate of 83.3%, with the lowest error rate, as shown in Table 12. However, AUC value is 0.5 then the class attribute is not correctly discriminated against. High dimensionality and low quantity of records would affect the model.

Attributes	Age	Sex	Time MD	Glucose	EFAM	DNS
Age	1	-0.410	0.326	0.264	0.065	-0.220
Sex	-0.410	1	-0.312	-0.109	-0.181	0.220
Time MD	0.326	-0.312	1	-0.376	0.043	-0.008
Glucose	0.264	-0.109	-0.376	1	0.060	-0.317
EFAM	0.065	-0.181	0.043	0.060	1	-0.230
DNS	-0.220	0.220	-0.008	-0.317	-0.230	1

Figure 6. Correlation between attributes.

Table 11. Correlations of time series features.

Dataset	Attribute 1	Attribute 2	Correlation	Level
General EO	Y_min	Y_median	0.93	High
	Y_max	Y_median	0.93	High
	X_min	X_median	0.89	High
	X_max	X_median	0.83	High
	Y_max	Y_min	0.81	High
General EC	Y_min	Y_median	0.88	High
	Y_max	Y_median	0.87	High
	X_min	X_median	0.82	High
	X_max	X_median	0.82	High
	Y_max	Y_min	0.59	Medium

Attributes	X_max	X_min	X_median	Y_max	Y_min	Y_median
X_max	1	0.540	0.831	0.158	-0.092	0.044
X_min	0.540	1	0.885	0.020	0.167	0.094
X_median	0.831	0.885	1	0.081	0.045	0.065
Y_max	0.158	0.020	0.081	1	0.810	0.929
Y_min	-0.092	0.167	0.045	0.810	1	0.934
Y_median	0.044	0.094	0.065	0.929	0.934	1

Figure 7. Correlation matrix on general EO dataset.

Attributes	X_max	X_min	X_median	Y_max	Y_min	Y_median
X_max	1	0.446	0.821	0.209	-0.074	0.092
X_min	0.446	1	0.823	-0.186	0.161	0.023
X_median	0.821	0.823	1	-0.002	0.019	0.037
Y_max	0.209	-0.186	-0.002	1	0.592	0.865
Y_min	-0.074	0.161	0.019	0.592	1	0.882
Y_median	0.092	0.023	0.037	0.865	0.882	1

Figure 8. Correlation matrix on general EC dataset.

Figure 9. Representation of generated Decision Tree.

Table 12. Performance of Decision Tree model.

Dataset	Accuracy	Error	Precision	Sensitivity	AUC
Excel spreadsheet	83.3%	16.7%	77.7%	77.7%	0.5

DISCUSSION OF RESULTS

Results were presented to the academic physical therapist, during the phases of the CRISP-DM

methodology and its development. The emphasis was on the analysis and transformation of the datasets, in addition to the modeling of solutions to meet the proposed objectives.

The results of the Excel spreadsheet dataset show that accuracy obtained was 67% (33% error), using the NB classifier, which it fits properly even though the high dimensionality and low quantity of records. With respect to the Excel spreadsheet dataset, it is agreed that the NSS exam, which obtained a higher score when applying the DT, is the most appropriate to be performed during the clinical evaluation of patients and to label them with DN or not DN. On the other hand, the tests that are performed with the tuning fork (SVAI, SVAD, SVBI and SVBD) are those with the greatest bias, because what is evaluated is the patient's perception of a stimulus, which makes it a very subjective measure, compared to survey type tests like EFAM, DNS or NSS. This assumption is verified in the DT model, since these do not appear in the final tree.

In time series analysis, with the KNN strategy, the model with worst performance accuracy and error rates are between 51% and 49% respectively and 86.5% with 13.5% of accuracy and error rate the one with best performance. In the case of the NN model, 60% and 40% of accuracy and error rate for the worst performance and 75.5% and 24.5% for the best performance. Nevertheless, results are influenced by the type of exam and the features extracted. This assumption is verified when separated models are applied to each dataset (time series files). The rest of the datasets got on average 69% of accuracy with KNN and 59% for NN. In general, models allow to classify the pathology, but are not achieving higher performance, thus different features from time series must be considered.

CONCLUSIONS AND FUTURE WORK

New combinations of attributes were not feasible to apply due to differences in quantities of each dataset, therefore it must be evaluated whether including or excluding attributes the accuracy rates are improved. Nevertheless, it is concluded that applied techniques achieve an ordered model that classify the DN in both datasets. On the other hand, the quantity of records must be reviewed in further analysis, since it is estimated that increasing the quantity of records improves performance. Dividing datasets in types, as indicated by the expert, is suitable since exams and the source where data is captured are different.

With respect to the use of the methodology in the health area, it is true that the activities are applicable. As the case study was in a reduced environment, activities where other clinical areas are involved were not necessary and did not affect the development of the research. However, the guide does not explicitly specify how to measure compliance of the model, since it is understood more like a guideline of good practices. For this reason, performance evaluation is carried out by verifying whether or not the activity was completed or if there are any observations. The strength of the guide is to encourage the use of tools for building models and datasets.

As future work, it is planned to increase the number of instances, in both types of datasets and to apply the same models developed, considering the reduction of attributes. The emphasis of the research will be on the transformation of time series. An alternative is to include the study of signal entropy. The focus will be on the study of the state of the art in signal transformations. The aim is to evaluate the feasibility of replicating the models in other similar problems.

REFERENCES

- [1] C.C. Castro. "Síntesis biblioteca digital dinámica para estudiantes y profesores de la salud". Universidad de Chile. 2017. Accessed: October 27, 2019.
- [2] M. Yamagata, T. Ikezoe, M. Kamiya, M. Masaki and N. Ichihashi. "Correlation between movement complexity during static standing and balance function in institutionalized older adults". *Clin Interv Aging*. Vol. 12, pp. 499-503. 2017. DOI: 10.2147/CIA.S132425.
- [3] A. Subasi. "Medical decision support system for diagnosis of neuromuscular disorders using DWT and fuzzy support vector machines". *Computers in Biology and Medicine*. Vol. 42 N° 8, pp. 806-815. 2012, doi: 10.1016/j.compbimed.2012.06.004.
- [4] P. Artameeyanant, S. Sultornsanee y K. Chamnongthai. "An EMG-based feature extraction method using a normalized weight vertical visibility algorithm for myopathy and neuropathy detection". *SpringerPlus*. Vol. 5 N° 2101. 2016. DOI: 10.1186/s40064-016-3772-2.
- [5] S.K. Saripalle, G.C. Paiva, T.C. Cliett, R.R. Derakhshani, G.W. King and C.T. Lovelace.

- “Classification of body movements based on posturographic data”. *Human Movement Science*, pp. 238-250. 2014. DOI: 10.1016/j.humov.2013.09.004.
- [6] D.A. Santos and M. Duarte. “A public data set of human balance”. *PeerJ Preprints* 4:e216v1. 2016. DOI: 10.7717/peerj.2648.
- [7] N. Kanekar, Y.-J. Lee and A.S. Aruin. “Frequency analysis approach to study balance control in individuals with multiple sclerosis”. *Journal of Neuroscience Methods*. Vol. 222, pp. 91-96, 2014. DOI: 10.1016/j.jneumeth.2013.10.020.
- [8] S. Russel y P. Norvig. “*Inteligencia Artificial Un Enfoque Moderno*”. Second Edition. Madrid, Spain. Pearson Educación S.A. 2004. ISBN: 978-84-205-4003-0.
- [9] J. Han, M. Kamber y J. Pei. “*Data Mining Concepts and Techniques*”. Third Edition, United States, Morgan Kaufmann. 2011. ISBN: 9780123814807.
- [10] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer y R. Wirth. “*CRISP-DM 1.0 Step-by-step data mining guide*”. August 2000. Accessed: October 22, 2019.
- [11] C.J. Krzysztof and W.G. Moore. “Uniqueness of medical data mining”. *Artificial Intelligence in Medicine*. N° 26, pp. 1-24. 2002. DOI: 10.1016/s0933-3657(02)00049-0.
- [12] O. Niaksu. “*CRISP Data Mining Methodology Extension for Medical Domain*”. *Baltic J. Modern Computing*. Vol. 3 N° 2, pp. 92-109. 2015.