applied sciences

MDPI

Article

Applying Machine Learning for Healthcare: A Case Study on Cervical Pain Assessment with Motion Capture

Juan de la Torre [1,*], Javier Marin [1], Sergio Ilarri [2,3] and Jose J. Marin [1,4]

[1] IDERGO-Research and Development in Ergonomics, Biomechanical Laboratory, I3A-University Institute of Research of Engineering of Aragon, University of Zaragoza, 50018 Zaragoza, Spain; 647473@unizar.es (J.M.); jjmarin@unizar.es (J.J.M.)

[2] Computer Science for Complex System Modelling (COSMOS), I3A-University Institute of Research of Engineering of Aragon, University of Zaragoza, 50018 Zaragoza, Spain; silarri@unizar.es

[3] Department of Computer Science and Systems Engineering, University of Zaragoza, 50018 Zaragoza, Spain

[4] Department of Design and Manufacturing Engineering, University of Zaragoza, 50018 Zaragoza, Spain

* Correspondence: 627471@unizar.es

Received: 9 July 2020; Accepted: 24 August 2020; Published: 27 August 2020

Abstract: Given the exponential availability of data in health centers and the massive sensorization that is expected, there is an increasing need to manage and analyze these data in an effective way. For this purpose, data mining (DM) and machine learning (ML) techniques would be helpful. However, due to the specific characteristics of the field of healthcare, a suitable DM and ML methodology adapted to these particularities is required. The applied methodology must structure the different stages needed for data-driven healthcare, from the acquisition of raw data to decision-making by clinicians, considering the specific requirements of this field. In this paper, we focus on a case study of cervical assessment, where the goal is to predict the potential presence of cervical pain in patients affected with whiplash diseases, which is important for example in insurance-related investigations. By analyzing in detail this case study in a real scenario, we show how taking care of those particularities enables the generation of reliable predictive models in the field of healthcare. Using a database of 302 samples, we have generated several predictive models, including logistic regression, support vector machines, k-nearest neighbors, gradient boosting, decision trees, random forest, and neural network algorithms. The results show that it is possible to reliably predict the presence of cervical pain (accuracy, precision, and recall above 90%). We expect that the procedure proposed to apply ML techniques in the field of healthcare will help technologists, researchers, and clinicians to create more objective systems that provide support to objectify the diagnosis, improve test treatment efficacy, and save resources.

Keywords: data mining; data anonymization; health; cervical injury; neck pain; inertial sensors

1. Introduction

In the field of healthcare, the exponential increase in the data that health centers must produce and manage is significant. The need has arisen to develop procedures that make this process easier and that take advantage of all the data generated [1], detecting unknown and valuable information in health data [2]. Thus, the volume of data generated is such that its processing and analysis by traditional methods is too complex and overwhelming [3]. To tackle this challenge, data mining (DM) can play a key role, as it allows the discovery of patterns and trends in large amounts of complex data and the extraction of hidden information to help in making decisions that can improve the quality of the care processes [4–7]. It is closely linked with the scientific discipline in the field of artificial intelligence called machine learning (ML), which "employs a variety of statistical, probabilistic and optimization techniques that allow computers to learn from past examples and to detect hard-to-discern patterns from large, noisy or complex data sets" [8].

Consequently, ML is generating growing interest in the field of healthcare (e.g., see [9,10]) for relevant special issues related to this topic), mainly derived from its possible applications, such as assessing the effectiveness of treatments, detecting fraud and abuse in health insurance, managing healthcare, making lower-cost medical solutions available to the patients, detecting symptoms and diseases [11], discovering treatment patterns from electronic medical records [12], detecting groups of incidents [13], and identifying medical treatment methods [2,3]. Likewise, ML also presents health benefits: (1) a potential reduction in the time and effort required for diagnosis and treatment, (2) the ability to examine multiple areas simultaneously, (3) a decreased potential for human error, and (4) data that are accessible anytime and anywhere [14]. Besides, DM and ML are key in the path towards personalized medicine, where the goal is to customize treatments to the specifics of each individual [15–18].

However, to take full advantage of the benefits offered by ML in the field of healthcare, several considerations are necessary. Among others, the following aspects can be highlighted:

- The data have to be structured and organized in order to properly process and transform them into suitable variables, which is essential in the development of any pattern recognition software and a highly problem-dependent task [19,20].
- Moreover, the secure treatment and management of data acquires special relevance in the field of healthcare, where the privacy of the patient must be ensured. The management of sensitive data contrasts with other fields of application of ML (fraud detection, stock prediction, etc.), where anonymization treatments may be sometimes not necessary or critical. Therefore, a specific treatment involving the anonymization and categorization of the data must be performed in order to ensure the privacy of the patients [21]. Due to privacy policies, on certain occasions if a suitable anonymization of the data is not performed and/or the required authorizations to access some data are not obtained, the needed health studies cannot be carried out. Therefore, data availability should also be considered as a key factor [21].
- Another difference with other ML applications is the existence of different costs of failures; in the health area, the cost of a false negative (e.g., failing to detect that a patient has a specific disease) is usually much higher than the cost of a false positive (e.g., if a person is initially considered to have a disease that he/she does not really have, additional tests will be performed to rule this out, which may be costly but usually less harmful than failing to diagnose an existing disease).
- It should also be considered that ML hardly ever follows a linear sequence ending at the first attempt; instead, it is rather an iterative feedback process where the stages interact with each other. Furthermore, in the field of healthcare, where the flow of data is continuous and constant, it is reasonable to assume that the model can be designed to be a "learning" model that must be continuously updated to improve its predictions over time. Therefore, the different stages needed to generate a reliable predictive model should be properly structured, from the acquisition of raw data to decision-making, which is essential to achieve the effectiveness of the model [22].

All this motivates the need to define the particularities of the application of ML techniques in the field of healthcare, where different stages in the ML workflow must be correctly defined and structured. The proper application of ML techniques would be beneficial for the clinicians, researchers, developers, and designers involved in the field of health, where the management of information acquires a transcendental role. It would favor the design of new products and services for improving healthcare access [23], creating truly accessible technological solutions [24], and enhance the relationship between health systems and people by providing adequate services at the right time [23].

Based on the above, the aims of this study are the following: (1) to show and develop the particularities of applying ML techniques in the field of healthcare, detailing all the stages that comprise this process, from the acquisition of raw data to the decision-making derived from the predictive model generated; and (2) to demonstrate and show its practical application in a real use case. Specifically, the ML process is applied in a cervical pain assessment study with patients affected by whiplash pathologies derived from traffic accidents or other causes. This case study shows the proposed methodology in action to solve a specific relevant problem. Moreover, the applied procedure can be used as an ML application guide for other similar studies in the field of healthcare. We believe that the combination of the use case study and the machine learning methodology is a relevant contribution of this paper. We do not remain in the theoretical/methodological part only or limit our work to apply different machine learning algorithms and compare the results, as many other works do; instead, we describe the whole machine learning process, highlighting the aspects that are more relevant for our use case but at the same time providing a general framework that could be used in other health-related projects. In this way, we think that the paper could be relevant both as a specific case study and also as a reference and guideline for other similar projects.

The structure of the rest of this paper is as follows. In Section 2, we present the use case scenario studied and the clinical methods used for data collection. In Section 3, we describe the proposed procedure to develop predictive models for healthcare, illustrating each step with our work on the case study. In Section 4, we present an overall discussion of the proposal and the lessons learnt. Finally, in Section 5, we summarize our conclusions, the limitations of the study, and some ideas for future work.

2. Use Case Scenario and Clinical Methods

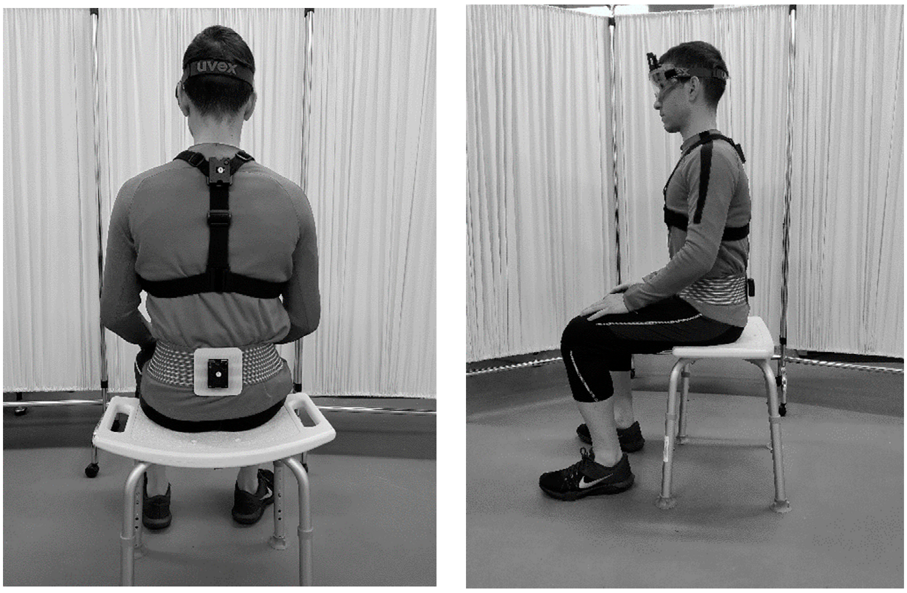
To illustrate the particularities of using ML techniques in the health area, a case study related to the detection of cervical pain is considered. The goal is to try to estimate automatically the presence of cervical pain, which can help to objectify a diagnosis and to clarify issues in case of insurance litigation. The selection of cervical pathology as a case study in this work is motivated by the fact that musculoskeletal disorders of the cervical spine have a high incidence and prevalence and are considered a public health problem, especially in developed countries [25,26]. Likewise, cervical injuries (usually due to whiplash after a traffic accident) are difficult to diagnose [26] because traumatic cervical spine injuries and their associated symptoms are diverse [25].

A real dataset was collected by evaluating the movement of the cervical spine in 151 patients (60 asymptomatic subjects, 42 with cervical pain resulting from a traffic accident, and 49 with neck discomfort due to other causes). Cervical movement assessment tests were performed by using an MH-sensor motion capture system [27,28] (see Figure 1). The participants performed a sequence of functional cervical Range of Motion (ROM) tests of the following movements: flexion-extension, rotation, and lateralization (Figure 2). The patients were collaborating subjects in order to avoid disturbances produced by non-collaborating subjects immersed in a judicial process with an insurance company [29]. The medical test was performed twice with each patient, giving a total of 302 samples.

Moreover, all the participants, who were either asymptomatic or had cervical pain, were also assessed with a clinical examination to verify that they met the inclusion criteria:

- age between 18 and 65 years;
- not immersed in a judicial process;
- no presence of surgery and/or cervical fracture.

The medical inspection, assessment by scales/clinical tests, and development of the clinical profile of the patients were conducted by clinicians. All the participants received information about the experiment and signed a consent agreement prior to the testing. The study received a favorable verdict from the Bioethics Committee of Aragón in Spain (CEICA) on 25 July 2017.



(**a**) (**b**)

Figure 1. Move Human (MH)-sensor motion capture system, cervical assessment. (**a**) Back view. (**b**) Lateral view.

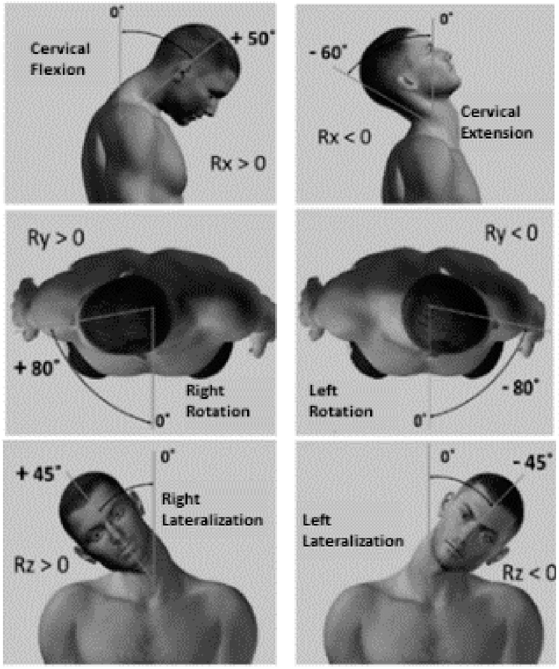


Figure 2. Cervical movements evaluated.

3. Proposed Procedure to Develop Predictive Models in Healthcare

A predictive model to support decision-making in the field of healthcare should be able to make predictions relative to relevant target clinical variables. The final goal is to deploy a system that can help in clinical decision-making (e.g., objectifying diagnoses, testing the efficacy of treatments, saving resources, providing suitable and customized treatments, etc.). The proposed procedure for continuous use in the field of healthcare is summarized and outlined in Figure 3.

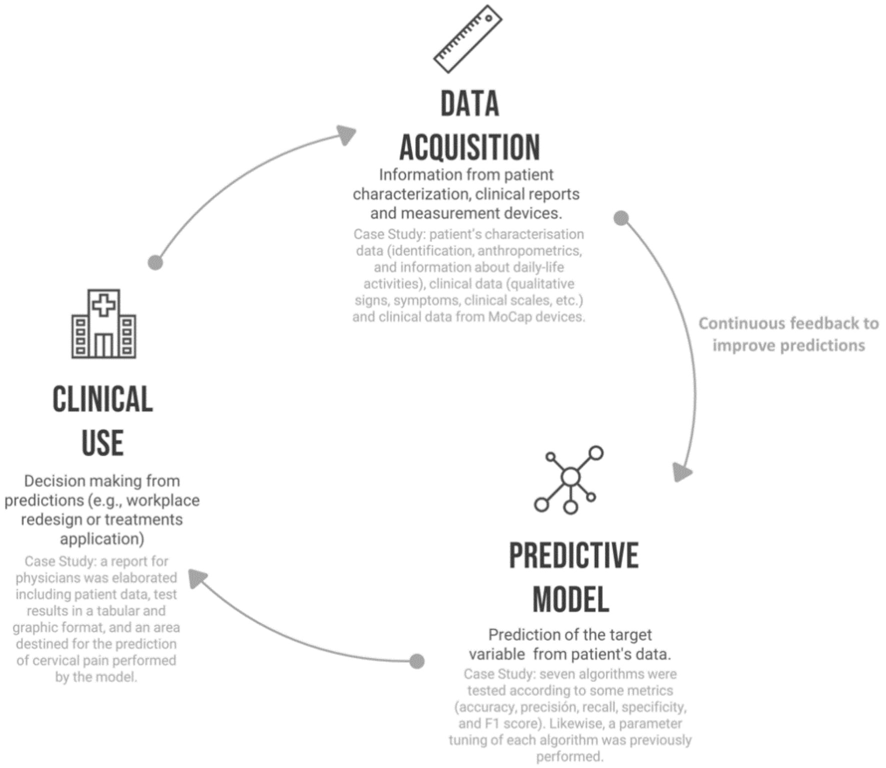

DATA ACQUISITION
Information from patient characterization, clinical reports and measurement devices.
Case Study: patient's characterisation data (identification, anthropometrics, and information about daily-life activities), clinical data (qualitative signs, symptoms, clinical scales, etc.) and clinical data from MoCap devices.
Continuous feedback to improve predictions
PREDICTIVE MODEL
Prediction of the target variable from patient's data.
Case Study: seven algorithms were tested according to some metrics (accuracy, precisión, recall, specificity, and F1 score). Likewise, a parameter tuning of each algorithm was previously performed.
CLINICAL USE
Decision making from predictions (e.g., workplace redesign or treatments application)
Case Study: a report for physicians was elaborated including patient data, test results in a tabular and graphic format, and an area destined for the prediction of cervical pain performed by the model.

Figure 3. Application of predictive models for clinical decision-making. Icons made by monkik, smashicons and mynamepong.

The complete ML process has been considered, with the particularities of its application in the healthcare area. It is based on seven stages that range from the definition of the target to the clinical use of the system, as shown in Figure 4. Each stage is explained in the following subsections. Besides, this paper is accompanied by electronic Supplementary Material to facilitate the understanding of the different stages of application in the case study considered; specifically, we provide sample data files obtained at the end of different stages of the process (anonymized datasets) and an example report that a health professional could obtain as the output of the process.

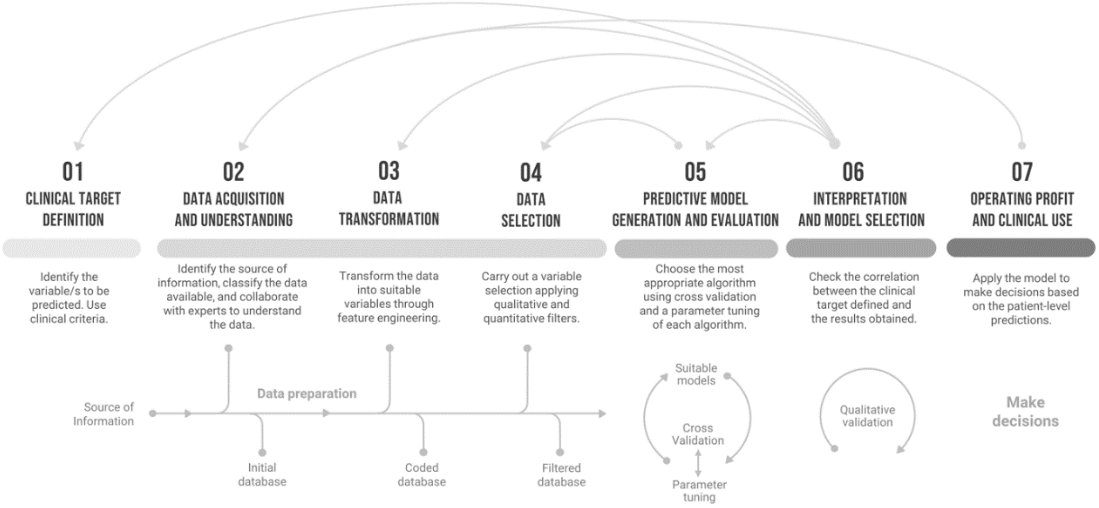

01
CLINICAL TARGET DEFINITION
Identify the variable/s to be predicted. Use clinical criteria.
02
DATA ACQUISITION AND UNDERSTANDING
Identify the source of information, classify the data available, and collaborate with experts to understand the data.
03
DATA TRANSFORMATION
Transform the data into suitable variables through feature engineering.
04
DATA SELECTION
Carry out a variable selection applying qualitative and quantitative filters.
05
PREDICTIVE MODEL GENERATION AND EVALUATION
Choose the most appropriate algorithm using cross validation and a parameter tuning of each algorithm.
06
INTERPRETATION AND MODEL SELECTION
Check the correlation between the clinical target defined and the results obtained.
07
OPERATING PROFIT AND CLINICAL USE
Apply the model to make decisions based on the patient-level predictions.
Source of Information
Data preparation
Initial database
Coded database
Filtered database
Suitable models
Cross Validation
Parameter tuning
Qualitative validation
Make decisions

Figure 4. Project management procedure proposed for the application of machine learning in healthcare.

In order to adapt and particularize the usual process of applying ML techniques in the healthcare field and develop the project management procedure, some outstanding studies such as [20,30], as well as the most widespread DM processes, such as knowledge discovery in databases (KDD) [31,32]; sample, explore, modify, model, and assess (SEMMA) [14,32]; and the cross-industry standard process for DM (CRISP-DM) [32–35], have been considered as a reference. Likewise, the project management procedure scheme proposed (shown in Figure 4) has been inspired by different outlines of clinical applications proposed by different authors [4,19,36–38] and adapted and extended according to our own experience and the work performed with clinicians in our case study and other related collaborations, such as the project "Mobile units for functional assessment of the musculoskeletal system" (CEICA reference of the project: OTRI-2019/0108) in collaboration with the hospital MAZ (Mutua de Accidentes de Zaragoza, Zaragoza, Spain), whose goal was to predict the degree of collaboration of patients in insurance litigation.

From the related proposals mentioned above, the CRISP-DM process has been our main inspiration to develop the project management procedure proposed in this paper. This is to be expected because CRISP-DM sets a general common framework that can be adapted to different scenarios. Thus, there are similarities between the six stages in CRISP-DM and our seven-stage proposal. For example, the CRISP-DM stage 6 "deployment" is closely related to our last stage, which is "operating profit and clinical use". As another example, stage 6 of CRISP-DM establishes that the creation of the model is not the end of the project and, similarly, in healthcare the knowledge provided to the clinician through the application of the predictive model is not the end of the process, since the system is continuously acquiring new data to improve the clinical performance. As the main difference between both procedures, we put more emphasis on data management aspects, since this is a key point in healthcare, and consider the whole process from the perspective of its application in a healthcare scenario. While only one stage for data management is considered in the CRISP-DM process (stage 3, "data preparation"), data management is the focus of three stages in our proposal (stage 2 "data acquisition and understanding", stage 3 "data transformation", and stage 4 "data selection").

The works mentioned in this section have inspired our proposal, which extends existing models by including a thorough analysis of all the data management challenges, as well as an illustration of each step through a real practical case study. Although the particularities of applying ML techniques in healthcare are exemplified in a specific case study, the procedure presented is flexible enough to adapt to any healthcare case.

3.1. Stage 1: Clinical Target Definition

In the first stage, the aim of the system is established—that is, the variables with clinical significance that the system should be able to predict are identified. Likewise, the final performance of the model and the statistical measures that will define its performance must also be defined. Measures such as the accuracy, precision, or recall are usual metrics used to assess the performance of a classification model, and metrics such as the Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE), to cite two examples can be used to evaluate the accuracy of a numeric prediction.

In predictive classification in a healthcare domain, it is usual that some metric must be highlighted in such a way that the model must always be generated with the aim to minimize it. This metric is usually the number of false negatives (affecting the recall metric). The reason for this is that, in healthcare, false negatives and false positives have no similar costs, which has always been an issue that clinicians have had to deal with [39]. Moreover, the clinical need must be identified (e.g., to classify a certain type of pathology, to predict a pattern or behavior, etc.). In addition, the sample size and viability of the project must be assessed prior to its realization [34].

Application to our Case Study

In our case study, the aim of the prediction model is to predict the presence of cervical pain (the target variable) in patients who have suffered whiplash or suffer from chronic cervical pathology. In our collected dataset, the cervical pain is a binary variable (the presence or absence of pain) which has been reported by the collaborating subjects, who were real patients undergoing an assessment process in a hospital.

Predicting the presence of cervical pain is of interest especially in the forensic field, as the incidence and prognosis of whiplash injury from motor vehicles is relevant to insurance litigations for pain and suffering [29]. The aim is to determine the presence or absence of pain with enough confidence to be able to aid clinicians to detect possible magnifications of the injury by the affected individuals and thus establish an unbiased compensation for the cervical pain [40]. It can help to identify and objectify pain in patients with a high degree of anxiety and in hypochondriac patients.

Without a loss of generality, the following target metrics have been determined for the purpose of this study, whose required threshold values have been stablished according to the criteria of clinical experts (for this particular case study, they considered that achieving this quality criteria would be enough for the system to be used as a decision-support system in production):

- Accuracy: greater than 85%.
- Precision: greater than 85%.
- Recall: greater than 90%.

The sample size is 302; although this is not a very large dataset, it contains a lot of variables with relevant information to characterize the presence of cervical pain during insurance litigation, which allows predicting the target variable, thus considering the project viable.

3.2. Stage 2: Data Acquisition and Understanding

The second stage implies identifying different sources that will allow access to the data necessary to feed the model, both in the initial phase of model design (initial training), as well as regarding a future continuous feedback when it reaches the operational stage of application in the field of healthcare. Likewise, the typology of these data will also be identified and selected. The level of accuracy that these sources of information can provide and the frequency of data collection (which may be determined by aspects such as the cost of the equipment needed, the availability of collaborating patients, etc.) will be considered in the choice [34].

In the field of healthcare, due to the diverse origins of data, their typology, their consistency, and even their veracity, the process of categorization of the data acquires special relevance for their correct structuring, understanding, and subsequent treatment.

When there are patient's personal data that make the identification of the patient possible, proper anonymization techniques must be applied. Apart from direct identification data (such as the name and last name of the patient, his/her history number, or his/her card ID), other data such as the age, nationality, height, weight, diagnosis, etc., can be used for indirect patient identification. In stage 3 "Data transformation" (see Section 3.3), certain techniques are presented to safeguard the privacy of patients and avoid the loss of useful information for the generation of the model, but these sensitive variables must be identified in this second phase.

Although there are classifications of information and data in the healthcare environment [30], alternative complementary classifications are proposed in this paper for a better understanding and structuring of the data. This has been motivated by the needs of medical staff, as well as by specialists of the medical legal/forensic field collaborating with us in our case study: a greater variety of classifications is of interest in order to include the perspectives of all the parties involved.

Clinical Data and Patient Data

Firstly, two types of data related to patients can be identified when we distinguish between clinical data and other data related to the patient:

- Patient characterization data. They are generally static data (although there may be small fluctuations in certain data values over large time intervals, such as in the case of the weight of the patient). They can be grouped in the following categories:

    - Identification data: name, age, gender, educational level, nationality, etc.
    - Temporary data: dates of control or highlighted clinical evolution, visits to the hospital, start and end of treatments, etc.
    - Anthropometric data: measurements of the size and proportions of the human body, such as the height, weight, percentage of fat and muscle, foot length, abdominal perimeter, etc., that usually require instrumentation to be obtained (scale, tape measure, etc.).
    - Daily life activities (DLA) data: data usually reported by the patient related to the habits and activities that he/she usually performs on a daily basis. In some cases, some of these data can be measured using wearables sensors or other devices deployed in smart homes.
- Clinical data: data of a medical nature that may require instrumentation and medical tests for their acquisition.

Data According to the Degree of Objectivity

Another possible classification is to categorize the data according to the degree of objectivity:

- Measures: objective data that do not require assessment by a clinician. These data are not affected by the reproducibility factor. Examples are test or clinical scales, test results, or data collected by medical instrumentation. Data recorded by sensor devices provide objective data on some measurable dimensions of the patient and can be of different types: motion capture (MoCap) sensors, surface electromyography (EMG), stabilometric platforms, dynamometers, etc.
- Assessed data: information that depends on the assessment of the clinician, such as diagnoses and treatments.
- Reported data: subjective information provided by the patient regarding his/her condition (perceived symptoms).

Data According to Clinical Considerations

We also present a classification that groups the collected data according to clinical considerations:

- Clinical profile: data about symptoms and clinical signs of the patient that can lead to a diagnosis by the clinician. We refer to symptoms when they are of subjective nature, reported by the patient, and to signs if they are objective and obtained by the clinician about the pathology. In addition, the signs can be qualitative (binary) or (discrete or continuous) quantitative (e.g., the temperature of a thermometer, image tests, other measurements, etc.).
- Treatment data: data about the treatment that the clinician has prescribed, such as the type of treatment, number of rehabilitation sessions, drugs received, surgery, etc.
- Clinical scales, tests, or surveys: data resulting from scales or validated and protocolized tests whose objective is to obtain objective information about the patient (e.g., the timed up and go test, the Unterberger test, psychological tests, etc.).
- Medical history: data concerning the patient's clinical history, ordered chronologically (e.g., first hospital visit, imaging test for diagnosis, treatment administration after diagnosis, etc.).

Data According to their Data Types

Finally, the different data variables can be grouped according to their types, independently of the specifics of the health area. For example, we could consider:

- Qualitative variables: also called categorical variables, they are variables that are not numerical. They describe data that fit into categories (e.g., educational level, the level of development of a disease, the level of invasiveness of a treatment, etc.).
- Quantitative variables: also called measurement or numerical variables, they represent quantities of different nature. They can be divided into discrete variables that can only take a finite number of values (e.g., the number of rehabilitation sessions, score on a clinical scale, age, etc.),

and continuous variables, which can take values in an infinite/continuous range of possible values (e.g., the temperature of a thermometer, weight, Body Mass Index (BMI), etc.).

- Textual data: data that are directly collected in text format, such as handwritten annotations in medical histories.

In this stage, data profiling [41] and cleaning [42] must be applied to detect potential problems and, if possible, fix them. By categorizing the data, using one of the proposed classifications (or another one that could be useful for a specific use case) independently, or several of them at the same time, the process of the understanding and assimilation of the data available and their scope is facilitated. This step must be carried out to obtain a preliminary database, containing the data collected, that will be called the initial database. For this task, having the support of both a clinical expert and a technologist is recommended. In some cases, when the amount of data to handle is large or coming from different data sources, a data warehouse can be created to integrate all the information and allow the easy and efficient analysis of the data stored [43,44].

During data collection for classification tasks, it is also important to collect enough instances/samples to represent in an appropriate way the different classes that must be predicted. In case there is imbalance regarding the number of samples in the different target classes, this should be identified as part of the data profiling, and some strategies could be applied to deal with this issue [45,46] (e.g., to try to prevent the majority class from dominating the predictions in a harmful way).

Application to Our Case Study

In this case study, the initial database was prepared according to the final goal, which was predicting the presence/absence of cervical pain. The categorization of the data was jointly agreed by physicians and technologists, considering the application of this research in the legal field and the degree of objectivity of the data. According to the four classifications presented, the data from our case study could be classified as shown in Table 1, following the exposed criteria:

Table 1. Possible classifications of the case study data.

1. Clinical Data and Patient Data	2. Data According to the Degree of Objectivity	3. Data According to the Clinical Considerations	4. Data According to Their Data Types
Patient characterization data: - Identification: name, age, gender, educational level, etc. - Temporary: accident date, visits to the hospital, date of the range of motion (ROM) test, etc. - Anthropometric: weight, height, body mas index, foot length, etc. - Daily life activities: physical activity intensity, workplace, etc.	Measures: all the data from the MoCap sensors or Whiplash scale (WDQ).	Clinical profile: - Symptoms: pain periodicity, feeling of instability, limitation of mobility, etc. - Signs: contracture, limitation of mobility, spinal column alterations, etc.	Qualitative variables: educational level, limitation of mobility, contracture, etc.
Clinical data: feeling of instability, surgery, all the data from the MoCap sensors, etc.	Assessed data: contracture, limitation of mobility, Jackson contraction, etc.	Treatment: n/a.	Quantitative variables: - Discrete: WDQ, age, etc. - Continuous: all the data from the MoCap sensors, weight, height, etc.
	Reported data: pain periodicity, feeling of instability, etc.	Clinical scales or tests: WDQ.	Textual data: n/a.
		Medical history: accident date, visits to the hospital, etc.	

The classifications shown in Table 1 illustrate different useful perspectives of the data collected in the case study. Specifically, and considering the final goal of objectively predicting the presence of cervical pain, the classification that best suited the point of view of the health professionals participating in our case study was the following (a combination of the first two classification approaches described):

- Patient characterization data: the data measured and reported, such as the identification information (e.g., ID, age, gender, etc.), anthropometrics (e.g., height, weight, BMI, etc.), and data relative to the activities of daily life (e.g., weekly physical activity and its intensity, workplace, etc.).
- Assessed and reported data: such as the characterization of a cervical accident (e.g., the time since the accident, type of impact of the traffic accident, position of the head, type of effect, etc.), qualitative signs (e.g., column alterations and contractures), clinical scales (e.g., whiplash scale—WDQ), and symptoms (e.g., instability, limitation of mobility, etc.).
- Measured data: such as data from a cervical assessment test with MoCap or from each movement studied, for example the angles reached and the speeds of the movements (e.g., maximum values, minimum values, average values, etc.).

For the purposes of medical legal/forensic assessments, a classification according to the degree of objectivity of the clinical data (assessed and reported) is of interest. Thanks to the understanding of an expert technologist in insurance litigation, it has been possible to identify objective information in the case study, such as the presence of a contracture, column alterations, the WDQ, the characterization of the traffic accident (in case such an event had occurred), etc.

The initial database is presented as Supplementary Material File S1, which collects the dataset that has been considered for the development of the model.

3.3. Stage 3: Data Transformation

Once the data to be considered in the initial database have been selected, certain transformations of the data must be performed in order to handle empty values, perform data transformations to define the required variables and adapt them to the required format, and ensure the anonymization of the data.

The information obtained from the different sources can correspond to variables already defined and structured or to raw data, and it can be presented as numerical, text, curves, images, etc. [47]. Transforming raw data into the format required for the application of specific ML algorithms is a common pre-processing step to be performed, and it could also be useful because the volume of data to be handled could be reduced, and its predictive power could be significantly increased, making it possible to have a significantly lower volume of data to achieve a reliable and stable predictive model [5,19]. To be exploited by traditional DM and ML algorithms, textual data can be transformed into structured variables and different text mining techniques can be applied [48,49]. Besides, depending on the purpose, unsupervised learning approaches can be applied on the texts—for example, for dimensionality reduction (e.g., using the Self-Organizing Map (SOM) method [50]), for clustering documents according to their similarity, or for discovering topics in documents (e.g., probabilistic topic modeling by using Latent Dirichlet allocation) [51].

This stage of transformation of raw data is known as feature engineering and is performed prior to modelling [52]. In some cases, several data mining techniques can be applied to extract features from raw data. For example, Principal Component Analysis (PCA), like the SOM method mentioned above, is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets; it can be used to tackle the problem of high dimensionality that appears in some projects when the number of variables is excessively high compared to the total number of samples (see Section 3.5) [39]. In the field of healthcare, the following types of transformations can be highlighted:

- Texts that become “positive or negative opinions”, concepts, dates, values, etc., through the application of text mining techniques [53,54].

- Images of medical tests that are converted into binary variables related to a pathology or other binary representations.
- Curves or other graphical representations from which information can be extracted as statistical variables, such as the mean, standard deviation, skewness, quartiles, etc.
- Different imputation techniques [18,55–57] that can be applied in order to fill empty values, either by means of interpolation or by using other procedures. Alternatively, some records may need to be discarded (e.g., if several key data values are missing).
- The potential addition or creation of variables based on the knowledge of clinical experts and technologists, either as a combination of existing variables or based on experience acquired in similar studies [30].
- Data anonymization, applied in order to preserve the privacy of patients [58]. The trade-off between privacy and information loss should be considered. A detailed analysis of data anonymization techniques for health data is out of the scope of this work but, for illustration purposes, some examples of transformations that can be applied to guarantee the privacy of sensitive information are:
    - Transformation of continuous variables (e.g., age, size, weight, income, etc.) into ordinal variables by defining suitable ranges. Normalization (e.g., min-max normalization) or standardization (z-score scaling) techniques could also be applied to transform the quantitative variables into variables in the range from 0 to 1.
    - Transformation of qualitative variables (e.g., diagnosis, treatment, education level, etc.), that could be classified, according to a scale, into ordinal variables (e.g., severity of diagnosis, treatment risks, range of studies (“high school”, ”BS”, ”MS”, “PhD”), etc.). In this way, these variables cannot be associated with a specific patient, thus preserving the patient’s privacy.
    - Transformation of qualitative variables (e.g., nationality, physical description, address, etc.), that could not be classified according to a scale into groups (e.g., by continents or country grouping, by groups according to a general description, by postal code, etc.) so that these variables cannot be associated with a specific patient.
    - The previous anonymization techniques are examples of the generalization of attribute values. Other possible techniques and privacy-preservation methodologies include adding noise [59], k-anonymity [60], differential privacy [61], etc.

Another important aspect to mention here is the need for the normalization of quantitative attributes; if we have quantitative variables with very different scales, the variables that can take larger values could end up dominating others when learning a predictive model. This is unsuitable because it mistakenly attributes more importance to those variables just because their usual values are higher than the values of other variables.

After applying the specified transformations, the initial database is transformed to an encoded (modified) database, which includes all the available information in the format of multiple variables.

Application to Our Case Study

In the case of the cervical pain assessment study, this process consisted of transforming the data from the MoCap sensors into variables (i.e., mean, deviation, maximum, minimum, etc., of the range of the movement) and transforming the clinical and characterization data of the patients into variables considering sensitive variables and their anonymization. Some ranges of variables were associated with a numeric coding to improve the anonymization process and the generation of the predictive model. The process was carried out jointly by a clinical expert and a technologist. As in our case study we have transformed all the quantitative variables into discrete variables, no additional normalization was needed. The following are examples of the transformations performed (the numeric coding used is indicated in brackets):

- Age transformation in a three-tier classification: under 30 (0), between 30 and 55 (1), and over 55 (2).
- Transformation of the level of weekly exercise into a classification of three levels as a function of the time required and the intensity of the exercise: slight (0), moderate (1), and intense (2).
- Transformation of the level of studies in a classification of three levels with their assigned numerical coding: basic/high school (0), medium/bachelor (1), and superior/university studies (2).
- Weight transformation in a three-tier classification: less than 60 Kg (0), between 60 and 85 Kg (1), and greater than 85 Kg (2).
- Height transformation in a three-level classification: less than 158 cm (0), between 158 and 185 cm (1), and greater than 185 cm (2).
- Transformation of the body mass index (BMI) into a three-tier classification: under 24 (0), between 24 and 30 (1), and over 30 (2).
- Grouping of the cervical pain periodicity to create a variable of three levels: sporadic (0), discontinuous (1), and frequent (2).

The completeness and quality of the data recorded is also a key aspect in any ML pipeline, and particularly in the health area. In our case study, some variables were initially incomplete, due to the lack of collected data from certain patients (five patients). These incomplete data were related to specific features (e.g., the workplace of the patient, his/her age, the dominant side of his/her body, the date of the traffic accident, etc.), and were later collected by a telephone call.

The encoded database is presented as Supplementary Material File S2, where the variables are obtained after the transformations that are carried out from the initial database are collected and the anonymized variables are highlighted (see Variable_View).

3.4. Stage 4: Data Selection

The next stage is to filter the encoded database obtained in the previous phase and select the most useful variables, applying different filters in a way that will lead to obtaining a filtered database, which will be the basis of the predictive model. Possible successive filters to be used include the following:

1. Filters due to ethical and legal issues: Discard personal or private variables and those that are unimportant for the purpose of the predictive model, such as names, clinical history numbers, telephone numbers, and addresses. The filtered database must be anonymous with respect to existing regulations on the protection of personal data, so a previous anonymization process becomes essential in order to keep as much important data as possible. Notice that during the previous step (data transformation, see Section 3.3), some data are transformed to increase privacy; in this step, privacy might need to be further increased by not selecting some sensitive data in case those data have not been properly transformed previously or in the case of other sensitive data that are irrelevant for predictions.
2. Manual selection: Screening based on the needs set by the target of the prediction, removing outliers or variables with a lot of missing data [5]. It is highly recommended that this filtering be conducted by an expert in the healthcare field.
3. Automated attribute selection: specific software and algorithms can be used for filtering, calculating the gain ratio for each of the variables and rejecting those with low predictive power—for example, using regression techniques.

Application to Our Case Study

In our case study, the encoded database initially included 230 variables in the initial dataset (the "Coded database"), which were reduced to 28 after the following consecutive filtering steps (see Table 2):

1. The removal of variables related to personal and ethical data not anonymized previously and with no predictive power. The variables name, surname, telephone number, address, and email were removed in this filtering.
2. Manual filtering performed by physicians and technologists, corresponding to non-objective or inappropriate variables in the medical legal/forensic field (e.g., the sensation of instability, mobility limitation, pain periodicity, etc.), as well as variables with missing data that could not be completed (e.g., the position of the head and type of effect in case of a traffic accident, intensity of work, etc.). In this filtering, 74 variables were removed according to the criteria indicated.
3. Finally, a filtering was applied based on the gain ratio of the variables. We used the IBM SPSS modeler software [62] (v. 18), discarding 123 variables with low predictive power (we selected those with a gain ratio higher than 0.95 out of 1). Variables such as the average angle, standard deviation, complementary angle, weight, height, etc., were selected. The selection of these 28 variables is consistent with the target variable (the presence or absence of pain) and its associated requirements, since it is a desirable situation for clinicians that these variables are objective, represent the main cervical movements, and correspond to clinical data objectively acquired by the clinicians.

Table 2. Final variables considered in the case study after feature selection.

Patient Characterization Data	**Gender**	**Age**	**Educational Level**	
Clinical data: assessed and reported data	Contracture	Traffic accident	Spinal column alterations	WDQ
Clinical data: data measured with sensors	Mean Speed [°/s] in:	Flex.-Ext.	Rotation	Lateralization
	Max Speed [°/s] in:	Flexion	Right Rotation	Right Lateral
		Extension	Left Rotation	Left Lateral
	Max Angle [°] in:	Flexion	Right Rotation	Right Lateral
		Extension	Left Rotation	Left Lateral
	Total Range [°] in:	Flex.-Ext.	Rotation	Lateralization
	Total Length [°] in:	Flex.-Ext.	Rotation	Lateralization

Table 2 shows the 28 final variables of the filtered database; the detailed information of each variable, with the different associated values for the different data instances, is included as Supplementary Material File S3.

3.5. Stage 5: Predictive Model Generation and Evaluation

In the next stage, a predictive model according to the established objective is designed based on the filtered database obtained in the previous stage. To do this, we must select those algorithms that are considered viable for the specific clinical project, such as decision trees, neural networks, or support vector machines (SVM), among others [63]. If the volume of data is very high, the use of specific support software can facilitate selecting a suitable algorithm by performing iterations before generating the full predictive model. For example, the IBM SPSS modeler classifier node (v. 18) can create several models and then compare them to select the best approach for a particular analysis.

In this stage, the performance of the selected algorithms should be tested to choose the most convenient one to implement in the predictive model. To evaluate their stability and effectiveness, different cross-validation approaches can be considered [19,30,64–66]. The simplest method consists of separating the sample into two sub-samples: one to train the model and another one to test it (holdout method). Other more advanced methods include dividing the sample into k sub-samples (k-fold cross validation), stratifying the sample with the same percentage of each class (stratified k-fold cross validation), or even making as many combinations as the number of data instances (leave-one-out cross validation). Furthermore, a validation set could be used (besides the "test set" and the "training set"), which is a set of examples used to tune the parameters of a classifier [67]. This is useful because the performance of the selected algorithms can be improved through parameter tuning, which consists of varying values of parameters of the algorithms in order to find the most suitable configuration for each of them. Once the suitable parameter configuration for each algorithm is selected, the performance of the different algorithms can be compared.

It must be stressed that the most appropriate prediction method depends on the data. Besides, overfitting (a phenomenon that occurs when the adjustment of the model to the training data is too strong and, as a consequence, finds difficulties in obtaining suitable conclusions about unobserved data) should be avoided, as this would lead to a model that will only be able to make predictions for the data with which it has been trained [52]. There are several techniques to deal with this, such as regularization (smoothing the models), data augmentation (increasing the training data), early stopping, etc. As an example, early stopping implies that the training process of the model must be stopped before overfitting [30,68,69], that is, before the model adjusts too much to the training data (i.e., before the performance of model gets worse on the validation data).

Over-parametrization is a recurrent problem in the application of ML techniques in healthcare, where there are cases where the ratio between variables and data is very high and overfitting effects can occur. The opposite can also happen when the volume of data is very high, but the number of variables is excessively reduced and/or has little predictive power. Therefore, if needed, depending on the results obtained we could come back to Stage 4 "Data Selection" to reduce the number of variables. As a guideline, to avoid over-parametrization the 1 to 10 ratio between variables (attributes or features) and data (number of instances) should not be exceeded [20,52]. If there was a smaller ratio between the variables and data, a more exhaustive screening of variables would have to be carried or a larger sample would have to be obtained. In terms of classification, having an excessive number of variables will lead to not solving the problem or not achieving the proposed objective because the model will only be able to classify the training data correctly [70].

To select the most suitable algorithm for the project, it is necessary to follow an organization strategy, storing each previous version and modification of the project [52]. In this way, the results are contrasted (training data, validation, errors, etc.). After the selection of the most effective algorithm, we will be able to generate the definitive predictive model that incorporates the already-existing filtered database. If the predictive model were evaluated positively, a continuous learning process could begin by receiving periodic information from the assessment tests performed on future patients, as shown in Figure 3.

Application to Our Case Study

In view of the target of our clinical study to classify the presence of cervical pain, only supervised learning algorithms must be selected, dismissing unsupervised learning algorithms. The algorithms selected in our study were logistic regression [71], decision trees [72], random forests [73], SVM [73], neural networks (MLP neural networks) [74], k-Nearest Neighbors (KNN) [73], and Gradient Boosting Algorithm (GBA) [75]. All these are popular supervised machine learning approaches. The main parameters selected to perform the parameter tuning of those algorithms are shown in Table 3. In our experimental evaluation, we combined the parameters of each algorithm presented in Table 3 using the software tool Weka [76] (v. 3.8); specifically, for the parameter tuning we used the Weka Experimenter user interface.

Table 3. ML approaches considered and their main parameters.

Approach	Main Parameters
Logistic regression	Ridge value in the log-likelihood: from 10^{-4} to 10^{-12} (parameter change every 10^{-2}).
Decision tree (C4 pruned)	Number of instances per leaf: 3/5/10/15/20. Confidence factor for pruning (Conf.): 0.15/0.25/0.35.
Random forest	Maximum depth of the tree: 3/4/5. Number of trees: 25/50/100/200.
Support vector machine (SVM)	Tolerance: 10^{-3}. Kernel function: radial basis function (RBF). Epsilon for round-off error: 10^{-12}. Complexity (C): 0.25/0.5/1/2/4. Gamma (kernel width): 0.01/0.25/0.5/1/2.
Neural Network (MLP neural network)	Type: multilayer perceptron (MLP). Learning Rate (LR, the amount the weights are updated): 0.2/0.3/0.4/0.5. Momentum (Mom., applied to the weights during updating): 0.1/0.2/0.3. Number of epochs for training: 500. Number of hidden layers: 15. Auto-built option in Weka set to true.
K-Nearest Neighbors (KNN)	Number of Neighbors (K): 1/3/5/7/9/11/13/15/20. Distance function: Euclidean distance, Manhattan distance.
Gradient Boosting Algorithm (GBA)	Iterations (Iter.): 10/20/50/100. Weight threshold (W.T.): 50/100/200. AdaBoost implementation provided by Weka.

In our case study, the predictive models generated by applying the previously selected algorithms are shown in Table 3. As mentioned previously, the DM software used in this study was Weka (v. 3.8). Considering the current availability of data for our case study, the performance of each parameter configuration of each algorithm was tested using a validation set which is the same as the test set; using the same set for validation and testing is not the ideal situation, but we decided not to reserve a part of the available dataset for validation because of the moderate size of our sample. We considered the accuracy metric (i.e., the percentage of instances correctly classified) for determining the most suitable algorithm configuration. Figure 5 shows the performance of the algorithms obtained during parameter tuning using a k-fold cross validation (k = 10).

Consequently, the parameter configuration selected for each algorithm is shown in the first row of Table 4. The effectiveness of the models generated by k-fold cross validation (k = 10) was evaluated. The following metrics were computed in order to determine the most suitable algorithms (see Table 4): the accuracy (the percentage of instances correctly classified), the precision (the percentage of patients with pain correctly classified over all the patients labelled by the algorithms as patients with pain), the recall/sensitivity (the percentage of patients with pain correctly classified over all the patients with real pain), the specificity (the percentage of healthy patients correctly classified over all the patients who are really healthy), and the F1-score (the harmonic average of the precision and recall).

We have noticed using the software Weka, which provides the attribute weights of each model in the results display section, that the variables with greater predictive power in all the predictive models evaluated are the following: maximum speed in all the movements, the existence of a traffic accident, and the presence of a contracture. In our case study, there is no indication of over-parametrization; as described in Section 3.4, we have a sample of 302 instances and 28 selected variables, which complies with the 1 to 10 ratio between the variables and data.

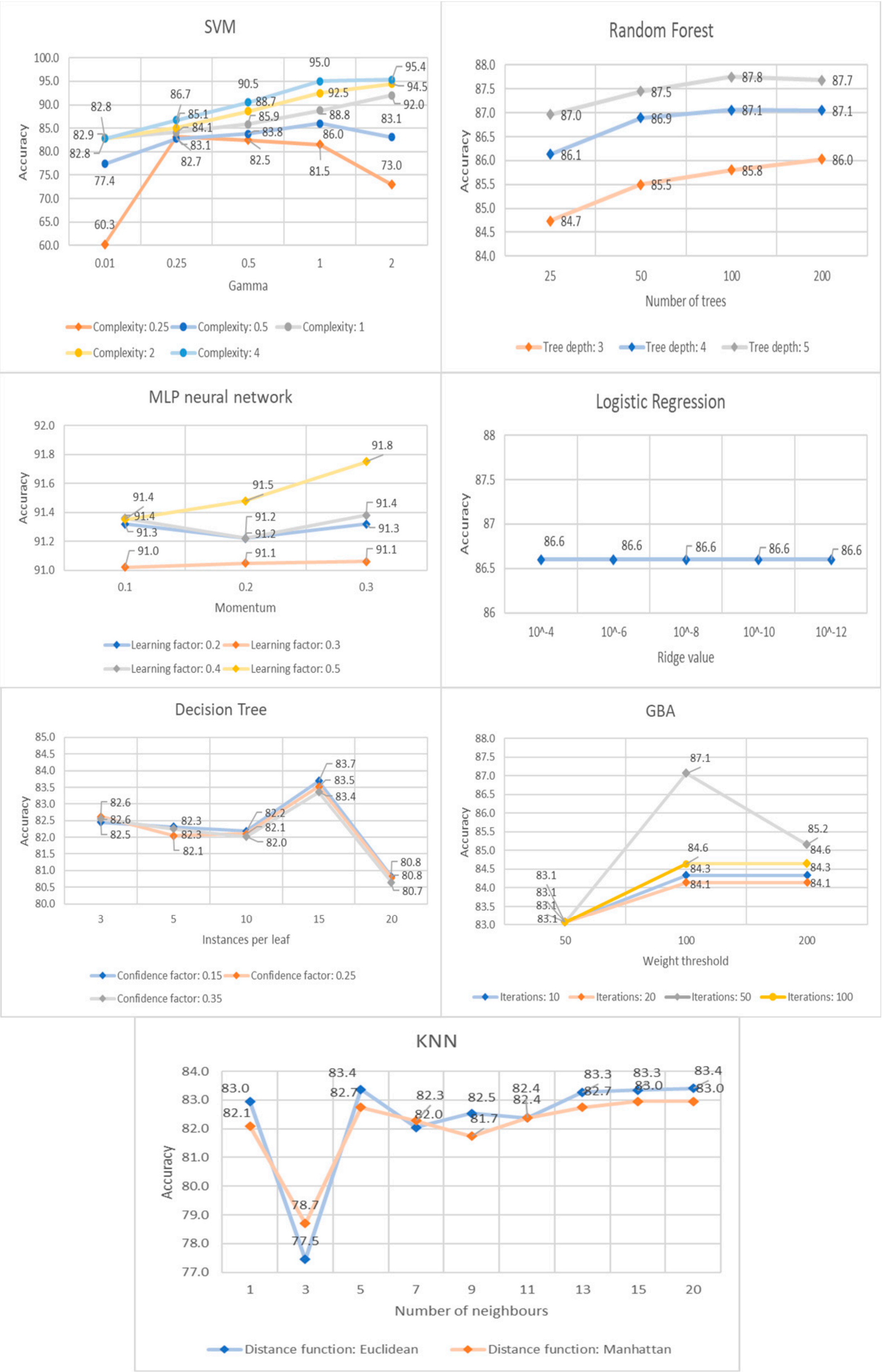


Figure 5. Parameter tuning of the seven algorithms selected.

Table 4. Metrics considered for algorithm selection and the results obtained for each algorithm.

	Logistic Regression	SVM	Decision Tree	Random Forest	MLP Neural Network	KNN	GBA
Parameter Selection	Ridge: 10^{-8}	C: 4; Gamma: 2	Instances per leaf: 15; Conf.: 0.15	Trees: 200 Depth: 5	LR: 0.5; Mom.: 0.3	K: 20; Euclidean	Iterat.: 50 W.T.:100
Accuracy	86.6%	95.4%	83.7%	87.7%	91.8%	83.4%	87.1%
Precision	88.6%	95.4%	86.9%	86.1%	93%	83.5%	87.1%
Recall/Sensitivity	89.6%	97.8%	84.1%	92.3%	92.5%	91.2%	92.9%
Specificity	82.5%	91.7%	80.8%	76.6%	90.5%	71.7%	78.3%
F1 Score	89.1%	95.3%	85.5%	88.9%	92.8%	83.2%	86.9%

3.6. Stage 6: Interpretation of Results and Model Selection

Once a statistically acceptable predictive model has been created, it can be deployed to be used in production, since it can provide predictions with good accuracy to support decision-making (Figure 4, stage 6). However, for the exploitation of the model, it is necessary first to evaluate the degree of correlation between the results obtained and the previously defined clinical target (in our case study, the prediction of cervical pain). Besides, some models (e.g., the tree-based classifiers) could be explored to analyze which particular features are the most decisive factors in the classification model.

If the evaluation of the correlation does not yield satisfactory results—that is, if the consonance of the results with the target is not achieved or the model does not have enough predictive power according to the minimum goals established initially—the previous stages must be repeated with the objective of improving and optimizing the process according to the target initially set, leading to an iterative process (see Figure 4) [20,77]. There can be several causes of low correlation or poor predictive power:

- The target is difficult to achieve or too complex;
- Patients are inadequately characterized;
- The sample is insufficient;
- The data include variables that are not necessary (e.g., irrelevant variables, confounding factors, or redundant variables) or do not include those that are;
- Problems exist in the predictive model generation stage regarding the selected algorithm [78], overfitting, or over-parameterization.

For effective deployment, not only the accuracy of the models but also the resources must be considered. For example, the scalability requirements are key when you have high volumes of data to avoid very long training times, lack of memory, etc. This is important in terms of productivity, portability, cost reduction, the minimization of staff involvement, etc. In cases where large volumes of data have been collected for many years (cancer studies, studies of discharge from the hospital and sick leaves, etc.), scalability acquires a transcendental role [30].

Application to Our Case Study

Based on the results collected in Table 4, a quite good performance of all the algorithms can be observed for most metrics. Considering the large number of variables available initially, the rigor with which all the data were obtained by the multidisciplinary team, and the adequate selection of variables carried out in previous phases (choosing the objective variables and with a greater predictive power), a suitable prediction performance was initially expected.

Considering the minimal target values of the statistical measures established in Stage 1 "Clinical target definition", several algorithms fulfil the requirements in terms of the target metrics:

- Accuracy (>85%): logistic regression, SVM, random forest, MLP neural network, and GBA.
- Precision (>85%): logistic regression, SVM, decision tree random forest, MLP neural network, and GBA.

- Recall (>90%): SVM, random forest, MLP neural network, kNN, and GBA.

According to the results, the four algorithms that fulfil the requirements are SVM, random forest, the MLP neural network, and GBA. From all these algorithms, SVM achieves the best results in all the metrics considered, and therefore it could be the model selected for production.

3.7. Stage 7: Operating Profit and Clinical Use

The last stage (Figure 4, stage 7) concerns the adaptation and organization of the acquired knowledge and the predictive capacity of the system to make it accessible to the physician [47]. At this point, it is important to emphasize the key role played by the expert medical professionals (final decision makers) in interpreting the results, in determining whether certain patterns observed make medical sense and are relevant, or in clearly distinguishing between correlation and causality (as studied for different heath topics [79,80]). The "intelligence" provided is not intended to replace the physician but to advise and guide his/her decisions, which will always prevail [20].

Application to Our Case Study

The results obtained directly by the model may involve difficulties when interpreted by the physicians. That is why the information presented to them must be intuitive, simple, and easily interpretable. In this regard, a concise graphic and clear report, where the results of the test and the prediction of the pathology/target variable to be predicted are presented, can help the clinical to more easily interpret the test performed and the results of the predictive model. This paper is accompanied by a report example as Supplementary Material File S4, showing a possible example of a report for our case study that includes patient data, test results in a tabular and graphic format, and an area showing the prediction of cervical pain obtained by the model.

Once the model enters production, it will be possible to add data regarding new patients both with or without cervical pain and verified diagnosis, which would increase the sample and thus the predictive power.

4. Discussion and Lessons Learnt

In this paper, the particularities of applying ML techniques in the field of healthcare are shown, developing all the stages that comprise it, to generate reliable and stable models. It has been exemplified through a case study of cervical pain evaluation, where we have been able to predict the presence of cervical pain with accuracy, precision, and recall above 85% with the approaches based on SVM, random forest, MLP neural networks, and GBA.

In order to clarify and structure the knowledge acquired during the development of the current study, a summary of some key aspects and lessons learnt regarding DM and ML in the field of healthcare is shown in Table 5. Every key aspect has been categorized in a general classification, followed by a description, the real situation exemplified in our case study, and some important related references.

Table 5. Summary of the key aspects and lessons learnt.

Category	Key Aspect	Description	Case Study	Sample References
Clinical target	Proper selection	Clinical target definition according to the aims and clinical needs. This facilitates the subsequent selection of data.	Presence of cervical pain. Only collaborating subjects and objective variables were selected.	[34,40]
	Definition of statistical measures	Minimum metrics to be fulfilled by the model according to the clinical target. Metrics are checked in stage 6 (interpretation of results) after the predictive model generation.	Performance required: accuracy: greater than 85%; precision: greater than 85%; recall: greater than 90%.	[39]

Table 5. *Cont.*

Category	Key Aspect	Description	Case Study	Sample References
Data	Identification and understanding	Diversity of the origins of data in healthcare (regarding typology, consistency, and veracity) and need to correctly understand the data. A health expert is required in this stage.	Prior to the field work, relevant clinical information was identified and the tests to perform were determined.	[34]
	Clear and concise structure	Categorization of data using appropriate variables/features, applying classifications motivated by medical needs. This is essential to carry out an adequate analysis of the information.	Data classification motivated by clinical staff and forensic experts: patient characterization data, assessed and reported clinical data, measured clinical data.	[30]
	Data transformations in healthcare	Feature engineering. Reducing the raw data to be handled and adapting them to the required format in order to increase the predictive power.	Variables such as the age, level of studies, weight, height, etc., were transformed into discrete variables. Ranges of variables were associated with a numeric code.	[5,19,47,52]
	Anonymization	Preservation of sensitive patient data by transforming values of data variables into scales or groups, thus avoiding patient identification. This is a key aspect in healthcare data management.	No quantitative variables remained after the anonymization process (through transformation into discrete variables and the removal of identifying attributes) that could be associated with patients.	[58]
	Selection	After data transformation, the selection of variables applying a filter according to the target:ethical and legal issues, manual selection, automated attribute selection.	The volume of data in the case study was reduced from 230 variables to 28 after applying the three successive aforementioned filters.	[19,21]
	Normalization	Normalization of quantitative attributes avoiding situations where variables that can take larger values could end up dominating others.	No quantitative variables remained after anonymization.	[55–57]
	Completeness	Completeness and quality of the data recorded as a key aspect in the health area.	Incomplete data from 5 patients related to specific features were collected through a telephone call.	[81]
	Over-parametrisation	Need not to exceed the 1 to 10 ratio between variables and data to avoid overfitting. The dimensionality is an issue in studies with a high number of variables compared to the total number of samples.	This was a real issue in our case study because of the volume of data provided by sensors. A sample of 302 and 28 variables was finally selected, which complies with the 1 to 10 ratio.	[20]

Table 5. *Cont.*

Category	Key Aspect	Description	Case Study	Sample References
	Scalability	Support for handling large amounts of data (efficient and effective collection, storage, management, and exploitation). Depending on the project duration, and especially if it is intended to have an adaptive character (projects with data collected for many years), scalability is a key issue to consider.	The current project is still in an initial stage, with no large-scale deployment. No scalability problems have been detected.	[30]
Predictive model	High recall and relatively high precision	Minimization of the number of false negatives (increasing recall). This is a key goal in healthcare, since false negatives and false positives have no similar costs in this area. The precision should also be suitable, as a high number of false positives would lead to false alarms, the performance of needless procedures, and increasing costs and discomfort for the patients.	The selected algorithms (SVM, random forest, MLP neural network, and GBA) have recall >90%.	[39]
Project work procedure	Multidisciplinary work as a key point	Composing teams involving technical people and diverse health professionals. This is required, but not always possible. Insufficient collaboration could be diminished by applying the stages assigned to each of the professionals in a concise and structured way.	There was interaction between professionals in almost all the stages. Nevertheless, more interventions could be encouraged because clinical experts were not present in stage 5.	[34,52]
	Continuous data collection	Improvement of the model performance thanks to a continuous learning process.	There is an intention to improve the current system by incorporating data of new collaborating patients.	[30]
MoCap	Sensors/devices in healthcare	Complementary objective tests to help physicians.	We expect the applicability of the proposal in the forensic field as an objective system of application to aid in judicial processes.	[19,82]

Data structuring. A clear and concise structuring of the data is essential to carry out an adequate analysis of the information as well as to make this information really useful for the purposes of the predictive model. In our case study, and prior to the field work, the relevant clinical information was identified and the tests to perform were determined (cervical ROM in three different planes with a MoCap system of inertial sensors), so that its processing and subsequent structuring were easier. It is essential to accurately structure the available information in a suitable way (using appropriate variables/features) when working with large volumes of data, and to classify the different variables in different groups (using appropriate categories) to facilitate access (for all the parties involved) in a more effective and useful way. Data management is so important that the preparation of the data covers three of the seven stages of the methodology (stages 2, 3, and 4).

Selection of variables. Likewise, the selection of the most adequate information to predict a certain characteristic, as well as its transformation in terms of variables, is essential. In relation to this adaptation, sensitive patient data must be anonymized for their use in the generation of a predictive model [19,21]. Although converting continuous predictors to discrete variables (specifically binary variables) is not always recommended [20], the necessary transformation of data for privacy reasons in the field of health conditions the stage of data transformation. The reduction in the volume of data in the case study was from 230 variables to 28 due to the large number of variables provided by the inertial sensors and the clinical data. The variable reduction applied (based on the gain ratio of the variables) after the corresponding selection of the data for ethical and legal issues and the screening made by an expert was necessary to fulfil the 1 to 10 ratio between the variables and data. This key point allowed us to ensure that the predictive models work properly, maximizing the predictive power and avoiding overfitting.

Selection of the predictive model. Regarding the selection of the predictive model, after performing a parameter tuning of the seven selected algorithms and comparing the most suitable configuration of each of them (see Figure 5), it was concluded that the models that meet the established requirements regarding accuracy, precision, and recall for the case study were SVM, random forest, MLP neural networks, and GBA. Our results highlight the low number of false negatives achieved (high recall), a fundamental aspect in healthcare studies [39]. These results are in agreement with other investigations of a similar nature, using the same software (Weka), in terms of the accuracy, precision, recall, and F1 score with the SVM and random forest algorithms [36].

Variability of the measures acquired. It has been detected that it is possible to assess the measurement capacity of our medical equipment in terms of the variability of the measures that it obtains. A *series* variable identifies whether the data are relative to the first measure of each subject or to the second (which was performed consecutively). The results obtained by the predictive models showed that there were no differences between the two series of cervical ROM (this variable had the lowest predictive power among all the variables introduced in the model). This result indicates that the measure has behaved stably in collaborating subjects. This result is interesting in the forensic field due to the following reason. If repeating the cervical ROM test in a patient results in significant differences, they would not be derived from the variability of the test, but by the type of pathology that prevents the patient from repeating the test normally. Alternatively, the patient may try to simulate or magnify the lesion by not showing consistent results between the first and second series. This aspect would be of relevance to judicial experts [29,83].

Data collection in production. Once the system is applied in its context and has been developed, continuous data collection must be planned in production in order to improve the prediction accuracy, resulting in a continuous learning system (Figure 3). In the case study, to increase the sample in the exploitation stage of the model it is possible to include those patients who perform cervical assessment tests in a care or rehabilitation setting; thus, their sincerity can be assumed regarding the degree of cervical pain as well as full collaboration in the performance of the tests. However, for the collection of training data it may be necessary to exclude those patients who are immersed in a judicial and indemnifying process and who report cervical pain because their degree of collaboration or sincerity is unknown. In the future, the system can be used to predict the cervical pain of non-collaborating patients (e.g., patients in a judicial process or patients with a high degree of anxiety or hypochondriacs) from the predictive model previously generated with collaborating patients, serving as objective evidence in judicial proceedings with insurance companies [40].

Multidisciplinarity. For a correct interpretation of the results, multidisciplinary work is a key point, since the contribution of each of the branches of knowledge is necessary in this type of project to optimize the possibilities offered by the model. In this way, it will be possible to assess the statistical quality of the results and their medical utility. For example, in our case study questions were raised regarding the way the data and the results should be represented (solved with the databases defined and the design of a report for physicians, presented as Supplementary Material), the possible interpretation and use of the system by the clinician (problems could have arisen if we had not been worked in close collaboration with the clinicians; however, the target variable and how to present the results were clarified from the beginning of the project), and the overlap with other possible decision-support systems (if other systems could also provide in the future a prediction indicating that a patient suffers cervical pain, both results would be presented to the clinician and he/she would take the final decision). Through collaboration with the medical experts, these issues have been solved. However, multidisciplinary work is not always possible, since professionals participate in different stages of the entire process according to their degree of knowledge, experience, and training, so in some cases there may be no direct or sufficient interaction between them [34]. This lack of interaction among professionals could be diminished by following the stages assigned to each of the professionals in a concise and structured way, thus avoiding problems that may lead to project failure [52]. In the case study, the rigor followed by the different professionals involved in the different stages have resulted in adequate results. Nevertheless, although there has been interaction between them, it could have been done in a more collaborative way, since clinical experts were not present in the generation of the predictive model and the interpretation of the results, which could have improved the quality of the study thanks to its specific medical knowledge.

Exploitation of sensor data. The use of sensors and devices with the use of ML could be implemented as a complementary objective test to help physicians. This type of test could constitute an aid to the decision-making in the diagnosis or treatment, or if there is doubt about the veracity of the information reported by the patient [19,82]. Although we wanted to show the clinical utility of this type of technology, the lack of studies on the application of ML techniques with motion capture sensors in healthcare, and specifically their applicability in the forensic field as an objective system of application in judicial processes, have further motivated the choice of the case study. Therefore, while our case study focuses on the medical legal/forensic field, the procedure proposed to use ML techniques could be applied in any study of the health field (cancer detection, studies of discharge from the hospital and sick leave, etc.).

Use of resources. Regarding the frequency and regularity of data acquisition, it is necessary to previously estimate it to limit the duration of the project [34], as well as to quantify the necessary storage size, which is a factor with high variability between projects. If the project is intended to have an adaptive character that can be applied or expanded for subsequent research, the scalability and magnitude should be considered. If the scalability of the project is not considered, and the intention is to continue acquiring data and adapting the model to the continuous growth of information, there may come a time when the project is no longer viable because it is unable to assimilate the corresponding increase in data size. This situation is common in epidemiological projects of data collection for large periods of time, where the scalability is transcendental for the future of the project [30]. On the other hand, it is important to consider that, in certain projects as in the case study presented in this paper, the ratio between available variables and data can be high. So, not exceeding the 1 to 10 ratio between variables and data is transcendental to avoid overfitting effects [20].

5. Conclusions and Future Work

Through a practical guide, the stages and particularities to consider for the application of ML techniques in the field of healthcare have been described, considering all the stages involved in the process. This procedure is shown through objective cervical functional assessment tests that use MoCap technology with inertial sensors and a predictive model whose goal is to estimate the presence of cervical pain from the data collected with the test. Four models (SVM, random forest, MLP neural network, and GBA) from the seven models initially generated obtained an accuracy and precision of more than 85% and a recall of more than 90% (i.e., the percentage of false negatives is smaller than 10%). The approach and the results obtained could help objectify diagnoses, improve test treatment efficacy, and save resources in healthcare systems. The procedure, which has been applied to data derived from a cervical assessment study for verification and evaluation, is also appropriate for any healthcare field regardless of the origin of the data. It can be useful, for example, in gait studies [82], balance studies [84], cardiac failure studies [85], the prediction of events [86], fertility tests [87], etc.

Despite the great usefulness of ML in the field of healthcare, some limitations have been detected in this field. First, a major limitation is how to achieve a suitable flow of data from health centers and hospitals, as well as the accessibility (in relation to privacy policies and authorizations) [21], gathering, and integration of the data [3]. If a global information collaboration policy were established between hospitals [6,69,88], the problem of access to information could be solved, and it would be possible to share more data and feed the predictive models applied to the field of healthcare more efficiently. The explainability of predictions [89,90] is an important issue in a health care domain, as it could increase the trust in the ML systems (for both clinicians and patients) and even lead to the acquisition of new knowledge; however, more research on how to achieve explainability while considering the potential trade-off with accuracy must be performed, especially in the health domain.

Regarding the limitations of the conclusions obtained with this case study, once the model is in the exploitation stage it would be advisable to carry out an external validation to verify the viability of the model in terms of geography, temporality, etc. [30,77]. Regarding the data sample used in our case study, its size has been large enough to obtain good results, but it would be relevant to see the impact of increasing it, as new data about patients becomes available, to enable a continuous learning process that could lead to better results over time. Besides, a study is currently being conducted to check the accuracy of the proposed models with a sample of non-collaborating patients.

Concerning the target variable (presence of cervical pain), which is a binary variable, it could be defined in a more granular way considering not only the presence of pain but also its intensity (as a continuous variable or as a discrete variable with several pain degrees). The problem with pain intensity is that pain scales are highly dependent on the subjectivity of the patient, and this issue could be further exacerbated with non-collaborating subjects. However, as a future goal of our research, it could be useful to tackle this issue and introduce some statistical techniques, such as the numerical measurement z-score, to normalize the subjective values from pain intensity scales (e.g., the Visual Analogue Scale) provided by the patients. The z-score could help to reduce the bias of patients, allowing us to include pain intensity as a target variable in our proposal.

Finally, as future work, it could also be interesting to extend the range of experiments performed and analyze the potential interest of other ML methods; for example, we could consider applying different classifiers applied over different categories of data proposed in the paper and combine them into an ensemble.

Supplementary Materials: The following materials are available online at http://www.mdpi.com/2076-3417/10/17/5942/s1, File S1: Initial database. File S2: Coded database. File S3: Filtered database. File S4: Report for physician.

Author Contributions: Conceptualization, J.d.l.T., J.M., S.I., and J.J.M.; data curation, J.d.l.T., J.M., and J.J.M.; formal analysis, J.d.l.T., J.M., S.I., and J.J.M.; funding acquisition, S.I. and J.J.M.; investigation, J.d.l.T.; methodology, J.d.l.T., J.M., and J.J.M.; project administration, J.J.M.; resources, S.I. and J.J.M.; software, J.d.l.T. and J.J.M.; supervision, S.I. and J.J.M.; validation, J.d.l.T.; visualization, J.d.l.T.; writing—original draft preparation, J.d.l.T., J.M., and J.J.M.; writing—review and editing, J.d.l.T., J.M., S.I., and J.J.M. All authors have read and agreed to the published version of the manuscript.

Funding: This project was co-funded by the Government of Aragon, the European Regional Development Fund and the University of Zaragoza. We thank the support of the project TIN2016-78011-C4-3-R (AEI/FEDER, UE) and the Government of Aragon (Group Reference T64_20R, COSMOS research group).

Acknowledgments: We thank Mapfre insurance company (Zaragoza, Spain) for allowing us to use their facilities. We also thank the I3A—University Institute of Research of Engineering of Aragon, University of Zaragoza, Zaragoza, Spain—for the materials and support they provided.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BMI	Body Mass Index
CEICA	Bioethics Committee of Aragón
CRISP-DM	CRoss-Industry Standard Process for Data Mining
DLA	Daily life activities
DM	Data Mining
EMG	Surface Electromyography
GBA	Gradient Boosting Algorithm
KDD	Knowledge Discovery in Databases
KNN	K-Nearest Neighbors
MAE	Mean Absolute Error
ML	Machine Learning
MLP	MultiLayer Perceptron
MoCap	Motion Capture
PCA	Principal component analysis
RMSE	Root Mean Squared Error
ROM	Range of Movement
SEMMA	Sample, Explore, Modify, Model, and Assess
SVM	Support Vector Machine
WDQ	Whiplash Scale

References

1. Kayyali, B.; Knott, D.; Van Kuiken, S. *The Big-Data Revolution in US Health Care: Accelerating Value and Innovation*; Mc Kinsey Co.: New York, NY, USA, 2013; Volume 2, pp. 1–13.
2. Tomar, D.; Agarwal, S. A survey on Data Mining approaches for Healthcare. *Int. J. Bio-Sci. Bio-Technol.* **2013**, *5*, 241–266. [CrossRef]
3. Koh, H.C.; Tan, G. Data mining applications in healthcare. *J. Healthc. Inf. Manag.* **2011**, *19*, 65.
4. Maity, N.G.; Das, S. Machine learning for improved diagnosis and prognosis in healthcare. In Proceedings of the 2017 IEEE Aerospace Conference, Big Sky, MT, USA, 4–11 March 2017; pp. 1–9.
5. Yoo, I.; Alafaireet, P.; Marinov, M.; Pena-Hernandez, K.; Gopidi, R.; Chang, J.-F.; Hua, L. Data Mining in Healthcare and Biomedicine: A Survey of the Literature. *J. Med. Syst.* **2012**, *36*, 2431–2448. [CrossRef]
6. Sen, I.; Khandelwal, K. Data Mining in Healthcare. 2018. Available online: https://www.researchgate.net/publication/322754945_DATA_MINING_IN_HEALTHCARE (accessed on 26 August 2020).
7. Clavel, D.; Mahulea, C.; Albareda, J.; Silva, M. A Decision Support System for Elective Surgery Scheduling under Uncertain Durations. *Appl. Sci.* **2020**, *10*, 1937. [CrossRef]
8. Cruz, J.A.; Wishart, D.S. Applications of Machine Learning in Cancer Prediction and Prognosis. *Cancer Inform.* **2006**, *2*, 59–77. [CrossRef]
9. Wang, F.; Stiglic, G.; Obradovic, Z.; Davidson, I. Guest editorial: Special issue on data mining for medicine and healthcare. *Data Min. Knowl. Discov.* **2015**, *29*, 867–870. [CrossRef]
10. Rosales, R.E.; Rao, R.B. Guest Editorial: Special Issue on impacting patient care by mining medical data. *Data Min. Knowl. Discov.* **2010**, *20*, 325–327. [CrossRef]
11. Alotaibi, S.; Mehmood, R.; Katib, I.; Rana, O.; Albeshri, A. Sehaa: A Big Data Analytics Tool for Healthcare Symptoms and Diseases Detection Using Twitter, Apache Spark, and Machine Learning. *Appl. Sci.* **2020**, *10*, 1398. [CrossRef]
12. Huang, Z.; Dong, W.; Bath, P.; Ji, L.; Duan, H. On mining latent treatment patterns from electronic medical records. *Data Min. Knowl. Discov.* **2015**, *29*, 914–949. [CrossRef]
13. Bentham, J.; Hand, D.J. Data mining from a patient safety database: The lessons learned. *Data Min. Knowl. Discov.* **2012**, *24*, 195–217. [CrossRef]
14. Obenshain, M.K. Application of Data Mining Techniques to Healthcare Data. *Infect. Control. Hosp. Epidemiol.* **2004**, *25*, 690–695. [CrossRef] [PubMed]
15. Zhang, P.; Wang, F.; Hu, J.; Sorrentino, R. Towards Personalized Medicine: Leveraging Patient Similarity and Drug Similarity Analytics. *AMIA Summits Transl. Sci. Proc.* **2014**, *2014*, 132. [PubMed]
16. Hamet, P.; Tremblay, J. Artificial intelligence in medicine. *Metab. Clin. Exp.* **2017**, *69*, S36–S40. [CrossRef] [PubMed]

17. Joyner, M.J.; Paneth, N. Seven Questions for Personalized Medicine. *JAMA* **2015**, *314*, 999–1000. [CrossRef]
18. Weiss, J.C.; Natarajan, S.; Peissig, P.L.; McCarty, C.A.; Page, D. Machine Learning for Personalized Medicine: Predicting Primary Myocardial Infarction from Electronic Health Records. *AI Mag.* **2012**, *33*, 33–45. [CrossRef]
19. Mannini, A.; Sabatini, A.M. Machine Learning Methods for Classifying Human Physical Activity from On-Body Accelerometers. *Sensors* **2010**, *10*, 1154–1175. [CrossRef]
20. Moons, K.; Kengne, A.P.; Woodward, M.; Royston, P.; Vergouwe, Y.; Altman, U.G.; Grobbee, D.E. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart* **2012**, *98*, 683–690. [CrossRef]
21. Wilkowska, W.; Ziefle, M. Privacy and data security in E-health: Requirements from the user's perspective. *Heal. Inform. J.* **2012**, *18*, 191–201. [CrossRef]
22. Dolley, S. Big Data Solution to Harnessing Unstructured Data in Healthcare. IBM Report. 2015. Available online: https://assets.sourcemedia.com/31/a6/cb1b019c4d6cb338fab539eea360/ims14428usen.pdf. (accessed on 26 August 2020).
23. Andersen, R.M.; Davidson, P.L.; Baumeister, S.E. Improving access to care. In *Changing the US Health Care System: Key Issues in Health Services Policy and Management*; John Wiley & Sons: Hoboken, NJ, USA, 2013; pp. 33–70.
24. Marin, J.; Blanco, T.; Marin, J.J. Research Lines to Improve Access to Health Instrumentation Design. *Procedia Comput. Sci.* **2017**, *113*, 641–646. [CrossRef]
25. Cassidy, J.D.; Carroll, L.; Côté, P.; Lemstra, M.; Berglund, A.; Nygren, Å. Effect of Eliminating Compensation for Pain and Suffering on the Outcome of Insurance Claims for Whiplash Injury. *N. Engl. J. Med.* **2000**, *342*, 1179–1186. [CrossRef]
26. Moreno, A.J.; Utrilla, G.; Marin, J.; Marin, J.J.; Sanchez-Valverde, M.B.; Royo, A.C. Cervical Spine Assessment with Motion Capture and Passive Mobilization. *J. Chiropr. Med.* **2018**, *17*, 167–181. [CrossRef] [PubMed]
27. Utrilla, G.; Marín, J.J.; Sanchez-Valverde, B.; Gomez, V.; JAuria, J.M.; Marin, J.; Royo, C. *Cervical Mobility Testing in Flexion-Extension and Protraction-Retraction to Evaluate Whiplash Syndrome Through Motion Capture*; Universidad de Zaragoza: Zaragoza, Spain, 2017.
28. Marín, J.J.; Pina, M.J.B.; Gil, C.B. Evaluación de Riesgos de Manipulación Repetitiva a Alta Frecuencia Basada en Análisis de Esfuerzos Dinámicos en las Articulaciones sobre Modelos Humanos Digitales. *Cienc. Trab.* **2013**, *15*, 86–93. [CrossRef]
29. Marín, J.; Boné, M.; Ros, R.; Martínez, M. Move-Human Sensors: Sistema portátil de captura de movimiento humano basado en sensores inerciales, para el análisis de lesiones musculoesqueléticas y utilizable en entornos reales. In Proceedings of the Sixth International Conference on Occupational Risk Prevention, Galicia, Spain, 14–16 May 2008.
30. Steyerberg, E.W. *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*; Springer: Berlin, Germany, 2009.
31. Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. *Advances in Knowledge Discovery and Data Mining*; AAAI Press: Menlo Park, CA, USA, 1996.
32. Azevedo, A.I.R.L.; Santos, M.F. KDD, SEMMA and CRISP-DM: A Parallel Overview. ISCAP—Informática—Comunicações em Eventos Científicos. 2008. Available online: https://recipp.ipp.pt/handle/10400.22/136 (accessed on 26 August 2020).
33. Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.; Wirth, R. *CRISP-DM 1.0 Step-by-Step Data Mining Guide*; SPSS Inc.: Chicago, IL, USA, 2000.
34. McGregor, C.; Catley, C.; James, A. A process mining driven framework for clinical guideline improvement in critical care. In Proceedings of the Learning from Medical Data Streams Workshop, Bled, Slovenia, 6 July 2011.
35. Catley, C.; Smith, K.; McGregor, C.; Tracy, M. Extending CRISP-DM to incorporate temporal data mining of multidimensional medical data streams: A neonatal intensive care unit case study. In Proceedings of the 2009 22nd IEEE International Symposium on Computer-Based Medical Systems, Albuquerque, NM, USA, 3–4 August 2009; pp. 1–5.
36. Araujo, F.H.; Santana, A.M.; Neto, P.S. Using machine learning to support healthcare professionals in making preauthorisation decisions. *Int. J. Med. Inform.* **2016**, *94*, 1–7. [CrossRef]
37. Bose, I.; Mahapatra, R.K. Business data mining—A machine learning perspective. *Inf. Manag.* **2001**, *39*, 211–225. [CrossRef]

38. Bhatla, N.; Jyoti, K. An analysis of heart disease prediction using different data mining techniques. *Int. J. Eng.* **2012**, *1*, 1–4.
39. Raschka, S.; Mirjalili, V. *Python Machine Learning*; Packt Publishing Ltd.: Birmingham, UK, 2017.
40. Schuller, E.; Eisenmenger, W.; Beier, G. Whiplash Injury in Low Speed Car Accidents: Assessment of Biomechanical Cervical Spine Loading and Injury Prevention in a Forensic Sample. *J. Musculoskelet. Pain* **2000**, *8*, 55–67. [CrossRef]
41. Naumann, F. Data profiling revisited. *ACM SIGMOD Rec.* **2014**, *42*, 40–49. [CrossRef]
42. Rahm, E.D.H. Data Cleaning: Problems and Current Approaches. *Bull. Tech. Comm. Data Eng.* **2000**, *23*, 3–13.
43. Jannot, A.-S.; Zapletal, E.; Avillach, P.; Mamzer, M.-F.; Burgun, A.; Degoulet, P. The Georges Pompidou University Hospital Clinical Data Warehouse: A 8-years follow-up experience. *Int. J. Med. Inform.* **2017**, *102*, 21–28. [CrossRef]
44. Evans, R.S.; Lloyd, J.F.; Pierce, L.A. Clinical Use of an Enterprise Data Warehouse. *AMIA Annu. Symp. Proc.* **2012**, *2012*, 189–198.
45. Batista, G.E.A.P.A.; Prati, R.C.; Monard, M.C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 20–29. [CrossRef]
46. Batista, G.E.A.P.A.; Prati, R.C.; Monard, M.C. Balancing Strategies and Class Overlapping. *Lect. Notes Comput. Sci.* **2005**, *3646*, 24–35. [CrossRef]
47. Bhardwaj, R.; Nambiar, A.R.; Dutta, D. A Study of Machine Learning in Healthcare. *2017 IEEE 41st Annu. Comput. Softw. Appl. Conf.* **2017**, *2*, 236–241. [CrossRef]
48. Dörre, J.; Gerstl, P.; Seiffert, R. Text mining: Finding nuggets in mountains of textual data. In Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Diego, CA, USA, 23–27 August 1999; pp. 398–401.
49. Aggarwal, C.C.; Zhai, C. *Mining Text Data*; Springer Science & Business Media: Berlin, Germany, 2012.
50. Janasik, N.; Honkela, T.; Bruun, H. Text mining in qualitative research: Application of an unsupervised learning method. *Organ. Res. Methods* **2009**, *12*, 436–460. [CrossRef]
51. Blei, D.M. Probabilistic topic models. *Commun. ACM* **2012**, *55*, 77–84. [CrossRef]
52. Henri, L. *Data Scientist y Lenguaje R Guía de Autoformación Para el uso de Big Data*; Francisco, J., Piqueres, J., Eds.; Colecciones Epsilon: Cornellá de Llobregat, Spain, 2017.
53. Stavrianou, A.; Andritsos, P.; Nicoloyannis, N. Overview and semantic issues of text mining. *ACM SIGMOD Rec.* **2007**, *36*, 23–34. [CrossRef]
54. Carlos, T.; Sergio, I.; Carlos, S. Text Mining of Medical Documents in Spanish: Semantic Annotation and Detection of Recommendations. In Proceedings of the 16th International Conference on Web Information Systems and Technologies (WEBIST 2020), Budapest, Hungary, 3–5 November 2020.
55. Donders, A.R.T.; Van Der Heijden, G.J.; Stijnen, T.; Moons, K.G. Review: A gentle introduction to imputation of missing values. *J. Clin. Epidemiol.* **2006**, *59*, 1087–1091. [CrossRef]
56. Farhangfar, A.; Kurgan, L.; Dy, J. Impact of imputation of missing values on classification error for discrete data. *Pattern Recognit.* **2008**, *41*, 3692–3705. [CrossRef]
57. Bennett, D.A. How can I deal with missing data in my study? *Aust. N. Z. J. Public Health* **2001**, *25*, 464–469. [CrossRef]
58. Arbuckle, L.; El Emam, K. *Anonymizing Health Data*; O'Reilly Media, Inc.: Newton, MA, USA, 2013.
59. Kargupta, H.; Datta, S.; Wang, Q.; Sivakumar, K. On the privacy preserving properties of random data perturbation techniques. In Proceedings of the Third IEEE International Conference on Data Mining, Melbourne, FL, USA, 19–22 November 2003; pp. 99–106.
60. El Emam, K.; Dankar, F.; Issa, R.; Jonker, E.; Amyot, D.; Cogo, E.; Corriveau, J.-P.; Walker, M.; Chowdhury, S.; Vaillancourt, R.; et al. A globally optimal k-anonymity method for the de-identification of health data. *J. Am. Med. Inform. Assoc.* **2009**, *16*, 670–682. [CrossRef]
61. Dankar, F.K.; El Emam, K. The application of differential privacy to health data. *Jt. EDBT/ICDT Workshops EDBT-ICDT* **2012**, *2012*, 158–166. [CrossRef]
62. IBM. SPSS Software. Available online: https://www.routledge.com/IBM-SPSS-Statistics-26-Step-by-Step-A-Simple-Guide-and-Reference/George-Mallery/p/book/9780367174354 (accessed on 16 February 2020).
63. Wu, X.; Kumar, V.; Quinlan, J.R.; Ghosh, J.; Yang, Q.; Motoda, H.; McLachlan, G.J.; Ng, A.; Liu, B.; Yu, P.; et al. Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **2007**, *14*, 1–37. [CrossRef]

64. Gupta, P. Cross Validation in Machine Learning. 2020. Available online: https://towardsdatascience.com/cross-validation-in-machine-learning-72924a69872f (accessed on 26 August 2020).
65. Arlot, S.; Celisse, A. A survey of cross-validation procedures for model selection. *Stat. Surv.* **2010**, *4*, 40–79. [CrossRef]
66. Refaeilzadeh, P.; Tang, L.; Liu, H. Cross-Validation. *Encycl. Database Syst.* **2009**, *5*, 532–538.
67. Xu, Y.; Goodacre, R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *J. Anal. Test.* **2018**, *2*, 249–262. [CrossRef] [PubMed]
68. McCaffrey, J. Neural Network Train-Validate-Test Stopping. 2020. Available online: https://visualstudiomagazine.com/articles/2015/05/01/train-validate-test-stopping.aspx (accessed on 26 August 2020).
69. Ferber, R.; Osis, S.T.; Hicks, J.L.; Delp, S.L. Gait biomechanics in the era of data science. *J. Biomech.* **2016**, *49*, 3759–3761. [CrossRef] [PubMed]
70. Reed, R.; MarksII, R.J. *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks*; Mit Press: Cambridge, MA, USA, 1999.
71. Christodoulou, E.; Ma, J.; Collins, G.S.; Steyerberg, E.W.; Verbakel, J.Y.; Van Calster, B.; Evangelia, C.; Jie, M. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J. Clin. Epidemiol.* **2019**, *110*, 12–22. [CrossRef]
72. Bae, J.M. The clinical decision analysis using decision tree. *Epidemiol. Health* **2014**, *36*. [CrossRef] [PubMed]
73. Noi, P.T.; Kappas, M. Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. *Sensors* **2018**, *18*, 18. [CrossRef]
74. Penny, W.; Frost, D.; Penny, W. Neural Networks in Clinical Medicine. *Med. Decis. Mak.* **1996**, *16*, 386–398. [CrossRef]
75. Zhang, Z.; Zhao, Y.; Canes, A.; Steinberg, D.; Lyashevska, O.; AME Big-Data Clinical Trial Collaborative Group. Predictive analytics with gradient boosting in clinical medicine. *Ann. Transl. Med.* **2019**, *7*. [CrossRef]
76. Witten, I.; Frank, E.; Hall, M.; Pal, C. Appendix B: The WEKA workbench. In *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed.; Morgan Kaufmann: Burlington, MA, USA, 2016.
77. Moons, K.; Kengne, A.P.; Grobbee, D.E.; Royston, P.; Vergouwe, Y.; Altman, U.G.; Woodward, M. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* **2012**, *98*, 691–698. [CrossRef] [PubMed]
78. Murphy, C.K. Identifying Diagnostic Errors with Induced Decision Trees. *Med. Decis. Mak.* **2001**, *21*, 368–375. [CrossRef] [PubMed]
79. Zhao, L. The gut microbiota and obesity: From correlation to causality. *Nat. Rev. Genet.* **2013**, *11*, 639–647. [CrossRef] [PubMed]
80. Dab, W.; Ségala, C.; Dor, F.; Festy, B.; Lameloise, P.; Le Moullec, Y.; Le Tertre, A.; Médina, S.; Quenel, P.; Wallaert, B.; et al. Air pollution and health: Correlation or causality? The case of the relationship between exposure to particles and cardiopulmonary mortality. *J. Air Waste Manag. Assoc.* **2001**, *51*, 220–235. [CrossRef] [PubMed]
81. Liu, C.; Talaei-Khoei, A.; Zowghi, D.; Daniel, J. Data completeness in healthcare: A literature survey. *Pac. Asia J. Assoc. Inf. Syst.* **2017**, *9*, 5. [CrossRef]
82. Mannini, A.; Trojaniello, D.; Cereatti, A.; Sabatini, A.M. A Machine Learning Framework for Gait Classification Using Inertial Sensors: Application to Elderly, Post-Stroke and Huntington's Disease Patients. *Sensors* **2016**, *16*, 134. [CrossRef]
83. Ramírez, P.C.; Ordi, H.G.; Fernández, P.S.; Morales, M.I.C. Detección de exageración de síntomas en esguince cervical: Pacientes clínicos versus sujetos análogos. *Trauma* **2014**, *25*, 4–10.
84. De La Torre, J.; Marin, J.; Marin, J.J.; Auria, J.M.; Sanchez-Valverde, M.B. Balance study in asymptomatic subjects: Determination of significant variables and reference patterns to improve clinical application. *J. Biomech.* **2017**, *65*, 161–168. [CrossRef]
85. Austin, P.C.; Tu, J.V.; Ho, J.E.; Levy, D.; Lee, D.S. Using methods from the data-mining and machine-learning literature for disease classification and prediction: A case study examining classification of heart failure subtypes. *J. Clin. Epidemiol.* **2013**, *66*, 398–407. [CrossRef]
86. Choi, E.; Bahadori, M.T.; Schuetz, A.; Stewart, W.F.; Sun, J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. *JMLR Work. Conf. Proc.* **2016**, *56*, 301–318.

87. Uyar, A.; Bener, A.; Ciray, H.N. Predictive modeling of implantation outcome in an in vitro fertilization setting: An application of machine learning methods. *Med. Decis. Mak.* **2015**, *35*, 714–725. [CrossRef] [PubMed]
88. Abdelaziz, A.; Elhoseny, M.; Salama, A.S.; Riad, A. A machine learning model for improving healthcare services on cloud computing environment. *Measurement* **2018**, *119*, 117–128. [CrossRef]
89. Hall, P.; Gill, N. *An Introduction to Machine Learning Interpretability*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2018.
90. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* **2018**, *51*, 1–42. [CrossRef]