UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE INFORMÁTICA

TESIS DOCTORAL

Metodologías de procesamiento de datos en el ámbito de e-Health para la categorización de respuestas terapéuticas en pacientes con migraña

Data procesing methodologies in the area of e-Health for categorizing therapeutic responses in patients with migraine

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Franklin Ricardo Parrales Bravo

DIRECTORES

Alberto Antonio del Barrio García
José Luis Ayala Rodrigo

Universidad Complutense de Madrid
Facultad de Informática

TESIS DOCTORAL

Metodologías de procesamiento de datos en el ámbito de e-Health para la categorización de respuestas terapéuticas en pacientes con migraña

---

Data processing methodologies in the area of e-Health for categorizing therapeutic responses in patients with migraine

MEMORIA PARA OPTAR AL GRADO DE DOCTOR
PRESENTADA POR:

FRANKLIN RICARDO PARRALES BRAVO

DIRECTORES:

PROF. ALBERTO ANTONIO DEL BARRIO GARCÍA
PROF. JOSÉ LUIS AYALA RODRIGO

Año 2020

Metodologías de procesamiento de datos en el ámbito de e-Health para la categorización de respuestas terapéuticas en pacientes con migraña

Tesis presentada a la Universidad Complutense de Madrid en cumplimiento de los requisitos para el grado de

Doctor en Ingeniería Informática

Presentado por:

Franklin Ricardo Parrales Bravo

Directores:

Profesor: Dr. Alberto Antonio del Barrio García

Profesor: Dr. José Luis Ayala Rodrigo

Facultad de Informática

Universidad Complutense de Madrid

2020

Data processing methodologies in the area of e-Health for categorizing therapeutic responses in patients with migraine

Thesis submitted to the Universidad Complutense de Madrid in fulfillment of the requirements for the degree of

Doctor en Ingeniería Informática

Presented by:

Franklin Ricardo Parrales Bravo

Advisors:

Professor: Dr. Alberto Antonio del Barrio García

Professor: Dr. José Luis Ayala Rodrigo

Facultad de Informática

Universidad Complutense de Madrid

2020

Metodologías de procesamiento de datos en el ámbito de e-Health para la categorización de respuestas terapéuticas en pacientes con migraña

Memoria presentada por Franklin Ricardo Parrales Bravo para optar al grado de Doctor por la Universidad Complutense de Madrid, realizada bajo la dirección de D. Alberto Antonio Del Barrio García y D. José Luis Ayala Rodrigo (Departamento de Arquitectura de Computadores y Automática, Universidad Complutense de Madrid).

Data processing methodologies in the area of e-Health for categorizing therapeutic responses in patients with migraine

Report presented by Franklin Ricardo Parrales Bravo at the Complutense University of Madrid in order to apply for the Ph.D. degree. This work has been supervised by Mr. Alberto Antonio Del Barrio García and Mr. José Luis Ayala Rodrigo (Department of Computer Architecture and Automation, Universidad Complutense de Madrid).

Madrid, 2020

DEDICATORIA

Dedico esta tesis a mi madre y mi padre,
y a toda mi familia y amistades.
La dedico a quienes hicieron posible que pueda dedicar
este trabajo, Dios primeramente y mis tutores,
por todo el tiempo dedicado para mi aprendizaje.
También la dedico a las personas que
sufren de esta dolencia crónica, espero
haber aportado con un grano de arena en la
mejora de calidad de vida para ellos.

Agradecimientos

A todos los que la presente vieren y entendieren.

Inicio de las Leyes Orgánicas. Juan Carlos I

Primeramente quiero agradecer a Dios por haberme cuidado dándome salud y recursos para la realización de este presente trabajo. Agradezco a mi padre, el Economista Franklin Parrales Marcillo y a mi madre Janeth Bravo Velez, quienes desde pequeño supieron inclucarme los valores de la perseverancia, empeño y amor en cada circunstancia de la vida. También agradezco a mis hermanas Ericka y Carolina, y a toda mi familia, quienes me dieron todo el apoyo durante mis horas de estudio.

Quiero agradecer también a quienes se han convertido en mi segunda familia, aquellas amistades que me acogieron en España e hicieron que mi estancia en este bello país se convierta en mi segundo hogar. En especial a Ruth Romero, Marjorie Romero, Javier Gandía, Luis García, Noemí Domínguez, Kiko Sánchez, María Pop, Montse García, Vasile, Doris, Trajano, Arthur, Sofía, y a todos quienes compartieron mis frustraciones y grandes momentos, y si me dejo a alguno le ruego me perdone.

También me gustaría agradecer a todas las personas involucradas en esta investigación, sin las cuales esta tesis no hubiera sido posible. En especial mi tutor, el Dr. Alberto Del Barrio García, quien siempre estuvo atento a llamarme la atención cuando era necesario y a dedicarme mucho tiempo para guiarme en la tesis. Trabajar con él ha sido muy desafiante, orientándome y motivándome cuando más lo necesitaba. Además agradezco al Dr. José Luis Ayala, quien supo dar pautas e ideas cuando no veía la luz en el túnel del doctorado. Fue gracias a él que pude contactarme con los centros donde realicé las estancias doctorales y también fue él quien gestionó la recopilación de la información clínica.

Finalmente, quiero agradecer también a mis asesores externos durante

mis estancias de investigación en Eslovenia y Portugal, Dr. Saso Dzeroski y Dr. Luis M. S. Russo, por su tiempo dedicado y por su ayuda en mejorar mis habilidades de investigación. Y, por supuesto, quiero agradecer al Departamento de Computación y Automática por facilitarme los recursos de investigación y por asumir los costes de publicación de artículos en las revistas científicas. Tampoco debo pasar por alto las ayudas económicas otorgada por la SENESCYT del Ecuador y por la "HiPEAC network". Sus becas económicas fueron vitales para que pueda dedicar el tiempo y esfuerzo necesario para llevar adelante la presente investigación.

A todos ellos dirijo mi agradecimiento sincero. El esmero y ayuda de todos ellos ha hecho posible que hoy puedas estar leyendo estas lineas.

* * *

Financial support

This work was supported in part by the Ministry of Education, Science, Technology and Innovation (SENESCYT), Government of the Republic of Ecuador, under Grant 8905-AR5G-2016, and by the European Union's Horizon 2020 Framework Programme for Research and Innovation under grants H2020-ICT-2015-687689 (HiPEAC collaboration grant-2017) and H2020-ICT-2017-779656 (HiPEAC collaboration grant-2019).

Abstract

This Ph.D. Thesis studies some data processing methodologies in the area of e-Health for categorizing therapeutic responses in patients with migraine. In a real e-Health scenario, this work focuses on the prediction of the response to the treatment of migraine through the use of retrospective medical records collected from *Hospital Clínico Universitario* in Valladolid and *Hospital Universitario de La Princesa*, in Madrid.

The goal of this research work is to pose and answer the following questions: is it possible to predict the response to every stage of the BoNT-A treatment for migraine? Does a pre-treatment prediction model for the BoNT-A treatment in migraine exist? How do these models respond under missing values? Is it possible to reveal those medical factors that make it possible a high response to the BoNT-A treatment? Are the medical factors used to predict the response of the treatment coherent with the knowledge of medical experts? To answer these questions, this work has explored and implemented different approaches for the training of the predictive models.

Three predictive approaches have been proposed, which are: panoramic, feedback and hierarchy prediction models. In addition, a data transformation is proposed for finding a better representation of the numeric labels while achieving high prediction accuracies without adding more columns to the dataset. Furthermore, in order to bridge the gap between the biomedical community and the data mining community, a consensus model technique has been proposed for unveiling relevant attributes from prediction models.

A significant improvement in accuracy due to the use of SAR encoding has been achieved, from close to 68% (baseline) to 75% with panoramic prediction, and up to around 88% when using feedback prediction. Furthermore , predictability of panoramic and feedback prediction models are improved when applying a hierarchy of models, obtaining accuracies close to 85% and 94% respectively. Regarding the runtime, the obtained results with the use of MOEAs show that training times are decreased from 8 to less than 2 hours when using 8 threads. In addition, this Ph.D. Thesis has made possible the extraction of relevant attributes that allow to know in advance the response

to the treatment. These are: “evolution of migraine time”, “unilateral pain”, “abuse of analgesics”, “days of headache” and the “retroocular component”. All these attributes have been consistent with the expert knowledge of doctors.

Resumen

La presente tesis doctoral estudia algunas metodologías de procesamiento de datos en el área de e-Health para clasificar las respuestas terapéuticas en pacientes con migraña. En un escenario real de e-Health, este trabajo se centra en la predicción de la respuesta al tratamiento de la migraña mediante el uso de registros médicos retrospectivos recopilados del *Hospital Clínico Universitario* en Valladolid y del *Hospital Universitario de La Princesa*, en Madrid.

El objetivo de este trabajo de investigación es plantear y responder las siguientes preguntas: ¿es posible predecir la respuesta a cada etapa del tratamiento para la migraña con BoNT-A? ¿existe un modelo predictivo para el tratamiento con BoNT-A en la migraña? ¿cómo responden estos modelos bajo registros incompletos? ¿es posible conocer aquellos factores médicos que hacen posible una alta respuesta al tratamiento con BoNT-A? ¿Los factores médicos utilizados para predecir la respuesta del tratamiento son coherentes con el conocimiento de los expertos médicos? Para responder a estas preguntas, este trabajo ha explorado e implementado diferentes enfoques para el entrenamiento de los modelos predictivos.

Se han propuesto tres enfoques predictivos, que son: modelos panorámicos, de retroalimentación y jerarquía de modelos. Además, se ha propuesto una transformación de datos para encontrar la mejor representación de las etiquetas numéricas mientras se alcanza una alta precisión de predicción sin agregar más columnas al conjunto de datos. Adicionalmente, para establecer nexos entre la comunidad biomédica y la comunidad de la minería de datos, se ha propuesto una técnica de consenso de modelos con la finalidad de extraer atributos relevantes de los modelos de predicción.

Se ha logrado una mejora significativa en la precisión debido al uso de la codificación SAR, desde cerca del 68% (baseline) al 75% con la predicción panorámica, y hasta alrededor del 88% cuando se usa la predicción por retroalimentación. Además, la precisión de los modelos de predicción panorámica y de retroalimentación se mejora al aplicar una jerarquía de modelos, obteniendo precisiones cercanas al 85% y 94% respectivamente.

Con respecto al tiempo de ejecución, los resultados obtenidos con el uso de MOEA muestran que los tiempos de entrenamiento se reducen de 8 a menos de 2 horas cuando se usan 8 hilos. Además, esta tesis doctoral ha hecho posible la extracción de atributos relevantes que permiten conocer de antemano la respuesta al tratamiento. Estos son: "evolución del tiempo de migraña", "dolor unilateral", "abuso de analgésicos", "días de dolor de cabeza" y el "componente retroocular". Todos estos atributos han sido coherentes con el conocimiento experto de los médicos.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Learning is not the product of teaching.
Learning is the product of the activity of learners.

John Holt

Contents

1.1 The migraine disease

The present research work focuses on the prediction of the treatment response to migraine by the use of medical records. Therefore, it is convenient to explain and motivate the study of the disease given its economic and social implications in the contemporary world.

Migraine is a common neurological disorder characterized by recurrent headaches. Migraine attacks usually last for 4-72 h and involve moderate or severe intensity headaches, which typically are worsened by routine physical activity, are of a pulsating nature, and are associated with nausea, vomiting, photophobia or phonophobia (IHS, 2013). In clinical terms, migraine can be classified into two types according to the frequency of pain: episodic migraine (less frequent headaches) and chronic migraine. Chronic migraine is defined as a headache occurring on 15 or more days per month for more than 3 months, and which has the attributes of a migraine headache on at least 8 days per month (IHS, 2013). The transformation of episodic migraine into chronic migraine occurs over months or years and involves atypical pain modulation and central sensitization triggered by repetitive inputs from sensitized peripheral sensory neurons (Diener et al., 2012).

Globally, approximately 10% of the population experiences chronic migraine (Natoli et al., 2010; Stovner et al., 2018). In fact, Stovner et al. (2018) have mentioned that around three billion people suffer migraine and tension type-headache together. Furthermore, their work manifest that the migraine prevalence in Europe is close to 15%. Moreover, they have mentioned that its derivative headache is the third most prevalent disorder (after dental caries and latent tuberculosis infection). Their work present a geographic distribution of the prevalence of headache and it is presented in Figure 1.1.

1.1.1 Socioeconomic cost

According to Linde et al. (2012), the economic consequences of the migraine represent €1,222 per patient per year in Europe. It implies almost €125,000 millions in this continent.

In addition to the increased use of analgesic medication, visits to doctors, and visits to the emergency services, chronic migraine has a high socioeconomic cost, with higher direct and indirect costs. In fact, some of the direct costs are due to absences at work or a low performance carrying out a job. Furthermore, chronic migraine sufferers are more prone to anxiety, depression, other chronic diseases (CDs) like respiratory, heart or circulatory diseases and more chronic pain, all of this associated with significant

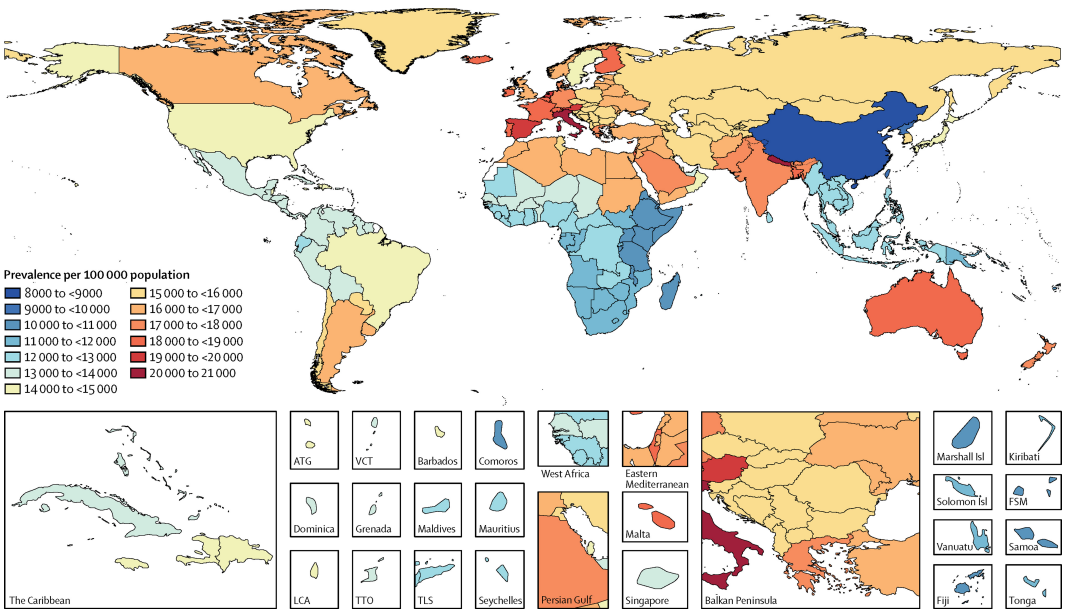


Figure 1.1: Age-standardised prevalence of tension-type headache per 100000 population by location for both sexes. Image taken from Stovner et al. (2018).

personal, societal, and economic burdens (Buse et al., 2010; Adams et al., 2015). In fact, in a survey conducted by the Eli Lilly and Company on 1,018 US adults[1] concludes that several people diagnosed with migraine usually suffer headache episodies for about half a month. They also point out that people who do not suffer from migraine, often underestimate the pain and average duration of migraine. On the other hand, migraine often adds stress and it can affect the relationship with their family and loved ones. Additionally, the professional potential of the person diagnosed with the disease may be affected. The conclusions of such survey are presented in Figure 1.2.

1.1.2 Assessment of migraine severity

In order to estimate the goodness of the treatment, it is necessary to define a metric, aka *severity index*, that indicates how efficient the treatment session has been. For this purpose, there is an abundant medical literature regarding various chronic diseases where different severity indexes are presented. However, each index is usually specific to any disease or symptom(s). For example, in the case of Parkinson Disease (PD), the HY scale (Hoehn et al.,

[1]Survey Reveals Many People with Migraine Live with Pain Nearly Half of Every Month `https://www.multivu.com/players/English/8259051-lilly-migraine-impact-report/` (accessed June 2019)

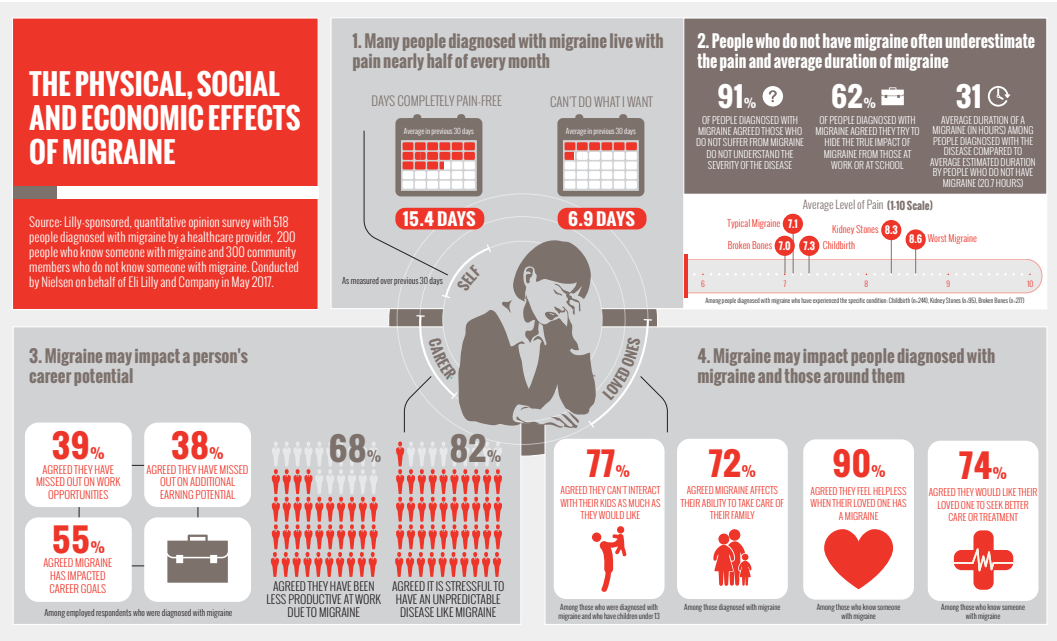


Figure 1.2: The physical, social and economic effects of migraine. Image taken from `https://www.multivu.com/` (accessed June 2019).

1998) is a classical instrument used to categorize patients according to PD stages. Other metric is the CISI-PD (Martínez-Martín et al., 2006), which extends the evaluated motor symptoms criteria to more complex aspects like the patients' cognitive state.

In the migraine scenario, the severity is typically evaluated through the HIT6 value (Yang et al., 2011), the intensity, the duration, the frequency of attacks (Gasbarrini et al., 1998), number of headache days (Schoenen et al., 1998), the Global Assessment of Migraine Severity (GAMS) (Sajobi et al., 2019; Dowson, 2001), the migraine severity scale (MIGSEV) (El Hasnaoui et al., 2003) or the Migraine Disability Scale (MIDAS) (Thompson et al., 2002). For measuring the associate depression due migraine, doctors use the 9-item Patient Health Questionnaire (PHQ-9) (Arroll et al., 2010) and the 14-item Hospital Anxiety and Depression Scale (HADS) (Sajobi et al., 2019; Zigmond & Snaith, 1983). Table 1.1 presents a more detailed information about these indexes.

Hence, the severity index is something inherent to the CD under consideration, but also depends on the number of medical records containing the index. For example, a dataset of 100 migraine patients can be considered. In it, the value of HIT6 may have been collected for only 10 patients, while 95 of them have values related to the "number days with headache" before and

after every stage of the treatment. In this example, the use of HIT6 would be ruled out given its absence in most dataset records. A better decision would be to use the "number days with headache" to measure the improvement to migraine for this specific dataset.

Table 1.1: Medical indexes used for measuring the severity level of migraine.

Severity index	Short description	Publication
HADS	A 14-item screening tool for depression and anxiety developed for use in populations with medical conditions.	Zigmond & Snaith (1983)
Intensity, duration and frequency of attacks	Indexes chosen based on available medical data	Gasbarrini et al. (1998)
Number of headache days	Index chosen based on available medical data	Schoenen et al. (1998)
PHQ-9	A 9-item questionnaire for screening, diagnosing, monitoring, and measuring the severity of depression.	Kroenke et al. (2001)
MIGSEV	Test for assessing the migraine attack at the level of an individual patient.	El Hasnaoui et al. (2003)
MIDAS	A high test-retest reliability in persons with migraine and correlates to clinical judgment regarding the need for medical care.	Thompson et al. (2002)
HIT6	A 6-item survey for discriminating headache impact across episodic and chronic migraine.	Yang et al. (2011)
GAMS	Test developed to assess patients' perception of their disease severity.	Sajobi et al. (2019)

1.1.3 The OnabotulinumtoxinA treatment

The pharmacological treatment of chronic migraine is based on two pillars: abortive treatment of acute migraine attacks (that are taken only in the acute pain phase) and preventive therapy. The latter is used to diminish the severity, frequency or duration of attacks. Preventive therapy includes additional

benefits such as reduction of disability and enhancement of response to acute treatments (Lipton & Silberstein, 1994). It may also result in a reduction in health care costs (Silberstein et al., 2003).

Many classes of medication are used for migraine prevention: antiepileptic drugs, antidepressants, betablockers, calcium channel antagonists, serotonin antagonists, and botulinum neurotoxins, among others. In the case of chronic migraine, although all preventive treatments for migraine may be useful, only topiramate (a type of antiepileptic) and OnabotulinumtoxinA (BoNT-A) (Frampton, 2012) have solid proven evidence for their use according to various works and clinical trials (Diener et al., 2007; Silberstein et al., 2007; Aurora et al., 2010, 2011; Diener et al., 2010; Dodick et al., 2010).

BoNT-A has been an extended use treatment for chronic migraine since its approval in 2010 by the Food and Drug Administration in the United States (FDA), having also shown a more sustained effect and better tolerability than topiramate in the few comparative studies performed (Mathew & Jaffri, 2009; Cady et al., 2011). BoNT-A can be injected under the skin (subcutaneous) or inside the muscles (intramuscular) in accordance with the so-called *The Phase III REsearch Evaluating Migraine Prophylaxis Therapy (PREEMPT)* paradigm. This injection method consists of using both fixed and follow-the-pain sites, with additional specific follow-the-pain sites considered depending on individual symptoms. Follow-the-pain refers to administering the rest of the medication in areas where patients particularly have pain. This procedure should be carried out in repeated patterns after several months.

Following the results of the initial clinical trials and subsequent published studies in real-life settings (Lipton et al., 2011; Oterino et al., 2011; Sandrini et al., 2011; Cernuda-Morollón et al., 2015), nowadays it is known that 70-80% of patients with chronic migraine show an improvement with this treatment (improvement defined as a reduction in migraine attack frequency or days with attacks by at least 50% within 3 months, leading to a significantly improved quality of life in patients). Moreover, there is evidence that patients with chronic migraine who do not show the desired treatment response after the first cycle of BoNT-A treatment may indeed experience clinical improvement after one or two additional treatment cycles (Silberstein et al., 2015). In the work presented by Lovati & Giani (2017) the importance of predicting whether BoNT-A treatment will be effective in a patient is pointed. Knowing the phenotype-response relationship may help in the development of new treatments for the 20-30% of patients that do not respond to the treatment. Besides the cost, it would avoid the patients suffering the pain associated with the treatment.

Several studies have looked at the clinical attributes of patients with

migraine which may be associated with a favorable response to BoNT-A treatment, although conclusive results are not yet available for use in clinical practice. In fact, the exact analgesic mechanism of action of BoNT-A is only partially known. The main hypothesis is that the toxin exerts its antinociceptive action inhibiting peripheral sensitization. BoNT-A lowers neuropeptide and neurotransmitter release from peripheral sensory neurons, thereby indirectly reducing central sensitization, the hallmark of chronic migraine (Aoki, 2005; Barbanti et al., 2015).

One of the most debated aspects in recent years has been the possible relationship between the clinical phenotype of migraine attacks and the response to BoNT-A. In this sense, the following possible predictors of a good response have been proposed in literature: allodynia (painful hypersensitivity to superficial stimuli) (Mathew et al., 2008b), the unilateral character of a migraine (Lainez et al., 2006; Mathew et al., 2008b), associated migraine aura (visual, language, motor or sensory alterations occurring prior to pain) (Grogan et al., 2013), or the build-up time to maximum pain (shorter time, better response to BoNT-A) (Schulman et al., 2008). Pain directionality also seems to be a possible clinical predictor. This attribute refers to whether the headache feels like it is exploding, imploding or ocular. The term exploding refers to when the discomfort is felt pushing from the inside out. Patients suffering from imploding or ocular pain tend to be relieved with the BoNT-A treatment than those with the exploding (Jakubowski et al., 2006). Pagola et al. (2014) studied a number of possible clinical predictive attributes in parallel, including unilateral location of headache, pericranial muscular tension, directionality of pain, duration of migraine history and medication overuse, comparing responders to BoNT-A treatment with non-responders, but no significant differences emerged. Other works suggest that the pharmacological response to BoNT-A might be better when the migraine headache is "trigeminal" in pain location and corresponds to reflex trigeminal-autonomic activation (Barbanti et al., 2015; Barbanti & Egeo, 2015). As a consequence, BoNT-A action may be more effective in migraineurs who overactivate peripheral trigeminal endings during the attack, and such patients may be identified by means of easily obtainable patient-reported clinical findings, such as pain location or direction (unilateral, implosive-retroocular), the presence of cranial autonomic symptoms (allodynia) and cortical spreading depression signs (aura) (Barbanti et al., 2015). Other data such as the response to anesthetic block of the greater occipital nerve (GON) or its local painful pressure (positive palpation) might suggest the same. Many authors believe that a therapy which blocks peripheral transmission of pain signals from extracranial areas prior to central sensitization will successfully disrupt

migraine headache propagation (Dodick et al., 2005; Olesen et al., 2009; Grogan et al., 2013). All in all, the reasons of a positive/negative response to BoNT-A treatment is not clearly understood yet.

1.2 Data mining

With the purpose of understanding the mechanisms that determine the effectiveness of medical treatments, doctors are considering different big data techniques (Kang, 2018). For this reason, the issue of data mining is addressed in this chapter.

The modern world is immersed in the era of data explosion (Kersting & Meyer, 2018). For every second, petabytes of data are generated (Fox, 2018). The omnipresent personal computers make it very easy to store things that we would have destroyed before (Chakrabarti et al., 2008). With the rapid growth of promising applications such as social networks, web, mobile services and other applications in various fields, an unprecedented generation of contents is observed and for which there is no end in sight (Idrees et al., 2018).

The field of medicine is not exempt from this phenomenon. Centers such as the European Institute of Bioinformatics, one of the world's largest biology-data repositories, are currently storing large amounts of petabyte of data and backup copies of genes, proteins and small molecules (Marx, 2013). Electronic medical records (EMRs) are also responsible for generating petabytes of data every second (Deshpande et al., 2018).

The problem lies not in the collection of information but in the interpretation that we give from those data collected, which means that quantity is as important as quality (Eisenstein, 2015). In fact, by increasing the data size the computational burden of this analysis increases (Tashkandi et al., 2018).

Data mining focuses on filling the growing gap between the generation of data and our understanding of it through finding patterns in data. That process can be carried out automatically. However, the most usual is to do it semiautomatically. The discovered patterns must be significant and must carry some advantage that usually is usually an economic advantage (Leventhal, 2018). The set of patterns obtained become part of the prediction model.

In the medical field, predictive models are tools, useful for helping decision-making by doctors, which, through the combination of two or more medical characteristics, allow to obtain the clinical outcomes (Wyatt, 1995). There

are two ways of representing the prediction models in data mining. One is called the black box model because we cannot get a direct or explicit interpretation of the predictive model and the other is called a white box model because we can access and visualize the structure of the patterns (decision structure) in an explicit way (Witten et al., 2016).

In the case of medical records, predictive models will be useful when they can be translated into knowledge (explanatory capability) and that knowledge will be useful if it can be used to improve the health of individual patients (Sacristán & Dilla, 2015). Black and white prediction models can achieve good predictions, but in the medical field it will be more useful to obtain predictive models that are represented in terms of a structure that is examined, reasoned and used to inform future decisions. Moreover, the capability of using the previous medical knowledge (background) in the data analysis process is well appreciated (Bellazzi & Zupan, 2008). That background knowledge must be understood as the essential medical information to comprehend a situation or a problem (Miller, 1998). This information does not need to be rediscovered from the data because it can be obtained from medical experts or medical literature (Bellazzi & Zupan, 2008).

1.2.1 Medical data

Predictive models of response to any CD treatment can leverage the use of digitally stored data, also called *electronic medical records* (EMRs). They are also called *electronic health records* (EHR). The rapid growth of the EMRs requires a combination between the traditional analysis of data manually collected by medical experts and the computational methods. It needs to be carried out in order to help in the decision making process of a specific treatment (Stone & Bornhorst, 2012). EMRs allow doctors the storage, retrieval and modification of medical records through the use of digital media instead of paper-based records systems, which often led to a loss of time and organizational problems (Kasthurirathne et al., 2015).

In this sense, the use of data mining techniques has allowed to approach the analysis of medical data and the construction of prediction models (Bellazzi & Zupan, 2008). Furthermore, some machine learning techniques have proved to be better suited for the analysis of medical databases because of the derivation of symbolic rules, the use of background knowledge, pattern-recognition and interpretation of time-ordered data (Lavrač, 1999). Hence, it is vitally important to explore, adapt and make use of those techniques, so that we can select the most appropriate ones. That is, those techniques that allow us to provide an accurate prediction and that the medical factors used

in their generated models are in accordance with the knowledge of medical experts.

Despite the great advantages offered by the use of information technologies for the collection of medical information in patients, dealing with the collected medical data is not an easy task at all. This is because some problems such as the heterogeneity of the data (Huddar et al., 2016) or directly the lack of values, typically happen in the EMRs (Lin & Haug, 2008).

1.2.2 Mining

In data mining, mining refers to the exploration of data with the purpose of finding repetitive patterns or rules that explain the behavior of the data in a given context. In this thesis, the medical context is addressed through the use of EMRs related to the treatment of migraine. There are different data mining methodologies that consider EMRs for the prediction of the therapeutic response on some CD treatments or responses to continuous treatments such as the case of oncological therapies. These methodologies predict the therapeutic response after several stages of the treatment but they do not consider the prediction of responses to several stages altogether. One example is the work presented by Kurosaki et al. (2011), who exposes the use of decision trees to model the prediction of the final outcome of the treatment of chronic hepatitis C after 48 weeks of PEG-IFN/RBV therapy treatment. Another methodology is presented by Lambin et al. (2013). Their work considers the prediction of the prognosis and the response to an oncological treatment based on radiation through the use of multifactorial decision support systems. Both methodologies discretize and normalize the data to avoid sensitivity to different orders of data scales. They also deal with the missing values, replacing them with calculated estimates.

In addition, there are some methodologies designed to reveal medical factors that influence the effectiveness of treatment. For example, Armañanzas et al. (2013) use the Feature Subset Selection (FSS) technique to reveal the most important attributes to predict the severity of a patient with Parkinson's disease (Parrales Bravo et al., 2017) using non-motor symptoms. In Armañanzas et al. (2012), authors propose to extract the most important attributes for a continuous CD treatment using a consensus model. Hence, based on the aforementioned works, it is desirable to incorporate the features of all these techniques into this thesis to be able to predict the response to several treatment stages as well as revealing the reasons that make them effective.

1.2.3 Life cycle of a medical data mining project

Figure 1.3 presents an adaptation of the CRISP-DM methodology (Wirth & Hipp, 2000), a life cycle of a data mining project, to the medical context defined according to techniques presented by Lan et al. (2018).

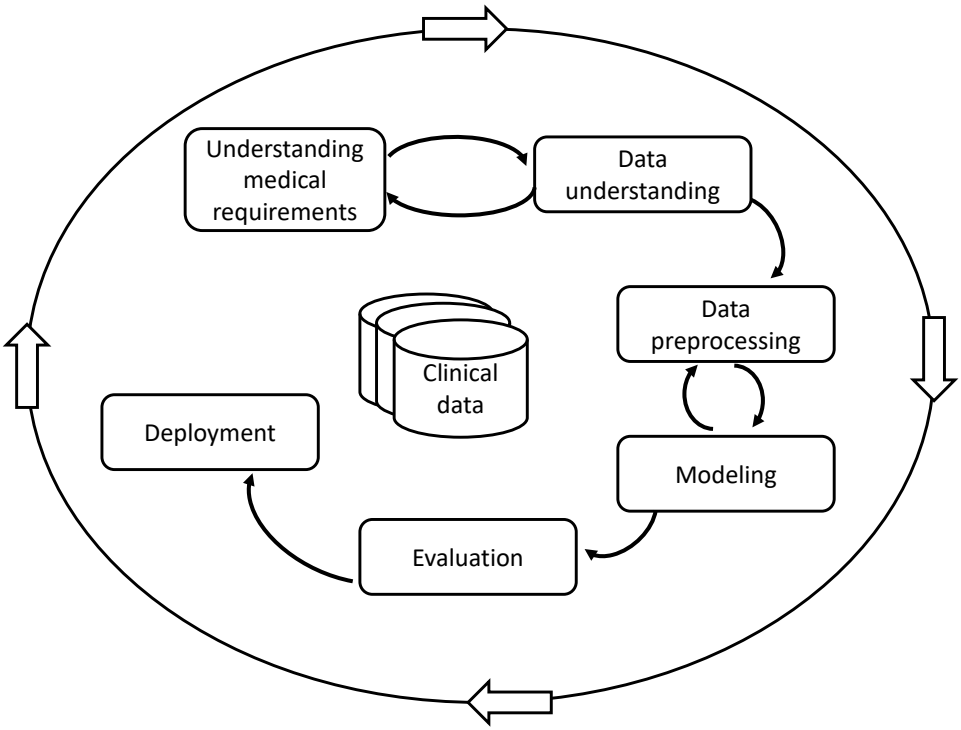


Figure 1.3: Life cycle of a data mining project, adapted for a medical context from Wirth & Hipp (2000).

Before applying the techniques offered by the field of data mining, we must understand what we are being asked to implement. Therefore, the first phase of the cycle consists of "Understanding medical requirements". In this stage, the objectives and requirements of the doctors and the feasibility of being solved by data mining techniques must be clarified. In addition, the available data should be reviewed and those that are possible to be collected.

"Data understanding" is the second stage and it begins with the review of the available data and determines if these are suitable to be processed by data mining. It will be necessary to collect more data based on more stringent criteria for the case when the data quality is poor. At this stage, we can also reconsider the benefit and need to carry out the application of

data mining to achieve the desired objective. The first and second stages are connected since the formulation of the medical project is based on the available data and the data collection is based on the stated objective of the medical project.

The "data preprocessing" stage takes care of preparing the data so that later it can be useful in the production of predictive models (next stage) by the data mining algorithms. The results obtained during the "modeling" stage may yield new knowledge. Thus, it may be necessary to return to this stage several times as new results can affect the selection of preprocessing techniques. According to Lan et al. (2018), the tasks of data cleaning, data integration, data transformation and data reduction should be considered in this stage.

In the "modeling" stage is where the predictive models are obtained with the use of the different data mining algorithms. To carry it out, it is necessary to specify the respective parameters and even some techniques may require specific data formats. In this stage, data problems can arise when modeling or we can figure out ideas to build new data, making it necessary to return to the previous "preprocessing" stage. According to Lan et al. (2018), available algorithms that allow us to get white box prediction models are the decision trees, rules, bayesian classifiers and logistic regression. Other techniques like k-nearest neighbor, support vector machine (SVM), neural networks and ensemble classifiers are more difficult to interpret.

The stage of "evaluation" of predictive models must be considered before deploying them in the medical context. For this, the steps executed for the construction of the model must be reviewed. It should be checked if we are having *overfitted* prediction models. The evaluation must be carried out first of all on the *training* dataset. There are techniques like k-fold cross validation that help us avoid overly optimistic results of the classification algorithms due to overfitting. Another validation method to work in the case of small datasets is an exhaustive cross-validation method called Leave-One-Out Cross-Validation (LOOCV) (McCarthy, 1976). After obtaining a desired precision of the model, it must be evaluated in the *validation* dataset in order to corroborate the obtained accuracy. Additionally, in the medical context, it must be validated if the predictive models have clinical factors that are or are not according to the medical literature. In the case that they are not, their relevance and causes must be studied with the purpose of unveiling new medical findings.

There is nothing worse to think that the project ends when obtaining the prediction model (s). While it is true that this is a good step forward, the knowledge acquired must be organized and presented in a useful way for the

medical environment. This will depend to a large extent on how demanding the medical requirements are. As mentioned by Lan et al. (2018), in most cases it is the user who carries out the implementation steps. In this sense, the guidelines to use to the obtained prediction models must be defined.

Each of the stages of the CRISP-DM methodology involves the development of several tasks. In the article where the methodology (Wirth & Hipp, 2000) is presented, a scheme with the tasks linked to each stage is added. Figure 1.4 presents a diagram with some changes made to the one presented by Wirth. This is done with the purpose of adapting it to the medical environment. In the following lines, an outline is presented with those stages and tasks that have been contemplated in the present research work, adding the section number in which the task is addressed.

- **Understanding medical requirements**
  - Determine the medical objectives:
    - Background: Continuous treatment prediction methodologies (Section 1.1.3), review of migraine severity indexes (Section 1.1.2).
    - Objectives and criteria of medical success: predictive models must be according to medical literature (Section 1.3).
  - Assess situation:
    - Inventory of Resources: EMRs (Section 1.2.1), data mining (Section 1.2) and other computational techniques described throughout the Chapter 2.
    - Costs and benefits: socio-economic cost (Section 1.1.1), advantage of predicting treatment responses (Section 1.3).
  - Determine data mining goals:
    - Data mining goals: Section 1.3 and 3.1.
    - Criteria for success in data mining: Accuracy, sensitivity and specificity (Section 2.3.1).
- **Data understanding**
  - Collect initial data: Section 3.2.1.
  - Describe data: Section 3.2.1.
  - Explore data: Sections 3.2.1 and 3.2.2.
- **Data preprocessing**
  - Dataset description: Section 3.2.1

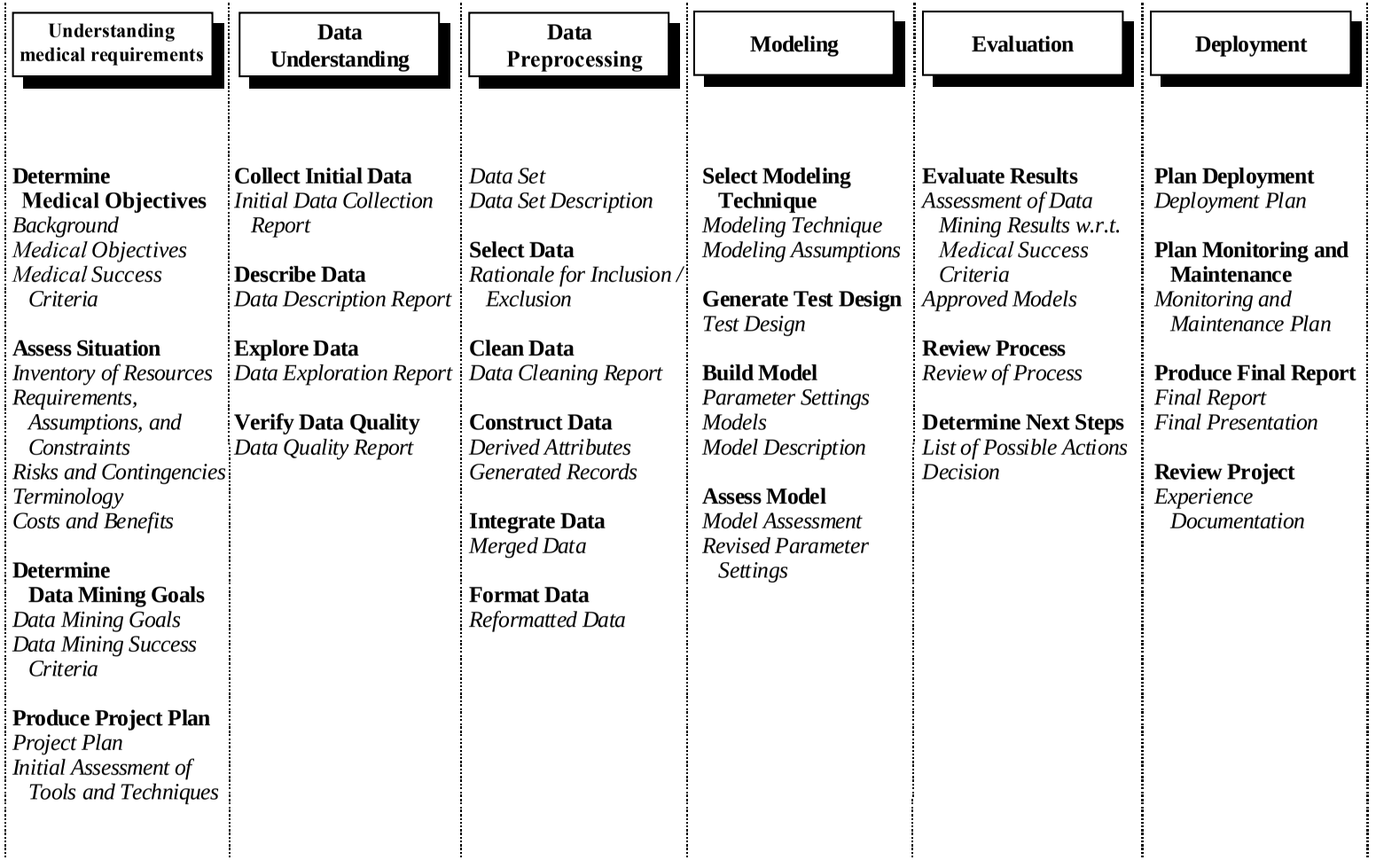


Figure 1.4: CRISP-DM: Overview of the tasks for each stage and its results. Adapted for a medical context from Wirth & Hipp (2000).

  - Select data: all dataset (Section 3.2.1) but dealing with missing values (Section 3.4).
  - Construct data: class attribute based on effects of reduction and adverse effects (Section 3.2.3).

- **Modeling**
  - Select modeling technique: Section 3.3, Section 4.2 and Section 4.3.
  - Build model:
    * Parameter settings models: Section 4.1 and more specifically in Section 4.2.
    * Model description: Section 4.4.
  - Assess model: accuracy values of the k-fold cross validation, sensitivity and specificity (Section 4.1), results (Section 4.2 and Section 4.3).

1.3 Purpose of this thesis

The present research work focuses on the prediction of the response to the treatment of migraine through the use of medical records in a real e-Health scenario. In this sense, retrospective medical data from *Hospital Clínico Universitario* in Valladolid and *Hospital Universitario de La Princesa*, in Madrid have been collected. Clinical data are used to develop predictive models to help doctors make better informed decisions about whether to administer treatment or not. More specifically, this thesis will focus on data derivated from the use of BoNT-A for diminishing the symptoms associated with migraine. As has been mentioned by Lovati & Giani (2017), it is very important to predict if the BoNT-A treatment will be effective in a patient. Knowing the phenotype-response relationship may help in the development of new treatments for the 20-30% of patients that do not respond to the treatment (Section 1.1.3).

Due to the importance of knowing in advance the therapeutic response to BoNT-A and avoiding unnecessary costs, the following questions are posed and answered within this thesis:

- Is it possible to predict the response to every stage of the BoNT-A treatment for migraine?
- Does a pre-treatment prediction model for the BoNT-A treatment in migraine exist?

- How do these models respond under missing values?
- Is it possible to reveal those medical factors that make possible a high response to the BoNT-A treatment?
- Are the medical factors used to predict the response of the treatment coherent with the knowledge of medical experts?

In order to answer these questions, we present throughout the thesis a methodology that considers:

- The preprocessing of the data, given that the situation that is most often found within the medical environment is the existence of missing values. Many attributes or medical factors and a low number of registers are also found very commonly in clinical datasets (Cabitza et al., 2019). Therefore, this thesis considers dealing with missing values when building predictive models for all stages of the BoNT-A treatment. In this sense, some data mining techniques such as imputation of data and feature subset selection (FSS) will be taken into account.
- Thus, a coarse-grained solution is considered when no session has been made yet. This approach is called the panoramic prediction and it will allow doctors to decide if the administration of the treatment will be beneficial without involving unnecessary treatments.
- Once the treatment has begun and the results of some stages are known, feedback prediction is proposed. It allows a more accurate prediction when considering the results of previous stages of the treatment.
- Finally, this thesis reviews some techniques in order to extract relevant medical attributes from the obtained predictive models. After that, those attributes are contrasted with expert medical knowledge. In this way, this research bridges the gap between biomedical community and data mining community thanks to the extraction of medical factors that make the treatment effective or not.

The objectives of the present manuscript are depicted in the Figure 1.5. In this figure, a framework to generate knowledge from real medical data is shown. This framework allows to improve the prediction accuracy through the "numeric label encoding" with two methods: SAR encoding (Section 3.2.5.1) and AMOR encoding (Section 3.2.5.2) for one-target and multi-target prediction models, respectively. They are an interesting contribution because

they allow to improve the representation of medical data that have been previously labeled by doctors. Moreover, this framework considers hierarchical models for dealing with missing values.

All the points presented in the scheme of Figure 1.5 will be presented throughout this PhD Thesis.

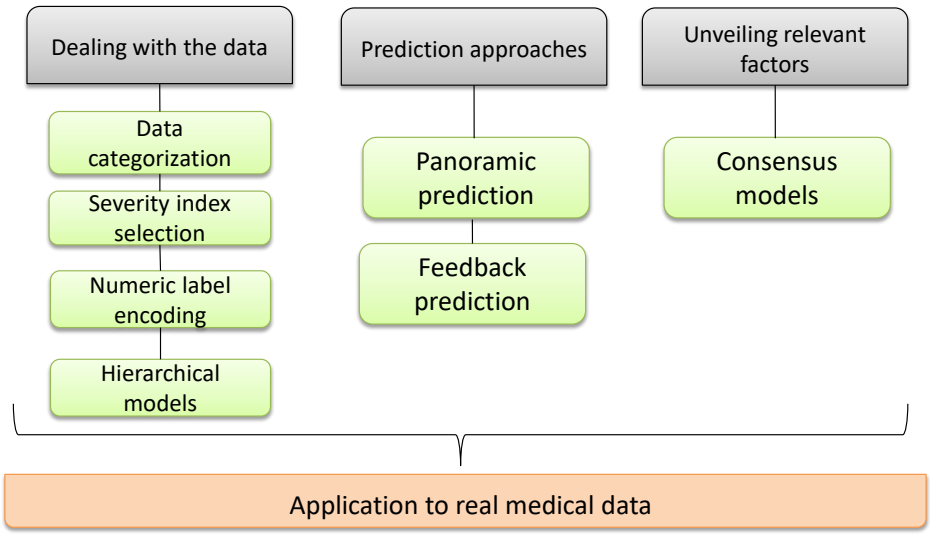


Figure 1.5: Research objectives. This scheme summarizes the different issues surrounding the research objectives.

1.4 Publications

The results of this thesis have lead to research publications in international conferences and journals. In the following lines these publications are shown with the detailed rankings of each one.

1.4.1 Journal papers

This thesis has generated the following articles in international journals:

1. Parrales, F., Del Barrio García, A., Gallego, M., Gago, A. V., Ruiz, M., Guerrero, A. P., Ayala, J. et al. Prediction of patient's response to OnabotulinumtoxinA treatment for migraine. *Heliyon*, vol. 5(2), pages e01043–e01043, 2019d
[CiteScore 2018=1.66]

2. PARRALES, F., DEL BARRIO, A. A., GAGO, A. B., GALLEGO, M. M., RUIZ, M., PERAL, A. G., DZEROSKI, S. & AYALA, J. L. SMURF: Systematic Methodology for Unveiling Relevant Factors in retrospective data on chronic disease treatments. *IEEE Access*, pages 1–1, 2019c. ISSN 2169-3536
[JCR 2018=Q1, IF=4.098]

1.4.2 Conference papers

This thesis has generated the following articles in national and international conferences:

1. PARRALES, F., DEL BARRIO, A. A. & AYALA, J. L. A study on the parallelization of moeas to predict the patient's response to the onabotulinumtoxina treatment. In *Proceedings of the Summer Simulation Multi-Conference*, page 12. Society for Computer Simulation International, 2019b
[CORE: B]

2. PARRALES, F., DEL BARRIO, A. A. & AYALA, J. L. Estudio sobre la paralelización de modelos MOEAs de predicción terapéutica con toxina botulínica tipo A en migraña. In *Actas de las Jornadas SARTECO 2019*, pages 96–101. Universidad de Extremadura, Servicio de Publicaciones, 2019a

1.5 Thesis structure

The rest of the thesis is organized as follows: Chapter 2 reviews the fundamental concepts related to the preprocessing, supervised classification and optimization metaheuristics. Chapter 3 presents the methodology proposed for processing the data related to the BoNT-A treatment. Chapter 4 describes the experiments carried out on the medical dataset for predicting treatment responses and extracting medical knowledge from data. The concluding remarks will be given in Chapter 5, as well as the future lines of work.

Finally, the Appendix A presents a copy of the ethical consent taken in this reseach study.

Chapter 2

Preliminaries

Success is neither magical nor mysterious. Success is the natural consequence of consistently applying the basic fundamentals.

Jim Rohn

Contents

2.1 Introduction

One of the biggest problems faced by CDs is their continuous treatment to mitigate or eliminate their symptoms. This must be considered when deciding if a continuous treatment can be beneficial to a specific patient. For example, patients suffering from Parkinson's disease usually discontinue the treatment due to its ineffectiveness when mitigating the pain (Beiske et al., 2009), which then involves wasting money. In order to avoid this, cost-benefit analyses have been applied, like those for patients with chronic kidney disease or hepatitis C (Klarman & Rosenthal, 1968; Leidner et al., 2015; Rein et al., 2015). The conclusions drawn after these studies are diverse. For some cases, doctors conclude that it is better to employ the treatment in short periods than in early phases of the CD (Rein et al., 2015). But, on the other hand, an earlier treatment has also been associated with a faster recovery (Wilkinson et al., 2004). Therefore, it is important to establish a prediction model of response to customize the treatment for each patient.

As briefly discussed in the literature reviewed in the previous chapter, the data mining techniques can be useful to reveal underlying patterns in the space of chemical and pharmacological attributes. These patterns can be decisive for the advancement of personalized medicine and more specifically, for the improvement of treatments in patients with migraine (Yosipof et al., 2018; Denny et al., 2018).

From the use of statistics and data mining, useful conclusions can be generated from data (Breiman et al., 2001). As expressed by the work of Larrañaga et al. (2018), these conclusions can be expressed in three different ways as those presented in Figure 2.1. These are: (1) Clustering, which aims to find groups of similar records; (2) Supervised classification, whose purpose is to forecast the response or output for future records; and (3) Discovery of associations, which means looking for (probabilistic) relationships between the input and output variables.

This thesis will focus on supervised classification techniques. In fact, this work will review the techniques that allow an easy medical interpretation.

Throughout this chapter, a brief collection of basic concepts around the data mining field will be presented in order to facilitate the understanding of subsequent chapters.

This chapter presents an overview about data mining fundamentals. Nonetheless, the examples and additional explanations are contributions of the author of this dissertation.

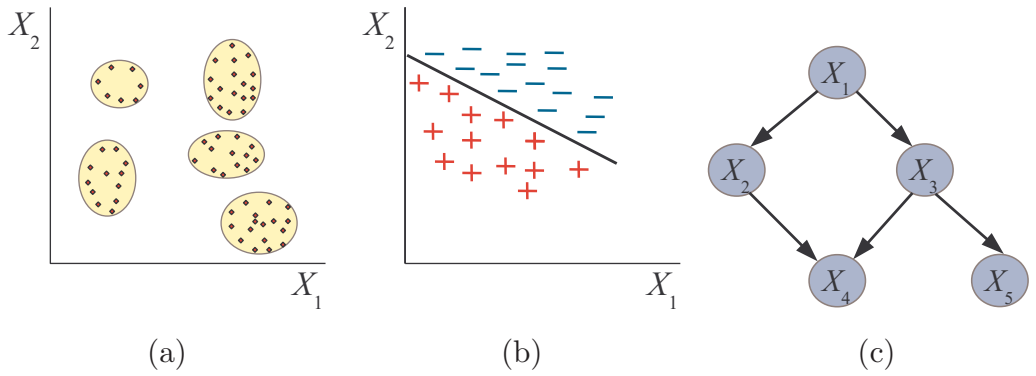


Figure 2.1: Three examples of tasks solved by statistics and machine learning methods. (1) Clustering. (2) Supervised classification. (3) Discovery of associations. Figure taken from Larrañaga et al. (2018).

2.2 Preprocessing

According to Yoo et al. (2012), the field of medicine differs considerably from other fields. The first difference is that the quality of the data within the biomedical and health fields is lower than that found in other fields due to many reasons. These are:

- Missing values are very commonly found in clinical records (Lin & Haug, 2008; Peterkova et al., 2018). This situation can even occur with patients of the same disease since they are not always subjected to identical laboratory tests (due to different ages, symptoms, family history and/or risk of complications).

- Obtaining high quality clinical data for data mining can be a somewhat difficult operation (Jollis et al., 1993; Dans, 1993). This situation is a consequence of the fact that the hospital information systems or their databases are designed mainly for financial or billing purposes and not for medical or clinical purposes.

- Some hospital centers still do not make full use of EMRs (Section 1.2.1). This situation is very common because part of the medical data (especially the results of laboratory tests) are not digitalized yet, which leads to obtain medical data that are often incomplete in terms of electronic or digital availability of the same (Prather et al., 1997). In addition, much of the patient's historical data is based on paper or scanned digital

format, so they cannot be used for data extraction without meaningful data preparation.

Therefore, it is important to study some data preprocessing methods in order to improve the quality of collected clinical data and the performance of classifications algorithms when applied to them to predict the responses to the BoNT-A treatment in all stages.

2.2.1 Dataset

To understand the preprocessing techniques that will be addressed throughout this section, it is necessary to begin by defining the key terms used in the data mining literature. The first point to mention is that the different techniques that allow to predict or classify results (medical labels in our case) work on a ***dataset***. This is defined as a collection of ***records***. A record represents an entity or concept. For the case of a medical database, the records can be patients. For the case of a database of a hospital, records can refer to rooms, departments, doctors, patients, treatments, among others. Every record is described by some ***attributes***, also called features, variables, features or columns. These records can be called samples, examples, instances or data objects. They are also called ***tuples*** when these records are stored within a database. In other words, the rows of a database correspond to the records and the columns correspond to the attributes of any dataset. Figure 2.2 shows each of the parts of which the dataset is composed and that will be described in detail below.

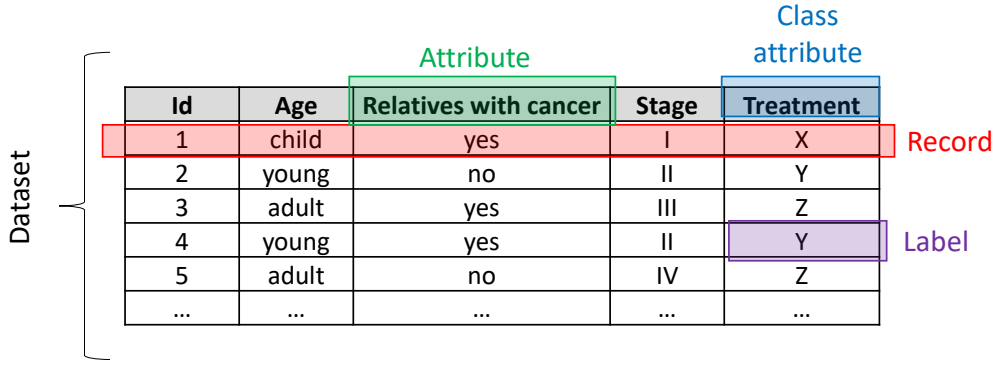


Figure 2.2: The different elements of a dataset.

An ***attribute*** is a data field, which represents a characteristic of a record. It is also commonly referred to as the dimension, feature, characteristic or

variable. The term *dimension* is commonly used in data storage. In the area of statistics, the term *variable* is commonly used. **Attribute** is the term that will be prefered to use in the present research work because it is usually employed in the areas of data mining and machine learning.

According to Han et al. (2011), *observations* are defined as the observed values for a given attribute. *Attribute vector* (or feature vector) refers to a set of attributes that are used to describe a given object. In addition, the distribution of data that involve only one attribute (or variable) is called *univariate*. A distribution that implies two attributes is called *bivariate*, and so on.

Within the different attributes in a dataset, the ***class attribute*** is/are the selected attribute(s) that indicate the response(s) of every record of the dataset.

2.2.1.1 HIT6

A key point to mention is how to measure the impact that headaches have on daily life of migraine patients. In this sense, one of the metrics mostly used as severity index for migraine according to the medical literature is the "Headache Impact Test" (HIT6) factor (Mathew & Jaffri, 2009; Silberstein et al., 2015; Grazzi & Usai, 2015).

The HIT6 (Kosinski et al., 2003) scale is a perceptional survey that is filled out by patients in order to measure their level of pain related with the migraine. In regular clinical practice, the BoNT-A response is considered successful by doctors if it reduces migraine attack frequency or days with attacks by at least 50% within 3 months. Response attributes such as the HIT6 score are reflected less consistently. Thus, in this thesis, where data were obtained retrospectively through the review of clinical histories, only a small set of patients have their HIT6 score. As a consequence, for the vast majority of cases an alternative way of determining the efficiency of treatment based on BoNT-A should be defined.

This value is obtained after patients fill out a standardized survey (Kosinski et al., 2003) consisting of six questions that capture the impact of headaches as well as their treatment. An example is shown in Table 2.1. These questions are:

1) When do you have headaches, how often is the pain severe?

2) How often do headaches limit your ability to perform usual daily activities including housework, your job, homework, or social activities?

3) When you have a headache, how often do you wish you could lie down?

4) In the past 4 weeks, how often have you felt too tired to do work or daily activities because of your headaches?

5) In the past 4 weeks, how often have you felt fed up or irritated because of your headaches?

6) In the past 4 weeks, how often did headaches limit your ability to concentrate on work or daily activities?

The values allowed for the answers are: never, rarely, sometimes, very often, and always. These values are graded with 6, 8, 10, 11 and 13 points, respectively. The HIT6 value is computed as the sum of all the individual scores. If the HIT6 value is 50 or higher, doctors interpret that the level of pain is enough to affect quality of life.

Table 2.1: HIT6 headache impact test example

	never	rarely	sometimes	very often	always
Question 1	X				
Question 2		X			
Question 3			X		
Question 4				X	
Question 5					X
Question 6	X				
Points added	6+6=12	8	10	11	13

2.2.2 Categorizing data

Data mining algorithms face some difficulties while evaluating heterogeneous data because they cannot infer a good model for predicting the outcome of the treatment (Tang et al., 2018). Medical data can come from images (X-rays, magnetic resonance, etc), interviews with the patients, laboratory data as well as the doctor's observations and interpretations (Cios & Moore, 2002). The homogeneity of the information can be achieved by simplifying and categorizing the data. For instance, this can be carried out through the transformation of heterogeneous clinical data to labels (Cimino et al., 1996; Huddar et al., 2016). Categorizing data is one of the techniques to explore in this work with the purpose of improving the prediction accuracy of the BoNT-A treatment responses. Due the heterogeneous clinical data provided by doctors from the two hospitals considered, it is necessary to categorize the data. For this purpose, the labels are previously defined and agreed by the experts in the disease to be analyzed in order to achieve an adequate

representation of the medical information (Cimino et al., 1996). However, the heterogeneity of data may still persist in medical factors previously categorized by doctors. Hence, as pointed by the aforementioned works, it is desirable that this PhD Thesis deals with heterogeneous data, leveraging the labelling provided by medical experts.

2.2.2.1 Equal-width-interval discretization

One type of data categorization is the one based on maximum and minimum values and intervals linearly defined between these bounds. This technique is also called *equal-width-interval discretization* (Liao & Lee, 2002). A uniform interval range width U can be defined when following the Equation 2.1, where I is the number of intervals to be obtained:

$$U = \frac{V_{max} - V_{min}}{I} \ . \tag{2.1}$$

It should be noted that V_{min} and V_{max} are the minimum and maximum values of the data, respectively, for a variable V.

For example, if it is necessary to obtain three intervals to refer to value 1, value 2 and value 3, respectively, the U value will take the value of $\frac{V_{max}-V_{min}}{3}$ and the intervals will be defined as presented in Figure 2.3.

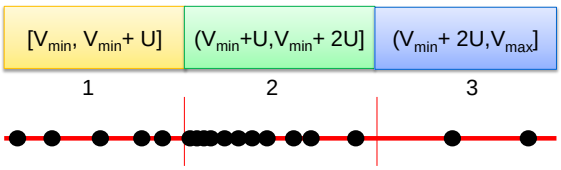


Figure 2.3: Defined intervals based on a uniform U for three categories.

The main problem of this approach is that many data could be concentrated in one of the categories (Muhlenbach & Rakotomalala, 2005), negatively affecting the training process of prediction models. In Figure 2.3, the black dots represent the various values of a continuous variable. Based on the linear categorization it can be observed how several points have been represented in the second category in contrast to the first and the third.

2.2.2.2 Categorization based on mean and standard deviation

Due to the aforementioned inconvenient, the method selected for the categorization of our medical data will be based on the mean and standard

deviation. Applying this method makes it possible to work with more homogeneous values. The mean and standard deviation categorization type centers the intervals around the mean (μ), and defines subsequent intervals by adding or subtracting the standard deviation (σ). For instance, if two categories are defined for a certain clinical attribute, the intervals to refer to value 1 and value 2, respectively will be as presented in Figure 2.4.

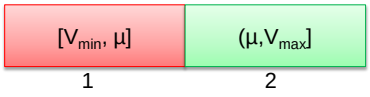


Figure 2.4: Defined intervals for two categories based on μ and σ.

If it is necessary to consider three categories, the intervals $[V_{min},\ \mu - \sigma]$, $(\mu - \sigma, \mu + \sigma]$ and $(\mu + \sigma,\ V_{max}]$ will be defined to refer to value 1, value 2 and value 3, respectively as shown in Figure 2.5.

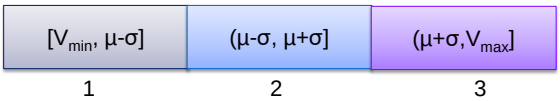


Figure 2.5: Intervals for three categories based on μ and σ.

By following a similar strategy it is possible to define multiple intervals (Parrales et al., 2019d). The pseudocode of the intervals generation for an attribute categorization is presented in Algorithm 1.

In this thesis, the categorization based on mean and standard deviation will be used given its better distribution of elements in each of its ranges than that performed by the equal-width-interval discretization.

2.2.3 Missing values

As mentioned in Section 2.2, the existence of missing values (NA) is very common within the medical environment. According to Barnard & Meng (1999), the treatment of missing data must be addressed, since otherwise they bring at least three main difficulties, which are:

1. Loss of information and efficiency.

2. Complication in the handling, calculation and analysis of data due to irregularities in data patterns and the non-application of standard software.

Algorithm 1: Intervals for categorizing attributes

Require: Number of intervals N, mean μ and standard deviation σ. An empty list of intervals $I = \emptyset$.

if $N = 1$ **then**
  $I = \{(V_{min}, V_{max})\}$
end if
if $N = 2$ **then**
  $I = \{(V_{min}, \mu), (\mu, V_{max})\}$
end if
if $N \geq 3$ **then**
  if isOdd(N) **then**
    $I = \{(\mu - \sigma, \mu + \sigma)\}$
    $\lambda = \frac{N-1}{2}$
  else
    $I = \{(\mu - \sigma, \mu), (\mu, \mu + \sigma)\}$
    $\lambda = \frac{N-2}{2}$
  end if
  $I_- = \{(V_{min}, \mu - \lambda\sigma)\}$
  $I_+ = \{(\mu + \lambda\sigma, V_{max})\}$
  for $j = \lambda - 1$ **downto** 1 **do**
    $I_- = I_- \cup \{(\mu - (j+1)\sigma, \mu - j\sigma)\}$
    $I_+ = I_+ \cup \{(\mu + j\sigma, \mu + (j+1)\sigma)\}$
  end for
  $I = I \cup I_- \cup I_+$
  sort(I)
end if
return I

3. Potentially very serious statistical bias due to the systematic differences between observed and unobserved data.

For example, consider the case of obtaining the mean of five values 1, 2, 7, 8 and 1 which is equal to 3.8. The problem begins when a missing value (NA) is found in the list, that is: 1, 2, 7, NA and 1. In this way, an undefined mean is obtained due to the unavailable data. The first solution to consider would be to eliminate all those records that take NA as values in any of their columns. This method of addressing missing values is called *listwise deletion* or *complete-case analysis* (Van Buuren, 2018). In this case, with the values 1, 2, 7, 1 an average of 2.75 is achieved, different from the obtained value of 3.8 with the complete data. Additionally, with this perspective, a large part of the available records would be eliminated, especially in the medical environment. As a consequence, other methods that consider how to replace missing values or how to work with them by grouping records according to the number of NAs should be studied.

2.2.3.1 Data imputation

One of the existing techniques in data mining to address the problem of missing data is the imputation of the data. It is responsible for completing the missing data with some plausible values. This has been a popular method for handling incomplete data problems (Barnard & Meng, 1999). This popularity is due in large part to the fact that once the missing values are completed, the standard data mining methods that operate on complete data sets can be easily applied to obtain predictive models, and thus avoiding the complication in the handling, calculation and data analysis due to irregularities in data patterns. However, to have an analysis based on datasets that are partially imputed, two requirements must be met. First, the imputation method or model must reasonably capture the real distributive relationships between the unobserved and the observed. Second, the analysis must take into account the uncertainty in the imputed values, because no matter how much effort one makes, the imputed values simply are not the actual observations.

According to Larrañaga et al. (2018) and Van Buuren (2018), there are different imputation approaches in the literature. These are:

- Single imputation (Allan & Wishart, 1930): refers to the imputation of a value for each missing data.

- Unconditional mean (or median) imputation (imputation based on the mean): replaces each missing value with the mean (or median) of the observed values of that variable (Yates, 1933).

- Regression imputation (Santos, 1981): in this case, the missing values for each variable are replaced with the predicted values from a regression of a variable.

- Imputation based on the Expectation-Maximization (EM) algorithm (Dempster et al., 1977): it is a model-based imputation approach. It groups records and estimates the missing values of a record based on its most similar records.

- Stochastic regression imputation (Schieber, 1978; Kalton & Kasprzyk, 1982): it is an extension of the imputation by regression. It tries to address the correlation bias by adding noise to the predictions.

- Multiple imputation (Rubin, 2004, 1996): It focuses on not creating a single dataset, but multiple datasets of imputed data in which different imputations. Each of the completed datasets are analyzed and the results are combined (for example, calculating their arithmetic mean) to produce the final imputed value.

According to Van Buuren (2018), the unconditional mean imputation is a quick and simple way to complete the missing data. However, the variance is usually underestimated, in addition to altering the relationships between the variables and biasing almost any estimate other than the mean. Additionally, it will skew the estimate of the mean when the missing data is not completely random. Van Buuren (2018) recommends applying this technique only as a quick way to impute some missing values. However, its use should be avoided in general.

Regarding regression imputation, it is necessary to emphasize that it incorporates the knowledge of other variables with the idea of producing more intelligent imputations. Nevertheless, Van Buuren (2018) warns that this approach is probably the most dangerous of all the methods described here because the relationships between the variables are not being preserved. In fact, the regression imputation artificially strengthens the relationships within the data. As a consequence, the correlations are biased, the variability is underestimated and the imputations are too good to become true. In summary, the use of regression imputation can become a recipe for obtaining false positives and spurious relationships.

The stochastic regression imputation represents an important conceptual advance. According to Van Buuren (2018), one may think that the data imputation is ruined by adding some random noise, but this is precisely what makes it suitable for this task. A well-executed stochastic regression

imputation preserves not only the regression weights, but also the correlation between variables. This technique estimates the intersection, slope and residual variance in the linear model. Afterwards, it calculates the missing value and adds a random value to it from the residual. Precisely, the idea of extracting a random value from the residuals is very powerful and forms the basis of more advanced imputation techniques.

Since imputing only one value (single imputation) for the missing value may not be correct in general, Rubin (Rubin, 1976) proposed the creation of multiple imputations that reflect the uncertainty of the missing data. Moreover, Rubin considers that a low number of imputations (five, for example) would be enough.

In this thesis, the multiple imputation method will be selected, since it is now accepted as the best general method to deal with incomplete data in many fields (Van Buuren, 2018). In this sense, several imputations will be made using the method of stochastic regression, since it preserves the correlation between variables.

2.2.4 Feature Subset Selection

The datasets for the analysis may contain hundreds of attributes, a situation that often occurs within the medical environment. The data mining task of *Feature Subset Selection* (FSS) (Lewis, 1962) is responsible for identifying and eliminating those attributes (features) that are considered irrelevant or redundant with the purpose of reducing the dimensionality of the dataset (n). The goal of applying this technique to the dataset is to improve the performance of the different classification algorithms in terms of time and accuracy. Additionally, it has the advantage of producing a more compact representation of the prediction model, helping to improve the patterns understanding by medical experts. However, such simplification is obtained in exchange for increasing the complexity of the modeling task due to the FSS process, especially if n is large.

This approach has certain advantages, such as offering a better understanding of the prediction model or a better generalization by reducing *overfitting*. This problem happens when a prediction model is very closely adjusted to the training data, so it does not perform well when predicting new observations (Molina et al., 2002). These methods have been applied to different neurological anomalies, for example: an attribute extraction and selection from EEG signals in combination with a sleep stages classifier (Şen et al., 2014), an automatic seizure detection system for newborns (Aarabi et al., 2006), or to assess the feasibility of employing accelerometers to charac-

terize the postural behavior of early Parkinson's disease subjects (Palmerini et al., 2011).

The FSS process can be highly expensive (Han et al., 2011). In fact, for a dataset with n attributes, 2^n possible subset combinations can be obtained. Therefore, it is necessary to make use of the heuristic search methods in order to explore promising regions of the search space. These methods usually apply a *greedy* approach. That is, while looking in the space of attributes, it always select what seems to be the best option at that time. Its strategy is to make an optimal choice locally with the hope that this will lead to an optimal global solution. Such greedy methods are effective in practice and can come close to estimating an optimal solution. The attributes are categorized as better and worse, typically based on tests of statistical significance. Many other attribute evaluation measures are also commonly used, such as the measure of *gain ratio* (Larrañaga et al., 2018). This metric is in turn based on the metric of mutual information, which is based on the entropy of Shannon (Shannon, 1948) that quantifies the uncertainty of the distribution of values in a random variable. For a discrete variable with l possible values, $x_1, \ldots, x_l$, its entropy is defined as:

$$H(X) = -\sum_{i=1}^{l} p(X = x_i) log_2\ p(X = x_i). \tag{2.2}$$

The mutual information $I(X, C)$ between any dataset attribute X and the class attribute C with m possible values is defined as:

$$I(X, C) = H(C) - H(C|X) = \sum_{i=1}^{l}\sum_{j=1}^{m} p(x_i, c_j)\ log_2\ \frac{p(x_i, c_j)}{p(x_i)p(c_j)}. \tag{2.3}$$

Thus, mutual information is interpreted as the reduction in uncertainty about C after observing X. According to Larrañaga et al. (2018), this metric has the disadvantage of preferring attributes with many different values to attributes with few different values. A fairer option is to use the information gain ratio defined as:

$$gain\ ratio = \frac{I(X_j, C)}{H(X_j)}. \tag{2.4}$$

Some FSS methods apply different types of filters, either univariate or multivariate. The univariate filtering evaluates each attribute with any attribute's relevance metric (for example, gain ratio), eliminating those that obtain low score. The selected attributes are used as input variables for the

classification algorithm. One of its disadvantages is that the dependencies among attributes are ignored, since they do not take into account the possible redundancy among them. This redundancy can be detrimental to the behavior of the classification model. Multivariate filtering techniques just address this problem by comparing different subsets of attributes and choosing a subset according to its relevance (with respect to the class attribute) and redundancy.

2.2.4.1 Correlation-based feature selection

Proposed by Hall (1999) in his doctoral thesis, it is one of the most widely used methods of multivariate attributes filtering. The goodness of a subset of attributes is defined in terms of its correlation with the class attribute (relevance) and the lack of correlation between feature pairs in the subset (redundancy). For a subset of attributes $\mathcal{S} \subseteq \mathcal{X} = \{X_1, \ldots, X_n\}$, CFS technique will look for that $\mathcal{S}^* = arg \max_{\mathcal{S} \subseteq \mathcal{X}} f(\mathcal{S})$, where:

$$f(\mathcal{S}) = \frac{\sum_{X_i \in \mathcal{S}} r(X_i, C)}{\sqrt{k + (k-1) \sum_{X_i, X_j \in \mathcal{S}} r(X_i, X_j)}}. \tag{2.5}$$

where k refers to the number of selected attributes, $r(X_i, C)$ is the correlation between the attribute X_i and the class attribute C, and $r(X_i, X_j)$ is the correlation between the attributes X_i and X_j. The correlations $r(X, Y)$ are given by the symmetric uncertainty coefficient defined as:

$$r(X, Y) = 2 \frac{I(X, Y)}{H(X) + H(Y)}. \tag{2.6}$$

The problem of maximization can be solved by using some heuristics such as the greedy approach.

In order to extend the CFS method for a multi-target classification problem, with z class attributes to classify, Fernandes et al. (2013) has proposed the following three approaches:

1. The union of subsets with higher-scoring attributes obtained by considering each class attribute separately.

2. The subset of attributes with the highest score of the composite class attribute that models all possible joint configurations of the class attribute.

3. The subset of attributes with the highest score of a modified metric, such that it rewards the correlation of each attribute in the subset with each of the z class attributes, defined as:

$$f(\mathcal{S}) = \frac{\sum_{X_i, C_z \in \mathcal{S}} r(X_i, C_z)}{\sqrt{k + (k-1) \sum_{X_i, X_j \in \mathcal{S}} r(X_i, X_j)}}. \tag{2.7}$$

This PhD Thesis will take into account the third approach since it considers in a better way the correlation of each attribute with each of the class attributes (Fernandes et al., 2013).

2.3 Supervised classification

Classification is one of the data mining tasks that allows to analyze the dataset by extracting models that describe its important attributes. For example, a ***classification or prediction model*** can be obtained for predicting medical insurance categories as normal, high or low risk. Another example may be the case of a medical researcher who needs to know which of the available and specific treatments a cancer patient should receive based on their collected clinical data. In each of these examples, the task of classification will build a model or classifier to predict the categorical response that will be given by different labels, such as "normal" or "low" or "high" risk for the data of the health insurance application or "treatment X", "treatment Y" or "treatment Z" for an oncological dataset. This PhD Thesis will focus on *supervised classifiers*, those classifiers that build models with labeled training data. It means that the label of every record of a training set is previously known. Afterwards, this model is applied to predict categorical labels of new records. These categorical labels can be represented by discrete values, where the order among the values has no meaning. For example, the values 1, 2 and 3 can be used to represent treatments X, Y and Z, where there is no implicit ordering among this group of treatments.

In the data mining field, many classification methods, pattern recognition and statistics have been proposed. Recent research in data mining addresses the generation of scalable prediction and classification techniques capable of handling large amounts of data resident on the disk or that incrementally accumulate (streaming). Nonetheless, the amounts of data collected in the employed dataset are still tractable enough for being accessible in RAM

memory. That is why in this PhD Thesis the proposed algorithms will focus on those techniques that use memory resident data.

2.3.1 Performance metrics

As Japkowicz & Shah (2011) and Larrañaga et al. (2018) have mentioned, performance evaluation measures of a classifier model are used as figures of merit for the supervised classifiers. There are several metrics or measurements and their choice depends on the objective and characteristics of the supervised classification problem, as well as the type of classifier used.

2.3.1.1 Confusion matrix

True/negative positives and true/false negatives are some of the most popular metrics used in medicine (Lavrač, 1999). To explain these metrics, it is necessary to start talking about what a confusion matrix is. As the name implies, it is a matrix that contains the key elements required in most performance measures in supervised classification algorithms. The contents of the cell of the position (i, j) presents the number of cases that really have the class label i and that the classifier assigns or predicts as the class label j.

Thus, the performance measures in the supervised classification are defined based on the entries in the confusion matrix. In binary classification problems the four counters of the confusion matrix are the number of true positives (TP), false positives (FP), false negatives (FN) and true negatives (TN). This confusion matrix is easily generalizable to problems of multiclass classification, that is, where there are more than two class labels. Figure 2.6 presents the contents of the matrix.

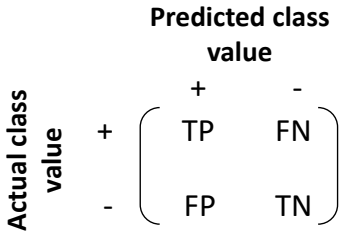


Figure 2.6: Confusion matrix for two class values

An example of the confusion matrix obtained when classifying the response to some medical treatment is presented in Figure 2.7. In this example, 15 and 14 medical records have obtained a correct classification as

positive and negative responses to the treatment, respectively. This implies that their predicted and real values match. In contrast, 6 medical records have obtained a bad classification, obtaining 2 responses predicted as "low" when their real values were "high". These 2 predicted responses are false negatives. Moreover, 4 responses were predicted as "high" when their real value were "low". These 4 predicted responses are false positives.

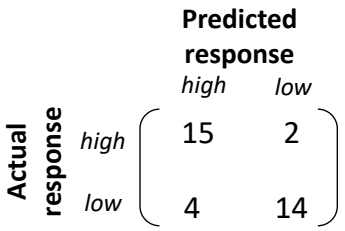


Figure 2.7: Example of a confusion matrix for high-low responses to any medical treatment.

2.3.1.2 Accuracy

The *accuracy* of a classifier given the test dataset and a learned classifier model, is the percentage of records of such dataset that are correctly classified by the model. The associated class label of each record of the test dataset is compared to the one that has been predicted by the classifier for that record. If the accuracy obtained from the classifier model is considered acceptable, the model can be used to classify future records for which the class label is not known.

Expressing it in the values of the confusion matrix, the equation that defines the value of accuracy is equal to:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \ . \tag{2.8}$$

For the example presented in Figure 2.7, the accuracy value of the classifier model will be equal to $\frac{15+14}{15+4+14+2} = 0.83$.

2.3.1.3 Sensitivity and specificity

Sensitivity and specificity values are often considered to be more important than high accuracy values in many medical problems (Lavrač, 1999). For example, let's consider that prediction of the response of some cancer treatment is needed. In this case, the detection of false positives is very expensive

in economic terms because it involves the application of treatments such as chemotherapy or immunotherapy unnecessarily, leading to their side effects. On the other hand, the detection of false negatives is delicate when it comes to an early stage of cancer, that is, when in theory there is a greater hope of cure with treatment. Another example could be the case of health insurers. These companies will prefer the detection of more erroneous cases of high-risk patients than a false detection of low-risk cases, given the economic implications involved.

Sensitivity measures the fraction of positive cases that are classified as positive, while specificity measures the fraction of negative cases classified as negative. Equations 2.9 and 2.10 define them based on the values of the confusion matrix.

$$Sensitivity = \frac{TP}{TP+FN} \, . \tag{2.9}$$

$$Specificity = \frac{TN}{FP+TN} \, . \tag{2.10}$$

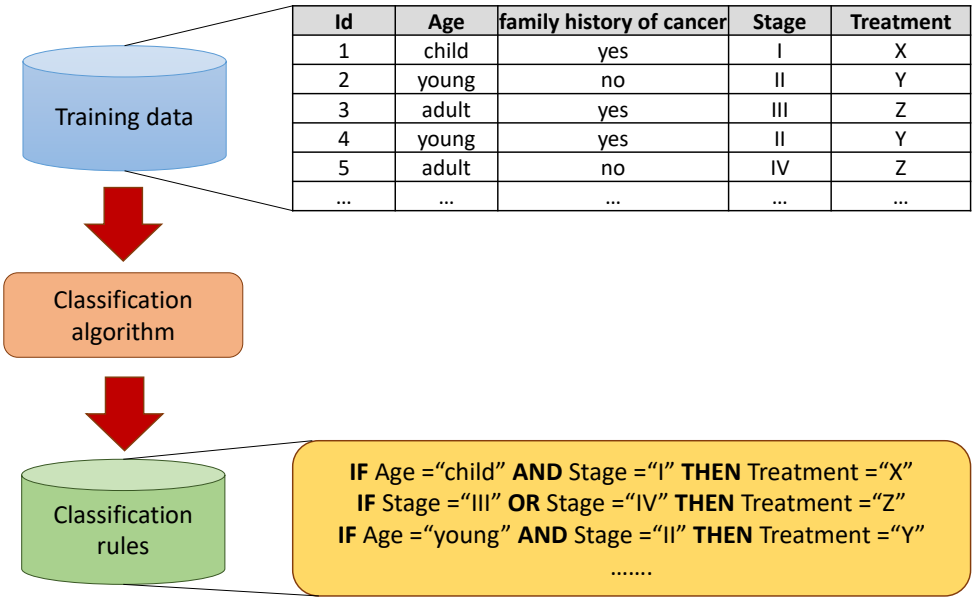
2.3.2 Classification process

Within the data mining field, the classification task refers to a process that includes two steps. The first is the learning step where a classification model is generated. The second step consists of a classification stage and it is where the model obtained in the previous phase is used in order to be able to predict the class labels for the new records.

In Figures 2.8 and 2.9 the classification process is presented, using for this purpose the aforementioned example of predicting the most appropriate oncological treatment for patients. It should be noted that the datasets presented in these figures have been simplified for the sake of clarity. It is noteworthy that especially in the field of medicine, the amount of attributes or medical factors that are considered within the dataset is enormous (Stewart et al., 2018).

Returning to the illustrative example, in the first step of the classification (Figure 2.8) a model is constructed. It describes a predetermined set of classes, factors, entities or data concepts. This is the learning stage or also known as the training stage of the predictive model. In this stage, some *classification algorithm* is used. This is who is responsible for the construction of a classifier model through learning from the training dataset. This dataset contains tuples of medical records and their associated class labels. Mathematically speaking, a tuple $\boldsymbol{X}$ is represented as a vector of n attributes,

$\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$. This vector represents the n measurements made in the tuple from n attributes of the database, respectively, $X_1, X_2, \ldots, X_n$. It is assumed that each tuple $\boldsymbol{X}$ belongs to a predefined class as determined by another attribute of the dataset called the ***class attribute***. This research work will focus on discrete and unordered class attributes. An attribute is called categorical when each of its values serves as a category. The individual tuples that are part of the training set are known as training tuples and they are taken randomly from the database that is analyzed.



Id	Age	family history of cancer	Stage	Treatment
1	child	yes	I	X
2	young	no	II	Y
3	adult	yes	III	Z
4	young	yes	II	Y
5	adult	no	IV	Z
...	...	...	...	...

Figure 2.8: Classification: Learning Stage. The training data is analyzed by a classification algorithm. Here, the class attribute is the type of oncological treatment, and the classification model is represented by the classification rules.

This first step is known as *supervised learning* since the learning is done knowing in advance the class label of each training tuple. This learning is opposed to unsupervised learning (or grouping), in which the class label of each training tuple is not known. For example, if the treatment applied to every patient of the training set is not known, clustering could be very useful in order to find out groups of cancer patients with similar values.

Mathematically, this first step of the classification process can be defined as the learning of a mapping function, $Y = f(\boldsymbol{X})$, which will allow to predict the class tag associated with a given $\boldsymbol{X}$ tuple. Usually, this function takes

the form of classification rules, decision trees or mathematical formulas. In Figure 2.8, the assignment function is represented by some classification rules that identify the most appropriate oncological treatment for each patient based on the data collected from it. The rules can be used to classify new medical records. Additionally, the rules allow to understand in a deeper way how the data are related.

The second step of the classification process is exemplified in Figure 2.9. As can be observed, the model extracted from the previous step is used to carry out the classification of tuples not used in the training stage. The aim is to estimate the accuracy (Section 2.3.1.2) of the classifier model. The training set should not be used to measure this value since it would be obtaining a too optimistic and unreal result. This topic will be expanded in Section 2.3.4.

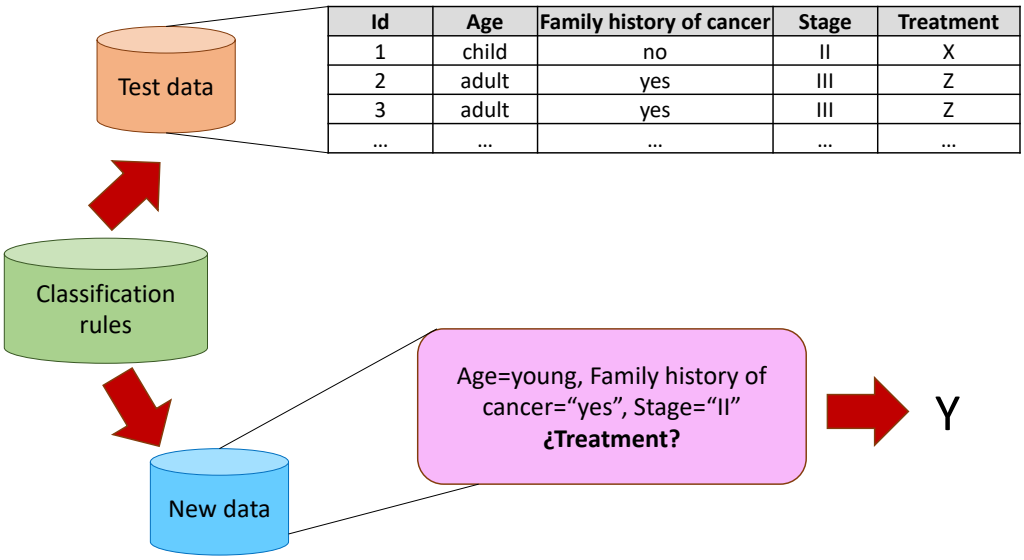


Figure 2.9: Classification: classification stage. The testing dataset is used to estimate the accuracy of the classification rules generated in the previous step. If the accuracy of the model is considered acceptable, these rules can be applied for the classification of new tuples or records.

2.3.3 Overfitting

When training the model, some particular anomalies can be incorporated into the training data that are not present in the general dataset. It is called *overfitting* and it implies the failure of the obtained model to generalize the knowledge that is intended to be acquired.

An example is presented in Figure 2.10, in which, a classification model is obtained from the training dataset in order to determine the most appropriate breast cancer treatment. Hence, this model is employed to classify new unseen records. Although the model's training classification error is zero, its error rate on the test set is 30%. In this example, the classification model is overadjusted to the training data, losing some generalization capability to classify new previously unseen records. Let's see as an example on one of the records of the testing dataset, more specifically, the clinical record with id=8. The most appropriate treatment for that patient is Y. However, if the classifier model is employed, the treatment X will be predicted since its age is "adult" and its stage is "II".

For avoiding the problem of overfitting, the complete dataset can be divided into two subsets. One for training and another for testing purposes. The testing dataset will not be used for model training. It is important to note that the testing dataset should have diverse samples with a sufficient quantity of them to be able to check the results once the model has been trained.

2.3.4 Honest estimation of accuracy

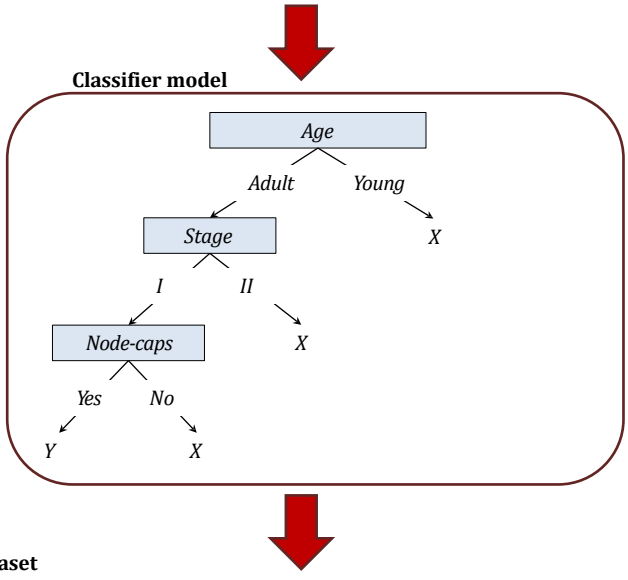
An important issue is knowing how to honestly estimate the accuracy of the prediction model. The first thing that comes to mind is to use the training set to learn the classifier model and then measure the accuracy of the model in the test set. However, as expressed by Larrañaga et al. (2018), this method (named *hold-out* in the literature) would be using only the training dataset to learn the final classifier model instead of learning from the whole dataset. In the work of Japkowicz & Shah (2011), various honest estimation methods are presented as hold-out, *k-fold* cross-validation or bootstrap. Among them, the *k-fold* cross-validation method will be considered in this thesis because it often outperforms the other estimators in some studies (Nakatsu, 2020; Borra & Di Ciaccio, 2010; Kohavi et al., 1995). This technique is implemented in the WEKA framework (Hall et al., 2009) that will be used for the experiments in this research work.

2.3.5 *k-fold* cross-validation

The *k-fold* cross-validation method (Kurtz, 1948) randomly divides the data set into k segments or folds of approximately equal size. To train the classifier model, the $k-1$ segments are used and the accuracy of the obtained model is evaluated in the remaining segment. This process is repeated about k times

Training dataset

Id	Age	Family history of cancer	Node-caps	Stage	Actual Treatment
1	Young	No	Yes	I	X
2	Young	Yes	No	II	X
3	Adult	No	No	II	X
4	Adult	No	No	I	X
5	Adult	No	Yes	I	Y



Testing dataset

Id	Age	Family history of cancer	Node-caps	Stage	Actual treatment	Predicted treatment
1	Adult	Yes	No	II	Y	X
2	Adult	No	No	II	X	X
3	Adult	Yes	Yes	II	Y	X
4	Young	Yes	No	II	X	X
5	Young	No	Yes	II	X	X
6	Young	No	No	II	X	X
7	Young	No	No	II	X	X
8	Adult	Yes	No	II	Y	X
9	Adult	No	Yes	I	Y	Y
10	Young	No	Yes	I	X	X

Figure 2.10: Overfitting example with clinical breast cancer records: the training dataset is used for building the classifier model. Afterwards, this model is used to classify the testing dataset. The incorrect predicted treatments are colored in red.

for each of the k segments. With them, k accuracy results are obtained. They are averaged to estimate the accuracy of the model obtained from the whole dataset. In this way, the classification model is learned on all the dataset. Figure 2.11 shows this entire process for 4 segments ($k = 4$) for illustrative purposes. The parameter k must be defined by the user. However, the literature of classification algorithms usually define k equal to 10 (Burman, 1989).

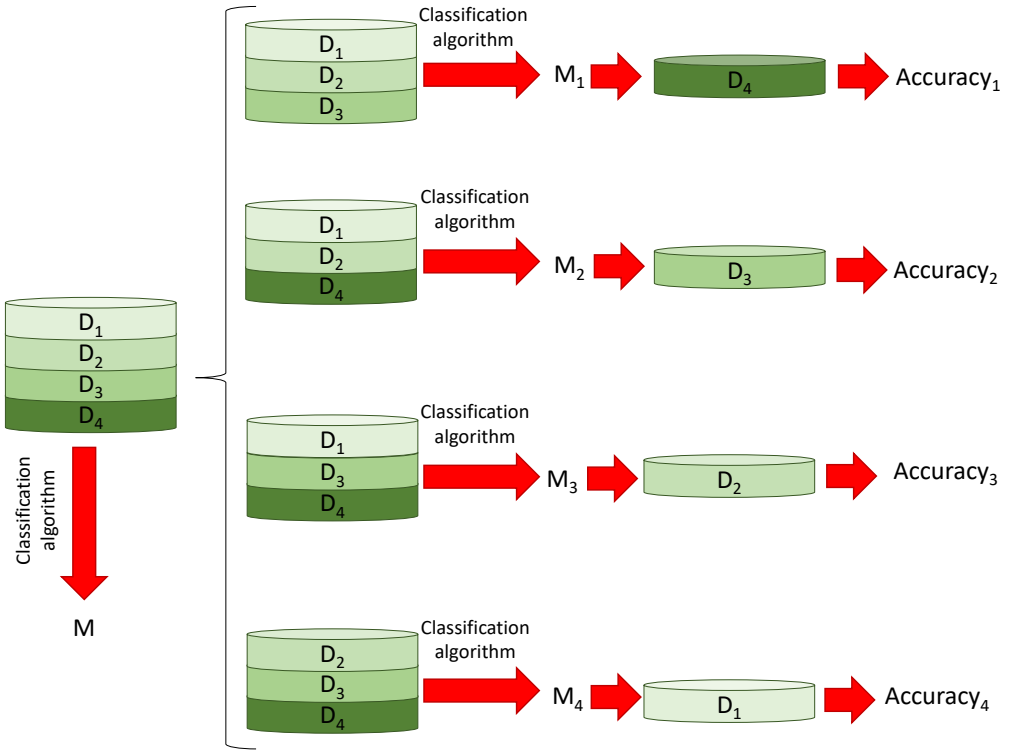


Figure 2.11: Example of k-fold cross validation with $k = 4$. An M model is obtained from the entire dataset on the left. To estimate the accuracy of such model, the dataset D is divided into four segments (D_1,D_2,D_3,D_4). A model is obtained from every of the four combinations of $k - 1$ segments (M_1,M_2,M_3,M_4). Each model is evaluated in its remaining segment to obtain the four values of accuracy to be averaged.

2.3.6 Multi-target classification

Section 2.3 has focused on describing the training and classification steps for predicting a single class attribute. However, in several applications, what

is really required is to learn classifying models that allow the prediction of several class attributes at the same time. This task of data mining is called *multi-target classification* or also called multi-objective or multi-label classification (Tsoumakas et al., 2009). With this, a tuple $\boldsymbol{X}$ represents a vector of n attributes, $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ from which a vector of s outputs $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_s)$ (with $s > 1$) is predicted (instead of a single output value) using a function $f(\boldsymbol{X})$ such that:

$$f(\boldsymbol{x}) : \boldsymbol{x} = (x_1, x_2, \ldots, x_n) \xrightarrow{f(\boldsymbol{x})} \boldsymbol{y} = (y_1, y_2, \ldots, y_s). \tag{2.11}$$

Each element of the output vector will be a binary value, indicating whether the corresponding label is relevant to the sample or not. Several tags can be active at the same time. Each different combination of labels is defined as *label set*.

For example, considering the example of the oncological treatment prediction, it may be necessary to find out the result of the treatment of five sessions of chemotherapy for a given patient. It could be obtained by applying the function $f(\boldsymbol{X})$ the following set of labels: {high, high, high, low, low}, which would indicate that the response of a certain oncological treatment will be highly effective only in the first three sessions. Multi-target classification algorithms are responsible for learning that function $f(\boldsymbol{X})$.

In the work presented by Madjarov et al. (2012), methods such as *predictive clustering trees* (PCT), *hierarchy of multi-label classifiers* (HOMER) and *binary relevance* (BR) have been recommended to carry out the learning of multi-target prediction models.

- In *predictive clustering trees (PCT)* (Blockeel & De Raedt, 1998), decision trees partition the set of examples into subsets in which the examples have similar values of the target variable, while clustering produces subsets in which the examples have similar values of the descriptive variables.

- In the *binary relevance (BR)* (Zhang & Zhou, 2007) method, the transformation of the multi-label prediction into z binary classification problems is considered. A prediction model is learnt for every target variable $(y_1, \ldots, y_z)$ independently. After that, all the results are combined to determine the predicted class set.

- In the *hierarchy of multi-label classifiers (HOMER)* (Tsoumakas et al., 2008), a hierarchy of multiple labels is built and a classifier is obtained for the label sets of each node of the hierarchy.

2.4 Optimization metaheuristics

As mentioned in Section 2.2.4, several heuristics can be used to find the subset of attributes that improve the performance of the classifying algorithms. For Martí et al. (2018), the term *heuristic* refers to the strategies that make use of readily accessible, loosely applicable information to control problem solving. For example, algorithms are a type of *heuristic*. On the other hand, the same authors mention that *metaheuristics* methods can coordinate the usage of several heuristics toward the formulation of a single method. All in all, *metaheuristic* algorithms are iterative procedures that guide a subordinate *heuristic*, intelligently combining different concepts to properly explore and exploit the search space (Glover & Kochenberger, 2006).

It is neccesary to mention that *metaheuristics* differ from the heuristics in that it can be applied to a large number of problems and not only to a specific field of application (Gendreau et al., 2010; Sörensen, 2015). For example, considers the search strategy. The hill climbing is an *heuristic* method employed to find local optimums, but it does not guarantee finding global optimum solutions (Skiena, 1998). For this reason, various *metaheuristic* methods have been proposed to improve local search heuristic in order to find better solutions (Blum & Roli, 2003). Some of these *metaheuristics* are SA, GRASP, variable neighborhood search (VNS), and the tabu search. However, the distinctions between *heuristic* and *metaheuristic* methods are inappreciable by some authors (Gandomi et al., 2013; Stojanović et al., 2017).

Previous to the use of any metaheuristic, it is very important to define what is the objective or *fitness* value that is necessary to optimize. For example, let's assume that it is necessary to find the selection of clinical data that allow obtaining a high percentage of accuracy for the prediction of a treatment response. In this case, the fitness value will be the percentage of accuracy and the target will be its maximization in order to achieve an accuracy close to 100%.

2.4.1 Simulated annealing

The simulated annealing method (SA) (Kirkpatrick et al., 1983) is a randomized search method for optimization. This technique is used in order to find those weights that improve the representation of the numeric labels encoded by doctors for each stage. SA is a stochastic, metaheuristic technique used in difficult optimization problems to approximate the global optimum of a given function in its search space. This approach has been widely employed to improve the performance of other algorithms. For example, SA has been

used to improve FSS in Sharma et al. (2012). Furthermore, SVM and SA have been combined to find the best selected attributes to increase the accuracy of anomaly intrusion detection in Lin et al. (2012), and for a hepatitis diagnosis method in Sartakhti et al. (2012).

The name has its origin from the phenomenon of the physical heating of a material such as steel. To heat it, this material is subjected to high temperature and then gradually cooled. Gradual cooling allows the material to cool to a state where there are few weak points. It achieves a kind of "global optimum" in which the whole object reaches a crystalline structure of minimal energy. If the material cools quickly, the object breaks easily in some parts because it would not have become strong in its entirety.

The SA method is an algorithm that begins with an initial solution that can be completely random, and every iteration makes slight changes in the solution (current solution) until it reaches a result close to the optimal solution. In this research, the percentage of classification error (100-accuracy) of the classifier model is the fitness value to optimize.

With the progress of SA, the current solution is altered, even if it is worse than the previous one. However, the probability of accepting a worse solution decreases with time (cooling process) and distance. A new solution is always accepted if it is better than the previous one. The probability of acceptance used is derived from the distribution proposed by Maxwell and Boltzmann. It is the classical distribution function for the distribution of an amount of energy between identical but distinguishable particles. Its value is equal to $e^{(-E_{diff}/T)}$, where T refers to its temperature and E_{diff} refers to the energy difference calculated as the fitness value distance between the initial solution and the minimum value reached so far (in the case of a minimization problem).

This algorithm has been implemented in different libraries. Hero library (Risco, 2016) has been selected because it implements the "natural optimization" (De Vicente et al., 2000), which means that the temperature does not need to be given because it is continuously tuned while running the SA algorithm through Equation 2.12.

$$T = \frac{K \times (C_{min} - C_{init})}{N} , \tag{2.12}$$

where N is the number of iterations, K is a constant that refers to the backward degree and time/quality trade-off and has been set to 1, and C_{min} and C_{init} refer to the current minimal cost and initial cost, respectively. The cost refers to the fitness value of the solution. The energy difference is defined

in Equation 2.13.

$$E_{diff} = C_{sol} - C_{min} \ , \tag{2.13}$$

where C_{sol} is the cost of the current solution. Finally, the probability (P) to compare with the random number (R) is given by Equation 2.14. This P value is the probability of changing to a new solution. This is calculated when C_{sol} is not lower than C_{min}. When $R \leq P$, SA moves the solution to another point within the search space to avoid being trapped in a local minimum.

$$P = e^{(-E_{diff}/T)} \ . \tag{2.14}$$

Figure 2.12 depicts a flowchart with the methodology proposed by De Vicente et al. (2000).

2.4.2 Multi-objective evolutionary algorithms

The goal of the multi-objective evolutionary algorithms (MOEAs) is to achieve a set of efficient solutions, non-dominated or *Pareto optimal solutions* (Zitzler et al., 2000). These set of solutions are called *Pareto optimal solutions* (denoted by X_P), when there is no other feasible solution that takes a lower value (in minimization problems) in some objective without causing a simultaneous increase in at least another. Mathematically, a multi-objective problem face the optimization of $n \geq 2$ objective functions. These functions are represented as: $f_1(x), f_2(x), \ldots, f_n(x)$. Their solutions are represented by $x_1, x_2, \ldots, x_n$. A solution $x_1 \in X$ dominates another solution $x_2 \in X$ if and only if $\forall i \in \{1, \ldots, n\}, f_i(x_1) \leq f_i(x_2)$ and $\exists j \in \{1, \ldots, n\}, f_j(x_1) < f_j(x_n)$. In this sense, A solution $x^* \in X$ is a non-dominated solution if and only if there is not another solution $x \in X$ such that x dominates x^*. X_P is the whole set of non-dominated solutions.

For example, doctors need to select the prediction model that achieves the best accuracy percentage when predicting the response of two stages of a migraine treatment, represented by f_1 and f_2 respectively. Let us the existence of three prediction models with the accuracy values represented by the tuples (f_1, f_2): (75%, 72%), (73%, 75%), (65%, 71%). These tuples represent the solutions x_1, x_2 and x_3 respectively. x_1 dominates x_3 when comparing their f_1 and f_2 values because $75\% > 65\%$ and $72\% > 71\%$. Also, x_2 dominates x_3 because $73\% > 65\%$ and $65\% > 71\%$. However, x_1 is not dominated by x_2 and viceversa because $75\% > 73\%$, but $72\% < 75\%$. Then, the resulting Pareto frontier would be composed of x_1 and x_2.

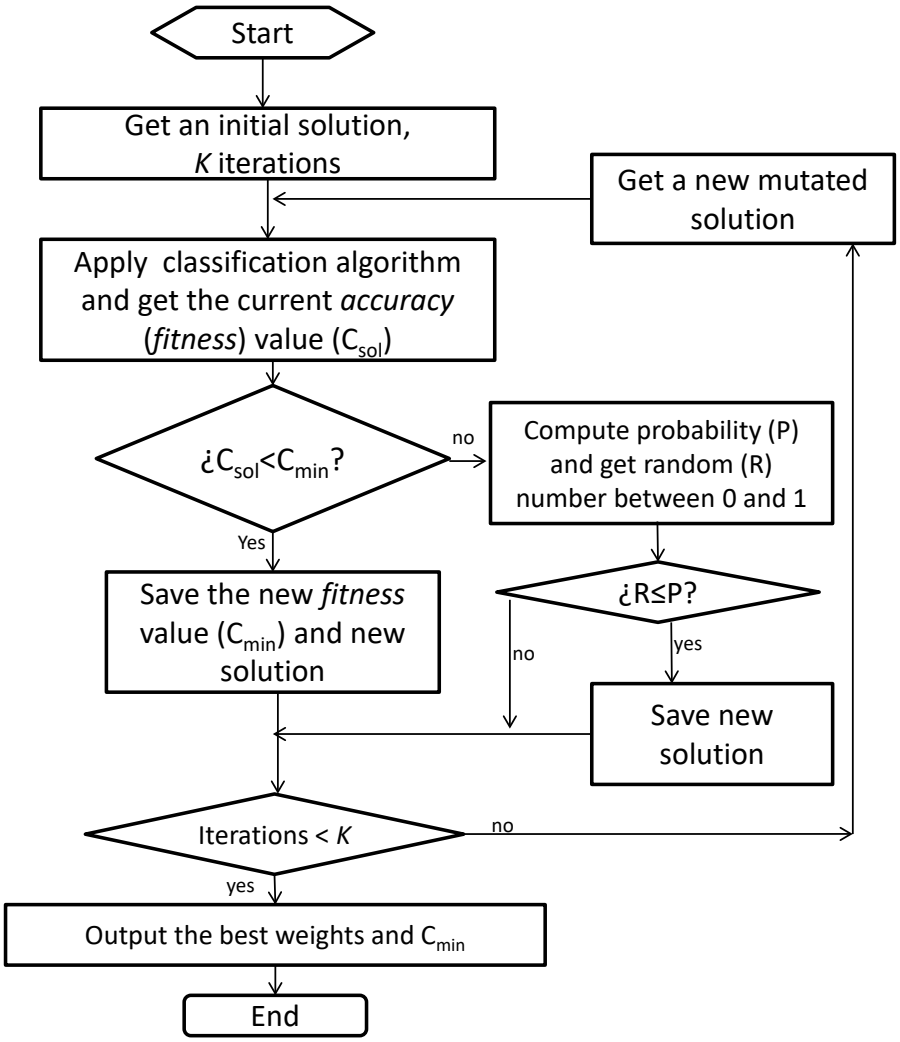


Figure 2.12: Flowchart with the Simulated Annealing-based methodology proposed by De Vicente et al. (2000)

The X_P set forms the Pareto frontier as depicted in Figure 2.13. An important point to consider is that MOEA algorithms handle a set of solutions (population) instead of a single solution as in the case of the SA. As a consequence of having more solutions, its computational cost is higher than algorithms with a single solution approach, specially when performing without parallelism (Durillo et al., 2008). The parallelization allows to distribute the computational load on different cores of the computer, making the execution of tasks efficiently. Parallel implementations of MOEAs can be used in order to achieve faster execution of algorithms and a superior numerical performance (Alba & Tomassini, 2002).

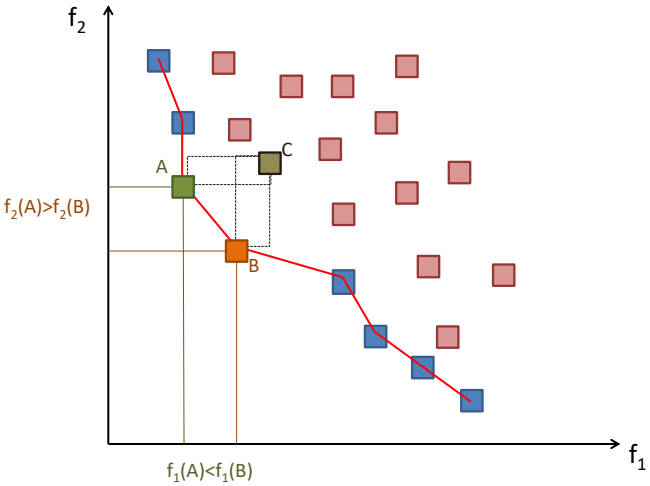


Figure 2.13: Example of a Pareto frontier (red line) formed by the set of Pareto optimal solutions. The boxed points represent feasible solutions, and smaller values are preferred to larger ones. Point C is not on the Pareto frontier because it is dominated by both points, A and B Points. A and B are not strictly dominated by any other, and hence do lie on the frontier.

Chapter 3

Methodology

We should be suspicious of any dataset
(large or small) which appears perfect.

David J. Hand

Contents

3.1 Introduction

D. Hand (Hand, 2018) mentioned that there is no perfect dataset. An analysis of its characteristics will allow to find the challenges to solve for obtaining a predictive model. In this chapter, an exhaustive study of the available clinical data is carried out. Based on this, the most convenient class attribute for measuring the efficiency of the migraine treatment is selected. Then, to get closer to medical needs, two prediction approaches are considered. They are: panoramic prediction and feedback prediction. Afterwards, some techniques are explored for knowing how to attack and solve the problem of missing values found in the medical records. Finally, in order to explore the medical characteristics described in the predictive models and reveal whether there are new medical findings, the use of consensus models is addressed.

3.2 Preprocessing

As the phrase at the beginning of this chapter mentions, there is no perfect dataset that is ready to obtain the prediction models of its class attributes. In order to achieve an efficient application of classification algorithms, it is necessary to explore and work on one of the most important components of this process, that is: the preprocessing of data.

In this section, the issues related to the treatment of data prior to learning prediction models are discussed.

3.2.1 Clinical data

The first step to take into account to visualize the problems to solve is the analysis of the available medical data. It has been collected in a retrospective way from the review of medical histories of patients with chronic migraine and under previous or current treatment with BoNT-A at the Headache unit of two tertiary-level hospitals. To this end, the approval of the ethics committee of both hospitals was obtained under the documents ANA-TOX-2015-1 and PI-17-832, which are provided as supplementary content.

As presented in Figure 3.1, a total of 173 patients were included (116 from *Hospital Clínico Universitario* in Valladolid and 57 from *Hospital Universitario de La Princesa*, in Madrid). Sixty-two baseline attributes were categorized. Therefore, they are the $X_1, \ldots, X_n$ attributes of the clinical dataset, where n is equal to 62. These attributes were related to the following points: clinical pain attributes, demographic attributes of patients,

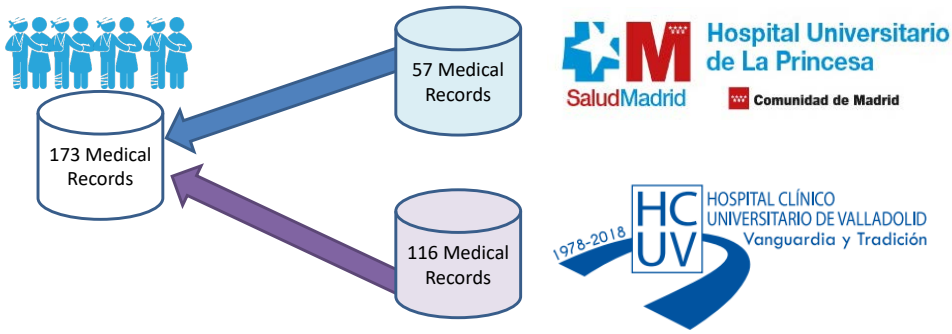


Figure 3.1: Source of our clinical data.

comorbidities, tested and concomitant preventive drugs, pain impact measures, and available analytical parameters. The latter were obtained from blood tests recorded in the clinical history that were performed for other reasons in the 3 months prior to, or 3 months after, the first stage, and included hemogram and liver, renal, thyroid, ferric, vitamin B12, folic acid and vitamin D profiles.

Figure 3.2 presents demographic data of patients. This figure shows that the majority of patients are between 30-55 years old. Also, most of the patients are women. This fact agrees with what is expressed in some studies (Gazerani & Cairns, 2020; Schwedt et al., 2019; Finocchi & Strada, 2014), where a high prevalence of migraine in women is found, and also falling within the aforementioned age range. In addition, Pelzer et al. (2019) found that migraine seems to be associated with a genetic predisposition (1st grade family) as in our patients.

The efficacy of BoNT-A was evaluated by comparing the baseline situation (before the first stage) and the situation after each of the three stages of treatment, through the following parameters: number of days of pain per month, percentage reduction in days with pain, subjective intensity of pain, number of days of disability due to pain per month, HIT-6 scale score, drug consumption for pain and adverse effects of stage. Since this was a retrospective study, not all the data could be obtained for each patient in a systematic way.

Some patients are *non-respondent*, while others respond after the i^{th} session. In order to predict the patients' behaviour after the stages, it is necessary to explore the patients' data before these take place. In other words, in order to predict the outcome after the i^{th} session, the clinical data of the patient as well as the outcome after the $(i-1)^{th}$ stage are required. Nev-

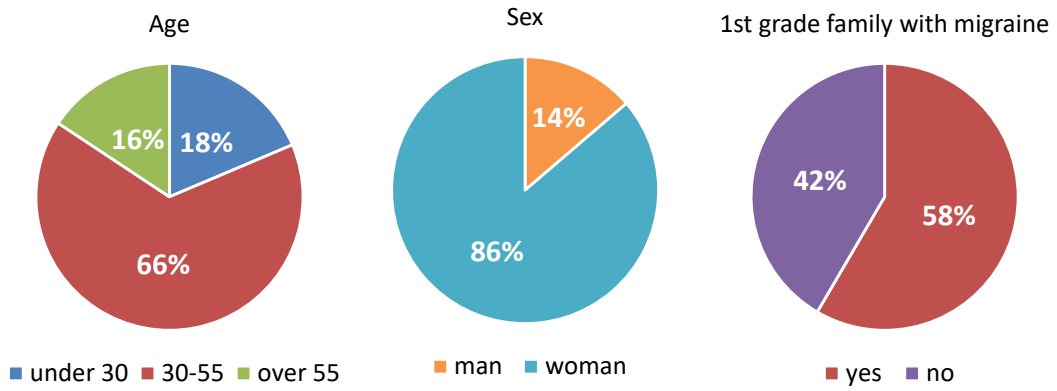


Figure 3.2: Demographic of patients.

ertheless, some problems are encountered while evaluating these data. For example, a small set of patients with many attributes is typically present in our medical databases. In addition, the incompleteness of data is another problem that must be dealt with. Some attributes are given as continuous numeric values while other attributes are categorized by doctors. All in all, it is hard to properly process all this information. As a consequence of these heterogeneous data, algorithms cannot infer a good model for predicting the outcome of the treatment. An example of these attributes can be observed in Table 3.1. In it, the age is represented by natural numbers while the body mass index and hemoglobin take decimal values. Furthermore, platelets can take very large values while creatinine can take very small values. Finally, the reduction effects take categorical values (from 1 to 4). This example shows the importance of categorizing the values prior to learning predictive models.

Table 3.1: Example of attributes in our clinical data.

Toxin-age of onset (years)	Body mass index (kg/m^2)	Hemoglobin (g/dL)	Creatinine (mg/dL)	Platelets (u/mcL)	Reduction effects (1-4)
51	20.39	13.4	0.71	213000	4
49	26.5	14.2	0.55	252000	2
36	23.15	13.5	0.44	304000	3
26	17.7	13.1	0.66	218000	2
31	NA	14.8	0.71	327000	1
50	NA	16.2	0.74	327000	3

3.2.2 Class attribute selection

Once the available data have been analyzed and with the purpose of measuring how efficient a treatment stage has been, it is neccesary to define the *class attribute*, also called *severity index* in the medical domain. In other words, the class attribute is the selected clinical attribute used to measure the effectiveness of treatment.

In the migraine scenario, the severity is typically evaluated through the HIT6 value (Yang et al., 2011), the intensity, the duration, the frequency of attacks (Gasbarrini et al., 1998) or the headache days (Schoenen et al., 1998). Hence, the severity index is something inherent to the CD under consideration, but also depends on the number of EMRs containing the index (Parrales et al., 2019d).

In this section, the HIT6 value and its limitations are exposed. In addition, a class attribute based on both the reduction and adverse effects is proposed to tackle the limitations imposed by the use of HIT6.

3.2.2.1 HIT6

As it has been mentioned in Section 2.2.1.1, HIT6 is a survey that allows to measure the level of pain associated with migraine episodes. As this metric is perceptional, this thesis has focused only on those medical records that contain the HIT6 value prior and after the stage (consecutive stages). By defining the class attribute as the difference between the two values, as Equation 3.1 indicates, the bias due to different perceptions from different patients is diminished. According to Silberstein et al. (2015), if the HIT6 value after the stage diminishes by more than 30%, the treatment is considered as successful, and unsuccessful otherwise. Hence, for this particular class attribute, only two categories have been defined, namely: successful and unsuccessful.

$$HIT6_{dif} = HIT6_b - HIT6_a \ . \tag{3.1}$$

A very hard limitation is that the HIT6 values are rarely found in our clinical database. In fact, only 18 out of 173 records from our clinical dataset had the perceptional HIT6 value before stages, and only 12 and 3 contain this value after the first and second BoNT-A treatment stages, respectively.

3.2.3 Reduction and adverse effects

As a consequence of the lack of HIT6 value in many of the collected clinical records, the reduction (R) and the adverse (A) effects, which are more fre-

quently found in collected records, have been considered to define the class attribute.

R and A are measurable values directly provided by doctors from an objective point of view based on definitions. R is a clinical objective value categorized from 1 to 4 according to the percentage of reduction of days of migraine (RD), being 1 when $RD \leq 25\%$, 2 when $25\% < RD \leq 50\%$, 3 when $50\% < RD \leq 75\%$ and 4 when $RD > 75\%$. A is equal to 1 when there are no adverse effects, 2 when there are mild adverse effects (easily tolerated), 3 when there are moderate adverse effects (interfere with usual activities and may require suspension of treatment) and 4 when there are serious adverse effects (incapacitate or disable usual activities, and require suspension of treatment as well as medical intervention).

A high level of R indicates good treatment results, while high levels of A point to many adverse effects. Hence, in order to obtain a directly proportional attribute, our class attribute (N_{AC}) has been determined by dividing R and A, as Equation 3.2 shows.

$$N_{AC} = \frac{R}{A} \,. \tag{3.2}$$

A similar approach to the one based on HIT6 (two response categories: high and low) (Silberstein et al., 2015) has been considered for class attribute categorization, instead of the three categories (low, medium and high) used for the rest of the clinical attributes. Thus, following this approach, two intervals (low and high) need to be defined before trying to predict the efficiency of the treatment when using N_{AC} as class attribute.

Responses are labeled as "high" for those N_{AC} values falling within the *[cut-off point, V_{max}]* interval, while responses are labeled as "low" when the N_{AC} value falls into the *[V_{min}, cut-off point)* interval. In this case, $V_{min} = 0.25$ occurs when $R = 1$ and $A = 4$, while $V_{max} = 4$ occurs when $R = 4$ and $A = 1$.

According to Silberstein et al. (2015), responses are considered as "high" when a patient obtains a decrease higher than 30%, as the criterion used for the HIT6 value in the PREEMPT clinical trial. For this reason, the cut-off point value ($V_{cut-off}$) is obtained with the Equation 3.3. A cut-off point of 1.40 is obtained when replacing $V_{min} = 0.25$ and $V_{max} = 4$. Thus, values lower than 1.40 represent the 30% of the values that N_{AC} can take. Then, the "low" and "high" categories are defined with the intervals [0.25; 1.40) and [1.40; 4], respectively. Table 3.2 depicts an instance of the N_{AC} computation

using different values provided by the hospitals.

$$30\% = \frac{V_{cut-off} - V_{min}}{V_{max} - V_{min}} \times 100\% \, . \tag{3.3}$$

Table 3.2: Class attribute categorization.

Reduction effects (R)	Adverse effects (A)	R/A	Categorized value
1	1	1	low
2	1	2	high
3	2	1.5	high
1	2	0.5	low

Applying the class attribute defined here for each of the three stages of the treatment, the distribution of high-low values is obtained over the records of the medical dataset, as shown in Table 3.3. This results in the following baseline values of accuracy: 56.64%, 58.95% and 51.44%. These values refer to classifying all the records with the most frequent label for each stage of the treatment. More in detail, 98, 71 and 89 patients have a high response for the first, second and third stage of the treatment. On the other hand, 75, 102 and 84 patients have a low response for the first, second and third stage of the treatment.

Table 3.3: Distribution of high-low categories through stages.

Response	Stage 1	Stage 2	Stage 3
high	98	71	89
low	75	102	84

3.2.4 Data categorization

Due to the heterogeneous clinical data provided by the doctors of the two hospitals, the data must be categorized. For this purpose, the labels are first defined and agreed by the experts in the disease. Then, they are analyzed to achieve an adequate representation of the medical information (Cimino et al., 1996). However, the heterogeneity of data may still persist in medical factors previously categorized by doctors. Therefore, as noted in Section 2.2.2, it is desirable that this work deals with heterogeneous data, taking advantage of the labeling provided by medical experts.

The method selected for the categorization of the collected medical data is based on the mean and standard deviation presented in Section 2.2.2.2. The number of categories to obtain is defined as 3 for each attribute that

contains numerical data. The categorization of class attributes will follow the approach addressed in Section 3.2.3.

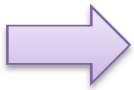
3.2.5 Numerical label encoding

At this point of the methodology, all the values have been categorized thanks to the mean and standard deviation based categorization. Although everything is categorized, there are data with non-numeric labels given by doctors. With the purpose of homogenizing the medical labels and working with numerical optimization algorithms as well as allowing a better representation of the labels with respect to the models, the nominal labels established by the doctors should be converted to numeric labels. This can be done by using consecutive natural numbers different to 0 for each label. Although this basic encoding method has the advantage of being simple, the numerical values can be misunderstood when applying and obtaining prediction models with the data mining algorithms. For example, a variable that identifies the sex of the patient will take values of M for men and W for women. When these nominal values are converted to numeric labels, they can be transformed into 1 and 2 respectively. However, this does not imply that one of them is greater than or lower than the other.

Another encoding approach is called one-hot encoding (Yu et al., 2019). The strategy of this method is to convert the different column labels of the original dataset into columns of a new dataset, defining a column for each different value. Then, the cell values of the new dataset will be filled with 1's or 0's (true/false) in the corresponding column according to the label value of the original dataset. An example can be seen in Figure 3.3. It has the benefit of not weighting a value improperly. However, its drawback consists of adding many new columns to the data set. Hence, this approach is not good for processing medical data, since the purpose is to reduce the number of medical factors to be contempled by the data mining algorithms.

Given the limitations of the one-hot approach, two approaches have been proposed to improve numerical coding in a dataset. The first is based on the use of SA. This approach is valid in the case of learning one-target prediction models, since SA optimizes a single objective. To extend it to a multi-target environment, the average accuracy of all targets can be considered as the value to be optimized. However, a better optimization can be achieved through the second approach based on the use of MOEAs, given that several optimization objectives are allowed. Below, the description of each of the proposed methods is presented.

Register	Feature 1
1	AA
2	AB
3	AA
4	AC
5	AC
6	AA



Register	F1-AA	F1-AB	F1-AC
1	1	0	0
2	0	1	0
3	1	0	0
4	0	0	1
5	0	0	1
6	1	0	0

Figure 3.3: One-hot example. It converts three different labels of the "Feature 1" column, creating a column for each different label value and adding the value of 1 or 0 depending on the label value that the record takes.

3.2.5.1 With SA

"SAR encoding" is proposed by Parrales et al. (2019c) and is designed for finding a better representation of the numeric labels. This technique produces a data transformation for achieving high prediction accuracies without adding more columns to the dataset. It is called SAR because this technique consider the SA algorithm (Kirkpatrick et al., 1983) and a rounding operation to perform small numeric label perturbations for each column of the medical dataset. The inputs of this method are the dataset to be analyzed, the number of decimal digits to consider (D) and the classification algorithm to be used to generate the predictive model. The outputs are the set of optimal column weights (W_{opt}), the optimal number of fractional digits for rounding (d_{opt}), and the optimized classification model (M_{opt}). The employed variables are defined in Table 3.4. The different steps in SAR are shown in Figure 3.4 and explained in the following lines:

- The input of the algorithm is an initial dataset O containing m clinical records, each containing the same set of n medical factors (columns) $c_1, c_2, c_3, \ldots, c_n$. The conversion of nominal labels to numbers is done following a consecutive order of integers begining with 1. It is done for the n columns of the dataset. The modified dataset will be called O' with $c'_1, c'_2, c'_3 \ldots, c'_n$ as modified columns.

- Once the O' dataset is generated, the next step is performing the attribute weighting task. For it, the SA algorithm will find the optimal weights w_j, $1 \leq j \leq n$, i.e. one for each column $c'_j \in O'$. The weights w_j will reflect the degree of relevance of a column c'_j for a problem to solve, where $w_j \in \mathbb{R}\{0, 1\}$. The values of every cell $o'_{i,j} \in O'$ are multiplied by the corresponding weight w_j through the $o'_{i,j} \times w_j$ operation,

Table 3.4: Description of variables employed in the SAR encoding.

Name	Description
O	Training dataset.
O'	Modified dataset.
m	Number of records of the initial dataset.
r	Number of records of the training dataset.
n	Number of columns of the training and test datasets.
c_i	Medical factor (column) of a dataset.
T	Threshold, number of columns that will be taken into account for grouping the dataset records.
W	Set of weights of the columns of a dataset.
w_j	Weight of the j^{th} column of a dataset.
$o_{i,j}$	Cell value of the i^{th} record and j^{th} column of a dataset O.
D	Number of decimal positions to analyze.
d	Decimal position.
O^d	O' modified dataset and rounded to the d decimal position.
c_j^d	j^{th} column of the O^d dataset.
$o_{i,j}^d$	Cell value of the i^{th} record and j^{th} column of a dataset O^d.
ACC_i^d	Accuracy of the O^d dataset in the i^{th} stage.
ACC_{AV}^d	Average accuracy of the O^d dataset.

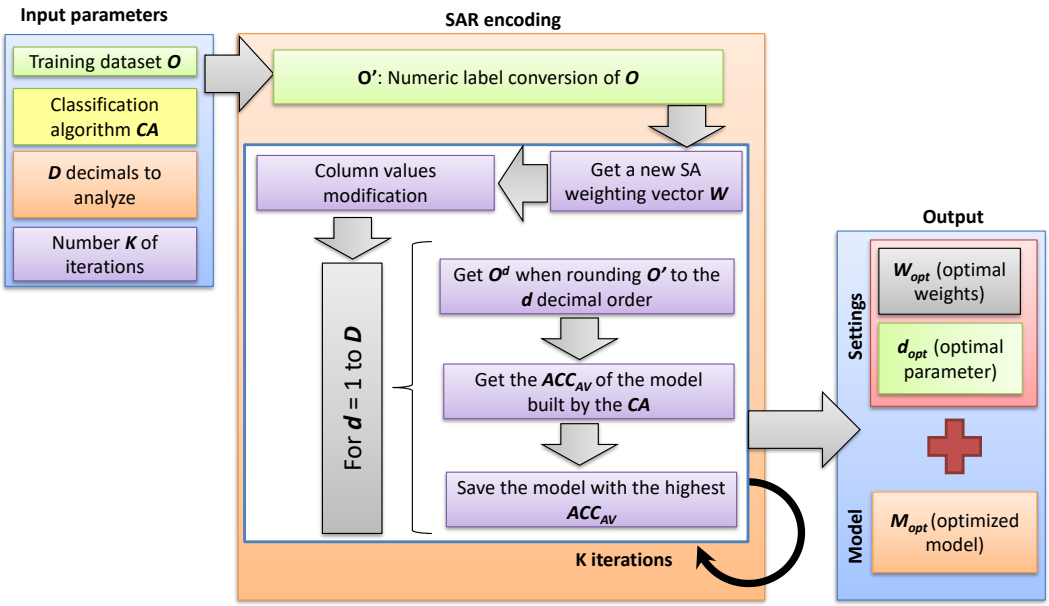


Figure 3.4: SAR encoding diagram.

$\forall i$, $1 \leq i \leq m$ and $\forall j$, $1 \leq j \leq n$. This multiplication is illustrated in Figure 3.5.

- The O' dataset is rounded to the nearest tenth, hundredth, thousandth, and other decimals in order to generate small perturbations among the different numeric labels in each column (Higham, 2002; García & Tarragona, 2010). These rounded labels will generate modifications in the prediction models learnt by the classification algorithms that work with real numbers. The number of decimals to be considered when rounding is defined by the parameter D, generating different datasets O^d with columns c_j^d and cells $o_{i,j}^d$, where $1 \leq d \leq D$, $1 \leq i \leq m$ and $1 \leq j \leq n$. For example, if D is equal to 3, three modified datasets O^1, O^2 and O^3 will be generated to train the SAR optimization, rounding to the nearest tenth, hundredth and thousandth for the first, second and third modified datasets, respectively. An example of this step is shown in Figure 3.5 when rounding O' to the hundredth (O^2).

- The prediction models are learnt by the classification algorithm when training it with each of the modified datasets O^d. The accuracy for each of the s stages, ACC_i^d, $1 \leq i \leq s$, of any modified dataset O^d, $1 \leq d \leq D$, is obtained through Equation 3.4. In the case of a one-target prediction model, s will be equal to 1. True positives (TP^d)

and true negatives (TN^d) refer to the correct prediction of positive and negative responses to treatment of the i^{th} class attribute, respectively. False positives (FP^d) and false negatives (FN^d) refer to the wrong prediction of positive and negative responses to treatment of the i^{th} class attribute, respectively.

$$ACC_i^d = \frac{TP_i^d + TN_i^d}{TP_i + FP_i^d + TN_i^d + FN_i^d} \ . \tag{3.4}$$

- The average accuracy of all the s stages ACC_{AV}^d (Equation 3.5) associated to the O^d modified dataset will be the value to be optimized. In this sense, $(1 - ACC_{AV}^d)$ is defined as the fitness value to be diminished by the SA algorithm. A number K of iterations is defined as input parameter in order to limit the execution of the SA algorithm.

$$ACC_{AV}^d = \frac{1}{s}(\sum_{i=1}^{s} ACC_i^d) \tag{3.5}$$

- The outputs of the SAR encoding are the transformation settings to apply in the initial O dataset and the optimized model M_{opt}. These settings are composed of the W_{opt} set of weights to apply to columns of the initial dataset and the d_{opt} number of fractional digits that achieved the best accuracy. These outputs will be used to transform the data and to apply the model M_{opt}.

Applying the proposed SAR encoding, the minimization of the prediction error for all stages is not solved simultaneously. This situation is a consequence of the fact that the SA technique does not consider the optimization of multiple objectives. This is the reason why an average error for all stages (100-ACC_{AV}) has been considered as the fitness value to be diminished in a multi-target prediction scenario.

3.2.5.2 With parallel MOEAs

Our label encoding problem can be considered as multiobjective when considering the improvement of accuracy of multi-target prediction models. Since SA is limited to optimize a single objective, the use of MOEAs (Section 2.4.2) will be considered for adapting the SAR encoding to a multi-target prediction scenario, while SA will be applied when improving numeric labels in one-target prediction scenario. This adaptation of the SAR encoding to

W (weights):

w_1	w_2	w_3
0.795269560373731	0.18469775	0.767716221

O' (dataset):

c'_1	c'_2	c'_3	y_1	y_2
1	2	3	low	high
2	3	1	low	low
3	1	2	high	high

$o'_{ij} \times w_j$:

$c'_1 \times w_1$	$c'_2 \times w_2$	$c'_3 \times w_3$	y_1	y_2
0.79526956	0.369395501	2.303148663	low	high
1.590539121	0.554093251	0.767716221	low	low
2.385808681	0.18469775	1.535432442	high	high

O^2 dataset:

c^2_1	c^2_2	c^2_3	y_1	y_2
0.80	0.37	2.30	low	high
1.59	0.55	0.77	low	low
2.39	0.18	1.54	high	high

Figure 3.5: Weighting dataset and rounding to the hundreth (d=2).

the multi-target scenario shall be named "A Multi-Objective and Rounding encoding" (**AMOR encoding**).

In a multi-target prediction scenario, it is possible to define the prediction error of each stage separately as a goal to minimize, instead of minimizing the average error of all stages. In this sense, the AMOR encoding will follow the SAR encoding approach for finding the optimal labels, replacing the SA metaheuristic by any of the MOEA methods. In addition, the fitness function need to be redefined. Thus, the new goal will be the minimization of the prediction error e_i for all the stages (s), described by Equation 3.6:

$$e_i = 100 - Acc_i \ , \quad \forall i \in [1, s] \ , \tag{3.6}$$

where Acc_i refers to the accuracy percentage of the corresponding prediction model. In the available clinical dataset, three stages are contemplated. This implies that there are three objectives to be minimized simultaneously, i.e. e_1, e_2 and e_3. The approach has been implemented using the MOEA framework presented in (Hadka, 2019). More specifically, the MOEAs that can be parallelized will be selected for diminishing the computational cost. Those selected algorithms are: GDE3 (Kukkonen & Lampinen, 2005), PESA2 (Corne et al., 2001), SMPSO (Nebro et al., 2009), NSGA-II (Deb et al., 2002), NSGA-III (Deb & Jain, 2014) and SPEA2 (Zitzler et al., 2001).

3.3 Prediction approaches

With the purpose of predicting the treatment response to the different stages of BoNT-A treatment, several classification algorithms have been considered when building the prediction models. In addition, two prediction approaches will be considered, those are: panoramic prediction and feedback prediction. Panoramic prediction makes it possible to decide whether the treatment will be beneficial without using previous treatment responses and without involving unnecessary treatments. On the other hand, the feedback prediction considers the results of previous stages of the treatment. Their application to the collected medical data set will be addressed in this section.

3.3.1 Panoramic prediction

The suitability of this method arises from studying the convenience of a certain drug for every stage involved in a CD treatment (Beiske et al., 2009). Predicting the response to every session of the treatment fits to what the personalized medicine is searching: allowing a cost-benefit analysis by the doctors and thus deciding whether the cost and other details inherent to the medication can be assumed by the patient (Suhrcke et al., 2006). Nevertheless, this first prediction approach aims to analyze a medical scenario of therapeutic response to a given treatment without any prior knowledge or *feedback*. This implies that the results of the first or subsequent treatment applications are not known yet. Hence, the responses cannot be used at different treatment stages to perfect the prediction model.

In contrast with the traditional one-target attribute prediction, this approach allows to carry out a simultaneous prediction of various responses from the described attributes of a medical record (Waegeman et al., 2019), where the prediction model is obtained with the use of multi-target classifier algorithms, as presented in Section 2.3.6.

It is important to incorporate this panoramic prediction approach to the proposed methodology due the goal is to present the prediction to treatment response in subsequent stages. This goal can be translated into a multi-target problem. This implies an important advantage because it is not necessary to obtain clinical data after a session of treatment to guess the response after the following session.

The selected methods for obtaining the multi-target prediction models are: predictive clustering trees (PCT), binary relevance (BR) and hierarchy of multilabel classifiers (HOMER). These methods have been selected following the recommendation done by Madjarov et al. (2012).

These techniques are implemented in two java frameworks: *MEKA* (Read et al., 2016) for multidimensional classification, and *CLUS* (Struyf et al., 1999) for predictive clustering. The MEKA framework implements the BR and HOMER methods while PCT is implemented in CLUS.

Multi-target prediction can be benefited with the use of the AMOR encoding in order to better represent the information coded by doctors and to improve the prediction of the therapeutic responses. Figure 3.6 presents the proposed approach. It has been exposed in the work presented by Parrales et al. (2019c) as the "panoramic prediction". However, AMOR encoding instead of SAR encoding will be applied in order to find the best label representation of the collected clinical data, since it addresses the generation of multi-target prediction models.

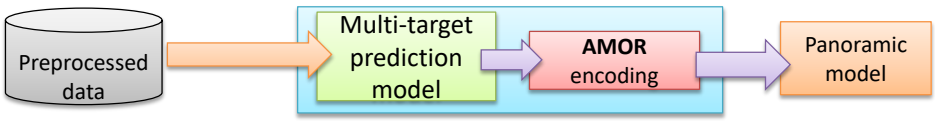


Figure 3.6: Panoramic prediction presented by Parrales et al. (2019c) and adapted to the multi-target scenario with AMOR encoding instead of SAR encoding.

3.3.2 Feedback prediction

This methodology seeks to improve the prediction of the therapeutic response by creating a predictive model for each treatment stage. This approach is called "feedback prediction" because it needs responses of previous treatment sessions for the generation of a later stage model. Therefore, in order to generate the prediction model for the i^{th} stage (M_i), the initial dataset O is required, as well as the feedbacks F_j, $1 \leq j < i$, corresponding to the $(i-1)$ previous stages of the treatment. The first stage of the treatment will not have any feedback. Figure 3.7 presents the proposed approach.

Feedback prediction implies the use of a one-target prediction approach instead of a multi-target prediction approach. SAR encoding is considered for improving the numeric labels in one-target prediction models. To carry out the learning of these models, several state-of-the-art classifiers (Wu et al., 2008) (e.g. TAN, RIPPER, C4.5 or NB tree algorithms) have been considered for comparing their prediction accuracy and to achieve a general idea of possible ways to improve the results. All these algorithms are described in Table 3.5. Applying this approach to the collected clinical data, a prediction model for each stage of the treatment will be obtained separately.

Table 3.5: Descriptions of one-target classifier algorithms selected for feedback prediction.

Method	Description
Naive Bayes	Numeric estimator precision values are chosen based on analysis of the training data.
IBk	k-nearest neighbours classifier.
RIPPER	Propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction.
C4.5	Generates a pruned or unpruned C4.5 classification tree.
Logistic	Builds and uses a multinomial logistic regression model with a ridge estimator.
AdaBoostM1	Meta classifier: Boosts a nominal class classifier.
Bagging	Meta classifier: Bagging a classifier to reduce variance.
LMT	Builds classification trees with logistic regression functions at the leaves.
NBTree	Generates a classification tree using Naive Bayes classifiers for the leaves.
Random forest (RF)	Builds a forest of Random trees (RTs).
Random tree (RT)	Builds a tree considering K randomly chosen attributes for each node. Performs no pruning.
REPTree	Builds a regression(classification) tree using information gain and variance and prunes it using reduced-error pruning.
DecisionStump	Builds a tree that make predictions based on the value of just a single input attribute (also called 1-rules).
SVM	Builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier.

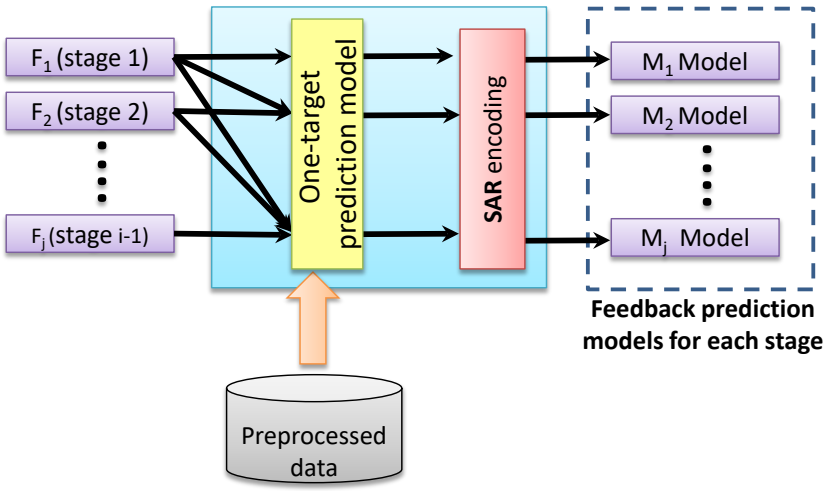


Figure 3.7: Feedback prediction presented by Parrales et al. (2019c).

3.4 Dealing with missing values

One of the most common problems involved in the collection of medical data is the presence of missing values (Lin & Haug, 2008), since doctors does not always have all the information of each patient or some medical characteristics are only necessary and relevant to collect for certain types of diseases.

In order to mitigate these problems, Lambin et al. (2013) propose replacing missing values by calculated estimates. On other hand, the absence of data can have value on its own information (Lin & Haug, 2008). For instance, Pagán et al. (2015) create a set of models to attack the lack of information due to the malfunction of medical sensors. This "lack of clinical information" needs to be considered in this proposed methodology because it can provide useful information to build a set of prediction models in order to adapt the prediction to the missing values appearing in the EMRs.

To address the aforementioned inconvenient of missing values, a *hierarchy of models* will be considered to group medical records that contain similar missing values. This approach also addresses data imputation to fill in the missing values based on records of each group.

Figure 3.8 presents the diagram of the Missing Value-Dependent Model Selection System (MVDMS^2). The purpose of this method is to generate a hierarchy of predictive models taking into account the missing values in the medical records. To do this, the preproccesed dataset composed of m records is split into two groups of records. One for training/validation and

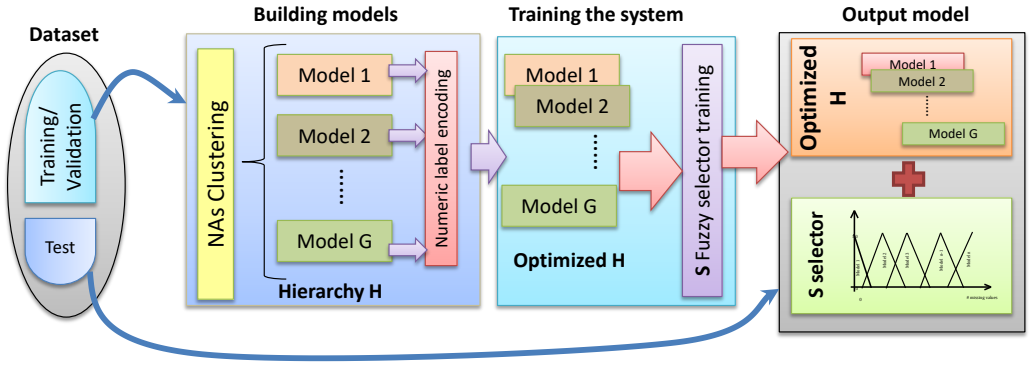


Figure 3.8: Missing value-dependent model selection system (MVDMS^2).

the other for testing purposes. Then, the training dataset is clustered first by their NA values. Afterwards, for each of the resulting groups, the initial predictive models are obtained using the classifier algorithm specified as input parameter. These models are optimized through the application of a numerical label encoding using SAR or AMOR encodings for one-target and multi-target prediction scenario, respectively. To take into account those patient records that do not meet the membership rule of the groups, a fuzzy selector is trained too. This selector establishes the membership rules of each record according to its number of NAs. The final product of this system is the hierarchy of models whose membership rules are governed by the fuzzy selector. In order to get the accuracy of the system, the test dataset is used to apply the fuzzy selector and the hierarchy models. Every process presented in Figure 3.8 will be described in detail though the following subsections.

3.4.1 Clustering of missing values

The preprocessed data are the input to the algorithm. Their missing values are represented by the label NA in the EMRs. Figure 3.9 shows an example of the medical dataset provided with many EMRs. O is the training/validation dataset composed of r records and n medical attributes (columns). The NA values of every cell $o_{i,j} \in O$, $1 \leq i \leq r$ and $1 \leq j \leq n$, are accumulated in the "Total NA" row for each column of the dataset. Afterwards, the n columns are sorted in descending order according to the number of NAs in Table A. Then, only the first T attributes will be considered to create Table B. The value of this threshold T should be defined by doctors depending on the number of attributes they wish to take into account. The cells $b_{i,j}$ of Table B, $1 \leq i \leq r$ and $1 \leq j \leq T$, are filled with 0's and 1's according to

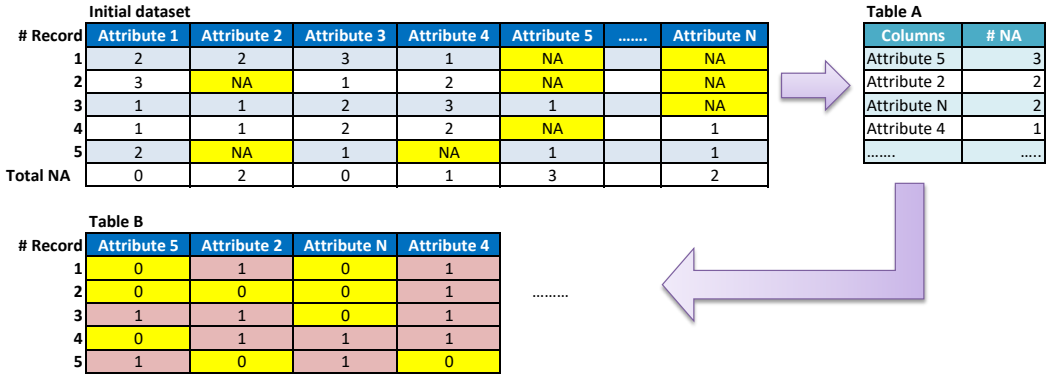


Figure 3.9: Data structure to analyze missing values found in medical records.

Equation 3.7. A pseudocode with these steps is presented in Algorithm 2.

$$b_{i,j} = \begin{cases} 1, & \text{if } o_{i,j} \neq \text{NA} \\ 0, & otherwise \end{cases} . \tag{3.7}$$

Algorithm 2: Selecting columns for Table B.

Require: Training dataset O composed of r records, columns c_j and cells $o_{i,j}, 1 \leq i \leq r$ and $1 \leq j \leq n$. Threshold T.

1: **for** $j = 0$, $j < n$, $j{+}{+}$ **do**
2: addToTableA(getColumnName(c_j),countColumnNAs(c_j))
3: **end for**
4: sortTableA(DESC)
5: **for** $i = 0$, $i < T$, $i{+}{+}$ **do**
6: addToTableB(c_i)
7: **end for**
8: fillValuesTableB(O)

Table B provides the information about the NA values of every record. The registers contained in Table B are then grouped using the k-medians clustering (Jain & Dubes, 1988) algorithm. The k-nearest neighbor algorithm (k-NN) has not been taken into account for this task, since k-medians is less sensitive to outliers. It must be noted that given a number of record groups G, the k parameter will be set to this value. Afterwards, it is important to generate rules for defining the membership of the medical records belonging to each group. For example, when considering a group named "model1" with

all the rows presented in Table B of Figure 3.9, their median of NAs will be equal to 2, as presented in Figure 3.10.

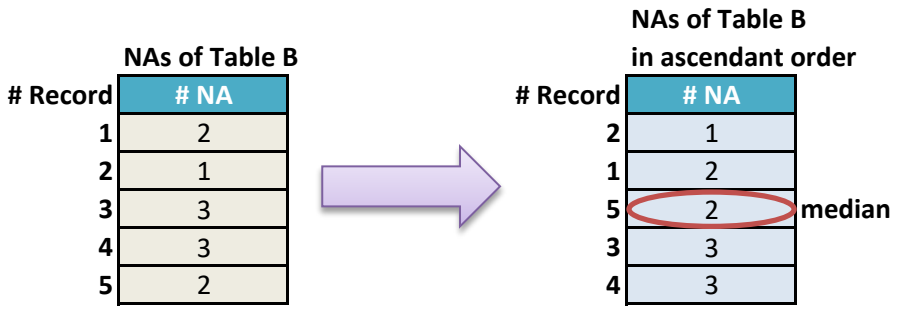


Figure 3.10: Number of NAs of every record of Table B (Figure 3.9). It is important to mention that not several groups are generated but a single group with all the records presented in Table B of Figure 3.9.

3.4.2 Initial set of models and numerical encoding

Once the groups of medical records have been found, each group is considered as a different dataset. Each of them is trained with a classification algorithm. In order to improve the numerical label encoding, each model is optimized by using the SAR or AMOR encoding proposed in Section 3.2.5. The output of this training phase will be a hierarchy of optimized models. For example, when considering the example of the "model1" group, the classification models will be trained with records located in rows 1 and 5 because they have only 2 NAs, as the median of the "model1" group.

3.4.3 Fuzzy model selector

Section 3.4.1 explains how to get different groups of medical records. Moreover, Section 3.4.2 describes how to get the membership rules for adding new records to each group. However, some medical records do not belong to any given group because they do not meet the group membership rules. To solve this issue, it is necessary to consider the use of fuzzy logic to soften the membership rules, selecting the most suitable model for each medical record.

For this purpose, a mapping table (T_{map}) is defined, as the one presented in Figure 3.11. It is composed of r records that present their number of NA values and also the model (cluster) assigned by the algorithm described in this

section. Then, this table is used for training-testing the fuzzy classifier and obtaining the fuzzy model selector. Afterwards, these rules will be applied to the $m - r$ remaining records (testing set), taking into account their number of NAs.

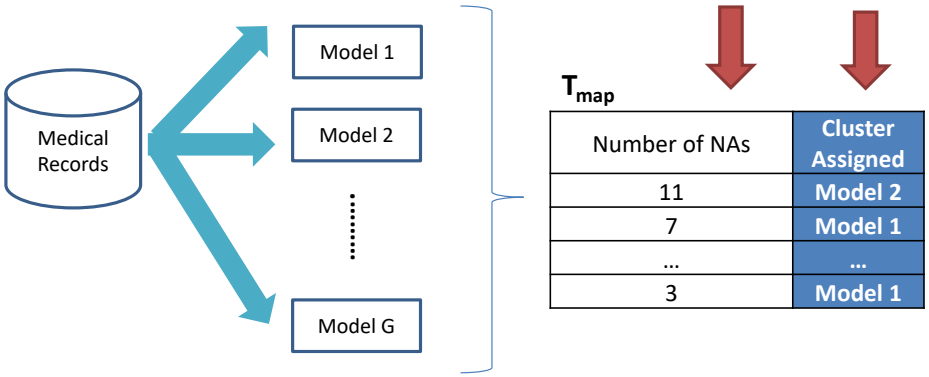


Figure 3.11: Mapping table (T_{map}) for training the fuzzy selection of models

The Fuzzy Unordered Rule Induction Algorithm (FURIA) (Hühn & Hüllermeier, 2009) has been selected as the algorithm to train our model selector because it combines the generation of classification rules with fuzzy logic, obtaining simple and compact sets of fuzzy classification rules.

This algorithm is implemented in the WEKA framework (Hall et al., 2009). The work presented by Hühn & Hüllermeier (2009) mentions that the FURIA fuzzy rules are obtained through replacing rule intervals by fuzzy intervals, namely fuzzy sets with trapezoidal membership function as shown in Figure 3.12. The fuzzy rule interval is specified by four parameters, as presented in Equation 3.8:

$$I^F = (\phi^{s,L}, \phi^{c,L}, \phi^{c,U}, \phi^{s,U}) \,, \tag{3.8}$$

whose parameters are defined in Equation 3.9:

$$I^F(v) = \begin{cases} 1, & \phi^{c,L} \leq v \leq \phi^{c,U} \\ \frac{v-\phi^{s,L}}{\phi^{c,L}-\phi^{s,L}}, & \phi^{s,L} \leq v \leq \phi^{c,L} \\ \frac{\phi^{s,U}-v}{\phi^{s,U}-\phi^{c,U}}, & \phi^{c,U} \leq v \leq \phi^{s,U} \\ 0, & otherwise \end{cases} \,, \tag{3.9}$$

where $\phi^{c,L}$ and $\phi^{c,U}$ are the lower and upper bound of the core respectively. On the other hand, $\phi^{s,L}$ and $\phi^{s,U}$ are the lower and upper bound of the support respectively. In the non-fuzzy case, a fuzzy interval can be open to one side, i.e. $\phi^{s,L} = \phi^{c,L} = -\infty$ or $\phi^{c,U} = \phi^{s,U} = \infty$.

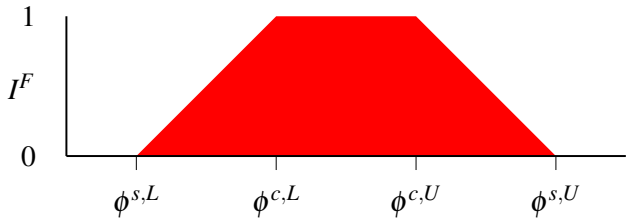


Figure 3.12: A fuzzy interval I^F (Image taken from Hühn & Hüllermeier (2009)).

Following the example of the "model1" group (Figures 3.9 and 3.10), an example of a FURIA rule that could define the membership of one medical record to that group would be the following:

IF NumNAs in [1,2,3,7] THEN model1, CF=0.85

The explanation would be: If the number of NAs (NumNAs) falls in the region defined by the trapezoidal membership function with [1, 2, 3, 7],then the selected model will be "model1" with a certainty factor (CF) of 0.85. Applying this rule to the example in Figure 3.11, those records with 7 and 3 NAs will belong to the "model1" group.

3.4.4 Data imputation

Once the clinical records are grouped by their missing values, these values are replaced taking into account the other records of their cluster through the imputation of data.

In this thesis, the multiple imputation method has been selected, following the recommendation presented by Van Buuren (2018) and commented in Section 2.2.3. In this sense, the stochastic imputation by regression has been selected to create different imputed datasets. Figure 3.13 presents an example of multiple imputations that creates 3 imputed datasets. The final imputed dataset is filled with the most frequent imputed values. In the collected clinical dataset, five multiple imputations will be carried out, following the Rubin recomendations commented in Section 2.2.3.1. The library MICE (van Buuren et al., 2015) will be used on the statistical software R.

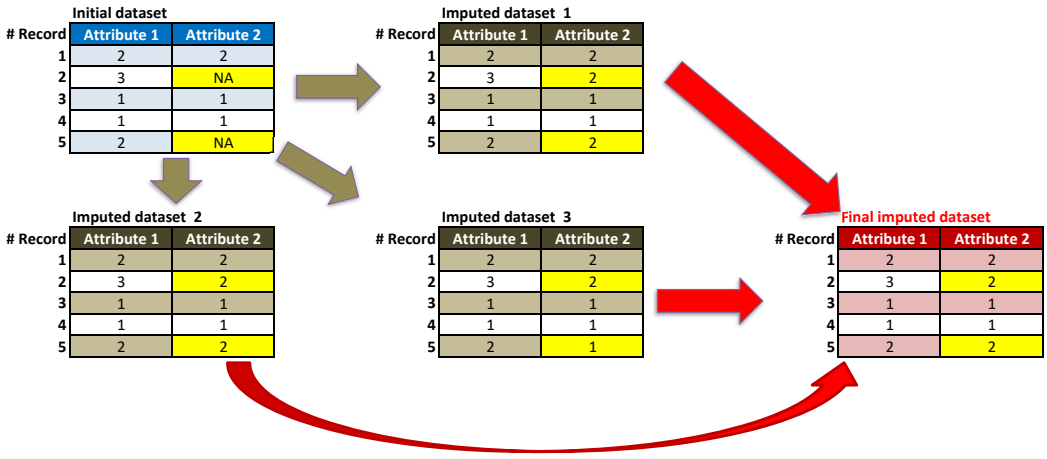


Figure 3.13: Example of a multiple imputation with 3 imputed datasets. The final imputed dataset is filled with the most frequent imputed values.

3.4.5 Integration of hierarchical models with panoramic and feedback prediction

In order to address the problem of missing data in the panoramic prediction and feedback approaches, the step of AMOR and SAR encoding from Figures 3.6 and 3.7 must not be carried out before. This is due to the numeric label will be performed by the MVDMS^2 instead, which includes SAR/AMOR encoding among its phases.

Figure 3.14 presents the inputs and outputs for integrating panoramic and feedback prediction approaches with MVDMS^2. The input and output of the MDVMS^2 will change because it depends on the prediction approach selected. The method for learning the predictive model will also depend on the chosen approach. Multi-target or one-target classification algorithms can be used for panoramic or feedback approaches, respectively.

3.5 Obtaining relevant medical factors

As was mentioned in Section 1.3, the pathophysiological attributes that determine the positive or negative response to the migraine treatment are not known yet (Ornello et al., 2015). In order to extract the attributes described by the predictive models of the response to treatment with BoNT-A, two approaches have been taken into account, namely: Feature Subset Selection and Consensus Models. They will be explained in this section.

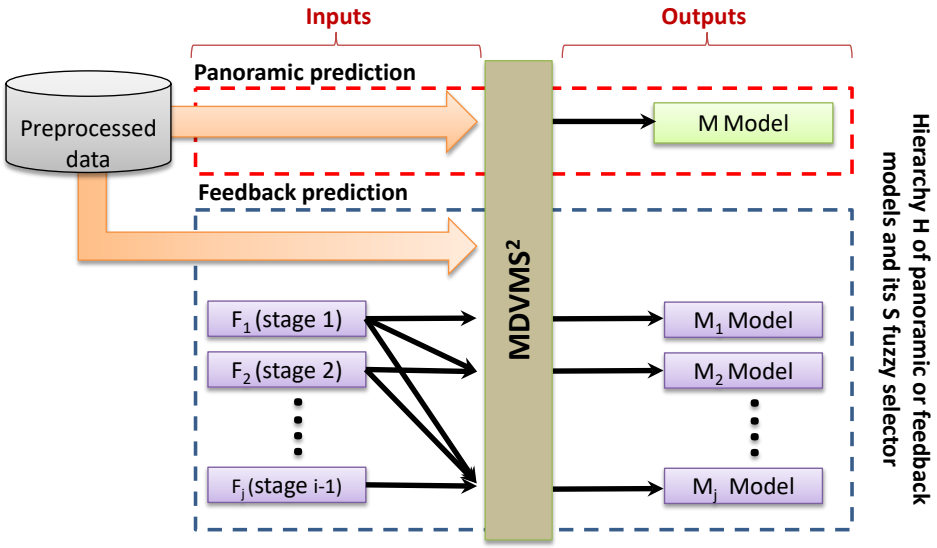


Figure 3.14: Integration between MVDMS^2 of Figure 3.8 and panoramic or feedback prediction.

3.5.1 Feature subset selection

This technique makes it possible to enhance the prediction efficiency of the classification methods, as it allows to consider the most influential attributes (features) when predicting the class attribute value. This approach has certain advantages, such as offering a better understanding of the prediction model and a better generalization by reducing overfitting (Witten et al., 2016). Several approaches have been designed to implement the FSS technique as the filter, wrapper or embedded method (Saeys et al., 2007). The filter type method selects attributes without considering the model. In this approach, the emphasis is placed on the general attributes such as the existent correlation with the class to predict. The wrapper method tries to find interactions between attributes by evaluating subsets of them. Finally, the embedded method considers certain search algorithms in order to combine the advantages of the first two methods.

As mentioned in Section 2.2.4.1, CFS has been the selected method in order to determine the most relevant clinical attributes when obtaining the treatment response prediction. C4.5 is the classifier selected to work together with the CFS method to measure the worthiness of the subset of attributes within the dataset. Moreover, CFS will select the subset of attributes with the highest score of the correlation of each attribute in the subset with each of the s class attributes (Section 2.2.4.1).

3.5.2 Consensus model

Ensemble techniques can help analyze attribute relations with the construction of consensus models to make new and relevant findings (Villoslada et al., 2009; Larrañaga et al., 2006). In this sense, Armañanzas et al. (2012) have proposed an ensemble interaction network for unveiling biological relations when analyzing Alzheimer's disease. In that study, many Bayesian k-dependence models are induced to output a gene interaction network composed of arcs (edges). An occurrence threshold t is defined to output the most frequent edges above a predefined confidence level (the 0.999 quantile is used in order to retain just the most important connections). The list of interaction networks and the associated list of highly relevant attributes are obtained to reveal or corroborate biological hypotheses in this disease. Other studies (Otaegui et al., 2009; Small et al., 2005) can be found in the literature with similar purposes.

Consensus models can be incorporated to this Ph.D. Thesis in order to reveal relevant attributes in the collected medical dataset. The idea is not to build a consensus predictor model, but to understand the most relevant clinical attributes that exist in the majority of the induced prediction models of the best classifier. Thus, this technique is applied in order to group different prediction models (classification trees) produced by the best classifier in terms of accuracy for all stages. This is done with the purpose of finding explicit attributes and relations between medical attributes that allow to predict the treatment response. In the FSS method, these attributes are selected before the construction of the prediction model by using different metrics. In the consensus model approach, the idea is to invert the attribute selection process of FSS, which means that the relevant attributes will be selected after, and not before, the construction of the prediction models.

A classification tree model is defined as a graph $G(V, E)$, where V represents the vertex list (attributes as vertices) of the model and E represents the list of edges (relations between vertices) of the model. The interactions in the classification tree consist of parent-child edge relations. Nodes are filled with the attribute values and edges represent the parent-child relation from the classification tree model.

Many classification trees will be induced by a resampling method (k-fold cross validation) together with the SAR or AMOR encoding. For each level of the classification tree, the most frequent clinical attributes will be taken into account. After this, a consensus model will be depicted with edges whose frequencies are higher than a reliability threshold t. Edges occurring more than t times for each level of the tree will be retained.

The methodology for building consensus models is presented in the Algo-

rithm 3. Moreover, Table 3.6 presents the functions and definition of variables used in the algorithm. Edges for the first level of the induced models ($E_{0\to1}$) will have a NULL value as vertex u in the edge tuple (u, v) because the roots of classification trees do not have parents. Edges will be sorted in descendant order, according to their frequency of appearance in a given level. In order to retain only one vertex as root of the consensus classification tree, only the head of the $E_{0\to1}$ list of the induced models will be retrieved. This step is carried out in the lines 3-6 of the Algorithm 3. For the rest of the levels, the lines 7-15 of the algorithm are applied. In these lines, the 0.9 quantile will determine the t value for retaining the most important edges. These quantile values have been defined by considering the 0.999 quantile applied by Armañanzas et al. (2012), but modified with the purpose of retaining multiple important child nodes in the consensus classification tree proposed. Moreover, only edges whose origin vertex is equal to any destination vertex of their previous level will be selected.

An example of the consensus model construction is presented in Figure 3.15. In it, the number of levels (L_{max}) has been defined as 3. For level 1, only the edge located in the head of the $E_{0\to1}$ list has been selected as root of the consensus model. For selecting the edges in level 2, they need to have a frequency greater than or equal than the t value of the level ($t = 850$). In addition, only edges whose origin vertex is equal to any destination vertex of their previous level has been selected. The same criteria is applied for level 3 with a t value of 642.

Table 3.6: Description of variables and functions employed in Algorithm 3.

Name	Description
v	Vertex.
$e(u, v)$	Edge $u \to v$, where u is parent of v.
$E_{(i-1)\to i}$	The edges list from level $i-1$ to i of the induced prediction models for a given stage.
$w(e, E_{(i-1)\to i})$	Weight of an edge e. $w(e, E_{(i-1)\to i}) = \lvert\{e \in E_{(i-1)\to i}\}\rvert$.
M	List of nodes that conform the consensus tree.
L_{max}	A defined maximum number of levels to explore for the consensus tree construction.
tvalue(q,$E_{(i-1)\to i}$)	Calculates the t value given the quantile (q) value and the $E_{(i-1)\to i}$ list.
head($E_{(i-1)\to i}$)	Returns and removes the first element of the $E_{(i-1)\to i}$ list.
add(e,$E_{(i-1)\to i}$)	Adds e to the $E_{(i-1)\to i}$ list.

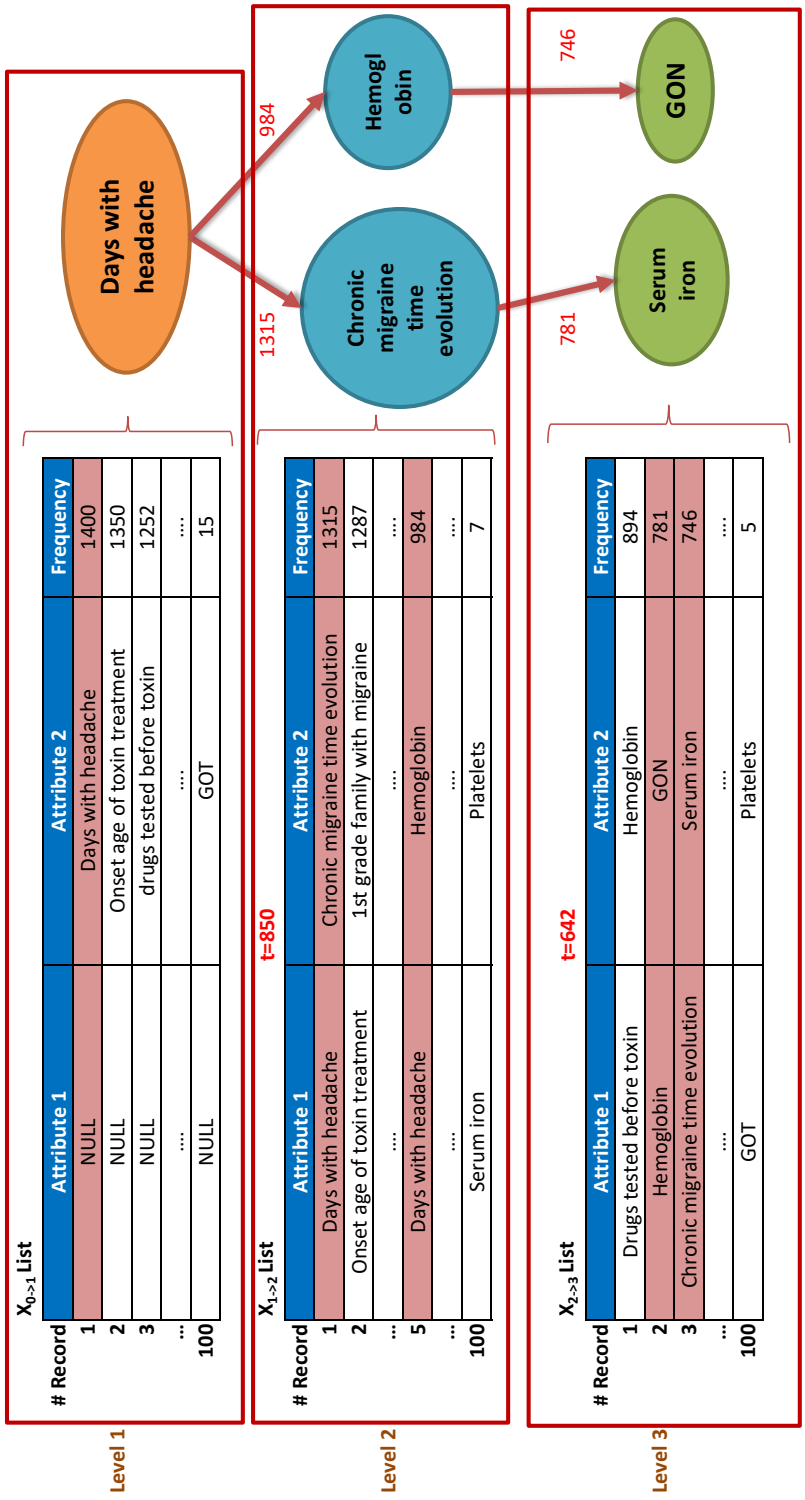


Figure 3.15: Example of a consensus model construction.

Algorithm 3: Relevant attributes in consensus trees.

Require: Lists $E_{0\to1}, \ldots, E_{(L_{max}-1)\to L_{max}}$ in descendant order and L_{max}.

$M = \emptyset$
t=tvalue(0.99, $X_{0\to1}$)
if $X_{0\to1} \neq \emptyset$ **then**
  $e(u, v)$=head($X_{0\to1}$)
  add(e, M)
end if
for $i = 2$, $i < L_{max}$, i++ **do**
  t=tvalue(0.9, $X_{(i-1)\to i}$)
  while $X_{(i-1)\to i} \neq \emptyset$ **do**
    $e(u, v)$=head($X_{(i-1)\to i}$)
    if $w(e, X_{(i-1)\to i}) \geq t$ and $\exists e' = (u', v') \in M : u = v'$ **then**
      add(e, M)
    end if
  end while
end for
return M

Chapter 4

Experiments

It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong.

Richard P. Feynman

Contents

4.1 Parameters

This section will present use of various environments, parameters and methods described in the previous chapters on the collected medical dataset for the classification of responses to migraine treatment with BoNT-A.

4.1.1 *k-fold cross-validation*

To honestly measure the performance of classification algorithms, k-fold cross-validation has been applied with k=10 in order to use the 90% of data for training and the 10% for testing in every loop of cross-validation. This method has been used to avoid reporting overly optimistic results of the classification algorithms that estimate their performance with replacement (using the training sample) (Section 2.3.3).

4.1.2 Sensitivity and specificity

The sensitivity and specificity values of prediction models are considered because these measurements are often used and are frequently more important than the accuracy of classification in some medical applications (Lavrač, 1999), as discussed in Section 2.3.1. Sensitivity measures the fraction of positive cases that are classified as positive, while specificity measures the fraction of negative cases classified as negative. In the collected medical dataset, the positive values will be patients who have a good therapeutic response (labeled "high") to treatment, while negative cases will be those who get a poor response (labeled "low"). High sensitivity values could be preferred when the goal is to improve the selection of patients on whom to apply BoNT-A treatment. In this way, doctors can assure that the economic investment to be made for the treatment will be beneficial for the patient. High specificity values will be preferred when it is desired to avoid unnecessary costs due to the ineffectiveness of migraine treatment with BoNT-A. For a clinic, better sensitivity values will be preferred, since they would ensure that the income to be obtained from migraine treatment with BoNT-A will correspond to customer satisfaction. On the other hand, patients will prefer a prediction model with a high specificity value because it could better ensure if the investment to be made will be worthy.

4.2 Obtaining classification models

4.2.1 One-target classification algorithms

It is convenient to start evaluating this type of classification algorithms because some multi-target classification methods need to define any of these one-target methods as parameter. Thus, the prediction of a single response to treatment will be compared between different one-target classification methods. In this sense, different models will be obtained for each of the treatment stages. To select the best method, accuracy, sensitivity and specificity will be compared for first, second and third stages of the treatment. For this experiment, the feedback prediction approach will not be considered. This implies that prediction models of a determined stage only will consider the response outputs of that stage and not from other stages in the training step. For example, treatment results of stages 1 and 2 will be ignored when obtaining a prediction model to the third stage of treatment.

One-target classifiers algorithms of Table 3.5 have been considered for this experiment. The parameters selected for these algorithms are described in Table 4.1. They have been defined following the work of Parrales et al. (2019d). Performance results will be obtained when using an imputed dataset. To impute the data from the collected clinical dataset, the guidelines addressed in Section 3.4.4 will be followed. After that, the results with the use of FSS and with the use of SAR encoding will be compared.

Table 4.2 presents accuracy, sensitivity and specificity values for all selected classifier algorithms over each stage of treatment. IBk, RIPPER and SVM methods have obtained the best accuracies for stages 1, 2 and 3, respectively. However, their mean accuracy values do not exceed 70%. These results mean that the response to the treatment will be correctly predicted in 6 out of 10 patients. These results are not far from baseline accuracy, i.e. close to 50% of accuracy, which means that the prediction of the response to the treatment will be correct in 5 out of 10 patients. Moreover, despite the fact that these methods have obtained high sensitivity or specificity values, their results are not good when both metrics are taken into account together. All in all, the results obtained could be the consequence of having irrelevant and redundant attributes among the 62 columns of the dataset. Irrelevant attributes do not affect the description of the class attribute. Redundant attributes provide nothing but noise towards the description of the class attribute. Removing irrelevant and redundant attributes could reduce the risk of overfitting while improving the predictability of classification models (Shilaskar & Ghatol, 2013; Chormunge & Jena, 2018). Therefore, it is convenient

Table 4.1: Parameters of one-target classifier algorithms selected for one-target prediction.

Method	Parameters
Naive Bayes	No parameters
IBk	k-NN=2, Linear Search algorithm
RIPPER	Pruning=true, Seed=1
C4.5	Confidence factor=0.25, Seed=1
Logistic	maxIts=-1, Ridge=$1 \cdot 10^{-8}$
AdaBoostM1	Classifier=Decision Stump, Iterations=10, Seed=1
Bagging	bagSizePercent=100, Classifier=Random tree or C4.5, Iterations=10, Seed=1
LMT	minNumInstances=15, numBoostingIterations=-1
NBTree	No parameters
Random forest (RF)	Number of trees=100, Seed=1
Random tree (RT)	minNum=1, Seed=1
REPTree	maxDepth=-1, minNum=2
DecisionStump	No parameters
SVM	cacheSize=40, cost=1, kernelType=radial

to explore the use of FSS on the collected clinical dataset in the next section.

In order to remove the correlated data, the use of FSS on the collected clinical dataset is considered. More specifically, it is important to consider the use of the CFS method for each stage to predict. Following it, the predictive attributes for each class attribute have been obtained independently. The selected attributes have been the following:

- First stage: Onset age of toxin treatment, Chronic migraine, Chronic migraine time evolution, Drugs tested before toxin, Tricyclic antidepressants, Vitamin B12.
- Second stage: GON, Preventive oral treatment at time of infiltration, Tricyclic antidepressants, Gastropathy, Pneumopathy, Dermopathy.
- Third stage: Calcium antagonists, Catamenial, Concomitant oral preventive treatment, Gastropathy, Headache days per month, Analgesic abuse.

In summary, the attributes selected for the first stage have been only used for obtaining classifier models of the first stage. In the same way, the attributes selected for the second and third stages have been used for building classifier models for the second and third stage, respectively. Their results

Table 4.2: Estimated performance metrics (mean ± standard deviation in percentage) of some classic classification methods without FSS or SAR encoding (baseline results). The best results are highlighted in bold.

Classification algorithm	First stage			Second stage			Third stage		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Naive Bayes	51.94±3.46%	60.01±1.27%	37.04±2.16%	55.84±4.07%	47.37±3.85%	64.10±3.71%	57.14±3.53%	29.17±2.67%	69.81±3.82%
IBk	**68.83±2.27%**	82.09±1.34%	44.44±2.53%	57.14±5.16%	52.63±3.24%	61.54±4.35%	57.14±4.05%	33.33±1.98%	67.92±4.73%
RIPPER	59.74±2.43%	92.10±0.52%	3.70±3.35%	**66.23±4.52%**	**63.16±4.17%**	69.23±3.28%	58.44±3.25%	16.67±2.13%	77.36±3.58%
C4.5	55.84±3.65%	68.19±3.57%	33.33±4.62%	50.65±5.41%	60.53±3.91%	41.03±4.28%	51.95±2.65%	29.17±2.24%	62.26±2.78%
Logistic	57.14±2.84%	64.73±2.65%	44.44±3.85%	50.65±4.41%	47.37±3.49%	53.85±4.62%	66.23±2.04%	50.03±0.52%	73.58±2.81%
AdaBoostM1 (DecisionStump)	54.55±2.64%	80.12±2.83%	7.41±3.24%	50.65±4.36%	52.63±3.86%	48.72±4.52%	61.04±0.51%	20.83±0.23%	79.25±1.07%
Bagging (RT)	53.25±2.76%	68.01±3.17%	25.93±2.82%	53.25±3.17%	44.74±2.63%	61.54±3.92%	62.34±2.05%	12.51±1.93%	84.91±2.14%
Bagging (C4.5)	55.84±2.57%	76.08±1.94%	18.52±2.83%	54.55±3.22%	52.63±2.40%	56.41±3.55%	54.55±0.47%	4.17±0.11%	77.36±0.78%
LMT	66.23±2.72%	90.16±2.91%	22.22±2.65%	50.04±2.45%	50.02±1.23%	66.67±2.71%	64.94±3.46%	25.11±1.71%	83.02±3.95%
NBTree	62.34±3.23%	76.17±3.48%	37.04±2.20%	54.55±2.22%	60.53±2.11%	48.72±2.35%	54.55±1.76%	20.83±1.81%	69.81±1.62%
RF	58.44±1.93%	88.25±3.65%	3.70±0.54%	50.65±2.61%	44.74±2.57%	56.41±2.82%	66.23±2.49%	8.33±1.82%	**92.45±3.12%**
RT	54.55±2.68%	58.03±1.27%	**48.15±0.93%**	58.44±3.86%	57.89±2.44%	58.97±2.91%	46.75±1.70%	29.17±1.35%	54.72±3.27%
REPTree	57.14±3.81%	84.24±2.57%	7.41±0.75%	54.55±2.15%	47.37±3.28%	61.54±3.07%	57.14±2.28%	4.12±0.08%	83.02±1.43%
DecisionStump	63.64±2.51%	**96.14±1.26%**	3.70±0.35%	62.34±2.76%	50.04±0.11%	**74.36±2.36%**	54.55±2.22%	**60.53±2.11%**	48.72±2.35%
SVM	64.93±3.48%	95.13±1.26%	3.70±0.35%	58.44±3.26%	55.26±1.74%	61.54±4.26%	**68.83±2.18%**	48.12±3.45%	69.41±1.16%

are presented in Table 4.3. It presents an improvement in the percentage of accuracy, exceeding 70% in most methods for each stage, which implies a correct prediction of the response to treatment for 7 out of 10 patients for each stage. However, the classification of high and low responses to the treatment is not well proportioned according to their sensitivity and specificity values. In fact, some methods obtain high percentages of specificity by sacrificing sensitivity. For example, the Bagging (C4.5) method presents an accuracy close to 70 % in the first stage when predicting a "high" response to the treatment for almost all patients (sensitivity close to 94%). Hence, it is necessary to obtain models that improve the prediction of high and low responses to the treatment. All in all, these results may indicate that a deeper review of correlated data is needed in order to get prediction models with high sensitivity and specificity values. In this sense, the SAR encoding method will be applied to the original dataset (without FSS) in the next section.

The previous selection of attributes (FSS) is not taken into account since SAR encoding performs an attribute weighting while optimizing numeric labels.

In this experiment, a number of 10^6 iterations for SAR encoding (K parameter) has been defined considering this as a sufficient number of iterations for the algorithm to converge to a good solution (Parrales et al., 2019d; Szűcs & Balázs, 2019). The D parameter has been defined in 2 for considering two orders of decimal magnitude. "SAR (d=1)" and "SAR (d=2)" will be the notation for prediction models improved with SAR encoding when using one and two decimals, respectively. Their results are presented in Table 4.4 and 4.5.

On the basis of the results, it can be observed that non-deterministic classifier algorithms (RT and RF) combined with SAR encoding perform the best in Tables 4.4 and 4.5. In fact, their accuracies are close to 85% when applying SAR with d=1. Previous results (Tables 4.2 and 4.3) show that the best classifiers were deterministic. Then, it can be concluded that SAR encoding becomes an important factor, as it helps to optimize non-deterministic algorithms. Looking for the lowest error percentage (100-accuracy percentage), the SA heuristic moves the solution within the search space to avoid being caught in a local minimum, benefiting from it mostly the non-deterministic algorithms.

Looking more closely at the results of sensitivity and specificity values of Tables 4.2, 4.3, 4.4 and 4.5, it can be observed an overall improvement in the classification of "high" and "low" responses to the treatment due to SAR encoding. This implies that SAR has been the best method for finding those correlated medical characteristics, which have not been taken into account

Table 4.3: Estimated performance metrics (mean ± standard deviation in percentage) of some classic classification methods with FSS. The best results are highlighted in bold.

Classification algorithm	First stage			Second stage			Third stage		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Naive Bayes	70.13±2.16%	92.12±1.56%	29.63±2.35%	64.94±3.78%	55.26±3.71%	74.36±2.43%	**76.62±2.36%**	54.17±2.52%	86.79±3.21%
IBk	49.35±3.58%	62.14±2.78%	25.93±3.74%	71.43±3.84%	65.79±3.57%	**76.92±3.25%**	76.62±2.50%	58.33±1.72%	84.91±4.15%
RIPPER	71.43±3.62%	96.03±0.12%	25.93±3.75%	63.64±3.91%	63.16±2.75%	64.10±3.88%	66.23±4.58%	29.17±3.71%	83.02±3.27%
C4.5	71.43±4.55%	96.12±1.02%	25.93±3.49%	63.64±5.21%	60.53±3.62%	66.67±4.76%	74.03±2.48%	**62.51±3.38%**	79.25±4.21%
Logistic	70.13±4.16%	92.14±2.49%	29.63±3.17%	72.73±2.56%	73.68±3.48%	71.79±4.15%	74.03±3.45%	54.17±2.56%	83.02±3.33%
AdaBoostM1 (DecisionStump)	70.13±3.78%	96.04±1.38%	22.22±0.14%	68.83±3.64%	71.05±2.85%	66.67±4.03%	68.83±2.94%	37.51±2.57%	83.02±2.93%
Bagging (RT)	57.14±2.34%	70.16±3.71%	**33.33±2.12%**	71.43±4.25%	71.05±2.71%	71.79±3.87%	74.03±2.11%	54.17±3.65%	83.02±4.21%
Bagging (C4.5)	**71.43±3.42%**	94.15±1.14%	29.63±2.32%	67.53±3.11%	68.42±2.40%	66.67±3.11%	71.43±2.21%	37.5±2.54%	86.79±2.84%
LMT	70.13±3.28%	92.44±3.64%	29.63±2.78%	72.73±2.17%	73.68±2.52%	71.79±1.05%	75.32±1.45%	54.17±2.54%	84.91±1.96%
NBTree	64.94±4.16%	96.03±0.12%	5.93±1.13%	67.53±3.26%	63.16±3.55%	71.79±2.84%	63.64±2.15%	8.33±1.06%	**88.68±1.53%**
RF	55.84±3.87%	72.45±3.58%	25.93±2.35%	70.13±2.77%	68.42±2.92%	71.79±2.43%	76.62±3.10%	58.33±1.21%	84.91±3.50%
RT	49.35±2.72%	60.16±1.21%	29.63±0.63%	68.83±3.61%	65.79±2.04%	71.79±2.67%	75.32±1.18%	58.33±3.74%	83.02±3.10%
REPTree	64.94±4.25%	**96.18±1.17%**	7.41±0.23%	59.74±3.67%	60.53±2.46%	58.97±2.98%	70.13±2.65%	33.33±1.82%	86.79±1.28%
DecisionStump	63.64±2.51%	96.14±1.26%	3.70±0.35%	62.34±2.76%	50.04±0.11%	74.36±2.36%	62.34±2.76%	16.64±3.15%	83.02±2.40%
SVM	64.93±3.48%	95.13±1.26%	3.70±0.35%	**74.03±3.54%**	**73.68±1.16%**	74.36±2.08%	70.13±3.73%	33.33±0.79%	86.79±1.85%

Table 4.4: Estimated performance metrics (mean ± standard deviation in percentage) of some classic classification methods with SAR (d=1). The best results are highlighted in bold.

Classification algorithm	First stage			Second stage			Third stage		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Naive Bayes	74.98±1.51%	77.46±2.14%	72.29±3.15%	71.22±1.67%	75.15±1.54%	68.54±2.35%	78.57±2.79%	72.32±1.67%	77.45±2.81%
IBk	79.91±2.14%	75.28±2.16%	**82.31±2.54%**	75.56±0.84%	74.25±0.78%	77.25±1.07%	78.43±3.84%	69.15±3.54%	76.13±3.16%
RIPPER	73.68±2.41%	75.48±1.23%	71.08±2.12%	77.46±0.84%	74.25±1.36%	78.07±0.84%	74.12±3.41%	78.24±1.20%	71.25±2.56%
C4.5	72.95±0.54%	78.25±2.31%	70.28±1.03%	70.64±1.45%	68.25±2.75%	72.05±2.84%	77.84±2.04%	73.29±2.07%	68.12±2.44%
Logistic	77.28±1.14%	75.94±1.24%	67.23±2.80%	76.49±1.25%	79.46±2.65%	70.56±3.18%	78.71±1.65%	65.49±1.27%	72.15±2.19%
AdaBoostM1 (DecisionStump)	69.24±1.41%	72.53±1.29%	64.18±3.89%	76.56±1.87%	74.28±1.67%	68.24±3.07%	75.89±2.23%	71.16±1.49%	83.02±2.93%
Bagging (RT)	79.86±2.13%	77.28±4.15%	81.48±2.13%	77.31±1.67%	74.12±1.84%	71.45±2.43%	80.42±2.04%	82.45±3.46%	78.16±2.25%
Bagging (C4.5)	73.22±2.14%	71.98±1.26%	75.32±1.58%	75.09±1.41%	65.82±2.63%	73.26±2.19%	75.89±2.64%	78.16±3.45%	71.18±3.16%
LMT	74.64±1.58%	76.48±3.59%	71.56±1.48%	74.32±2.26%	73.58±1.85%	77.14±2.16%	78.14±1.07%	75.49±1.36%	78.29±1.84%
NBTree	71.25±3.46%	68.27±2.37%	74.69±1.18%	75.34±1.18%	72.14±2.54%	65.08±2.73%	74.36±1.62%	71.28±1.35%	64.58±3.61%
RF	84.12±1.74%	85.26±1.82%	81.64±2.25%	82.23±1.15%	80.71±2.64%	84.26±2.78%	81.95±2.84%	78.16±3.27%	86.26±2.54%
RT	**84.93±1.25%**	**87.56±1.85%**	81.45±3.65%	**85.74±2.17%**	**83.24±2.54%**	**88.14±2.72%**	**83.29±1.07%**	**84.68±2.27%**	**86.54±2.28%**
REPTree	75.14±2.23%	70.25±1.43%	78.18±3.76%	72.71±1.56%	71.08±3.04%	65.57±2.18%	75.45±1.98%	78.16±2.85%	67.14±2.82%
DecisionStump	70.38±2.16%	74.29±2.14%	65.72±3.16%	71.25±0.56%	68.47±1.23%	73.51±3.82%	71.51±1.84%	65.17±1.37%	68.12±2.39%
SVM	77.45±1.61%	71.38±2.05%	77.61±3.40%	76.29±2.46%	74.89±1.38%	78.19±3.64%	74.19±1.45%	76.58±2.38%	72.94±3.89%

Table 4.5: Estimated performance metrics (mean ± standard deviation in percentage) of some classic classification methods with SAR encoding and d=2. The best results are highlighted in bold.

Classification algorithm	First stage			Second stage			Third stage		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Naive Bayes	72.95±0.54%	62.51±3.38%	**74.36±2.36%**	68.83±3.61%	55.26±3.71%	74.36±2.43%	74.32±2.15%	54.17±2.52%	86.79±3.21%
IBk	66.23±2.49%	62.14±2.78%	25.93±3.74%	72.95±3.91%	65.79±3.57%	**76.92±3.25%**	75.14±3.10%	58.33±1.72%	84.91±4.15%
RIPPER	72.73±2.17%	96.03±0.12%	25.93±3.75%	66.23±2.49%	63.16±2.75%	64.10±3.88%	68.83±3.61%	29.17±3.71%	83.02±3.27%
C4.5	74.19±1.45%	73.68±1.24%	65.15±2.14%	67.41±2.49%	60.53±3.62%	66.67±4.76%	72.73±3.52%	**62.51±3.38%**	79.25±4.21%
Logistic	73.25±3.64%	69.31±2.54%	74.18±2.14%	73.22±1.04%	73.68±3.48%	71.79±4.15%	71.25±2.75%	54.17±2.56%	83.02±3.33%
AdaBoostM1 (DecisionStump)	71.43±2.51%	96.04±1.38%	22.22±0.14%	69.56±4.18%	71.05±2.85%	66.67±4.03%	65.08±3.24%	37.51±2.57%	83.02±2.93%
Bagging (RT)	76.49±3.18%	70.16±3.71%	33.33±2.12%	72.71±3.05%	71.05±2.71%	71.79±3.87%	75.89±3.57%	54.17±3.65%	83.02±4.21%
Bagging (C4.5)	70.38±2.78%	94.15±1.14%	29.63±2.32%	71.51±2.78%	68.42±2.40%	66.67±3.11%	72.73±3.11%	37.5±2.54%	86.79±2.84%
LMT	71.43±2.37%	96.03±0.12%	5.93±1.13%	70.25±3.05%	63.16±3.55%	71.79±2.84%	71.18±3.21%	61.57±2.69%	68.13±2.84%
NBTree	67.14±3.54%	96.03±0.12%	5.93±1.13%	68.05±1.76%	63.16±3.55%	71.79±2.84%	70.05±1.49%	61.57±2.69%	68.13±2.84%
RF	80.27±3.19%	72.45±3.58%	25.93±2.35%	78.37±2.27%	68.42±2.92%	71.79±2.43%	78.51±1.68%	58.33±1.21%	84.91±3.50%
RT	**82.23±3.91%**	60.16±1.21%	29.63±0.63%	**80.35±1.67%**	65.79±2.04%	71.79±2.67%	**81.47±2.96%**	58.33±3.74%	83.02±3.10%
REPTree	69.72±2.65%	**96.18±1.17%**	7.41±0.23%	67.28±3.29%	60.53±2.46%	58.97±2.98%	72.15±2.54%	33.33±1.82%	**86.79±1.28%**
DecisionStump	67.14±2.36%	64.14±2.21%	69.52±2.64%	68.35±3.94%	60.15±2.80%	71.21±2.05%	68.74±2.14%	16.64±3.15%	83.02±2.40%
SVM	73.81±3.15%	95.13±1.26%	3.70±0.35%	75.54±2.51%	**73.68±1.16%**	74.36±2.08%	72.84±2.45%	33.33±0.79%	86.79±1.85%

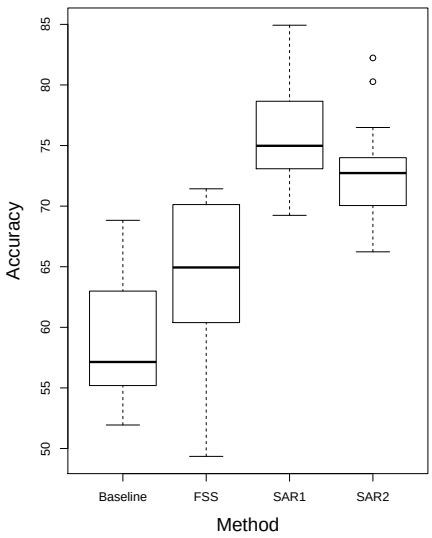
in predictive models.

With the purpose of statistically validating if the improvement in classification mean accuracies due to the FSS and SAR methods (d=1 and d=2) is significant, the Kruskal-Wallis (non-parametric) test was carried out between the accuracy values of Tables 4.2, 4.3, 4.4 and 4.5 for all stages. The Kruskal-Wallis test consist of a non-parametric test that can be used to determine the existence of statistically significant differences between three or more groups of an independent variable on a continuous or ordinal dependent variable. Therefore, because there are 4 techniques to compare, this test will be applied. This test gave us the results of $p = 1.791 \cdot 10^{-8}$, $p = 4.143 \cdot 10^{-9}$ and $p = 2.224 \cdot 10^{-8}$ for the first, second and third stages, respectively. These values, being less than 0.05, guarantee us that there is a significant difference in the distributions of accuracy mean values among groups. The distribution of classification mean accuracy obtained under the baseline, FSS and SAR methods (d=1 and d=2) used in Tables 4.2, 4.3 4.4 and 4.5 for all stages are presented in Figure 4.1. It can be observed a global improvement in accuracy due to the application of SAR (d = 1) with respect to FSS and baseline for all stages. In addition, SAR (d = 2) and FSS improve baseline results. However, there is no great improvement in results with respect to FSS after using SAR with two decimals (d = 2) in the second and third stages.

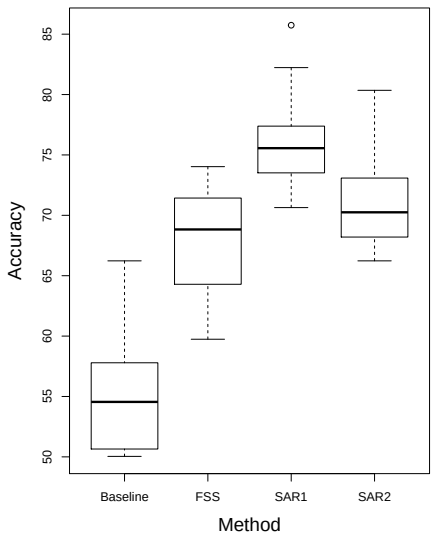
For detecting which pairs of methods are significantly different, the criteria exposed by García & Herrera (2008) has been considered. It consists in the use of Nemenyi's (post-hoc) test in order to know which group pairs differ after a statistical test of multiple comparisons. In this experiment, the Kruskal-Wallis was the selected test for performing multiple comparisons. The adjusted p-values are compared against a significance level of $\alpha = 0.05$ to reject or accept the null hypothesis that a pair of methods perform equally.

Table 4.6 shows the results of the Nemenyi post-hoc test. According to this test, the classifiers improved with SAR (d=1) had a highly significant difference ($p < 0.01$) in comparison to baseline classifiers for all stages. Moreover, SAR (d=1) had a highly significant difference ($p < 0.01$) in comparison to classifiers improved with FSS with the exception of a significant difference ($p < 0.05$) achieved for the third stage. It should also be noted that although SAR (d = 1) improves the results achieved by SAR (d = 2), the test has not achieved significant differences between both methods for all stages. However, SAR (d = 2) has only significantly improved the FSS results for the first stage ($p < 0.05$). Regarding FSS, it can be contemplated that its use significantly improves baseline results for the second and third stages ($p > 0.05$). All in all, it can be concluded with this experiment that the use of SAR (d = 1) is recommended, since it has achieved significant improvements with

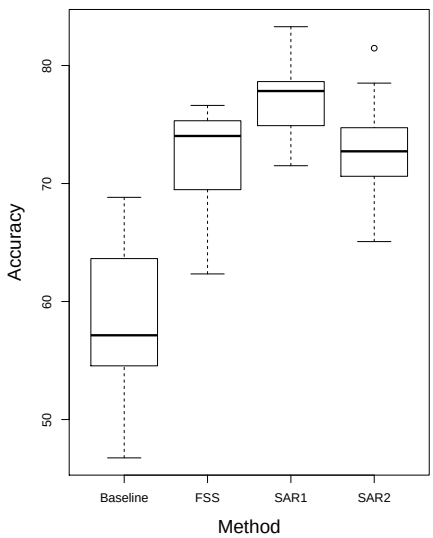
respect to the FSS and baseline results.



(a) First stage.



(b) Second stage.



(c) Third stage.

Figure 4.1: Distribution of classification mean accuracy values obtained under the Baseline, FSS and SAR methods used in Tables 4.2, 4.3, 4.4 and 4.5 for all stages.

Table 4.6: Nemenyi post-hoc test for accuracies of Tables 4.2, 4.3, 4.4 and 4.5.

Pair-methods comparison	First stage		Second stage		Third stage	
	Mean rank difference	p	Mean rank difference	p	Mean rank difference	p
Baseline-FSS	-10.16667	0.3819	-20.10000	0.0088	-22.73333	0.00206
SAR (d=1)-FSS	25.40000	0.0004	20.06666	0.0090	16.43333	0.04896
SAR (d=1)-Baseline	35.56667	$1.5 \cdot 10^{-7}$	40.16667	$1.8 \cdot 10^{-9}$	39.16667	$4.9 \cdot 10^{-9}$
SAR (d=2)-FSS	17.70000	0.0282	7.766667	0.6154	1.233333	0.99744
SAR (d=2)-Baseline	27.86667	$7.3 \cdot 10^{-5}$	27.866667	$7.3 \cdot 10^{-5}$	23.966667	0.00098
SAR (d=2)-SAR (d=1)	-7.70000	0.6221	-12.3000	0.2159	-15.2000	0.0802

4.2.2 Parallel MOEAs

At this point, the RT algorithm has achieved the best accuracy results when SAR (d = 1) is applied for improving accuracy in prediction models. However, a study towards the multiobjective optimization must be considered because a MOEA instead of SA heuristic must be selected in order to apply the panoramic prediction approach. MOEA implies the optimization of all stage accuracies in the predictive models.

Regarding the training of the predictive models for this experiment, it is important to note that the multi-target classification methods will not be considered here. The reason is because this experiment will look for that MOEA that achieves the best performance. Thus, one prediction model for each stage of the treatment will be obtained. Thus, MOEA methods will be in charge of looking for that set of weights for the attributes that optimize the accuracy of the three predictive models simultaneously.

Vectors of attribute weights will be the candidate solutions to be found by the use of MOEAs. These weights will be multiplied by the numerical labels of the collected medical dataset and their result will be rounded as presented in Figure 3.5. The one decimal rounding (d = 1) will be considered given the good results shown in the previous experiment when SAR (d = 1) was applied together with RT.

For determining which of MOEAs has the best performance, two criteria will be considered: (1) the execution time and (2) the accuracy achieved by prediction models with solutions provided by MOEAs. In this sense, the MOEA framework presented by Hadka (2019) will be used. More specifically, those MOEAs that can be parallelized for diminishing the computational cost when optimizing accuracies in prediction models will be examined. Those selected algorithms are: GDE3 (Kukkonen & Lampinen, 2005), PESA2 (Corne et al., 2001), SMPSO (Nebro et al., 2009), NSGA-II (Deb et al., 2002), NSGA-III (Deb & Jain, 2014) and SPEA2 (Zitzler et al., 2001). For completing the comparisons, SAR encoding results of Table 4.4 will be presented. It is important to note that that the SA implementation used in SAR encoding was not implemented with parallel execution support (De Vicente et al., 2000). The number of threads that has been considered in parallel MOEAs has been: 1, 2, 4, 6 and 8. The machine used to perform the experiments consists of an Intel Core i7-4790 CPU running at 3.60GHz with 4 cores and 2 threads/per core and 16GB of RAM. The number of iterations of the experiment was established in 10^6 as in Parrales et al. (2019d); Szűcs & Balázs (2019). The population size for MOEAs was established in 100. This has value has been selected in order to guarantee the diversity of solutions while avoiding a slow convergence of individuals (Chen et al., 2012; Zitzler

& Thiele, 1999).

4.2.2.1 Runtime

Table 4.7 presents the execution time of the previously employed algorithms following the hour:minutes:seconds format. It is important to note that the applied SA method (De Vicente et al., 2000) does not perform a multiobjective optimization. The SA time executions presented in Table 4.7 refer to the time taken by the feature weighting task for each stage of the BoNT-A treatment and not for all stages at the same time. SA results are presented only for comparison purposes.

Table 4.7: Runtime achieved by SA and MOEA parallel algorithms with RT.

Methods	**Number of threads**				
	1	**2**	**4**	**6**	**8**
SPEA2	8:16:53	4:06:08	2:30:37	2:02:08	1:54:20
NSGAIII	8:12:20	4:05:14	2:28:52	2:02:17	1:55:32
NSGAII	8:12:04	4:05:12	2:28:48	2:02:05	1:54:35
SMPSO	8:13:02	4:06:16	2:29:04	2:07:27	1:59:15
PESA2	8:16:53	4:07:01	2:29:27	2:05:33	1:58:19
GDE3	8:13:02	4:06:08	2:29:45	2:02:39	1:54:38
SA-stage 1	4:13:05	NA	NA	NA	NA
SA-stage 2	4:32:31	NA	NA	NA	NA
SA-stage 3	4:19:25	NA	NA	NA	NA

According to the results of Table 4.7, apparently, MOEAs have a longer execution time than SA when only one thread is used. However, SA only performs accuracy optimization for a single stage. Therefore, the real total time employed by SA is the sum of the times of the first, second and third stages. This value is around 4 hours higher than the employed by the MOEAs with 1 thread.

Furthermore, it can be observed that parallel MOEAs executed on two or more threads have required less time than the SA algorithm. Parallel MOEAs are benefited from the use of more threads to distribute the computational load in the feature weighting task. However, it is important to note that the time difference between 6 and 8 threads is much smaller than the difference between 1, 2 and 4 threads. Because of that, it is important to detect any significant difference between the time spent using 6 and 8 threads. In this sense, the Wilcoxon (non-parametric) test was carried out between the time in seconds achieved when using 6 and 8 threads. The Wilcoxon test has been

considered because there are two groups of results to be contrasted, which are the obtained results when using 6 and 8 threads. The adjusted p-value is compared against a significance level of $\alpha = 0.05$ to reject or accept the null hypothesis that MOEAs perform with a significant difference in time. The obtained p-value of 0.003948, being less than 0.05, guarantee that there is a significant difference in runtime due to the use of 8 threads. It can be concluded that the number of threads improves runtime.

From Table 4.7, it can be seen that SPEA2 is the MOEA that has had the best execution time when using 8 threads. However, when applying the Wilcoxon test between 8 thread MOEA's values in seconds, a p-value of 0.4159 is obtained. This value, being more than 0.05, does not guarantee that there is a significant difference between execution times of MOEAs. This fact indicates that there is no MOEA that significantly improves others' time. Therefore, it is necessary to review their accuracy results on predictive models to select the best MOEA method.

4.2.2.2 Accuracy

Table 4.8 presents the best accuracy percentages values when predicting the treatment response to BoNT-A for the first, second and third stages. To present the results of this table, only non-dominated solutions that have the highest accuracies (lowest errors) have been selected for each algorithm.

According to the results, it can be observed that high values of accuracy, sensitivity and specificity are obtained both when the MOEA methods are applied and when SA is used. According to the results shown in this table, SA achieves the best accuracy (84.93%) for stage 1 while NSGAII achieve the best performance for stage 2 and 3. It means accuracies of 85.96% and 84.88% for stages 2 and 3, respectively. In all these best results, percentages higher than 80% were obtained as values of sensitivity and specificity, indicating a low number of false positives and false negatives.

To compare the runtime and the error obtained by each of the MOEAs contained in Table 4.8, Figure 4.2 is presented. Figures 4.2a and 4.2b depict the errors and runtimes produced during the feature weighting task for the first and second stages of the BoNT-A treatment, respectively. In addition, the solutions provided by SA have been considered in both figures, since they present the results of each stage separately. In the figures, the best points in terms of accuracy are marked with red circles for MOEAs. Blue circles are used to present best results achieved by SA. It is important to note how the charts show a better performance for MOEAs when using 6 and 8 threads than when using 1, 2 and 4 threads (both in runtime and accuracy), since the error for each stage decreases when each one is considered sepa-

Table 4.8: Accuracy percentage of SA and parallel MOEAs in combination with RT algorithm.

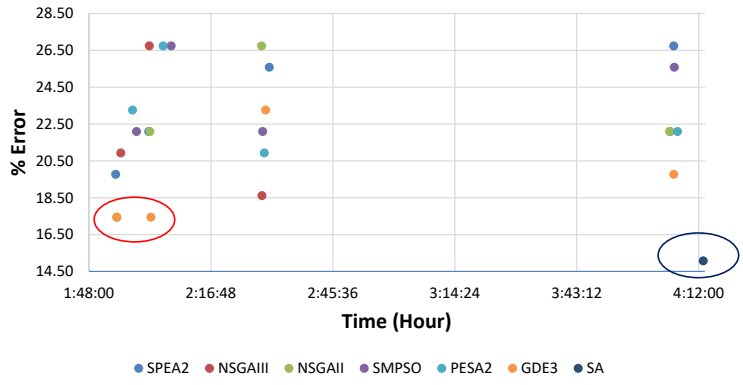
Classification algorithm	First stage			Second stage			Third stage		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
SPEA2	77.91%	75.94%	81.45%	79.06%	83.24%	75.56%	81.39%	82.31%	77.28%
NSGAIII	81.39%	77.61%	83.29%	82.56%	81.45%	78.14%	81.39%	75.89%	84.68%
NSGAII	82.56%	83.29%	78.71%	**85.96%**	86.26%	81.95%	**84.88%**	88.14%	82.45%
SMPSO	79.06%	82.31%	75.09%	82.56%	81.45%	84.26%	80.23%	77.14%	80.71%
PESA2	76.74%	82.31%	71.25%	84.88%	76.56%	82.31%	75.58%	68.27%	75.28%
GDE3	82.56%	85.74%	77.31%	81.39%	76.29%	83.29%	83.71%	78.07%	84.26%
SA	**84.93%**	87.56%	81.45%	85.74%	83.24%	88.14%	83.29%	84.68%	86.54%

rately. In Figure 4.2a, it can be observed that SA achieves the best accuracy (84.93%) when only stage 1 is considered. However, SA implementation performed (De Vicente et al., 2000) does not support multiobjective optimization or parallelism, as it has been commented in Section 3.2.5.2. Thus, it takes more time in the feature weighting task (close to 4 hours) without being able to minimize errors for both stages at the same time, while the MOEAs are able to do this. One of them, GDE3, achieves an error of 17.44% for stage 1, but it gets an error of 18.60% for stage 2 (see Figure 4.2b), being surpassed by NSGAII and PESA2 in that stage. These last two obtain the best error minimizations for stage 2 (errors of 14.04% and 15.12%, respectively) and stage 3 (errors of 15.12% and 16.29%, respectively), but they are surpassed by SA in stage 1. In stages 2 and 3, SA cannot achieve an error as low as GDE3 and NSGAII, despite its low error achieved in stage 1.

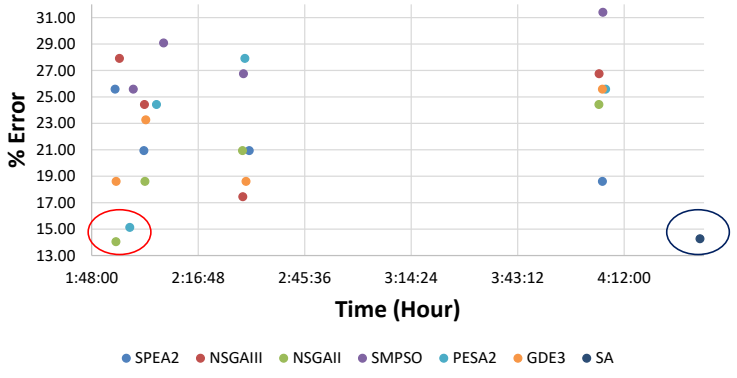
Given that it is difficult for us to visualize which method maximizes the accuracy (i.e., minimize the error) for both stages simultaneously, Figure 4.3 is presented considering only the MOEAs. It is important to note that SA is not taken into account in Figure 4.3 since it does not minimize both stages simultaneously. As can be seen in this figure, the best tradeoff is the one that minimizes the error for both stages, which is achieved with NSGAII when performing on 8 threads (marked with a red circle). To see if there is a significant difference between the accuracy values of each stage obtained by the MOEAs and SA methods, the Wilcoxon test is applied. The adjusted p-value is compared against a significance level of $\alpha = 0.05$ to reject or accept the null hypothesis that MOEAs and SA perform with a significant difference in accuracy. The p-values of 0.3168, 0.4115 and 0.5091 are obtained when comparing the accuracy values of first, second and third stages, respectively. These p-values have exceeded the 0.05 threshold, which means that the hypothesis of having a significant difference in accuracy has been rejected. It can be concluded that NSGAII is the MOEA that achieves the best performance when minimizing the prediction error. However, there is no significant time and accuracy difference with other MOEAs.

4.2.3 Panoramic prediction

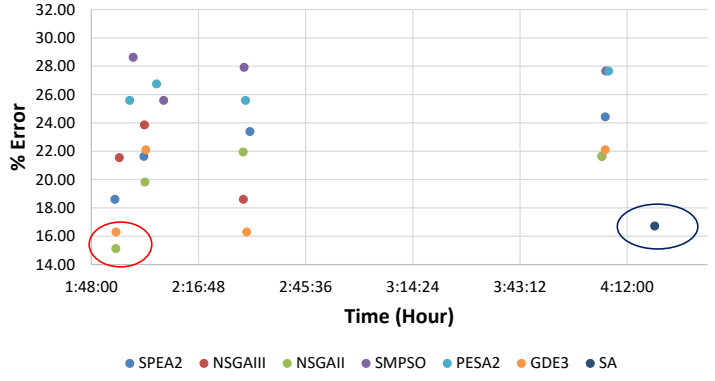
The purpose of this experiment is to obtain a panoramic model with the best accuracy. As mentioned in Section 3.3.1, the panoramic prediction approach makes use of multi-target algorithms. In this sense, PCT, BR and HOMER have been employed, following the recommendation given by Madjarov et al. (2012). Regarding their one-target classification parameter, the RT method (RT) has been selected due to the accurate results achieved in Table 4.4.



(a) First stage.

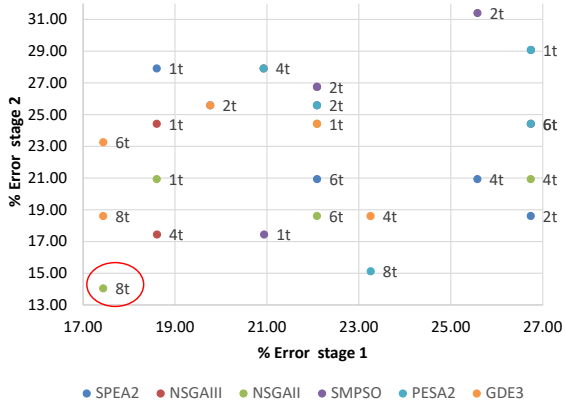


(b) Second stage.

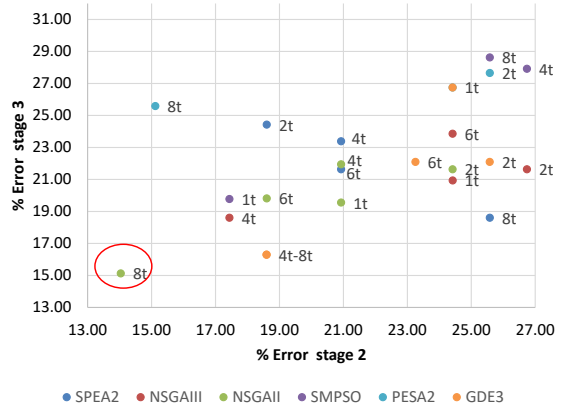


(c) Third stage.

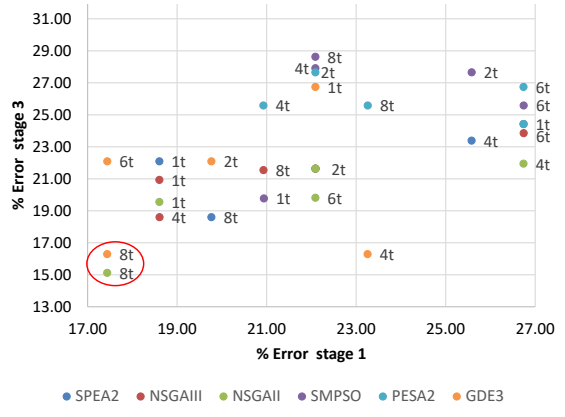
Figure 4.2: Time vs Error in 1st, 2nd and 3rd stages.



(a) First stage vs second stage



(b) Second stage vs third stage



(c) First stage vs third stage

Figure 4.3: Best points for each thread setting and MOEA method.

NSGAII has been selected as heuristic method for AMOR encoding due to the good results achieved in Table 4.8. More details about the parameters of the multi-target classifiers used in the experiments are presented in Table 4.9. The baseline results (without any rounding) are shown in Table 4.10. Also, FSS has been used for evaluation, applying the CFS method with the following parameters: 1 as the number of threads to use, 1 as the size of the thread pool, and best-first as the search method. The parameter D is set to 3 because three decimal magnitude orders have been considered as in Parrales et al. (2019c).

Table 4.9: Description of the multi-target classifier parameters used in experiments.

Classification algorithm	Parameters
Predictive clustering tree (PCT)	classifier=RT, Heuristic = Gain ratio
Binary relevance (BR)	classifier=RT
HOMER	type=Random, classifier=RT, Multi-label learner=BR

With the purpose of comparing the performance of the methods presented in Table 4.10, the Nemenyi's test procedure will be selected to conduct all pairwise comparisons in a multiple comparison analysis. The idea is to detect which technique has a statistically significant difference when outperforming the other methods. By observing the p-values of the tests from Tables 4.11, 4.12 and 4.13, the conclusions are: (1) The use of AMOR encoding with $d = 1$ produces a significant improvement in the accuracy values of the PCT, BR and HOMER methods without FSS (baseline), and with it for the first, second and third stage with the exception of BR with FSS. (2) The use of AMOR encoding in PCT with $d = 1$ produces significant improvements in the baseline and FSS values of the BR and HOMER methods for the first, second and third stages. (3) The use of AMOR encoding with $d = 2$ and $d = 3$ does not produce significant improvements in the accuracy values of PCT, BR and HOMER baseline and FSS for the first, second and third stages. According to these results, the highest accuracies are obtained when there are more perturbations in the numerical labels, e.g. rounding to the tenth ($d = 1$) instead of the hundredth ($d = 2$). It can be inferred that AMOR encoding then becomes an important factor, as it helps to optimize the prediction models, moving the solution within the search space to avoid being caught in a local minimum.

The aforementioned tables show that 73.26%, 75.58% and 74.61% are the best mean values of accuracy for the first, second and third stages of treatment, and they were obtained when performing PCT with $d = 1$. These results imply that without applying the treatment, it can be predicted how it

Table 4.10: Estimated performance metrics (mean ± std deviation) of panoramic prediction with $D = 3$ using 10-fold cross validation. The best results are highlighted in bold.

Method	Setting	First stage			Second stage			Third stage	
		Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
PCT	baseline	61.63%±3.75%	0.65±0.02	0.25±0.02	63.95%±2.74%	0.75±0.05	0.57±0.03	62.79%±4.16%	0.65±0.02	0.61±0.02
	FSS	63.07%±2.41%	0.61±0.03	0.66±0.04	64.61%±2.13%	0.56±0.02	0.72±0.02	60.76%±2.81%	0.52±0.05	0.69±0.02
	$d = 1$	**73.84%±0.91%**	0.72±0.02	0.75±0.01	**75.38%±1.24%**	0.73±0.03	0.76±0.02	**74.61%±2.14%**	0.73±0.03	0.75±0.02
	$d = 2$	70.93%±1.27%	0.74±0.02	0.61±0.02	66.92%±0.95%	0.68±0.03	0.65±0.03	68.46%±1.66%	0.71±0.03	0.67±0.02
	$d = 3$	65.12%±1.05%	0.74±0.03	0.48±0.02	66.28%±1.11%	0.65±0.01	0.67±0.02	67.69%±2.42%	0.60±0.02	0.54±0.04
BR	baseline	62.79%±2.37%	0.75±0.05	0.56±0.04	54.65%±4.18%	0.60±0.03	0.54±0.03	67.44%±3.56%	0.75±0.04	0.63±0.03
	FSS	63.95%±2.17%	0.75±0.03	0.57±0.02	68.60%±3.42%	0.75±0.04	0.64±0.03	66.27%±3.75%	0.68±0.02	0.61±0.03
	$d = 1$	70.93%±1.01%	0.74±0.02	0.61±0.02	72.09%±1.32%	0.75±0.03	0.70±0.02	73.84%±2.41%	0.72±0.02	0.75±0.03
	$d = 2$	67.44%±0.95%	0.68±0.03	0.66±0.04	69.76%±2.07%	0.68±0.03	0.70±0.03	72.09%±2.71%	0.72±0.02	0.71±0.03
	$d = 3$	63.95%±1.26%	0.79±0.02	0.56±0.04	65.12%±1.62%	0.74±0.03	0.48±0.04	68.60%±2.23%	0.75±0.02	0.64±0.02
HOMER	baseline	55.81%±2.14%	0.55±0.03	0.56±0.04	56.97%±3.86%	0.55±0.04	0.57±0.03	61.62%±4.13%	0.62±0.02	0.61±0.03
	FSS	56.97%±1.81%	0.55±0.03	0.57±0.03	58.13%±2.15%	0.58±0.02	0.57±0.03	60.46%±3.84%	0.58±0.04	0.61±0.02
	$d = 1$	68.60%±1.04%	0.72±0.02	0.66±0.03	72.09%±2.36%	0.72±0.03	0.71±0.03	73.25%±3.24%	0.72±0.02	0.73±0.02
	$d = 2$	63.95%±0.72%	0.65±0.02	0.63±0.01	66.27%±1.96%	0.65±0.02	0.66±0.02	67.44%±2.26%	0.72±0.03	0.64±0.04
	$d = 3$	59.30%±0.87%	0.58±0.02	0.59±0.02	60.46%±1.71%	0.58±0.02	0.61±0.03	65.11%±2.34%	0.65±0.03	0.64±0.03

will work at each stage for three out of four patients. In addition, 0.72, 0.73 and 0.73 are the mean sensitivity values obtained when $d = 1$ for the first, second and third stages of treatment, respectively, indicating a good detection of patients who respond positively to treatment. Moreover, the model obtained with PCT obtains mean specificity values of 0.75, 0.76 and 0.75 for the first, second and third stage of the treatment, which indicates that this model is good when detecting patients who respond negatively to all stages of treatment. Results allow to conclude that panoramic prediction allows the doctor to provide an insightful preliminary criterion for the response to the treatment and make the respective medical decisions.

4.2.4 Feedback prediction

The purpose of this experiment is to improve the prediction of the therapeutic response to BoNT-A through the use of known information. This information is the response that the patient has had to the previous stages of treatment. Hence, it is not known before the first stage.

As explained in Section 3.3.2, this approach of prediction implies the use of one-target classification algorithms. In this sense, SAR encoding will be used to improve accuracy results in predictive models due to the good results achieved in Table 4.4. In this phase of the methodology, only rounding to the tenth will be considered, given the good results obtained in the previous experiment. To do this, the parameter D has been set to 1. Also, CFS will be the FSS method used to compare with SAR encoding results in this experiment due the significant differences with baseline results achieved in 4.6. It is set in the same way as described in Section 4.2.3. RT will be the one-target classification algorithm to consider given the high accuracies obtained in Table 4.4. Results will be obtained after imputing the initial clinical dataset following the guidelines addressed in Section 3.4.4. The obtained results are presented in Table 4.14. To make comparisons with results obtained by RT with SAR rounded to one decimal (RT+SAR with d = 1) and FSS, their results of Tables 4.4 and 4.3 have been included in Table 4.14 as "Single" prediction approach.

With the purpose of verifying whether the improvement in classification due to the feedback prediction approach is statistically significant, the Wilcoxon (non-parametric) test was carried out between the accuracy values of RT+SAR and the RT+FSS methods under feedback and single prediction approaches when using 10-fold cross validation. The Wilcoxon test has been considered because there are two methods to be contrasted, which are the feedback and single prediction approaches. The adjusted p-values are

Table 4.11: Nemenyi-test p-values on the 10-fold cross validation accuracy values of methods used in Table 4.10 for the first stage.

		PCT					BR					HOMER			
		Baseline	FSS	d=1	d=2	d=3	Baseline	FSS	d=1	d=2	d=3	Baseline	FSS	d=1	d=2
PCT	FSS	0.99999	-	-	-	-	-	-	-	-	-	-	-	-	-
	d=1	$2.20{\cdot}10^{-5}$	0.00077	-	-	-	-	-	-	-	-	-	-	-	-
	d=2	0.0008	0.0154	0.99999	-	-	-	-	-	-	-	-	-	-	-
	d=3	0.68523	0.98252	0.14702	0.60169	-	-	-	-	-	-	-	-	-	-
BR	Baseline	1	1	0.00033	0.00782	0.95201	-	-	-	-	-	-	-	-	-
	FSS	0.98661	1	0.01324	0.13567	0.99999	0.99993	-	-	-	-	-	-	-	-
	d=1	0.00094	0.01755	0.99999	1	0.62872	0.00898	0.149	-	-	-	-	-	-	-
	d=2	0.08485	0.44299	0.81558	0.99623	0.99941	0.31935	0.8961	0.99722	-	-	-	-	-	-
	d=3	0.98953	1	0.01158	0.12329	0.99998	0.99996	1	0.13567	0.88079	-	-	-	-	-
HOMER	Baseline	0.89397	0.41693	$2.70{\cdot}10^{-10}$	$3.80{\cdot}10^{-8}$	0.00601	0.55099	0.0836	$4.80{\cdot}10^{-8}$	$5.00{\cdot}10^{-5}$	0.0927	-	-	-	-
	FSS	0.96316	0.59003	$1.40{\cdot}10^{-9}$	$1.60{\cdot}10^{-7}$	0.01456	0.72134	0.157	$2.00{\cdot}10^{-7}$	0.00016	0.1718	1	-	-	-
	d=1	0.02662	0.21445	0.95786	0.99993	0.98661	0.13752	0.6852	0.99996	1	0.6592	$7.10{\cdot}10^{-6}$	$2.50{\cdot}10^{-5}$	-	-
	d=2	0.9922	1	0.00991	0.11024	0.99996	0.99998	1	0.1216	0.86171	1	0.10413	0.18995	0.62872	-
	d=3	0.99983	0.94559	$9.50{\cdot}10^{-8}$	$7.10{\cdot}10^{-6}$	0.115	0.9795	0.5549	$8.70{\cdot}10^{-6}$	0.00301	0.5822	0.99993	1	0.00061	0.61331

Table 4.12: Nemenyi-test p-values on the 10-fold cross validation accuracy values of methods used in Table 4.10 for the second stage.

		PCT					BR					HOMER			
		Baseline	FSS	d=1	d=2	d=3	Baseline	FSS	d=1	d=2	d=3	Baseline	FSS	d=1	d=2
PCT	FSS	1	-	-	-	-	-	-	-	-	-	-	-	-	-
	d=1	0.00101	0.00276	-	-	-	-	-	-	-	-	-	-	-	-
	d=2	0.97305	0.99429	0.20441	-	-	-	-	-	-	-	-	-	-	-
	d=3	0.99941	0.99997	0.055	1	-	-	-	-	-	-	-	-	-	-
BR	Baseline	0.5081	0.34259	$9.90{\cdot}10^{-10}$	0.00736	0.04094	-	-	-	-	-	-	-	-	-
	FSS	0.78511	0.90222	0.52756	1	0.99972	0.0009	-	-	-	-	-	-	-	-
	d=1	0.03456	0.07183	0.9999	0.80968	0.46954	$3.50{\cdot}10^{-7}$	0.9789	-	-	-	-	-	-	-
	d=2	0.30646	0.46572	0.93095	0.99789	0.94692	$3.40{\cdot}10^{-5}$	1	0.99997	-	-	-	-	-	-
	d=3	1	1	0.00565	0.99876	1	0.23826	0.9556	0.11826	0.59781	-	-	-	-	-
HOMER	Baseline	0.78826	0.62872	$1.40{\cdot}10^{-8}$	0.03063	0.13026	1	0.0048	$3.50{\cdot}10^{-6}$	0.00024	0.4965	-	-	-	-
	FSS	0.86907	0.73534	$3.70{\cdot}10^{-8}$	0.04991	0.18995	1	0.0086	$8.00{\cdot}10^{-6}$	0.00049	0.6094	1	-	-	-
	d=1	0.03456	0.07183	0.9999	0.80968	0.46954	$3.50{\cdot}10^{-7}$	0.9789	1	0.99997	0.1183	$3.50{\cdot}10^{-6}$	$8.00{\cdot}10^{-6}$	-	-
	d=2	0.99826	0.99986	0.07636	1	1	0.02856	0.9999	0.55099	0.96972	1	0.09688	0.14508	0.55099	-
	d=3	0.99357	0.97058	$1.20{\cdot}10^{-6}$	0.23014	0.55099	0.99894	0.0605	0.00015	0.00554	0.9294	0.99999	1	0.00015	0.46954

Table 4.13: Nemenyi-test p-values on the 10-fold cross validation accuracy values of methods used in Table 4.10 for the third stage.

		PCT				BR				HOMER			
		Baseline	FSS	d=1	d=2	d=3	Baseline	FSS	d=1	d=2	d=3	Baseline	FSS	d=1	d=2
PCT	FSS	0.99998	-	-	-	-	-	-	-	-	-	-	-	-	-
	d=1	$8.30{\cdot}10^{-5}$	$1.10{\cdot}10^{-6}$	-	-	-	-	-	-	-	-	-	-	-	-
	d=2	0.51198	0.09004	0.42062	-	-	-	-	-	-	-	-	-	-	-
	d=3	0.81264	0.2637	0.17182	1	-	-	-	-	-	-	-	-	-	-
BR	Baseline	0.83547	0.28767	0.15497	1	1	-	-	-	-	-	-	-	-	-
	FSS	0.9899	0.69257	0.02856	0.99947	1	1	-	-	-	-	-	-	-	-
	d=1	0.00026	$4.10{\cdot}10^{-6}$	1	0.59392	0.29076	0.26663	0.0605	-	-	-	-	-	-	-
	d=2	0.00614	0.00018	0.9998	0.96416	0.79758	0.77229	0.3632	0.99999	-	-	-	-	-	-
	d=3	0.42062	0.06245	0.51198	1	1	1	0.9982	0.68523	0.98252	-	-	-	-	-
HOMER	Baseline	1	1	$1.30{\cdot}10^{-5}$	0.26958	0.56663	0.59781	0.9261	$4.40{\cdot}10^{-5}$	0.00139	0.2044	-	-	-	-
	FSS	0.99999	1	$1.30{\cdot}10^{-6}$	0.09688	0.27854	0.30328	0.7107	$4.80{\cdot}10^{-6}$	0.00021	0.0675	1	-	-	-
	d=1	0.00108	$2.20{\cdot}10^{-5}$	1	0.80669	0.50422	0.47337	0.1451	1	1	0.8715	0.00021	$2.60{\cdot}10^{-5}$	-	-
	d=2	0.87384	0.33586	0.12674	1	1	1	1	0.22483	0.72134	1	0.65538	0.3528	0.41693	-
	d=3	0.99999	0.9746	0.00237	0.93858	0.99571	0.99696	1	0.00614	0.07521	0.8961	0.99918	0.97819	0.0196	0.99849

Table 4.14: Estimated performance metrics (mean ± std deviation) of feedback and single prediction approach with SAR (d=1) and FSS using 10-fold cross validation. The best results are highlighted in bold.

Prediction approach	Method	First stage			Second stage			Third stage		
		Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Feedback	RT+SAR	**87.86%±1.38**	85.16±1.74	**90.37±1.61**	**88.95%±2.72**	**86.58±2.31**	**89.62±2.05**	**87.14%±1.64**	**92.23±1.95**	84.48±1.78
	RT+FSS	68.25±2.91%	67.40±3.14%	65.19±2.19%	76.51±3.24%	74.15±2.81%	77.42±2.31%	79.43±2.37%	75.26±2.48%	72.65±3.57%
Single	RT+SAR	84.93±1.25%	**87.56±1.85%**	81.45±3.65%	85.74±2.17%	83.24±2.54%	88.14±2.72%	83.29±1.07%	84.68±2.27%	**86.54±2.28%**
	RT+FSS	49.35±2.72%	60.16±1.21%	29.63±0.63%	68.83±3.61%	65.79±2.04%	71.79±2.67%	75.32±1.18%	58.33±3.74%	83.02±3.10%

compared against a significance level of $\alpha = 0.05$ to reject or accept the null hypothesis that prediction approaches performs equally. The obtained p-values were 0.009409, 0.0186 and 0.01383 for the first, second and third stages of the treatment prediction, respectively. These values, being less than 0.05, guarantee that there is a significant difference in the distributions of values between the two prediction approaches. Therefore, results allow to conclude that the feedback prediction approach improves the results obtained by incorporating into the predictive model the therapeutic responses that are already known so far.

4.3 Dealing with missing values

Until now, the experiments have made use of imputed data in general, considering all the records to carry out the imputation. The purpose of this section is to experience the use of a more specific imputation. That is, by using a hierarchy of models that takes into account the number of missing values (NAs) of each record. Since not all records in the group will have the same number of NAs, the use of a fuzzy selector to establish the membership of a medical record to a certain model has been proposed in Section 3.4. Thus, NAs are imputed taking into account only the medical records of their respective groups. The use of the MVDMS^2 method for dealing with NAs has been proposed for it. Regarding its parameters, 3 has been defined as the number of groups (G parameter) in order to categorize the records according to their low, medium and high level of NA cells. The dataset has been split using 75% for training and 25% for testing the hierarchical model. The training dataset has been split into training and validation when using the k-fold validation approach. The table B described in Figure 3.9 is generated from the training dataset. This table is used for clustering the records by their NA values when applying the k-medians clustering with $k = G$.

With the purpose of building a fuzzy selector that considers the number of NAs in new records when assigning the correspondent model, the FURIA algorithm has been applied to the T_{map} table described in Section 3.4.3 with the following parameters: 3 folds for pruning (the rest for growing the rules), 2 as the number of optimization runs, 2 as the minimum total weight of the instances in a rule and 2 as the number of decimal places to be used for the output of numbers in the model. One rule per model with an accuracy of 85.52% has been obtained with the FURIA algorithm. Regarding accuracy, it is necessary to clarify that the purpose of these rules is not to classify treatment responses, but to build a fuzzy selector that assigns the corresponding model. The rules R1, R2 and R3 are defined according to the number of

missing values in the interval [0,14], where:

- R1: If the number of NAs falls in the region defined by the trapezoidal membership function with [0, 0, 3, 4], then the selected model will be "model1" with a CF of 0.83,
- R2: If the number of NAs falls in the region defined by the trapezoidal membership function with [3, 4, 14, 14], then the selected model will be "model2" with a CF of 0.85,
- R3: If the number of NAs falls in the region defined by the trapezoidal membership function with [0, 0, 11, 12], then the selected model will be "model2" with a CF of 0.78,

where 14 is the maximum number of NAs found in the medical registers, CF is the certainty factor and [a,b,c,d] represents the boundaries of the trapezoidal region (Hühn & Hüllermeier, 2009). These functions are graphically represented in Figure 4.4. If the number of NAs falls in the middle of three regions as in the case of NAs = 3, the selected model will be the model with the highest CF value. After that, the missing values are replaced with the values obtained from multiple imputation within their group.

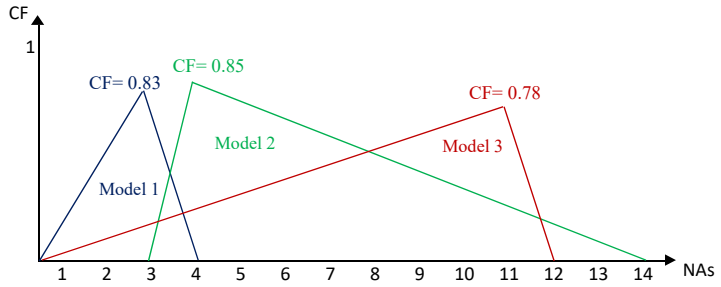


Figure 4.4: Membership functions of the fuzzy model selector

The purpose of the next experiments is to demonstrate whether dealing with missing values helps to improve the results obtained when a general imputation of the medical records has been carried out. In this sense, panoramic and feedback prediction approaches will be taken into account due to the good accuracy results achieved in the previous section. This implies that the groups obtained when applying the fuzzy rules on medical records and after imputing them will be trained using panoramic and feedback prediction approaches.

4.3.1 Panoramic prediction

The fuzzy selector of Figure 4.4 has been applied to the records. The groups obtained are trained using the PCT+AMOR (d=1) combination due to its good results achieved as shown in Table 4.10. Accuracies achieved by the three hierarchical models (models 1, 2 and 3) generated when considering the NA number are presented in Table 4.15. The best results are obtained when performing a hierarchy of PCT+AMOR prediction models instead of only PCT+AMOR combination of Table 4.10, with a mean accuracy of 85.14%, 88.35% and 85.73%, as shown in the row labelled "Hierarchy" in Table 4.15. Moreover, the high values of sensitivity and specificity indicate the goodness of the hierarchy model when predicting the "high" and "low" responses to treatment.

With the purpose of verifying if the improvement in classification due to the use of a hierarchy of models in panoramic prediction is statistically significant, the Wilcoxon test was carried out between the accuracy values of the hierarchical model and panoramic prediction results of the PCT+AMOR (d=1) combination presented in Table 4.10. The Wilcoxon test has been selected because there are only two methods to be tested, which are the accuracy values with and without a hierarchy of models. The results of models 1, 2 and 3 are not taken into account in the statistical validation since they are part of the final hierarchical model built in the MVDMS^2 process presented in Figure 3.14. The adjusted p-values are compared against a significance level of $\alpha = 0.05$ to reject or accept the null hypothesis that a pair of methods perform equally. The p-values of 0.0001571, 0.0001571 and 0.0001571 were obtained for the first, second and third stages of the treatment prediction, respectively. These values, being less than 0.05, guarantee that there is a significant difference in the distributions of values between the two methods. With the results obtained, it can be concluded that the use of a hierarchy of models helps to improve the accuracy in panoramic prediction models, since it takes into account the medical information available of each patient.

4.3.2 Feedback prediction

As in the previous experiment, the fuzzy selector of Figure 4.4 has been applied to the medical records. The groups obtained are trained using the RT+SAR combination due to its good results achieved in Table 4.14. Accuracies achieved by the three hierarchical models (models 1, 2 and 3) generated when considering the NA number are presented in Table 4.16. The best results are obtained when performing a hierarchy of RT+SAR prediction

Table 4.15: Estimated performance metrics (mean ± std deviation) of hierarchy models with $D = 1$ and $G = 3$ using panoramic prediction and 10-fold cross validation. The hierarchy results are highlighted in bold.

Algorithms	Model	First stage			Second stage			Third stage		
		Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
PCT+AMOR	Model 1	81.25±1.23%	85.71±1.05%	77.78±2.35%	84.37±2.69%	86.67±2.33%	82.35±1.89%	85.16±2.52%	75.14±3.45%	93.75±2.26%
	Model 2	87.87±2.68%	82.35±2.13%	93.75±2.47%	90.90±1.91%	94.11±1.34%	87.51±0.48%	84.85±2.25%	77.78±2.35%	84.21±1.13%
	Model 3	81.08±1.42%	76.15±2.14%	84.34±1.29%	83.78±2.62%	88.47±1.35%	78.94±1.05%	81.25±0.94%	83.32±0.85%	77.78±2.35%
	Hierarchy	**85.14±1.29%**	88.69±2.05%	81.98±1.89%	**88.35±2.17%**	86.67±1.95%	89.82±2.14%	**85.73±1.53%**	85.92±1.86%	84.97±2.23%

models instead of only RT+SAR combination of Table 4.14, with a mean accuracy of 92.11%, 94.23% and 96.05%, as shown in the row labelled "Hierarchy" in Table 4.16. Moreover, the high values of sensitivity and specificity indicate the goodness of the hierarchy model when predicting the "high" and "low" responses to treatment.

In order to check whether the improvement in classification due to the hierarchical models in feedback prediction is statistically significant, the Wilcoxon (non-parametric) test was carried out between the accuracy values of the hierarchical model and feedback prediction results of the RT+SAR combination presented in Table 4.14. This test has been selected because there are only two prediction approaches to be compared, the hierarchical and the feedback approaches, both with RT+SAR. The results of models 1, 2 and 3 are not taken into account in the statistical validation since they are part of the final hierarchical model built in the MVDMS^2 process presented in Figure 3.14. The adjusted p-values are compared against a significance level of $\alpha = 0.05$ to reject or accept the null hypothesis that a pair of methods perform equally. The p-values of 0.0006697, 0.0003811 and 0.0001571 have been obtained for the first, second and third stages of the treatment prediction, respectively. These values, being less than 0.05, guarantee that there is a significant difference in the distributions of values between the two methods. It is concluded that the use of a hierarchy of models helps to improve the accuracy in feedback prediction models since it is better suited to the medical information available in the medical dataset.

4.4 Obtaining relevant medical attributes

Section 3.5.2 discusses the importance of studying a consensus model with the prediction models built for the first, second and third stage of the BoNT-A treatment. In this research work, the hierarchical models have proved to be the best classifiers for all stages of treatment when using RT as classification algorithm and AMOR and SAR encoding for panoramic and feedback prediction approaches, respectively.

4.4.1 Extracting relevant attributes

With the purpose of extracting relevant attributes from prediction models, many prediction models will be induced. Moreover, only the most frequent attributes for each level of the studied models will be taken into account. An important point to emphasize is that the obtained ensemble trees are not intended to be a prediction model of the treatment response for each

Table 4.16: Estimated performance metrics (mean ± std deviation) of hierarchy models with $D = 1$ and $G = 3$ using feedback prediction and 10-fold cross validation. The hierarchy results are highlighted in bold.

Algorithms	Model	First stage			Second stage			Third stage		
		Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
RT+SAR	Model 1	87.51±2.14%	86.67±1.15%	88.23±0.78%	93.75±1.05%	87.51±0.48%	98.12±0.24%	93.33±0.92%	88.36±1.26%	97.24±1.05%
	Model 2	90.91±2.38%	88.23±1.59%	93.75±2.13%	93.93±1.45%	94.11±1.23%	93.61±1.78%	90.91±2.05%	93.75±0.63%	88.23±1.83%
	Model 3	91.89±1.28%	88.88±1.45%	94.73±0.78%	94.59±1.78%	97.16±0.23%	89.47±1.05%	97.29±0.32%	94.44±1.02%	98.12±0.24%
	Hierarchy	**92.11±2.12%**	93.49±2.16%	90.65±2.09%	**94.23±2.17%**	92.36±2.13%	95.94±2.21%	**96.05±2.21%**	98.12±0.24%	94.53±0.71%

treatment stage. On the contrary, they will allow the study of the most frequent clinical attributes and the relations that appear in the majority of the selected prediction models (only prediction models with the highest accuracies).

Many classification trees will be induced by the resampling method (using k-fold cross validation with k=10) with the AMOR and SAR encoding (used for the experiments in Section 4.3). These relevant attributes are contrasted with the important features obtained when using the FSS methodology in Section 4.2.1. The prediction models selected for induction will be the models (1, 2 and 3) that are part of the hierarchical models presented in Tables 4.15 and 4.16 for panoramic and feedback prediction approaches, respectively. 5000 prediction models for each treatment stage will be generated from 50 solutions found by SA and MOEA heuristic methods (weighted attribute vectors), which make it possible to achieve the highest accuracies for all treatment stages. Regarding the root vertex of the ensemble tree, the 0.99 quantile will be applied as the t value.

Table 4.17: Top-10 clinical attributes for the first level (root) of feedback prediction model 1 on the first stage.

Feature	Frequency
Platelets	1673
Hemoglobin	1012
Emergency days by month	752
Migraine days by month	516
History of migraine status	500
1st grade family with migraine	482
Creatinine	464
Unilateral pain	348
GON	303
Onset age of toxin treatment	241

For example, inducing many RTs from feedback prediction model 1 of Table 4.16, this value was equal to 1449.08 for the first treatment stage. In this way, Platelets was selected as the root of the consensus tree for the first stage because of its high frequency (1673 times). In a similar way, t is defined as the 0.9 quantile from the empirical observation on the edge frequency distributions in the other levels of the ensemble tree for all treatment stages. In this way, only the attributes with occurrences higher than 0.9 are retained.

Figure 4.5 presents the most frequent clinical attributes from feedback prediction model 1 of Table 4.16. An important aspect to note is that the L_{max} value has been defined as 3 for all treatment stages. This value was established by considering the comprehension of the resultant consensus tree

as a primordial criterion. Higher values of this parameter would allow to see more attributes, but comprehension could decrease when contrasting these attributes with those obtained with the FSS method. In this sense, a consensus tree with a low number of leaves is more understandable.

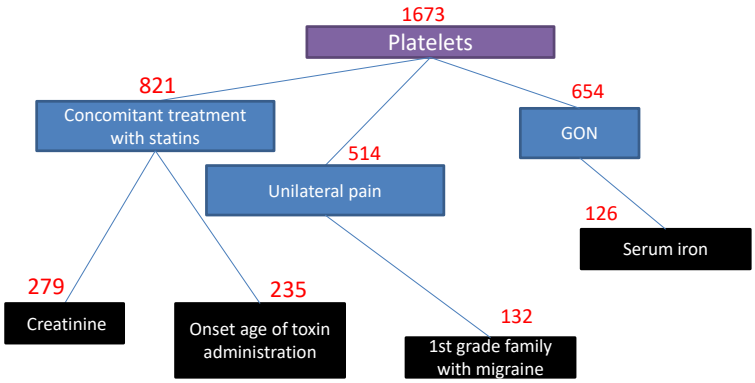


Figure 4.5: Consensus tree using RTs from feedback prediction model 1 of first stage of Table 4.16.

For the hierarchy of panoramic prediction models (PCT + AMOR encoding with $d = 1$), the medical attributes are the same in the three stages, given that a consensus model is obtained from the multi-target prediction models. All the relevant medical attributes from the 13 consensus models are presented in Table 4.18. The selected features when applying FSS in Section 4.3.1 are also presented in Table 4.18. The "GON", "Analgesic abuse", "1st grade family with migraine" and "Drugs tested before toxin" medical attributes have been selected for the consensus models 10, 9, 8 and 6 times, respectively. Moreover, "1st grade family with migraine" and "Chronic migraine time evolution" have also been selected for the majority of the feedback consensus models for the first and second treatment stage, respectively. Furthermore, "Headache days per month", "Unilateral pain" and "Migraine days per month" are present in the majority of the third stage consensus models. Finally, "GON", "Drugs tested before toxin" and "Chronic migraine time evolution" have also been selected by FSS, which was described in Section 4.2.1.

To summarize, the medical attributes that appear at least once for each stage in the consensus models of the feedback prediction and appear at least one of the panoramic consensus models are: "Chronic migraine time evolution", "GON", "Hemoglobin", "Analgesic abuse", "Serum iron", "1st grade family with migraine", "Retroocular component", "Chronic migraine time evolution", "Headache days per month", "Unilateral pain", "Platelets", "Anxiety", "Concomitant oral preventive treatment" and "Onset age of toxin treatment".

Table 4.18: Relevant attributes from hierarchical models of panoramic and feedback prediction approaches of Tables 4.15 and 4.16 and FSS.

Model		First stage	Second stage	Third stage
Consensus models	Feedback prediction model 1	Platelets, Concomitant treatment with statins, Unilateral pain, GON, 1st grade family with migraine, Onset age of toxin treatment, Serum iron, Creatinine	Analgesic abuse, Preventive oral treatment at time of infiltration, Neuromodulator, Concomitant antidepressant treatment, GON, Chronic migraine time evolution, Retroocular component	Hemoglobin, Platelets, History of migraine status, Headache days per month, 1st grade family with migraine, GON, Preventive oral treatment at time of infiltration, Unilateral pain, Pneumopathy, Catamenial, Depression
	Feedback prediction model 2	Chronic migraine time evolution, Hemoglobin, Analgesic abuse, Retroocular component, GON, Anxiety, Onset age of toxin treatment, 1st grade family with migraine	GON, Chronic migraine time evolution, Headache days per month, Hemoglobin, Serum iron, Concomitant oral preventive treatment, Creatinine, Onset age of toxin treatment, Anxiety, Depression	GON, Drugs tested before toxin, Retroocular component, Concomitant oral preventive treatment, Serum iron, Onset age of toxin treatment, Migraine days per month
	Feedback prediction model 3	Headache days per month, GGT, Migraine days per month, Drugs tested before toxin, Neuromodulator, Concomitant oral preventive treatment, Enolism, Analgesic abuse, 1st grade family with migraine	Unilateral pain, GON, Drugs tested before toxin, Chronic migraine time evolution, Chronic migraine, Anxiety, 1st grade family with migraine, Analgesic abuse, Platelets	Headache days per month, Unilateral pain, Migraine days per month, GON, Chronic migraine time evolution, Analgesic abuse, Onset age of toxin treatment, 1st grade family with migraine, Platelets, Anxiety
	Panoramic prediction model 1	GOT, Drugs tested before toxin, Chronic migraine, Onset age of toxin treatment, Unilateral pain, Analgesic abuse, Headache days per month, 1st grade family with migraine, Anxiety		
	Panoramic prediction model 2	Hemoglobin, Analgesic abuse, GPT, Serum iron, GON, 1st grade family with migraine, Calcium antagonists, Chronic migraine time evolution, GOT, Retroocular component		
	Panoramic prediction model 3	Migraine days per month, Drugs tested before toxin, Dermopathy, Analgesic abuse, Pneumopathy, Serum iron, Concomitant oral preventive treatment, GPT, Triptans per month		
FSS (Section 4.2.1)		Onset age of toxin treatment, Chronic migraine, Chronic migraine time evolution, drugs tested before toxin, tricyclic antidepressants, vitamin B12	GON, Preventive oral treatment at time of infiltration, Tricyclic antidepressants, Gastropathy, Pneumopathy, Dermopathy	Calcium antagonists, Catamenial, Concomitant oral preventive treatment, Gastropathy, Headache days per month, Analgesic abuse

These are the most important medical attributes among all the relevant attributes from the consensus models presented in Table 4.18.

4.4.2 Medical discussion

Some predictors of response to treatment with BoNT-A are in agreement with current publications, namely: migraine time evolution (Eross et al., 2005; Domínguez et al., 2018), unilateral pain (Domínguez et al., 2018; Mathew et al., 2008a), analgesic abuse (Freitag, 2010), days of headache (Domínguez et al., 2018) and the retroocular component (Lin et al., 2014). Moreover, these articles continue supporting the approach of not delaying treatment with BoNT-A in those patients who have a diagnosis of chronic migraine, who will improve more than those with a shorter evolution time and with a profile of lesser severity of the migraine. Following this line of thought, it is not strange to find that the presence of status, the number of triptans per month or the number of previous tested drugs, are also predictors of response.

A interesting fact not assessed so far is the predictive nature of the response in patients who take concomitant oral preventive treatment. Although it is not described in the literature, it is possible that the variables such as relatives in the first degree, the catamenial component and the presence of sensory alterations such as sono or photophobia, are predictive, either because they really assure that we are dealing with a patient with chronic migraine, a fact whose diagnosis is not always easy when a patient presents daily headaches and the semiological profile is no longer so pure.

A clinical relation with the analytical parameters (liver profile, iron, platelets, creatinine, hemoglobin) and associated pathologies such as dermopathy, gastropathy, dyslipidemia, hypertension and lung disease has not been found. But these points open up future lines of research with more targeted prospective studies. Relevant attributes also agree with the literature that neither gender nor nausea or vomiting (Jakubowski et al., 2006) have been predictive.

To conclude, several of the medical attributes that are relevant to predict the treatment response to BoNT-A are coherent with the medical literature. Those are: migraine time evolution, unilateral pain, analgesic abuse, days of headache and the retroocular component. Other medical attributes revealed as relevant by the consensus models as "Concomitant oral preventive treatment" or "Platelets" have no medical explanation yet. Therefore, they should be studied in the future with more specific prospective studies.

Chapter 5

Conclusions and future work

I am slowly coming to the conclusion
that it's more important to learn to work
with what you've got, under the
circumstances you've been given, than
wishing for different ones.

Charlotte Eriksson

5.1 Conclusions

This Ph.D. Thesis has explored some data processing methodologies in the area of e-Health for categorizing therapeutic responses in patients with migraine. In a real e-Health scenario, this work has focused on the prediction of the response to the treatment of migraine through the use of retrospective medical records collected from *Hospital Clínico Universitario* in Valladolid and *Hospital Universitario de La Princesa*, in Madrid.

In this research we pose and answer the following questions: is it possible to predict the response to every stage of the BoNT-A treatment for migraine? Does a predictive model for the BoNT-A treatment in migraine exist? How do these models respond under missing values? Is it possible to reveal those medical factors that make it possible a high response to the BoNT-A treatment? The medical factors used to predict the response of the treatment are coherent with the knowledge of medical experts? To answer these questions, a methodology has been developed, which considers the following issues:

1. The preprocessing of the data in order to mitigate some limitation problems that are commonly found in clinical datasets, like the presence of many attributes or medical factors present in a low number of

registers (Cabitza et al., 2019).

2. A panoramic prediction for allowing doctors to decide whether the administration of the treatment will be beneficial without involving unnecessary treatments.

3. A feedback prediction for those situations when the treatment has begun and the results of some stages are known.

4. The extraction of relevant medical attributes for allowing to verify whether prediction models are coherent with the knowledge of medical experts.

To address the heterogeneous clinical data provided by the doctors of the two hospitals, a numerical encoding approach is considered. In this sense, the SAR encoding is proposed for finding a better representation of the numeric labels. This technique considers the Simulated Annealing (SA) algorithm and a rounding operation to perform small numeric label perturbations for each column of the medical dataset, producing a data transformation for achieving high prediction accuracies without adding more columns to the dataset.

Because the proposed SAR encoding considers a SA implementation that does not allow the optimization of multiple objectives, the minimization of the prediction error for all stages is not solved simultaneously. For solving this issue, the use of MOEAs metaheuristics has been considered for adapting the SAR encoding to a multi-target prediction scenario. In this way, the SA metaheuristic is applied only when improving numeric labels in one-target prediction scenario, while MOEAs are applied for a multi-target prediction scenario. This adaptation of the SAR encoding to the multi-target scenario has been called AMOR encoding.

To address the existence of missing values, the imputation of data has been considered in this Thesis. For this reason, a hierarchy of models has been proposed in order to handle that the lack of clinical information, because it can provide useful information to build a set of prediction models. The purpose is to adapt the prediction to the missing values appearing in the collected clinical records. This technique considers the clustering of records that contain similar missing values. This approach also addresses data imputation to fill in the missing values based on records of each group.

The results show a significant improvement in accuracy due to the use of SAR encoding, from close to 68% (baseline) to 75% with panoramic prediction, and up to around 88% when using feedback prediction. Moreover, predictability of panoramic and feedback prediction models are improved

when applying a hierarchy of models, obtaining accuracies close to 85% and 94% respectively. Regarding the runtime, the obtained results with the use of MOEAs show that training times are decreased from 8 to less than 2 hours when using 8 threads.

Through the use of the proposed methodology it has been possible to extract the relevant attributes that allow to know in advance the response to the treatment. These are: "evolution of migraine time", "unilateral pain", "abuse of analgesics", "days of headache" and the "retroocular component". All these attributes have been consistent with the expert knowledge of doctors. However, other medical attributes revealed as relevant by consensus models such as "Concomitant oral preventive treatment" or "Platelets" still have no medical explanation. Therefore, they should be studied by doctors in the future with more specific prospective studies in order to find out the medical relevance of such attributes.

As it can be seen, a functional predictive methodology of therapeutic responses for the treatment of chronic migraine based on retrospective data has been presented. This opens the research to many other areas, which the author believes would lead to other relevant innovative solutions.

5.2 Future work

This work marks a starting point, as well as a very promising future to the prediction of treatment response to multi-stage treatments in chronic diseases. However, further studies will require the collection of a high number of medical records to give statistical rigor to the technical work done here. A deeper study will allow the following issues:

1. The training of predictive models considering the economic cost of the attributes. This cost could be expressed in terms of time (time for collecting surveys, interviews or medical tests) or money (medical tests).

2. The clustering of migraine patients according to their phenotype values for creating prediction models adjusted to their common characteristics.

3. Carrying out specific prospective studies that consider those medical attributes that, being indicated as relevant by the methodology presented in this Ph.D. Thesis, they have not been found relevant by doctors yet.

All these points will enhance the effort made throughout this research work and will contribute with a grain of sand to establish closer links between the medical and computer community.

Appendix A

Ethical consent

A.1 Description

Data were collected retrospectively from the review of the clinical histories of patients with chronic migraine and in previous or current treatment with BoNT-A with follow-up in the headache unit of two hospitals, the *Hospital Clínico Universitario* in Valladolid and *Hospital Universitario de La Princesa* in Madrid.

To this end, the approval of the ethics committee of both hospitals was obtained under the code documents **ANA-TOX-2015-1** and **PI-17-832** that are provided as complementary content. These documents authorize the investigation of the evaluation of the evolutionary characteristics of chronic migraine at different levels of care and its influence on the response to BoNT-A.

Both documents consider that the suitability requirements of the protocol in relation to the objectives of the study and the foreseeable risks and inconveniences are justified by the researchers.

The contents of both documents are presented below:

Hospital Universitario de La Princesa

Comunidad de Madrid

SaludMadrid

INFORME DEL COMITÉ ÉTICO DE INVESTIGACIÓN CLÍNICA

Dña. Dolores Ochoa Mazarro, vocal-secretaria en funciones del Comité Ético de Investigación Clínica del Hospital Universitario de la Princesa

Certifica

Que este Comité ha evaluado la propuesta del investigador principal la **Dra. Ana Beatriz Gago Veiga (Servicio de Neurología, Hospital Universitario de La Princesa)**, para que se realice el estudio EPA-OD con código de protocolo **ANA-TOX-2015-1**, titulado**: Evaluación de los factores evolutivos de la Migraña crónica en los diferentes niveles asistenciales y su influencia en la respuesta a Onabotulinumtoxin A (OnabotA)** y considera que:

Se cumplen los requisitos necesarios de idoneidad del protocolo en relación con los objetivos del estudio y están justificados los riesgos y molestias previsibles para el sujeto.

La capacidad del investigador y los medios disponibles son apropiados para llevar a cabo el estudio.

Son adecuados tanto el procedimiento previsto para obtener el consentimiento informado como la compensación prevista para los sujetos por daños que pudieran derivarse de su participación en el ensayo.

El alcance de las compensaciones económicas previstas no interfiere con el respeto a los postulados éticos.

Y que este Comité acepta que dicho estudio posautorización sea realizado por la **Dra. Ana Beatriz Gago Veiga (Servicio de Neurología)** como investigador principal, en el **Hospital Universitario de La Princesa.**

Lo que firmo en **Madrid** a **09** de **julio** de **2015**

Fdo: Dra. Dolores Ochoa Mazarro
Vocal-Secretaria en funciones del C.E.I.C.

CEIC Hospital Universitario La Princesa. C/ Diego de León 62, MADRID (28006). Tel.: 91 520 24 76/Fax: 91 520 25 60
ceic.hlpr@salud.madrid.org

HOSPITAL CLINICO UNIVERSITARIO
Avda. Ramón y Cajal, 3
Telf. 983 42 00 00
47003 - VALLADOLID

Sacyl
SANIDAD DE CASTILLA Y LEÓN

CONFORMIDAD DE LA DIRECCIÓN DEL CENTRO

Don Francisco Javier Vadillo Olmo,

Director Gerente del
Hospital Clínico Universitario de Valladolid,

CODIGO HOSPITAL	TITULO	INVESTIGADOR PRINCIPAL SERVICIO PROMOTOR
PI 17-832	EVALUACIÓN DE LAS VARIABLES MÁS PREDICTORAS DE LA RESPUESTA DEL PACIENTE CON MIGRAÑA AL TRATAMIENTO CON TOXINA BOTULÍNICA (ONABOTA)	I.P.: ANGEL L. GUERRERO PERAL NEUROLOGIA RECIBIDO: 24-10-2017

En relación con el citado Proyecto de Investigación, de acuerdo a la evaluación favorable a su realización en este Hospital por parte del CEIC Área Valladolid Este en su sesión del 26-10-2017.

Se Informa favorablemente la realización del dicho estudio en el Hospital Clínico Universitario de Valladolid,

Lo que firma en Valladolid, a 26 de octubre de 2017

EL DIRECTOR GERENTE

HOSPITAL CLÍNICO UNIVERSITARIO
DIRECCIÓN GERENCIA
SACYL - VALLADOLID

D. Francisco Javier Vadillo Olmo

Bibliography

A person who cites his source brings deliverance to the world.

Avot 6:5

Aarabi, A., Wallois, F. & Grebe, R. Automated neonatal seizure detection: a multistage classification system through feature selection based on relevance and redundancy analysis. *Clinical Neurophysiology*, vol. 117(2), pages 328–340, 2006.

Adams, A. M., Serrano, D., Buse, D. C., Reed, M. L., Marske, V., Fanning, K. M. & Lipton, R. B. The impact of chronic migraine: The Chronic Migraine Epidemiology and Outcomes (CaMEO) Study methods and baseline results. *Cephalalgia*, vol. 35(7), pages 563–578, 2015.

Alba, E. & Tomassini, M. Parallelism and evolutionary algorithms. *IEEE transactions on evolutionary computation*, vol. 6(5), pages 443–462, 2002.

Allan, F. & Wishart, J. A method of estimating the yield of a missing plot in field experimental work. *The Journal of Agricultural Science*, vol. 20(3), pages 399–406, 1930.

Aoki, K. Review of a proposed mechanism for the antinociceptive action of Botulinum toxin type A. *Neurotoxicology*, vol. 26(5), pages 785–793, 2005.

Armañanzas, R., Bielza, C., Chaudhuri, K. R., Martinez-Martin, P. & Larrañaga, P. Unveiling relevant non-motor parkinson's disease severity symptoms using a machine learning approach. *Artificial intelligence in medicine*, vol. 58(3), pages 195–202, 2013.

Armañanzas, R., Larrañaga, P. & Bielza, C. Ensemble transcript interaction networks: A case study on Alzheimer's disease. *Computer Methods and Programs in Biomedicine*, vol. 108(1), pages 442–450, 2012.

ARROLL, B., GOODYEAR-SMITH, F., CRENGLE, S., GUNN, J., KERSE, N., FISHMAN, T., FALLOON, K. & HATCHER, S. Validation of phq-2 and phq-9 to screen for major depression in the primary care population. *The Annals of Family Medicine*, vol. 8(4), pages 348–353, 2010.

AURORA, S., DODICK, D. W., TURKEL, C., DEGRYSE, R., SILBERSTEIN, S., LIPTON, R., DIENER, H. & BRIN, M. OnabotulinumtoxinA for treatment of chronic migraine: results from the double-blind, randomized, placebo-controlled phase of the PREEMPT 1 trial. *Cephalalgia*, vol. 30(7), pages 793–803, 2010.

AURORA, S. K., WINNER, P., FREEMAN, M. C., SPIERINGS, E. L., HEIRING, J. O., DEGRYSE, R. E., VANDENBURGH, A. M., NOLAN, M. E. & TURKEL, C. C. OnabotulinumtoxinA for treatment of chronic migraine: pooled analyses of the 56-week PREEMPT clinical program. *Headache*, vol. 51(9), pages 1358–1373, 2011.

BARBANTI, P. & EGEO, G. Pharmacological trials in migraine: it's time to reappraise where the headache is and what the pain is like. *Headache*, vol. 55(3), pages 439–441, 2015.

BARBANTI, P., EGEO, G., FOFI, L., AURILIA, C. & PIROSO, S. Rationale for use of Onabotulinum toxin A (Botox) in chronic migraine. *Neurological Sciences*, vol. 36(1), pages 29–32, 2015.

BARNARD, J. & MENG, X.-L. Applications of multiple imputation in medical studies: from aids to nhanes. *Statistical methods in medical research*, vol. 8(1), pages 17–36, 1999.

BEISKE, A., LOGE, J., RØNNINGEN, A. & SVENSSON, E. Pain in parkinsonâs disease: prevalence and characteristics. *PAIN®*, vol. 141(1-2), pages 173–177, 2009.

BELLAZZI, R. & ZUPAN, B. Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics*, vol. 77(2), pages 81–97, 2008.

BLOCKEEL, H. & DE RAEDT, L. Top-down induction of first-order logical decision trees. *Artificial intelligence*, vol. 101(1-2), pages 285–297, 1998.

BLUM, C. & ROLI, A. Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM computing surveys (CSUR)*, vol. 35(3), pages 268–308, 2003.

Borra, S. & Di Ciaccio, A. Measuring the prediction error. a comparison of cross-validation, bootstrap and covariance penalty methods. *Computational statistics & data analysis*, vol. 54(12), pages 2976–2989, 2010.

Breiman, L. et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, vol. 16(3), pages 199–231, 2001.

Burman, P. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, vol. 76(3), pages 503–514, 1989.

Buse, D., Manack, A., Serrano, D., Turkel, C. & Lipton, R. Sociodemographic and comorbidity profiles of chronic migraine and episodic migraine sufferers. *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 81(4), pages 428–432, 2010.

van Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L. & Jolani, S. Package âmiceâ. *Computer software]. Retrieved from: http://cran. r-project. org/web/packages/mice/mice. pdf*, 2015.

Cabitza, F., Ciucci, D. & Rasoini, R. A giant with feet of clay: on the validity of the data that feed machine learning in medicine. In *Organizing for the Digital World*, pages 121–136. Springer, 2019.

Cady, R. K., Schreiber, C. P., Porter, J. A., Blumenfeld, A. M. & Farmer, K. U. A multi-center double-blind pilot comparison of OnabotulinumtoxinA and Topiramate for the prophylactic treatment of chronic migraine. *Headache*, vol. 51(1), pages 21–32, 2011.

Cernuda-Morollón, E., Ramón, C., Larrosa, D., Alvarez, R., Riesco, N. & Pascual, J. Long-term experience with onabotulinumtoxinA in the treatment of chronic migraine: What happens after one year? *Cephalalgia*, vol. 35(10), pages 864–868, 2015.

Chakrabarti, S., Cox, E., Frank, E., Güting, R. H., Han, J., Jiang, X., Kamber, M., Lightstone, S. S., Nadeau, T. P., Neapolitan, R. E. et al. *Data mining: know it all*. Morgan Kaufmann, 2008.

Chen, T., Tang, K., Chen, G. & Yao, X. A large population size can be unhelpful in evolutionary algorithms. *Theoretical Computer Science*, vol. 436, pages 54–70, 2012.

CHORMUNGE, S. & JENA, S. Correlation based feature selection with clustering for high dimensional data. *Journal of Electrical Systems and Information Technology*, vol. 5(3), pages 542–549, 2018.

CIMINO, J. J. ET AL. Coding systems in health care. *Methods of information in medicine*, vol. 35, pages 273–284, 1996.

CIOS, K. J. & MOORE, G. W. Uniqueness of medical data mining. *Artificial intelligence in medicine*, vol. 26(1-2), pages 1–24, 2002.

CORNE, D. W., JERRAM, N. R., KNOWLES, J. D. & OATES, M. J. Pesa-ii: Region-based selection in evolutionary multiobjective optimization. In *Proceedings of the 3rd Annual Conference on Genetic and Evolutionary Computation*, pages 283–290. Morgan Kaufmann Publishers Inc., 2001.

DANS, P. E. Looking for answers in all the wrong places. *Annals of internal medicine*, vol. 119(8), pages 855–857, 1993.

DE VICENTE, J., LANCHARES, J. & HERMIDA, R. Adaptive FPGA placement by natural optimisation. In *Rapid System Prototyping, 2000. RSP 2000. Proceedings. 11th International Workshop on*, pages 188–193. IEEE, 2000.

DEB, K. & JAIN, H. An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part i: Solving problems with box constraints. *IEEE Trans. Evolutionary Computation*, vol. 18(4), pages 577–601, 2014.

DEB, K., PRATAP, A., AGARWAL, S. & MEYARIVAN, T. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, vol. 6(2), pages 182–197, 2002.

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39(1), pages 1–22, 1977.

DENNY, J. C., DRIEST, S. L., WEI, W.-Q. & RODEN, D. M. The influence of big (clinical) data and genomics on precision medicine and drug development. *Clinical Pharmacology & Therapeutics*, vol. 103(3), pages 409–418, 2018.

DESHPANDE, P., RASIN, A., BROWN, E., FURST, J., RAICU, D. S., MONTNER, S. M. & ARMATO, S. G. Big data integration case study for radiology data sources. In *2018 IEEE Life Sciences Conference (LSC)*, pages 195–198. IEEE, 2018.

Diener, H., Bussone, G., Oene, J. V., Lahaye, M., Schwalen, S. & Goadsby, P. Topiramate reduces headache days in chronic migraine: a randomized, double-blind, placebo-controlled study. *Cephalalgia*, vol. 27(7), pages 814–823, 2007.

Diener, H., Dodick, D. W., Aurora, S., Turkel, C., DeGryse, R., Lipton, R., Silberstein, S. & Brin, M. OnabotulinumtoxinA for treatment of chronic migraine: results from the double-blind, randomized, placebo-controlled phase of the PREEMPT 2 trial. *Cephalalgia*, vol. 30(7), pages 804–814, 2010.

Diener, H.-C., Dodick, D. W., Goadsby, P. J., Lipton, R. B., Olesen, J. & Silberstein, S. D. Chronic migraineâclassification, characteristics and treatment. *Nature Reviews Neurology*, vol. 8(3), pages 162–171, 2012.

Dodick, D. W., Mauskop, A., Elkind, A. H., DeGryse, R., Brin, M. F. & Silberstein, S. D. Botulinum toxin type A for the prophylaxis of chronic daily headache: Subgroup analysis of patients not receiving other prophylactic medications: A randomized double-blind, placebo-controlled study. *Headache*, vol. 45(4), pages 315–324, 2005.

Dodick, D. W., Turkel, C. C., DeGryse, R. E., Aurora, S. K., Silberstein, S. D., Lipton, R. B., Diener, H.-C. & Brin, M. F. OnabotulinumtoxinA for treatment of chronic migraine: pooled results from the double-blind, randomized, placebo-controlled phases of the PREEMPT clinical program. *Headache*, vol. 50(6), pages 921–936, 2010.

Domínguez, C., Pozo-Rosich, P., Torres-Ferrús, M., Hernández-Beltrán, N., Jurado-Cobo, C., González-Oria, C., Santos, S., Monzón, M., Latorre, G., Álvaro, L. et al. Onabotulinumtoxina in chronic migraine: predictors of response. a prospective multicentre descriptive study. *European journal of neurology*, vol. 25(2), pages 411–416, 2018.

Dowson, A. J. Assessing the impact of migraine. *Current medical research and opinion*, vol. 17(4), pages 298–309, 2001.

Durillo, J. J., Nebro, A. J., Luna, F. & Alba, E. A study of master-slave approaches to parallelize nsga-ii. In *2008 IEEE International Symposium on Parallel and Distributed Processing*, pages 1–8. IEEE, 2008.

Eisenstein, M. The power of petabytes. *Nature*, vol. 527(7576), page S2, 2015.

EL HASNAOUI, A., VRAY, M., RICHARD, A., NACHIT-OUINEKH, F., BOUREAU, F., GROUP, M. ET AL. Assessing the severity of migraine: development of the migsev scale. *Headache: The Journal of Head and Face Pain*, vol. 43(6), pages 628–635, 2003.

EROSS, E. J., GLADSTONE, J. P., LEWIS, S., ROGERS, R. & DODICK, D. W. Duration of migraine is a predictor for response to botulinum toxin type a. *Headache: The Journal of Head and Face Pain*, vol. 45(4), pages 308–314, 2005.

FERNANDES, J. A., LOZANO, J. A., INZA, I., IRIGOIEN, X., PÉREZ, A. & RODRÍGUEZ, J. D. Supervised pre-processing approaches in multiple class variables classification for fish recruitment forecasting. *Environmental modelling & software*, vol. 40, pages 245–254, 2013.

FINOCCHI, C. & STRADA, L. Sex-related differences in migraine. *Neurological Sciences*, vol. 35(1), pages 207–213, 2014.

FOX, C. Big data. In *Data Science for Transport*, pages 147–164. Springer, 2018.

FRAMPTON, J. E. OnabotulinumtoxinA (Botox). *Drugs*, vol. 72(6), pages 825–845, 2012.

FREITAG, F. G. Importance of botulinum toxin for prevention of migraine. *Expert review of neurotherapeutics*, vol. 10(3), pages 339–340, 2010.

GANDOMI, A. H., YANG, X.-S., TALATAHARI, S. & ALAVI, A. H. Metaheuristic algorithms in modeling and optimization. *Metaheuristic applications in structures and infrastructures*, pages 1–24, 2013.

GARCÍA, M. I. & TARRAGONA, S. Perturbación de los valores propios simples de matrices de polinomios dependientes diferenciablemente de parámetros. In *2nd Meeting on Linear Algebra Matrix analysis and applications*, pages 1–7. Servicio de publicaciones de la UPV, 2010.

GARCÍA, S. & HERRERA, F. An extension on"statistical comparisons of classifiers over multiple data sets"for all pairwise comparisons. *Journal of Machine Learning Research*, vol. 9(Dec), pages 2677–2694, 2008.

GASBARRINI, A., DE, A. L., FIORE, G., GAMBRIELLI, M., FRANCESCHI, F., OJETTI, V., TORRE, E., GASBARRINI, G., POLA, P. & GIACOVAZZO, M. Beneficial effects of helicobacter pylori eradication on migraine. *Hepatogastroenterology*, vol. 45(21), pages 765–770, 1998.

GAZERANI, P. & CAIRNS, B. E. Sex-specific pharmacotherapy for migraine: A narrative review. *Frontiers in Neuroscience*, vol. 14, page 222, 2020.

GENDREAU, M., POTVIN, J.-Y. ET AL. *Handbook of metaheuristics*, vol. 2. Springer, 2010.

GLOVER, F. W. & KOCHENBERGER, G. A. *Handbook of metaheuristics*, vol. 57. Springer Science & Business Media, 2006.

GRAZZI, L. & USAI, S. Onabotulinum toxin a (botox) for chronic migraine treatment: an italian experience. *Neurological Sciences*, vol. 36(1), pages 33–35, 2015.

GROGAN, P. M., ALVAREZ, M. V. & JONES, L. Headache direction and aura predict migraine responsiveness to rimabotulinumtoxin B. *Headache*, vol. 53(1), pages 126–136, 2013.

HADKA, D. Moea framework. `http://moeaframework.org/`, 2019.

HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P. & WITTEN, I. H. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, vol. 11(1), pages 10–18, 2009.

HALL, M. A. Correlation-based feature selection for machine learning. 1999.

HAN, J., PEI, J. & KAMBER, M. *Data mining: concepts and techniques*. Elsevier, 2011.

HAND, D. J. Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 181(3), pages 555–605, 2018.

HIGHAM, N. J. *Accuracy and stability of numerical algorithms*, vol. 80, page 54. Siam, 2002.

HOEHN, M. M., YAHR, M. D. ET AL. Parkinsonism: onset, progression, and mortality. *Neurology*, vol. 50(2), pages 318–318, 1998.

HUDDAR, V., DESIRAJU, B. K., RAJAN, V., BHATTACHARYA, S., ROY, S. & REDDY, C. K. Predicting complications in critical care using heterogeneous clinical data. *IEEE Access*, vol. 4, pages 7988–8001, 2016.

HÜHN, J. & HÜLLERMEIER, E. Furia: an algorithm for unordered fuzzy rule induction. *Data Mining and Knowledge Discovery*, vol. 19(3), pages 293–319, 2009.

IDREES, S. M., ALAM, M. A. & AGARWAL, P. A study of big data and its challenges. *International Journal of Information Technology*, pages 1–6, 2018.

IHS. The international classification of headache disorders, 3rd edition (beta version). *Cephalalgia*, vol. 33(9), pages 629–808, 2013. PMID: 23771276.

JAIN, A. K. & DUBES, R. C. Algorithms for clustering data. 1988.

JAKUBOWSKI, M., MCALLISTER, P. J., BAJWA, Z. H., WARD, T. N., SMITH, P. & BURSTEIN, R. Exploding vs. imploding headache in migraine prophylaxis with botulinum toxin A. *Pain*, vol. 125(3), pages 286–295, 2006.

JAPKOWICZ, N. & SHAH, M. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011.

JOLLIS, J. G., ANCUKIEWICZ, M., DELONG, E. R., PRYOR, D. B., MUHLBAIER, L. H. & MARK, D. B. Discordance of databases designed for claims payment versus clinical information systems: implications for outcomes research. *Annals of internal medicine*, vol. 119(8), pages 844–850, 1993.

KALTON, G. & KASPRZYK, D. Imputing for missing survey responses. In *Proceedings of the section on survey research methods, American Statistical Association*, vol. 22, page 31. American Statistical Association Cincinnati, 1982.

KANG, S. Personalized prediction of drug efficacy for diabetes treatment via patient-level sequential modeling with neural networks. *Artificial intelligence in medicine*, vol. 85, pages 1–6, 2018.

KASTHURIRATHNE, S. N., MAMLIN, B., KUMARA, H., GRIEVE, G. & BIONDICH, P. Enabling better interoperability for healthcare: lessons in developing a standards based application programing interface for electronic medical record systems. *Journal of medical systems*, vol. 39(11), page 182, 2015.

KERSTING, K. & MEYER, U. From big data to big artificial intelligence? 2018.

KIRKPATRICK, S., GELATT, C. D. & VECCHI, M. P. Optimization by simulated annealing. *science*, vol. 220(4598), pages 671–680, 1983.

KLARMAN, H. E. & ROSENTHAL, G. D. Cost effectiveness analysis applied to the treatment of chronic renal disease. *Medical care*, vol. 6(1), pages 48–54, 1968.

KOHAVI, R. ET AL. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, vol. 14, pages 1137–1145. Montreal, Canada, 1995.

KOSINSKI, M., BAYLISS, M., BJORNER, J., WARE, J., GARBER, W., BATENHORST, A., CADY, R., DAHLÖF, C., DOWSON, A. & TEPPER, S. A six-item short-form survey for measuring headache impact: The HIT-6. *Quality of Life Research*, vol. 12(8), pages 963–974, 2003.

KROENKE, K., SPITZER, R. L. & WILLIAMS, J. B. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, vol. 16(9), pages 606–613, 2001.

KUKKONEN, S. & LAMPINEN, J. Gde3: The third evolution step of generalized differential evolution. In *Evolutionary Computation, 2005. The 2005 IEEE Congress on*, vol. 1, pages 443–450. IEEE, 2005.

KUROSAKI, M., TANAKA, Y., NISHIDA, N., SAKAMOTO, N., ENOMOTO, N., HONDA, M., SUGIYAMA, M., MATSUURA, K., SUGAUCHI, F., ASAHINA, Y. ET AL. Pre-treatment prediction of response to pegylated-interferon plus ribavirin for chronic hepatitis c using genetic polymorphism in il28b and viral factors. *Journal of hepatology*, vol. 54(3), pages 439–448, 2011.

KURTZ, A. K. A research test of the rorschach test. *Personnel Psychology*, 1948.

LAINEZ, M., GIL, R., SALVADOR, A., PIERA, A. & LOPEZ, B. Unilateralism as a predictor of response in treatment of chronic headache patients with botulinum toxin. *Headache*, vol. 46(5), page 846:F12, 2006.

LAMBIN, P., VAN STIPHOUT, R. G., STARMANS, M. H., RIOS-VELAZQUEZ, E., NALBANTOV, G., AERTS, H. J., ROELOFS, E., VAN ELMPT, W., BOUTROS, P. C., GRANONE, P. ET AL. Predicting outcomes in radiation oncologyâmultifactorial decision support systems. *Nature reviews Clinical oncology*, vol. 10(1), page 27, 2013.

LAN, K., WANG, D.-T., FONG, S., LIU, L.-S., WONG, K. K. & DEY, N. A survey of data mining and deep learning in bioinformatics. *Journal of medical systems*, vol. 42(8), page 139, 2018.

LARRAÑAGA, P., ATIENZA, D., DIAZ-ROZO, J., OGBECHIE, A., BIELZA, C. & PUERTO-SANTANA, C. *Industrial Applications of Machine Learning*. Chapman and Hall/CRC Data Mining and Knowledge Discovery Series. CRC Press, 2018. ISBN 9780815356226.

LARRAÑAGA, P., CALVO, B., SANTANA, R., BIELZA, C., GALDIANO, J., INZA, I., LOZANO, J. A., ARMAÑANZAS, R., SANTAFÉ, G., PÉREZ, A. ET AL. Machine learning in bioinformatics. *Briefings in Bioinformatics*, pages 86–112, 2006.

LAVRAČ, N. Selected techniques for data mining in medicine. *Artificial intelligence in medicine*, vol. 16(1), pages 3–23, 1999.

LEIDNER, A. J., CHESSON, H. W., XU, F., WARD, J. W., SPRADLING, P. R. & HOLMBERG, S. D. Cost-effectiveness of hepatitis c treatment for patients in early stages of liver disease. *Hepatology*, vol. 61(6), pages 1860–1869, 2015.

LEVENTHAL, B. *Predictive Analytics for Marketers: Using Data Mining for Business Advantage*. Kogan Page Publishers, 2018.

LEWIS, P. The characteristic selection problem in recognition systems. *IRE Transactions on information theory*, vol. 8(2), pages 171–178, 1962.

LIAO, S.-C. & LEE, I.-N. Appropriate medical data categorization for data mining classification techniques. *Medical informatics and the Internet in medicine*, vol. 27(1), pages 59–67, 2002.

LIN, J.-H. & HAUG, P. J. Exploiting missing clinical data in bayesian network modeling for predicting medical problems. *Journal of biomedical informatics*, vol. 41(1), pages 1–14, 2008.

LIN, K.-H., CHEN, S.-P., FUH, J.-L., WANG, Y.-F. & WANG, S.-J. Efficacy, safety, and predictors of response to botulinum toxin type a in refractory chronic migraine: A retrospective study. *Journal of the Chinese Medical Association*, vol. 77(1), pages 10–15, 2014.

LIN, S.-W., YING, K.-C., LEE, C.-Y. & LEE, Z.-J. An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection. *Applied Soft Computing*, vol. 12(10), pages 3285–3290, 2012.

LINDE, M., GUSTAVSSON, A., STOVNER, L., STEINER, T., BARRÉ, J., KATSARAVA, Z., LAINEZ, J., LAMPL, C., LANTÉRI-MINET, M., RASTENYTE, D. ET AL. The cost of headache disorders in europe: the eu-

rolight project. *European journal of neurology*, vol. 19(5), pages 703–711, 2012.

LIPTON, R., VARON, S., GROSBERG, B., MCALLISTER, P., FREITAG, F., AURORA, S., DODICK, D. W., SILBERSTEIN, S., DIENER, H., DEGRYSE, R. ET AL. OnabotulinumtoxinA improves quality of life and reduces impact of chronic migraine. *Neurology*, vol. 77(15), pages 1465–1472, 2011.

LIPTON, R. B. & SILBERSTEIN, S. D. Why study the comorbidity of migraine? *Neurology*, vol. 44(10 supl.(7)), pages S4–S5, 1994.

LOVATI, C. & GIANI, L. Action mechanisms of Onabotulinum toxin-A: hints for selection of eligible patients. *Neurological Sciences*, vol. 38(1), pages 131–140, 2017.

MADJAROV, G., KOCEV, D., GJORGJEVIKJ, D. & DŽEROSKI, S. An extensive experimental comparison of methods for multi-label learning. *Pattern recognition*, vol. 45(9), pages 3084–3104, 2012.

MARTÍ, R., PARDALOS, P. M. & RESENDE, M. G. *Handbook of Heuristics*. Springer, 2018.

MARTÍNEZ-MARTÍN, P., FORJAZ, M. J., CUBO, E., FRADES, B., DE PEDRO CUESTA, J. & MEMBERS, E. P. Global versus factor-related impression of severity in parkinson's disease: a new clinimetric index (cisi-pd). *Movement Disorders*, vol. 21(2), pages 208–214, 2006.

MARX, V. Biology: The big challenges of big data. 2013.

MATHEW, N. T. & JAFFRI, S. F. A. A double-blind comparison of OnabotulinumtoxinA (Botox) and Topiramate (Topamax) for the prophylactic treatment of chronic migraine: a pilot study. *Headache*, vol. 49(10), pages 1466–1478, 2009.

MATHEW, N. T., KAILASAM, J. & MEADORS, L. Botulinum toxin type a for the treatment of nummular headache: four case studies. *Headache: The Journal of Head and Face Pain*, vol. 48(3), pages 442–447, 2008a.

MATHEW, N. T., KAILASAM, J. & MEADORS, L. Predictors of response to botulinum toxin type A (BoNTA) in chronic daily headache. *Headache*, vol. 48(2), pages 194–200, 2008b.

MCCARTHY, P. J. The use of balanced half-sample replication in cross-validation studies. *Journal of the American Statistical Association*, vol. 71(355), pages 596–604, 1976.

MILLER, G. *WordNet: An electronic lexical database*. MIT press, 1998.

MOLINA, L. C., BELANCHE, L. & NEBOT, À. Feature selection algorithms: A survey and experimental evaluation. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 306–313. IEEE, 2002.

MUHLENBACH, F. & RAKOTOMALALA, R. Discretization for continuous attributes. In *Encyclopedia of Data Warehousing and Mining*, pages 397–402. IGI Global, 2005.

NAKATSU, R. T. An evaluation of four resampling methods used in machine learning classification. *IEEE Intelligent Systems*, 2020.

NATOLI, J., MANACK, A., DEAN, B., BUTLER, Q., TURKEL, C., STOVNER, L. & LIPTON, R. Global prevalence of chronic migraine: a systematic review. *Cephalalgia*, vol. 30(5), pages 599–609, 2010.

NEBRO, A. J., DURILLO, J. J., GARCIA-NIETO, J., COELLO, C. C., LUNA, F. & ALBA, E. Smpso: A new pso-based metaheuristic for multi-objective optimization. In *Computational intelligence in miulti-criteria decision-making, 2009. mcdm'09. ieee symposium on*, pages 66–73. IEEE, 2009.

OLESEN, J., BURSTEIN, R., ASHINA, M. & TFELT-HANSEN, P. Origin of pain in migraine: evidence for peripheral sensitisation. *The Lancet Neurology*, vol. 8(7), pages 679–690, 2009.

ORNELLO, R., LISI, S. V., DEGAN, D., TISEO, C., PISTOIA, F., CAROLEI, A. & SACCO, S. O059. predictors of response to botulinum toxin for the treatment of chronic migraine: data from a Headache Center. *The Journal of Headache and Pain*, vol. 16(S1), page A179, 2015.

OTAEGUI, D., BARANZINI, S. E., ARMAÑANZAS, R., CALVO, B., MUÑOZ-CULLA, M., KHANKHANIAN, P., INZA, I., LOZANO, J. A., CASTILLO-TRIVIÑO, T., ASENSIO, A. ET AL. Differential micro RNA expression in PBMC from multiple sclerosis patients. *PloS One*, vol. 4(7), page e6309, 2009.

OTERINO, A., RAMÓN, C. & PASCUAL, J. Experience with onabotulinumtoxinA (Botox) in chronic refractory migraine: focus on severe attacks. *The Journal of Headache and Pain*, vol. 12(2), pages 235–238, 2011.

PAGÁN, J., ORBE, D., IRENE, M., GAGO, A., SOBRADO, M., RISCO-MARTÍN, J. L., MORA, J. V., MOYA, J. M. & AYALA, J. L. Robust

and accurate modeling approaches for migraine per-patient prediction from ambulatory data. *Sensors*, vol. 15(7), pages 15419–15442, 2015.

PAGOLA, I., ESTEVE-BELLOCH, P., PALMA, J., LUQUIN, M., RIVEROL, M., MARTINEZ-VILA, E. & IRIMIA, P. Predictive factors of the response to treatment with onabotulinumtoxinA in refractory migraine. *Revista de Neurologia*, vol. 58(6), pages 241–246, 2014. PMID:9925225.

PALMERINI, L., ROCCHI, L., MELLONE, S., VALZANIA, F. & CHIARI, L. Feature selection for accelerometer-based posture analysis in Parkinson's disease. *IEEE Transactions on Information Technology in Biomedicine*, vol. 15(3), pages 481–490, 2011.

PARRALES, F., DEL BARRIO, A. A. & AYALA, J. L. Estudio sobre la paralelización de modelos MOEAs de predicción terapéutica con toxina botulínica tipo A en migraña. In *Actas de las Jornadas SARTECO 2019*, pages 96–101. Universidad de Extremadura, Servicio de Publicaciones, 2019a.

PARRALES, F., DEL BARRIO, A. A. & AYALA, J. L. A study on the parallelization of moeas to predict the patient's response to the onabotulinumtoxina treatment. In *Proceedings of the Summer Simulation Multi-Conference*, page 12. Society for Computer Simulation International, 2019b.

PARRALES, F., DEL BARRIO, A. A., GAGO, A. B., GALLEGO, M. M., RUIZ, M., PERAL, A. G., DZEROSKI, S. & AYALA, J. L. SMURF: Systematic Methodology for Unveiling Relevant Factors in retrospective data on chronic disease treatments. *IEEE Access*, pages 1–1, 2019c. ISSN 2169-3536.

PARRALES, F., DEL BARRIO GARCÍA, A., GALLEGO, M., GAGO, A. V., RUIZ, M., GUERRERO, A. P., AYALA, J. ET AL. Prediction of patient's response to OnabotulinumtoxinA treatment for migraine. *Heliyon*, vol. 5(2), pages e01043–e01043, 2019d.

PARRALES BRAVO, F., DEL BARRIO GARCÍA, A., GALLEGO DE LA SACRISTANA, M., LÓPEZ MANZANARES, L., VIVANCOS, J. & AYALA RODRIGO, J. Support system to improve reading activity in parkinson's disease and essential tremor patients. *Sensors*, vol. 17(5), page 1006, 2017.

PELZER, N., LOUTER, M. A., VAN ZWET, E. W., NYHOLT, D. R., FERRARI, M. D., VAN DEN MAAGDENBERG, A. M., HAAN, J. & TERWINDT,

G. M. Linking migraine frequency with family history of migraine. *Cephalalgia*, vol. 39(2), pages 229–236, 2019.

PETERKOVA, A., NEMETH, M. & BOHM, A. Computing missing values using neural networks in medical field. In *2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)*, pages 000151–000156. IEEE, 2018.

PRATHER, J. C., LOBACH, D. F., GOODWIN, L. K., HALES, J. W., HAGE, M. L. & HAMMOND, W. E. Medical data mining: knowledge discovery in a clinical data warehouse. In *Proceedings of the AMIA annual fall symposium*, page 101. American Medical Informatics Association, 1997.

READ, J., REUTEMANN, P., PFAHRINGER, B. & HOLMES, G. Meka: a multi-label/multi-target extension to weka. *The Journal of Machine Learning Research*, vol. 17(1), pages 667–671, 2016.

REIN, D. B., WITTENBORN, J. S., SMITH, B. D., LIFFMANN, D. K. & WARD, J. W. The cost-effectiveness, health benefits, and financial costs of new antiviral treatments for hepatitis c virus. *Clinical infectious diseases*, vol. 61(2), pages 157–168, 2015.

RISCO, J. L. Hero, a collection of optimization algorithms developed in java. `https://github.com/jlrisco/hero`, 2016. Accessed: 2019-06-18.

RUBIN, D. B. Inference and missing data. *Biometrika*, vol. 63(3), pages 581–592, 1976.

RUBIN, D. B. Multiple imputation after 18+ years. *Journal of the American statistical Association*, vol. 91(434), pages 473–489, 1996.

RUBIN, D. B. *Multiple imputation for nonresponse in surveys*, vol. 81. John Wiley & Sons, 2004.

SACRISTÁN, J. A. & DILLA, T. No big data without small data: learning health care systems begin and end with the individual patient. *Journal of evaluation in clinical practice*, vol. 21(6), pages 1014–1017, 2015.

SAEYS, Y., INZA, I. & LARRAÑAGA, P. A review of feature selection techniques in bioinformatics. *Bioinformatics*, vol. 23(19), pages 2507–2517, 2007.

SAJOBI, T. T., AMOOZEGAR, F., WANG, M., WIEBE, N., FIEST, K. M., PATTEN, S. B. & JETTE, N. Global assessment of migraine severity

measure: preliminary evidence of construct validity. *BMC neurology*, vol. 19(1), page 53, 2019.

SANDRINI, G., PERROTTA, A., TASSORELLI, C., TORELLI, P., BRIGHINA, F., SANCES, G. & NAPPI, G. Botulinum toxin type-A in the prophylactic treatment of medication-overuse headache: a multicenter, double-blind, randomized, placebo-controlled, parallel group study. *The Journal of Headache and Pain*, vol. 12(4), pages 427–433, 2011.

SANTOS, R. L. Effects of imputation on complex statistics. *Survey Research Center, Institute for Social Research, University of Michigan*, 1981.

SARTAKHTI, J. S., ZANGOOEI, M. H. & MOZAFARI, K. Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA). *Computer Methods and Programs in Biomedicine*, vol. 108(2), pages 570–579, 2012.

SCHIEBER, S. J. A comparison of three alternative techniques for allocating unreported social secrity income on the survey thelow income aged and disabled. In *American Statistical Association, Proceedings of the Section on Survey Research Methods*, pages 212–218. 1978.

SCHOENEN, J., JACQUY, J. & LENAERTS, M. Effectiveness of high-dose riboflavin in migraine prophylaxis a randomized controlled trial. *Neurology*, vol. 50(2), pages 466–470, 1998.

SCHULMAN, E. A., LAKE, A. E., GOADSBY, P. J., PETERLIN, B. L., SIEGEL, S. E., MARKLEY, H. G. & LIPTON, R. B. Defining refractory migraine and refractory chronic migraine: proposed criteria from the Refractory Headache Special Interest Section of the American Headache Society. *Headache*, vol. 48(6), pages 778–782, 2008.

SCHWEDT, T., LIPTON, R., ALAM, A., DODICK, D., MUNJAL, S., FANNING, K., REED, M. & BUSE, D. Impact of migraine headache day frequency on associated health: Results from the 2017 migraine in america symptoms and treatment (mast) study (p3. 10-001). 2019.

ŞEN, B., PEKER, M., ÇAVUŞOĞLU, A. & ÇELEBI, F. V. A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms. *Journal of Medical Systems*, vol. 38(3), page 18, 2014.

SHANNON, C. E. A mathematical theory of communication. *Bell system technical journal*, vol. 27(3), pages 379–423, 1948.

SHARMA, M. C., SHARMA, S. & BHADORIYA, K. S. QSAR analyses and pharmacophore studies of tetrazole and sulfonamide analogs of imidazo [4, 5-b] pyridine using simulated annealing based feature selection. *Journal of Saudi Chemical Society*, vol. 10, page 1016, 2012.

SHILASKAR, S. & GHATOL, A. Feature selection for medical diagnosis: Evaluation for cardiovascular diseases. *Expert Systems with Applications*, vol. 40(10), pages 4146–4153, 2013.

SILBERSTEIN, S., LIPTON, R., DODICK, D., FREITAG, F., RAMADAN, N., MATHEW, N., BRANDES, J., BIGAL, M., SAPER, J., ASCHER, S. ET AL. Topiramate chronic migraine study group efficacy and safety of topiramate for the treatment of chronic migraine: a randomized, double-blind, placebo-controlled trial. *Headache*, vol. 47(2), pages 170–180, 2007.

SILBERSTEIN, S. D., DODICK, D. W., AURORA, S. K., DIENER, H.-C., DEGRYSE, R. E., LIPTON, R. B. & TURKEL, C. C. Per cent of patients with chronic migraine who responded per onabotulinumtoxinA treatment cycle: PREEMPT. *J Neurol Neurosurg Psychiatry*, vol. 86(9), pages 996–1001, 2015.

SILBERSTEIN, S. D., WINNER, P. K. & CHMIEL, J. J. Migraine preventive medication reduces resource utilization. *Headache*, vol. 43(3), pages 171–178, 2003.

SKIENA, S. S. *The algorithm design manual: Text*, vol. 1. Springer Science & Business Media, 1998.

SMALL, S. A., KENT, K., PIERCE, A., LEUNG, C., KANG, M. S., OKADA, H., HONIG, L., VONSATTEL, J.-P. & KIM, T.-W. Model-guided microarray implicates the retromer complex in Alzheimer's disease. *Annals of Neurology*, vol. 58(6), pages 909–919, 2005.

SÖRENSEN, K. Metaheuristicsâthe metaphor exposed. *International Transactions in Operational Research*, vol. 22(1), pages 3–18, 2015.

STEWART, J., SPRIVULIS, P. & DWIVEDI, G. Artificial intelligence and machine learning in emergency medicine. *Emergency Medicine Australasia*, vol. 30(6), pages 870–874, 2018.

STOJANOVIĆ, I., BRAJEVIĆ, I., STANIMIROVIĆ, P. S., KAZAKOVTSEV, L. A. & ZDRAVEV, Z. Application of heuristic and metaheuristic algorithms in solving constrained weber problem with feasible region bounded by arcs. *Mathematical Problems in Engineering*, vol. 2017, 2017.

Stone, A. & Bornhorst, J. Chapter 6 - an introduction to personalized medicine. In *Therapeutic Drug Monitoring* (editado por A. Dasgupta), pages 121 – 142. Academic Press, Boston, 2012. ISBN 978-0-12-385467-4.

Stovner, L. J., Nichols, E., Steiner, T. J., Abd-Allah, F., Abdelalim, A., Al-Raddadi, R. M., Ansha, M. G., Barac, A., Bensenor, I. M., Doan, L. P. et al. Global, regional, and national burden of migraine and tension-type headache, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet Neurology*, vol. 17(11), pages 954–976, 2018.

Struyf, J., Aho, T., Fromont, E., Gjorgjioski, V., Kocev, D., Schietgat, L., Slavkov, I., Vens, C. & Zenko, B. Clus, a framework for predictive clustering. `http://clus.sourceforge.net/doku.php`, 1999. Accessed: 2019-06-18.

Suhrcke, M., Nugent, R. A., Stuckler, D. & Rocco, L. Chronic disease: an economic perspective. 2006.

Szűcs, J. & Balázs, P. An improved simulated annealing approach for reconstructing binary images with fixed number of strips. In *International Conference on Image Analysis and Recognition*, pages 174–185. Springer, 2019.

Tang, H., Cui, F., Liu, L. & Li, Y. Predictive models for tyrosinase inhibitors: Challenges from heterogeneous activity data determined by different experimental protocols. *Computational biology and chemistry*, vol. 73, pages 79–84, 2018.

Tashkandi, A., Wiese, I. & Wiese, L. Efficient in-database patient similarity analysis for personalized medical decision support systems. *Big data research*, vol. 13, pages 52–64, 2018.

Thompson, D., Jenkinson, C., Roebuck, A., Lewin, R., Boyle, R. & Chandola, T. Development and validation of a short measure of health status for individuals with acute myocardial infarction: the myocardial infarction dimensional assessment scale (midas). *Quality of life research*, vol. 11(6), pages 535–543, 2002.

Tsoumakas, G., Katakis, I. & Vlahavas, I. Effective and efficient multilabel classification in domains with large number of labels. In *Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMDâ08)*, vol. 21, pages 53–59. sn, 2008.

TSOUMAKAS, G., KATAKIS, I. & VLAHAVAS, I. Mining multi-label data. In *Data mining and knowledge discovery handbook*, pages 667–685. Springer, 2009.

VAN BUUREN, S. *Flexible imputation of missing data*. Chapman and Hall/CRC, 2018.

VILLOSLADA, P., STEINMAN, L. & BARANZINI, S. E. Systems biology and its application to the understanding of neurological diseases. *Annals of Neurology*, vol. 65(2), pages 124–139, 2009.

WAEGEMAN, W., DEMBCZYŃSKI, K. & HÜLLERMEIER, E. Multi-target prediction: a unifying view on problems and methods. *Data Mining and Knowledge Discovery*, vol. 33(2), pages 293–324, 2019.

WILKINSON, T. M., DONALDSON, G. C., HURST, J. R., SEEMUNGAL, T. A. & WEDZICHA, J. A. Early therapy improves outcomes of exacerbations of chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, vol. 169(12), pages 1298–1303, 2004.

WIRTH, R. & HIPP, J. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pages 29–39. Citeseer, 2000.

WITTEN, I. H., FRANK, E., HALL, M. A. & PAL, C. J. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

WU, X., KUMAR, V., QUINLAN, J. R., GHOSH, J., YANG, Q., MOTODA, H., MCLACHLAN, G. J., NG, A., LIU, B., PHILIP, S. Y. ET AL. Top 10 algorithms in data mining. *Knowledge and Information Systems*, vol. 14(1), pages 1–37, 2008.

WYATT, A. A. Prognostic models: clinically useful or quickly forgotten? *Br Med J*, vol. 311, pages 539–541, 1995.

YANG, M., RENDAS-BAUM, R., VARON, S. F. & KOSINSKI, M. Validation of the headache impact test (hit-6â) across episodic and chronic migraine. *Cephalalgia*, vol. 31(3), pages 357–367, 2011.

YATES, F. The analysis of replicated experiments when the field results are incomplete. *Empire Journal of Experimental Agriculture*, vol. 1(2), pages 129–142, 1933.

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. *Journal of medical systems*, vol. 36(4), pages 2431–2448, 2012.

Yosipof, A., Guedes, R. C. & García-Sosa, A. T. Data mining and machine learning models for predicting drug likeness and their disease or organ category. *Frontiers in Chemistry*, vol. 6, 2018.

Yu, Z., Niu, Z., Tang, W. & Wu, Q. Deep learning for daily peak load forecastingâa novel gated recurrent neural network combining dynamic time warping. *IEEE Access*, 2019.

Zhang, M.-L. & Zhou, Z.-H. Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, vol. 40(7), pages 2038–2048, 2007.

Zigmond, A. S. & Snaith, R. P. The hospital anxiety and depression scale. *Acta psychiatrica scandinavica*, vol. 67(6), pages 361–370, 1983.

Zitzler, E., Deb, K. & Thiele, L. Comparison of multiobjective evolutionary algorithms: Empirical results. *Evolutionary computation*, vol. 8(2), pages 173–195, 2000.

Zitzler, E., Laumanns, M. & Thiele, L. Spea2: Improving the strength pareto evolutionary algorithm. *TIK-report*, vol. 103, 2001.

Zitzler, E. & Thiele, L. Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. *IEEE transactions on Evolutionary Computation*, vol. 3(4), pages 257–271, 1999.

–¿Qué te parece desto, Sancho? – Dijo Don Quijote –
Bien podrán los encantadores quitarme la ventura,
pero el esfuerzo y el ánimo, será imposible.

Segunda parte del Ingenioso Caballero
Don Quijote de la Mancha
Miguel de Cervantes

–Buena está – dijo Sancho –; fírmela vuestra merced.
–No es menester firmarla – dijo Don Quijote–,
sino solamente poner mi rúbrica.

Primera parte del Ingenioso Caballero
Don Quijote de la Mancha
Miguel de Cervantes