UNIVERSIDADE D
COIMBRA

Maria João Coelho de Sousa

On the Explainability of Multiple Sclerosis Disease Progression Models

Thesis submitted to the Faculty of Science and Technology of the University of Coimbra for the degree of Master in Biomedical Engineering with specialization in Clinical Informatics and Bioinformatics, supervised by Prof. Dr. César Alexandre Domingues Teixeira and MSc Mauro Filipe da Silva Pinto.

November of 2021

1 2 9 0

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D COIMBRA

Maria João Coelho de Sousa

On the Explainability of Multiple Sclerosis Disease Progression Models

Dissertion presented to the University of Coimbra in order to complete the necessary requirements to obtain the Master's degree in Biomedical Engineering.

Supervisors:
Prof. Dr. César Alexandre Domingues Teixeira (CISUC)
MSc Mauro Filipe da Silva Pinto (CISUC)

Coimbra, 2021

This work was developed in collaboration with:

CISUC - Center for Informatics and Systems of the University of Coimbra

CHUC - Coimbra Hospital and University Centre

Agradecimentos

Antes de mais, gostaria de agradecer aos meus orientadores por todo o apoio e orientação dados ao longo da realização deste projeto. Ao Professor Doutor César Teixeira por todo o conhecimento e disponibilidade que sempre demonstrou no decorrer do ano letivo. Ao Mauro, por estar sempre disponível, apesar de todos os seus projectos paralelos. Por a toda a motivação que me deu, pela constante confiança e otimismo e por todas as aprendizagens que me proporcionou.

Quero também agradecer ao Centro Hospitalar Universitário de Coimbra, nomeadamente à Doutora Sónia Batista e ao serviço de neurologia do hospital por disponibilizarem os dados clínicos necessários para a realização deste trabalho.

Aos meus colegas de laboratório, pela ajuda e disponibilidade oferecida na concretização do projeto.

A todos os meus amigos que fizeram parte desta aventura. À Rocha, Rita, Rafa, e 12 , à minha família académica, Marli e Beatriz, e à Sofia e Ana Maria, que são muito mais do que simples colegas de casa. Todas foram o meu suporte em Coimbra e continuarão a sê-lo por onde o futuro me levar, 99.9% de certeza. Obrigada pelos magníficos momentos que me deram, que fizeram de Coimbra um capítulo tão bonito da minha vida. Obrigada também por me darem na cabeça quando era necessário. À Rocha, por me aturar como parceira de trabalhos todos estes anos, sei que nem sempre foi fácil! À Daniela, Carolina e Zé, amigos de longa data e maiores confidentes, que estiveram sempre presentes, mesmo quando a perda de contacto parecia inevitável.

Por último quero agradecer às pessoas que fizeram de mim um ser humano decente, toda a minha família. Não há forma de exprimir toda a gratidão que tenho a vocês, pais. Obrigada por me apoiarem sempre, sem exigências, apesar de todo os sacrifícios que passaram para eu e a mana não passarmos dificuldades. Deram-nos muito muito mais do que aquilo que precisamos. Dizer que temos os melhores pais do mundo é um grande cliché mas não deve estar longe de ser verdade. À minha irmã, a minha

primeira professora, obrigada por toda a amizade e ajuda. Apesar de longe, sei que posso contar sempre com os teus conselhos e palavras de conforto. Não vou agradecer pela paciência porque ambas sabemos que isso é mais uma característica minha... O papel de uma irmã é muito maior que o de um excelente amigo e nem sempre é fácil mas tu consegues sempre desempenhar essa responsabilidade como ninguém.

Vocês os três são o meu pilar. Muito muito obrigada pela força que me dão constantemente. Essa força move montanhas e conclui cursos.

"There's a million things I haven't done. Just you wait."

LIN MANUEL MIRANDA, *Hamilton*

Resumo

Esclerose múltipla é a doença neurológica mais predominante em jovens adultos globalmente. A complexidade e heterogeneidade da sua progressão, e o consequente desafio de um prognóstico adequado, levaram à criação de modelos de Machine learning (ML), capazes de fornecer um prognóstico auxiliado por computador. No entanto, os modelos desenvolvidos podem não oferecer garantias de confiança ou segurança que promovam a sua aplicação num contexto clínico.

Explicabilidade é um conceito recente que visa criar explicações compreensíveis sobre os modelos de ML, para ajudar a mitigar a desconfiança associada à falta de informação sobre a lógica dos mesmos.

O objetivo deste projeto é compreender se os modelos desenvolvidos por Pinto et al. [81], que preveem a progressão da doença, podem ser aplicados em ambiente clínico, e que tipo de explicações adicionais podem ajudar a atingir esse objetivo.

Nesta dissertação de mestrado, vários métodos de explicabilidade foram desenvolvidos para gerar explicações humanamente compreensíveis sobre os modelos de previsão. As explicações continham informações gerais sobre os modelos, e o estudo de previsões de doentes específicos. Os resultados foram avaliados qualitativamente, com base na teoria fundamentada, através de entrevistas com cientistas de dados.

As explicações mostraram que, geralmente, a escala de quantificação da condição neurológica (EDSS) e os *Scores* de alguns sistemas funcionais, nomeadamente os sistemas *piramidal, cerebelar* e *mental*, tiveram maior relevância nas previsões. A análise dos cientistas de dados sugeriu que os métodos de explicabilidade mais adequados para apoiar os modelos de previsão eram o poder preditivo de Pinto et al. [81], a *permutation feature importance*, os *partial dependence plots (PDPs)*, e os valores de *Shapley*.

Keywords: Esclerose Múltipla, Progressão, Previsão, Explicabilidade, Machine Learning

Abstract

Multiple Sclerosis (MS) is the neurological disease most prevalent in young adults worldwide. The complexity and heterogeneity of its progression and consequent challenge of an adequate prognosis have led to the creation of ML models capable of providing a computer-aided prediction. However, the developed models may not offer trust or safety guarantees that promote their application in a clinical context.

Explainability is a recent field of study that aims to create human-comprehensible explanations about ML models to help mitigate the doubts associated with the lack of information about the models' logic.

The goal of this project is to understand if the models developed by Pinto et al. [81], that predict the progression of the disease, are able to be applied in a clinical environment, and what type of additional explanations can help to achieve that objective.

In this master thesis, several explainability methods were developed to produce human-comprehensible explanations about the prediction models. The explanations contained general information about the framework and analysis of specific patient predictions. Then, these results were qualitatively evaluated through interviews with data scientists that were analysed based on the Grounded Theory (GT).

The explanations showed that, in general, the Expanded disability status scale (EDSS) and the scores of some Functional System (FS), namely the pyramidal, cerebellar, and mental systems, had the most predictive relevance. The analysis by the data scientists suggested that the explainability methods most suited to support the prediction models were the predictive power by Pinto et al. [81], the permutation feature importance, the PDP, and the Shapley Values.

Keywords: Multiple Sclerosis, Progression, Prediction, Explainability, Machine Learning

Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial Intelligence. 15
ALE Accumulated Local Effects. 26
ALS Amyotrophic Lateral Sclerosis. 49
ANN Artificial Neural Network. 49
AUC Area Under the Curve. 20, 38, 39, 41, 67

CHUC Centro Hospitalar e Universitário de Coimbra. 41
CIS Clinically Isolated Syndrome. 8, 11, 12, 14, 38, 39, 40
CNN Convolutional Neural Network. 38, 39, 46, 49
CNS Central Nervous System. 2, 5, 7, 8, 11, 12, 39, 44
CSF Cerebrospinal Fluid. 7, 8, 43, 57
CV Cross Validation. 18, 41

DCNN Deep Convolutional Neural Network. 47
DIS dissemination in space. 7
DIT dissemination in time. 7
DNA Deoxyribonucleic Acid. 6
DNN Deep Neural Network. 49

EBV Epstein-barr Virus. 6
EDSS Expanded disability status scale. ix, xi, xiii, xvii, xviii, xix, 8, 9, 10, 11, 38, 39, 40, 42, 44, 45, 47, 61, 62, 63, 64, 65, 67, 68, 69, 70, 71, 72, 73, 74, 75, 83, 90, 91, 92, 93
EHR Electronic Health Record. 49

FDA Food and Drug Administration. 14
FLAIR Fluid-attenuated Inversion Recovery. 46, 47
FN False Negatives. 19
FP False Positives. 19, 20
FPR False Positive Ratio. 20
FS Functional System. xi, 8, 9, 10, 39, 40, 62, 63, 83, 93

GDPR General Data Protection Regulation. 2, 21
GT Grounded Theory. xi, xiv, xvii, 5, 34, 35, 59, 93

ICE Individual Conditional Expectation. xvii, 26, 27, 54
IM Intramuscular. 14
IV Intravenous. 14

KNN K-nearest neighbors. 18, 22, 29, 39, 40

LASSO Least Absolute Shrinkage and Selection Operator. 17, 41, 54
LIME Local Interpretable Model-Agnostic Explanations. xviii, xix, 28, 29, 48, 49, 56, 68, 69, 70, 77, 83, 85
LRP Layer-wise Relevance Propagation. 46, 47
LSTM Long Short-term Memory. 49

MAR Missing at Random. 16
MCAR Missing Completely at Random. 16
MCI Mild Cognitive Impairment. 49
MEP Motor Evoked Potential. 38
ML Machine learning. ix, xi, xvii, xxi, 1, 2, 3, 4, 5, 13, 15, 16, 18, 19, 21, 22, 23, 24, 25, 30, 31, 32, 37, 38, 39, 40, 42, 48, 50, 52, 58, 82, 84, 88, 92
MLP Multi-layer Perceptron. 49
MMD Maximum Mean Discrepancy. 31
MNAR Missing Not at Random. 16
MRI Magnetic resonance imaging. xvii, 7, 8, 13, 38, 39, 43, 46, 47, 49, 57
MS Multiple Sclerosis. xi, xiv, xvii, xviii, xxi, 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 16, 20, 33, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 50, 51, 52, 57, 58, 61, 66, 68, 72, 76, 78, 79, 81, 83, 84, 85, 86, 88, 89, 90, 91, 92, 93, 94

NN Neural Network. 39
NPv Negative predictive value. 39

PDP Partial Dependence Plot. ix, xi, xvii, xviii, 25, 26, 27, 33, 49, 53, 54, 64, 65, 66, 77, 83, 87, 90, 93
PET Positron Emission Tomography. 49
PP Primary Progressive. xvii, 11, 12, 13, 14, 39, 41
PPv Positive predictive value. 39

ROC Receiver Operating Characteristic. xvii, 20, 21
RR Relapse-remitting. xvii, 2, 3, 11, 12, 13, 14, 39, 40, 41

SC Subcutaneous. 14
SHAP SHapley Additive exPlanations. 48, 49
SP Secondary Progressive. xvii, 2, 3, 12, 13, 14, 39, 40, 41, 51
SPECT Single-photon Emission Computed Tomography. 49
SVM Support Vector Machine. xxi, 18, 38, 39, 41, 49, 55, 67, 78, 87

TCN Temporal Convolutional Network. 49
TN True Negatives. 19, 20
TP True Positives. 19, 20
TPR True Positive Ratio. 20

XGBoost Extreme Gradient Boosting. 39, 48, 49

1

Introduction

This chapter is divided into four sections. The section 1.1 explains the motivation of the developed work, while the section 1.2 presents the context of the disease. The third section shows the context of explainability in Machine learning (ML) models. The main goals of this master thesis appear in section 1.3. Lastly, the last section presents the outline of the document.

1.1 Motivation

Multiple Sclerosis (MS) is the neurological disease that most affects young adults. Currently, more than 2 million people are diagnosed with MS worldwide. This disease is more prevalent in adults between 20 to 45 years old. It is generally defined by reversible events of neurological problems, called relapses, that regularly alter to a progressive state of deterioration. The cause of such modification in the disease's characteristics is still unknown [37].

With severe physical and cognitive impairment, MS can cause a significant loss on quality of life for the patient and people around them, since many family members become lifelong caregivers as the condition worsens. Additionally, the treatment expenses tend to increase over time drastically. Nonetheless, an early diagnosis results in the delay of the progression of MS and reduction of disability, which leads to a better quality of life in the long run [36].

The symptoms of this disease and its severity development depend from patient to patient, which contributes to its unpredictability and, consequently, makes it a challenge to identify how MS will evolve over the years. This can cause some doubts on the best approach of treatment and medication administrated.

A ML model capable of predicting MS progression in the early years of follow-up could help the physicians significantly. It would help to inform the patients of their prospects and their expectations in the long term. Most importantly, it would be beneficial as a resource tool to identify early-stage cases that need a more aggressive

therapeutic intervention, to prevent the rapid development of the disease that would result in severe disability in the long term otherwise [8].

However, these developed models must be robust, without any bias, not only to fulfil the General Data Protection Regulation (GDPR) of 2018 [24], but to give data scientists reinsurance of their work and clinicians reasons to trust them. The goal is to mitigate the scepticism surrounding ML, and, most of all, to secure the safety of the patient. Therefore, an algorithm that gives human-comprehensible explanations is essential to the acceptance of these models in the clinical community [35].

1.2 Prognosis Context

MS is characterized by the demyelination of the axons in the Central Nervous System (CNS). It is an inflammatory autoimmune disease that affects millions of people by inducing neurological deterioration associated with long term disability.

People with MS may experience physical, sensory and cognitive dysfunction, often with memory and attention deficit, a decline in mobility, spasticity, and paralysis, and vision loss in some cases. Fatigue, pain, anxiety, and depression are also prevalent symptoms [49].

Although the MS development is quite unpredictable, and its symptoms vary from patient to patient, most patients are firstly diagnosed with a Relapse-remitting (RR) MS course, defined by relapses followed by a partial or complete recovery. Over time these patients might develop a Secondary Progressive (SP) state, characterized by an inhibition of recovery that results in progressive neurological deterioration regardless of relapse episodes [18].

The prediction of a SP course continues to be challenging, since there is not a clear criterion of this alteration with clinical and imaging data analysis. Precise identification of this modification has a determinant role on the efficiency of the therapy prescribed. Therefore, the delay of this diagnosis is a significant factor for permanent disability [87].

The disease is frequently evaluated on its severity, which is an informal classification. There is a distinction between benign and malignant cases. Benign MS is associated with low tissue damage, high capacity of repair after relapses, and slight physiological disability with the absence of functional problems. Malignant MS is a more severe stage with a fast progression of the disease, an evident difficulty of compensation after an attack, and serious body's impairment. So far, there is not a universal definition to determine what type of severity level a patient has, as experts usually use different criteria to identify mild cases. Nonetheless, the recogni-

tion of the disorder's severity offers essential information on the therapeutic decision making, since a patient with benign MS needs medication less aggressive and fewer monitoring appointments [19, 86].

Explainability Context

Having in mind all the complexity associated with the MS progression, prediction ML models have been in development recently, since this type of models can achieve high performances and be effective in larger data sets.

However, there is incoherence in these studies about the most significant predictors and the definitions used to identify benign cases of the disease, which leads to inconsistency and heterogeneity. Thus, validating the results is a challenge.

In addition, considering that the success of ML models depends considerably on the data set's quality, the existing diversity in the characteristics analyzed may also contribute to this discrepancy. Furthermore, the different levels of patients' diversity and the prevalence of RR patients, which leads to imbalanced data, may contribute to this heterogeneity as well.

To help combat the different disease's courses problematic, our lab created a framework [81] that predicts the progression of MS based on two predictions, the development of SP and the disease's severity, with good performance results. The partition of data was created with the K-fold method, to explore all information available and assure the absence of bias, which results in the development of several runs and a final performance, resultant of the different models mean. However, the used data set consists of a low number of patients (145 in the prediction of the disease's severity and 187 in the prediction of SP development) [81]. With a low amount of samples, the results might be limited to a possible over-fit and biases, due to lack of representation. As the data was also retrospective, the risk of exhausting the database with several methodologies by the authors is also possible.

Therefore, there are still some doubts if the developed work is effectively robust and can be trusted. In addition, the absence of information about the relations between features that trigger a prediction is a problem towards their applicability.

The absence of human-comprehensible explanations, often associated with ML models that deal with several variables, is a big obstacle that needs to be addressed. It contributes to the distrust of the models and uncertainty regarding the patients' safety, which hinders their real-world applicability in a clinical environment. These explanations may also contribute to understanding if they are logically in agreement with clinical observations and, therefore, diminish the doubt surrounding the

generalization power of the models for new data.

1.3 Main goals

This theses aims to answer the following essential questions: 'Can the framework in [81] be applied in a clinical setting?' and 'What are the human-comprehensible explanations more suited to provide trust?'. This goal can then be divided into:

- Explore distinct methods of explanation's formulation from the previously created models;
- Evaluate the produced explanations with data scientists;
- Analyse the applicability of these models in future evaluation context with clinicians;
- Contribute to a better understanding of the disease's dynamics.

1.4 Structure

This document contains six chapters beyond the introduction.

Chapter 2 presents background information about MS, ML, explainability and grounded theory terminology that will be referred to through all the document.

Chapter 3 presents the state of art concerning explainability in MS progression models and related problems.

Chapter 4 describes the experimental procedure used in this thesis.

Chapter 5 contains the results obtained.

Chapter 6 presents the discussion about the methodology used and the results obtained with this work.

Chapter 7 presents the conclusion and future plans about the developed work.

2

Background Concepts

This chapter introduces the main concepts necessary to understand this document. Section 2.1 is a summary of the definitions regarding Multiple Sclerosis (MS) with additional information about its risk factors and available therapies. The section 2.2 presents a description of Machine learning (ML), while section 2.3 summarizes the explainability concepts with a description of some methods used to achieve it. Lastly, the section 2.4 gives a brief description about the Grounded Theory (GT).

2.1 Multiple Sclerosis

Multiple Sclerosis (MS) is a chronic autoimmune neurological disease identified as the demyelination and loss of axons in the Central Nervous System (CNS). The patient's immune system attacks and destroys an electrical surface axon layer called myelin responsible for providing fast communication between neurons [41]. With the destruction of the myelin sheath, the transmission of electrical impulses between nerve cells significantly decreases velocity and, consequently, efficacy, leading to a neurological deficit. This causes mental damage and physical disability and induces a substantial loss of quality of life.

Around 2.8 million people are diagnosed with MS globally [68]. This irreversible and debilitating disease is predominant in young adults and affects more females than males [14, 15]. Patients with MS often suffer from fatigue, mobility issues, a higher risk of depression and anxiety, pain, and cognitive decline, with relapses that can last weeks. It reduces their ability of day-to-day activities, capacity to work, and life in general [77]. Yet, the progression of this disease can be inhibited and its symptoms prevented if it is identified early and treated properly according to each patient's needs.

2.1.1 Risk factors

MS is a very complex disease in terms of etiopathogenesis. The cause of this disease continues unknown. However, it is possible to say that there is a genetic predisposition to MS that, combined with a triggering environment, induces MS development [76]. This thesis does not consider risk factors, despite all this influence, due to the lack of data available to explore this aspect. Still, it is important to mention the main risk factors to understand some limitations about the developed work in this project associated with the lack of analysis of such information.

Its incidence worldwide has increased throughout the years, probably because people have a higher life expectancy and there is an increase of access to medical facilities capable of dealing with MS [52]. The cases of MS in females have also been significantly rising, with the prevalence now almost three times more frequent in women than in men [25, 42]. This increment might be associated with the better accessibility of medical care to females over the years and that women tend to look for a medical consult for minor symptoms more often than men [58].

It exists a strong relation between genetics and developing MS, since the concordance increases with the amount of shared Deoxyribonucleic Acid (DNA). Still, it is safe to say that it is not the only factor associated with the disease. Several studies found that the environment has a significant influence on the level of risk of contracting the disease as well [66].

The prevalence of MS in geographic terms has some distinctions, being higher in northern Europe, North America, south of Australia, and New Zealand, and lower in South America and Asia, as observed in Figure 2.1. These differences may occur because of the disparities in diagnosis criteria and methods countries adopt. However, some studies relate the higher disease's risk in higher latitude regions with the levels of sun exposure [58] and, directly associated, vitamin D deficiency [52, 66]. However, this is not verified all over the world, since diet can also provide good levels of vitamin D [25, 58].

The increase of MS risk is also associated with unhealthy habits in adolescence, and young adulthood, such as obesity and smoking [76].

After several studies performed about infectious agents and the disease risk, it was determined that the presence of an Epstein-barr Virus (EBV) infection in childhood and adolescence also contributes to a higher risk of MS development [58, 76].

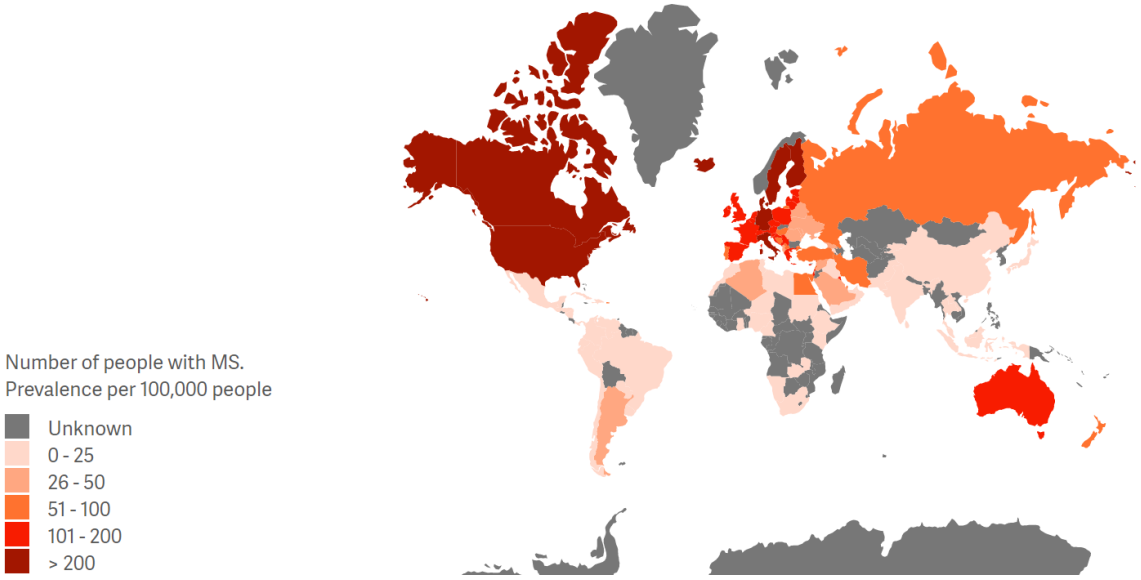

Figure 2.1: Prevalence of Multiple Sclerosis worldwide [67].

2.1.2 Diagnosis

Although this aspect is not studied in this thesis, it is important to note that the diagnosis does not occur in the same moment of life for every patient. This may influence the prognosis quality of the disease, as expected, which is a reason for the complexity associated with the prediction of the MS progression. Since this disease is characterized by heterogeneity in clinical and imaging traits, the diagnosis of multiple sclerosis is not fulfilled with a simple exam, but rather with the analysis of imaging, laboratory, and physical examinations [91].

To simplify the identification of MS, clinicians generally apply the McDonald Criteria 2017 [91], a standard that aims to give a reliable diagnosis as soon as possible to start therapy that can inhibit the effects of the disease. The presence of inflammation in distinct zones of the CNS i.e. dissemination in space (DIS), and dissemination in time (DIT), the recurring inflammation of the CNS, is the base to diagnose an individual with this disorder. Thus, this criteria use examinations of the Cerebrospinal Fluid (CSF) to identify CSF-specific oligoclonal bands, a possible indicator, but not exclusive to MS, Magnetic resonance imaging (MRI) findings, and the number of relapses [43].

The application of the McDonald Criteria 2017, represented in Table 2.1, results in three different outcomes:

- Confirmed Multiple Sclerosis - if the criteria are fulfilled, and there is an absence of a better diagnosis for the clinical evidence;
- Possible Multiple Sclerosis - if the 2017 McDonald Criteria are only partially

Table 2.1: McDonald criteria 2017 for MS diagnosis [91].

Number of relapses	Number of with objective clinical evidence	Additional data needed to the MS diagnosis
≥ 2	≥ 2	No further data is needed
≥ 2	1 (as well as clear-cut historical evidence of a previous attack involving a lesion in a distinct anatomical location)	No further data is needed
≥ 2	1	Dissemination in space demonstrated by an additional clinical attack implicating a different CNS site or by MRI
1	≥ 2	Dissemination in time demonstrated by an additional clinical attack or by MRI OR demonstration of CSF-specific oligoclonal bands
1	1	Dissemination in space demonstrated by an additional clinical attack implicating a different CNS site or by MRI AND Dissemination in time demonstrated by an additional clinical attack or by MRI OR demonstration of CSF-specific oligoclonal bands

fulfilled, even if Clinically Isolated Syndrome (CIS) leads to the suspicious of the disease;

- Not Multiple Sclerosis - if it exists clinical evidence of a better diagnosis.

2.1.3 Expanded disability status scale (EDSS)

The progression of disability for MS patients is usually evaluated by the EDSS [55], a scale based on the 8 different functional systems (FS) with values between 0 (healthy) and 10 (death by MS) in steps of 0.5. The systems considered are:

- Pyramidal - Associated with muscular weakness and lack of voluntary control of movements;
- Cerebellar - Related to deficiency of balance and coordination;
- Brain Stem - Can cause speech problems as well as difficulties on swallowing and breathing;
- Sensory - Related to loss of feeling under the head;
- Bowel and Bladder - Associated with urinary retention and incontinence;

- Visual - Responsible for loss of visual keenness;
- Cerebral - Involved with memory and concentration deficit, as well as thinking and mood disturbances;
- Other - Covers all the symptoms not comprehended in the other functional systems, e.g. pain or fatigue.

All the functional systems are evaluated on a scale of 0 to 5/6 with a direct increase of impairment, except the last one, defined as 0 if there are no other symptoms and one if there are specific neurological findings. With all the values attributed, the EDSS is then calculated with the combination of those FS scores for the values between 0 to 3.5. The next steps also regard the level of mobility incapacity [55]. Therefore, represented in Figure 2.2, the different levels of the EDSS are:

- **EDSS 0** - Normal neurological exam with all FS graded 0 with the possibility of grade 1 in the cerebral system;
- **EDSS 1** - No disability with minor signs in one FS with grade 1, except the cerebral system that can also grade 1;
- **EDSS 1.5** - No disability with more than one FS grade 1 except cerebral grade 1;
- **EDSS 2** - Minimal disability in one FS with grade 2 and the others graded 0 or 1;
- **EDSS 2.5** - Minimal disability in two FS (grade 2) with the rest FS graded with 0 or 1;
- **EDSS 3** - Fully ambulatory but with moderate disability in one FS (grade 3) and the others graded 0 or 1 OR slight disability in three or four FS graded with 2 with the others graded with 0 or 1;
- **EDSS 3.5** - Fully ambulatory but with moderate disability in one FS (grade 3) and mild disability in one or two FS (grade 2), or two FS graded 3, or five FS graded 2; all the others graded 0 or 1;
- **EDSS 4** - Fully ambulatory without aid and self-sufficient but with severe disability in one of the FS (grade 4) and the others graded 0 or 1, or a combination of lower grades that go beyond the limits of the previous levels; capable of walking without aid or rest approximately 500 meters;
- **EDSS 4.5** - Fully ambulatory, able to do a full workday, may have slight limitations of complete activities or need some assistance and is capable of walking without resting or aid some 300 meters; one FS graded as 4 or combinations of lesser grades that exceed the limits of the prior steps;
- **EDSS 5** - Disability capable of impairing full daily activities with the ability

to walk approximately 200 meters without rest or aid; associated frequently with one FS grade 5 alone or combination of lower grades that exceed the limits of the level 4;

- **EDSS 5.5** - Disability capable of precluding full daily activities with the ability to walk about 100 meters without aid or rest; only one FS graded 5 and other 0 or 1 or combinations of lower grades that pass the limits of the previous steps;
- **EDSS 6** - Intermittent or unilateral aid (canes, crutch or brace) required to walk about 100 meters independently of possible resting; a combination of more than two FS graded 3 or higher;
- **EDSS 6.5** - Necessary constant bilateral assistance with canes, braces or crutches to walk approximately 20 meters without resting; a combination of more than two FS graded 3 or higher;
- **EDSS 7** - Impossibility of walking more than 5 meters regardless of assistance, restricted to a wheelchair with the capacity to transfer alone and wheel self in a standard wheelchair. More than one FS is graded as 4 or higher or, sporadically, the pyramidal system is graded 5 alone;
- **EDSS 7.5** - Inability to take more than a few steps, restricted to wheelchair without the capacity to be in a standard wheelchair a full day (may need a motorized wheelchair). More than one FS has the value of 4 or higher;
- **EDSS 8** - Essentially restricted to bed or passively in a wheelchair with the possibility to be out of bed much of the day and ability to do most self-care functions with efficient use of arms. Frequently, several FS are graded with 4 or higher;
- **EDSS 8.5** - Restricted to bed the majority of the day with some effective use of arms and ability to perform some self-care activities. Several systems usually are graded with four or more.
- **EDSS 9** - Extreme incapacity but able to communicate and eat with most of the FS with values equal to four or higher;
- **EDSS 9.5** - Entirely helpless patient without the capacity to effectively communicate, eat nor swallow with almost all FS graded with four or higher;
- **EDSS 10** - Death by MS.

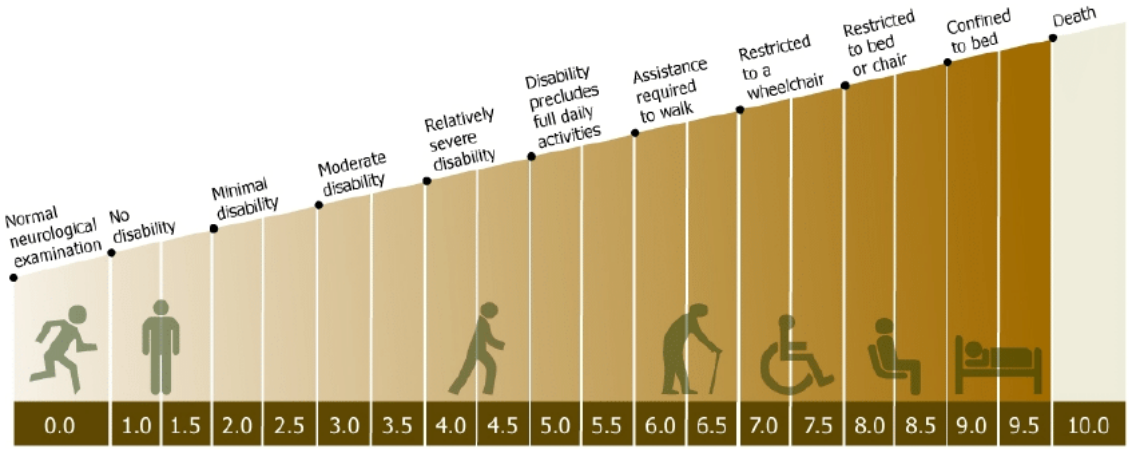

Figure 2.2: Representation of the EDSS [13].

.

Although a standard in the evaluation of MS disability, EDSS is not perfect. The application of this scale largely depends on the subjective opinion of the specialists. It prompts disparities in the attribution of a score, regarding different neurologists or even the same clinician. It is also mainly based on motor function evaluation, without much importance to other affected functions, such as cognitive capacity. Furthermore, the analysed impairment is not always the same from step to step, and the difference between the steps is not homogeneous [32].

2.1.4 Courses

Considering all the diversity of the expressions of the disease, the different courses of MS were defined to improve the communication between specialists and provide better treatments by identifying the patients included in each course [61].

Established by the 2017 McDonald Criteria, there are three different courses of the disease [69]. The most easily identifiable course is the Primary Progressive (PP). It is represented as a gradual increase of neurological deterioration from the onset with a possibility of occasional stagnation of the disease's development and a temporarily minor improvement of its effects. It is predominant in men and usually occurs in patients older at onset [70].

Since only around 15% of patients have PP, generally, people diagnosed with MS have previously a CIS, defined as the first episode of neurological problem associated with inflammation or demyelination of the CNS. The CIS is a part of the spectrum of the Relapse-remitting (RR) course. The RR is characterized by various relapse episodes (neurological attacks or new symptoms) followed by periods of remission with full or partial recovery [59]. Therefore, initially, a patient can only be diagnosed

with PP or RR, depending on if there is already deterioration of the CNS.

RR often develops a state where significant recovery of the relapses is not possible anymore, where there is disease progression independently of possible relapses or minor recovery, and moments of stagnant periods of disease's development [93]. This course is the Secondary Progressive (SP). The characteristics of the different courses can be observed in Figure 2.3.

Besides the phenotype described, the disease's courses can also be classified in their activity and progression, as referred to in Table 2.2.

Table 2.2: Definitions of disease's activity and progression [60].

Disease activity	**Active**	Presence of gadolinium-enhancing or new/larger T2 lesions or manifestations of episodes of new or advancing neurological impairment through a specific time period.
	Not active	No demonstration of disease's activity
Disease progression	**Progressive**	Evidence of an increase in the neurological disability without an apparent recovery, with the possibility of some periods of stability, during at least one year.
	Nor progressive	No demonstrations of the worsening of the disease during at least one year

All courses can be characterized by their activity (CIS, RR, SP and PP). A patient with an active CIS can be diagnosed with MS if all the requirements of the McDonald Criteria of 2017 are fulfilled, where the CIS belongs to the spectrum of the RR course [61].

If a patient shows evidence of a progressive state (SP or PP), those courses can also be characterized by progressive or not progressive. Thus, these MS courses can be associated with four distinct classifications over time [61]:

1. Active with progression: The patient is gradually worsening and has attacks throughout time;
2. Active without progression: The patient has relapses but the condition state is stagnant;
3. Not active but with progression: The patient does not suffer from relapses, but there is evidence of neurological worsening;
4. Not active without progression: The patient has a stable form of MS.

In this project, only the forms RR, SP and PP of the disease were analysed to include a reasonable amount of data. Patients with a PP course suffer from a gradual deterioration soon after diagnosis, easily identifiable. These patients' data is therefore not relevant for this study. The focus of this thesis is to help understand the evolution of unexpected courses starting with apparent RR, by implementing

explainability procedures over ML models.

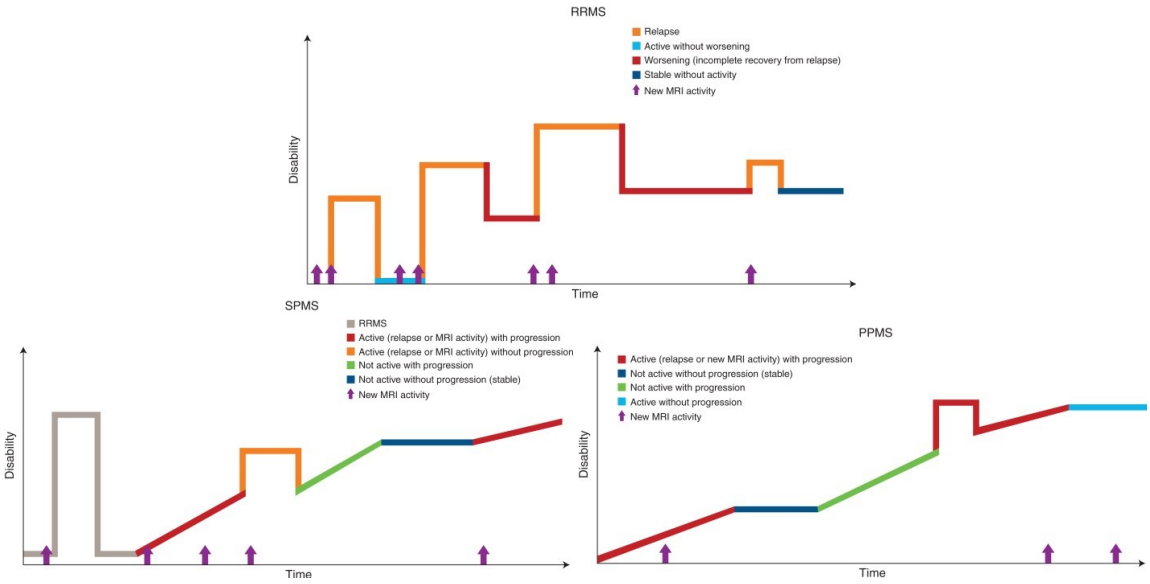

Figure 2.3: MS courses [50]. RRMS defines a RR course, PPMS defines a PP course, and SPMS represents a SP course of the disease. the purple arrows represent when a new MRI was performed.

Disease's severity

The terminologies benign and malignant are often used as indicators of the expected severity of MS over time, although they are not phenotyping descriptors of the disease [61]. Benign MS represents the lack of impairment and symptoms after several years of onset (approximately 15 years) with all neurological systems fully functional. It is mostly associated with young female patients, that were initially diagnosed with RR [18].

Since this disease can worsen unpredictably after a long time of apparent stability, it is necessary to use the term benign with caution, since it may not be a definitive characteristic [60, 61]. In addition, the boundaries that recognise mild cases are not precisely identified due to lack of consensus between specialists [18].

It is also important to note the difference between the concepts of worsening and progressing. The term worsening is related to patients whose disease is advancing because of recurrent relapses or lack of total recovery. In contrast, the term progressing refers to evidence of gradual worsening over time in a progressive course of the disease [60].

2.1.5 Therapy

The basis of the therapy of MS lies on disease-modifying agents that decrease the frequency and duration of the neurological attacks, suppress the progression of the disease, and may inhibit or even modestly reduce disability [44].

In the last decades, there was significant progress in discovering highly effective treatments that reduce the relapse rate. However, there is still a deficiency of therapies capable of combat progressive courses of the disease [44]. Table 2.3 provides information about several Food and Drug Administration (FDA) approved therapies for MS.

Table 2.3: Summary of Approved Disease-Modifying Therapies used in MS treatment [44].

Name	Type and frequency of administration	Disease's Course	Action	Side effects
Ocrelizumab	Intravenous (IV) infusion, every 6 months	RR and PP	Reduction of annualized relapse's rate and disability's progression	Infusion-related reaction, nasopharyngitis, urinary and upper respiratory tract infection, headache and oral herpes infection
Ofatumumab	Subcutaneous (SC) injection, every 4 weeks	RR	Reduction of annualized relapse's rate	Injection-related reaction, nasopharyngitis, urinary and upper respiratory tract infection and headache
Natalizumab	IV infusion, every 4 weeks	RR	Reduction of annualized relapse's rate and disease's progression	Fatigue and allergic reaction
Alemtuzumab	IV infusion, once daily	RR	Reduction of annualized relapse's rate	Headache, rash, nausea and pyrexia
Mitoxantrone	IV infusion, every month or 3 months	RR and SP	Reduction of relapses	Dose-related cardiomyopathy and promyelocytic leukemia
Fingolimod	Oral, once daily	RR	Reduction of annualized relapse rate	Bradycardia, atrioventricular conduction block, macular edema, elevated liver-enzyme levels and mild hypertension
Siponimod	Oral, once daily	CIS, RR and active SP	Reduction of disability's progression	Headache, nasopharyngitis, urinary tract infection and falls
Ozanimod	Oral, once daily	CIS, RR and active SP	Reduction of annualized relapse's rate	Headache and elevated liver aminotransferase
Dimethyl fumarate and diroximel fumarate	Oral, twice daily	RR	Reduction of annualized relapse's rate	Flushing, diarrhea, nausea, upper abdominal pain, decreased lymphocyte counts and elevated liver aminotransferase
Cladribine	Oral, 4-5 days over 2-week treatment courses	RR	Reduction of annualized relapse's rate	Headache, lymphocytopenia, nasopharyngitis, upper respiratory tract infection and nausea
Teriflunomide	Oral, once daily	RR	Reduction of annualized relapse's rate	Nasopharyngitis, headache, diarrhea and alanine aminotransferase increase
Glatiramer acetate	SC injection, once daily or 3 times weekly	RR	Reduction of annualized relapse's rate	Injection-site reactions
Rebif (IFN-β-1a)	SC injection, 3 times weekly	CIS and RR	Reduction of annualized relapse's rate	Injection-site inflammation, flu-like symptoms, rhinitis and headache
Avonex (IFN-β-1a)	Intramuscular (IM) injection, once weekly	CIS and RR	Reduction of disability's progression	Flu-like symptoms, muscle aches, asthenia, chills and fever
Plegridy (IFN-β-1a)	SC injection, every 2 weeks	CIS and RR	Reduction of annualized relapse's rate	Injection-site erythema, influenza-like illness, pyrexia and headache
Betaseron (IFN-β-1a	SC injection, every other day	CIS and RR	Reduction of annualized relapse's rate	Lymphopenia, flu-like symptoms and injection-site reactions

It is possible to verify that the prescription of these therapies depends on each clinical condition, since different medications have different objectives (to prevent relapses, disability, or MS progression). Additionally, they are associated with several adverse effects that potentially cause a decrease in the patient's quality of life. It is also important to note that some medications are being applied on the MS treatment, even though they still are not FDA approved (e.g. IV immunoglobulin)

[44].

Since the administration of medications depends on the physicians' judgement and may alter the development of the disease, the use of therapy's information may lead to a biased outcome and, therefore, prediction models may not use this type of data. In addition, the purpose of these models is to identify the disease's courses to provide adequate therapy before any medication is prescribed that could lead to a potential risk to the patient. However, the data used in the developed models have information from several years of follow up, where therapies are administrated to the patients. Therefore, the medication is regulating other variables of the disease, which indirectly influences the disease's course and possibly the results, even though it is not directly considered. It is an essential limitation of this study that can not be combated, as it is not ethical to deprive patients of the treatment that makes them better.

2.2 Machine learning

Machine learning is a form of Artificial Intelligence (AI) focused on providing computer learning skills based on experience and inductive methods that mimic the human approach to learning [46]. It can explore complex data without specific knowledge, that otherwise would turn out to be a challenge or even impossible, with the intuit to learn behaviours or to be used as a prediction tool [72].

ML can be divided into unsupervised learning, supervised learning and reinforcement learning. Unsupervised learning characterizes models that only have information about the input. Supervised learning defines models with knowledge about the input and output that find the best function representing their relation. Reinforcement learning is also influenced by rewards and penalties from the environment [74].

ML models, specifically supervised learning, start with the acquisition of raw data that may then be pre-processed and transformed to create inputs with important information for prediction, the features. These inputs are used afterwards in different classifiers that find relations between features, and predict the outcome of new data. Lastly, the predictions' performance is evaluated with several metrics to measure the model's generalization ability [21].

The focus of this thesis is supervised learning, since this was the type of learning used in the prediction models. With the input as the information of the first years of follow-up, models were generated to predict the outcome, medical annotations of several years after the input data acquisition.

2.2.1 Data preparation

Since ML is greatly dependent on the input data characteristics, models that perform efficiently are only possible if the data set is a good representation of the problem studied, i.e., if it is representative.

Therefore, firstly, it is important to analyse if the input will be raw or transformed data that may have more relevant characteristics to study. Since raw data generally presents inadequate formats with unnecessary information, the latter hypothesis is usually more suitable. Thus the raw data is processed to obtain the best possible features [75].

2.2.1.1 Missing data

As models can not use data with missing values, it is necessary to eliminate samples with missing values, which possibly causes loss of relevant characteristics for learning and may lead to bias results, or impute the values.

With the application of imputation of missing values, it is necessary to take into account if the values are Missing Completely at Random (MCAR), Missing at Random (MAR), or Missing Not at Random (MNAR). MCAR represents values where the probability of missing values does not depend on any variable. MAR defines incidents where the probability of missing values on one variable may depend on the values of other variables. Lastly, MNAR consists of when the probability of missing values relies on unobserved data [22, 78].

Single imputation is frequently used in this problem. The imputation of values is associated with the use of estimations from the self variable value (e.g. the variable's mean), or by considering information from other variables. Multiple imputation can also be applied, e.g, by a multivariate regression as a resource to add diverse values that guaranty a level of uncertainty and maintain variables connections [78].

Missing data is often observed when dealing with the progression of MS problems [81, 88]. Generally, the data sets are created in a natural hospital environment by manually imputing the different variables' values, which may occasionally lead to some unfilled fields.

Single imputation overcame this problem in the mentioned models [81], by filling the missing data with the mean value of the specific variable in the training set, and with the mean of the training set in the testing group.

2.2.1.2 Feature Engineering

The creation of features that best represent the studied problem by transforming data is called feature engineering [84]. This transformation may be associated with the combination or segmentation of information. Furthermore, the features must be normalized to prevent heterogeneity in scale.

It is also important to make a selection of the most significant features, not only to avoid redundancy and irrelevance in data, but also to prevent overfitting, the lack of generalization ability caused by noise and detail learning [95]. There are three types of feature selection methods:

- Filter methods - Based on ranking criteria, these methods often select features with the highest scores regarding several metrics, e.g. correlation criteria. It is independent of the learning algorithm and ignores features' interactions [11];
- Wrapper methods - They act by analysing all features combinations and evaluating the resultant performance through model training until its maximization. Wrapper methods consider feature interactions and offer the best accuracy but are computationally expensive and may lead to overfitting [16];
- Embedded methods - These methods apply feature selection when training the model, to spend less computational time than wrapper methods but maintaining a better performance than filter methods [95]. They may be tree-based (decisions trees) or regularization (Least Absolute Shrinkage and Selection Operator (LASSO) or Ridge regression) methods. The last incorporates penalties depending on the model's dimensionality, reducing its levels of freedom and consequently increasing the generalization power and robustness of the model [16].

The developed framework to select the most significant features included two forms of filter methods and one embedded method. Firstly, Pearson's linear correlation coefficient method was applied to choose the best 100 features, and then the area under the curve method was used to select the best 50 features. Lastly, to find the optimal set of features and eliminate redundancy, the LASSO method was applied [81].

2.2.2 Classification

Supervised learning generally starts the division of data into three distinct groups: one for training the models by learning interactions between features and labels, the training set; one used to evaluate the trained models' performance, the validation set; and the last to confirm the results of the final model after all im-

provements made, the testing set [54].

An often used technique to data partition is Cross Validation (CV) methods, and specifically K-fold CV for studies with small data sets, as used in Pinto et al. [81]. K-fold CV consists of dividing the data into k groups, k-1 parts of training the model and the other for validation. It operates by doing k iterations to guaranty the analysis of the complete data set [83].

The existence of considerable differences in the number of the class examples in classification influences negatively model training, as it learns the patterns from the class that has more examples (majority) to the detriment of the minority class. This leads to underestimating classification errors related to the lack of detection of the minority class and, therefore, to a false sense of good results. Imbalanced data generally is found in real-world problems [33].

It is possible to balance classes by undersampling, oversampling or weight balancing. Undersampling consists of eliminating (often randomly) samples of the majority class from the training set. On the contrary, oversampling duplicates samples from the minority class and adds them to the training set. Moreover, weight balancing is a method that gives increased weight to samples from the minority class, since usually are the classes with greater importance in prediction problems [33].

2.2.2.1 Classifiers

There are several classification methods available in ML that function in different techniques to suit specific needs about the studied problem and data set.

A very simple classifier is the K-nearest neighbors (KNN). This instance-based algorithm decides which classes to attribute to an instance, based on the similarities with the training data points. This method assigns to the sample in the study the class that is most predominant among the k number nearest neighbours [74].

Decision trees are logic-based algorithms with a tree-like structure where each node is a feature instance, and each branch possible values for the node. It starts at the root and, in each node, the model decides which branch represents better the data until it reaches a classification, the leaf nodes [90].

Linear regression is also frequently used in ML. It considers linear relationships between features and the output. Therefore, in the training phase, these models obtain a combination of linear functions representative of the mentioned relationship and use it to predict new data [2].

It is important to highlight support vector machines (SVMs). These algorithms

map each instance into a feature dimension space and find a hyperplane that separates the classes with the most significant separation margins possible. New samples are then classified depending on what side of the hyperplane they are positioned [74].

Lastly, deep learning is increasingly popular to manage large amounts of data in ML, often getting valuable knowledge unobtainable by other methods [71, 74]. It consists of neural networks with multiple nonlinear layers associated with multiple weight parameters, optimised to perform the most efficiently possible. It is, therefore, a very complex algorithm [71].

Since the data set available for this thesis was very limited [81], it was not possible to apply deep learning classifiers. Neural networks need significant amounts of data to learn feature relations correctly and consequently offer a good prognosis. Therefore, with the used data set, the models would overfit. However, when there is sufficient data, these models generally provide higher performance values, due to their ability to analyse and transform the input through their layers. It is possible to find essential hidden relations with the target with neural networks, which makes this type of classifiers frequently used in similar prognosis problems [88].

2.2.3 Performance evaluation

Various metrics can evaluate the performance of the models. For binary classification, one class is frequently considered positive and the other negative. These metrics are generally based on the number of samples correctly classified as positive, the True Positives (TP) and the number of instances identified correctly as negatives, the True Negatives (TN). They also consider the samples classified as positives but are negatives, the False Positives (FP) and vice-versa, the False Negatives (FN) [64].

Thus, the prediction's performance can be evaluated by [64]:

1. **Accuracy** - Ratio of correctly classified samples considering all data. This measure is not suited to evaluate imbalanced data since a FN has minimal weight in this measure. The accuracy may be high but the model not adequate;
2. **Sensitivity** - Also called as recall, it is the proportion of correctly samples classified as positive given all real positive samples:

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.1}$$

3. **Specificity** - Proportion of correctly samples classified as negative given all

real negative samples:

$$Specificity = \frac{TN}{TN + FP} \tag{2.2}$$

4. **G-mean** - It calculates the geometric mean of sensitivity and specificity:

$$G\text{-}mean = \sqrt{sensitivity * specificity} \tag{2.3}$$

   This metric is very popular to evaluate the performance of imbalanced data problems, since it is not significantly influenced by the heterogeneous distribution [6]. Therefore it gives a better notion of the general classification's performance than accuracy in MS prediction models, which use data sets largely imbalanced.

5. **F1-score** - It is the harmonic mean of precision, the ratio of TP considering all samples classified as positive, and recall [6].

$$Precision = \frac{TP}{TP + FP} \tag{2.4}$$

$$F1\text{-}Score = \frac{2 * precision * recall}{precision + recall} \tag{2.5}$$

   It is greatly related to the prediction of rare cases and the sensitivity. A large amount of false positives in MS predictions can cause serious adversities on the patient's health caused by unnecessary aggressive medication, a problem that can be evaluated by this important metric. Considering the high percentage of imbalanced data, often the F1-score has lows values in MS problems [81, 88].

6. **Area Under the Curve (AUC)** - It is the area under the Receiver Operating Characteristic (ROC) curve, a plot graphic that represents the True Positive Ratio (TPR), or sensitivity, dependent of the False Positive Ratio (FPR), or 1-specificity, depending on different thresholds to evaluated which class a sample belongs. An AUC equal to 1 represents an ideal model, capable of correctly predicting every instance, and an AUC equal to 0 demonstrates that the model is inversely identifying the samples. If the AUC has a value of 0.5, the model follows a random classification.

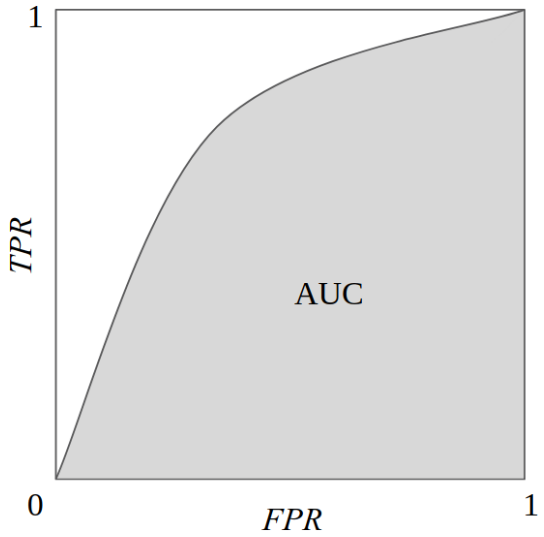

Figure 2.4: ROC curve [79].

Other ML problems in clinical problems

To study this type of problems is associated with some challenges. In these clinical problems, it is necessary to assess data relative to several years of patient monitoring. Since acquiring such data can be a difficult task, generally it is used retrospective data. The analysis of retrospective data, specially associated with a low number of samples, can lead to over-fitting or overestimation, due to the exhaustion of the data with various implemented methods.

Thus, it is important to apply, step by step, the methodologies to prospective data. However, it takes several years of patient monitoring, and consequently, several years to collect this data.

2.3 Explainability

With the rising use of ML systems in real-world applications, specific criteria need to be assured when dealing with critical safety activities. However, only general performance metrics may completely meet them, e.g. accuracy or sensitivity [23].

The concern regarding this problem has been rising in recent years, ideally reinforced by the creation of the right to explanation in the General Data Protection Regulation (GDPR) 2018 [38].

The addition of the regulation demonstrates the need to guarantee the users' safety. Criteria such as fairness, i.e. the reassurance that there is no discrimination associated with the predictions towards protected groups, need to be assured. Additionally, it is vital to guaranty that robustness, i.e. the ability to maintain the

same performance levels when submitted to variations of the input or parameters, is fulfilled [23]. It is also fundamental to trust the used models from the human users and guarantee that the results are not effects of randomness. To assure causality in the model's behaviour is essential, i.e. that the model prediction perturbations will also be observed in the real-world system [23].

It might be possible to achieve these requirements by evaluating the model's logic if, indeed, they can explain it [23].

Interpretability represents the ability of a system to present its logic so that a human can easily predict the output with the analysis of the input. Methods like KNN or decision trees are considered interpretable models, due to the simplicity of their mechanisms. However, with the increase of features, the gain of complexity is guaranteed, and even those models become less and less comprehensible, thus loosing this capacity [34].

Interpretability may not respond to all requirements needed, since generally interpretable models lack performance compared to complex models, but explainable ML might. Interpretability offers the simplicity of understanding the model's logic so that it is effortless to know the result of new predictions easily. However, an essential condition for the success of prediction models is high performance. Interpretable models may not achieve that. Due to their absence of complex learning, it may be possible to loose essential relations between the input and outcome, resulting in unsatisfactory performance values.

Explainability is the ability of a model to explain its reasoning and behaviour in human terms, without the necessity to comprehend the underlying mechanisms of the models fully [34]. The techniques that implement explainability may only work in particular machine learning models, such as neural networks or linear regressions, or applied after training any model, with the analysis of the system's internals knowledge [72].

These models can support their actions with explanations, fighting the incompleteness of ML models and, consequently, gaining validation and trust from the scientific community [35]. The challenge is to know where interpretability is necessary, or when explainability is enough to ensure their applicability in a clinical environment. The difference relies on the type of medical problems analysed. If grounded knowledge about the studied problem and risk score models are already being applied, the best option is to create interpretable models. However, explainability is fundamental to dealing with issues without a high underlying knowledge about its real dynamics, which requires models with complex internal structures.

2.3.1 Taxonomy

There are several methods used for ML interpretability and to produce explanations that can be analysed having in mind different concepts.

Firstly, it is possible to analyse the models if the interpretability is intrinsic, by limiting the model's complexity, or, post hoc, by using methods that produce explanations after the training step of the ML models [72]. It is also possible to distinguish the methods if the explanations are associated with the system's global behaviour, i.e. if the explanation is global, or if they are related to specific predictions and groups of instances, defined as local explanations [23].

The results can differ in the type of explanations desired. It is possible to analyse summaries about the features' characteristics with statistics (e.g. importance of each element) or visualise graphic plots. There is also the possibility to understand the model's internals, such as linear weights or the structure of the trees, as it happens with intrinsic interpretability. Another way to understand a classification model is to study the samples of the data set and find specific characteristics to compare new data points to their dynamics [72].

This is still a very recent field of study. Thus, there is still no consensus on many issues, namely, how to group types of explanations, evaluate them, and specify the definitions of interpretability and explainability.

2.3.2 Explainability Evaluation

It is necessary to explore three levels of evaluation to evaluate the interpretability of a model: application, human, and function level of evaluation [34], as observed in Figure 2.5.

Application-level evaluation lies on the creation of human experiments with specialists regarding the real end task in study [34]. The foundation is how efficiently human-created explanations help other humans to complete the tasks. Although, this evaluation is challenging. It requires a significant time and cost to analyse the methods, since it requires people exceptionally trained in the subject difficult to enter in contact and need to be compensated for their work [23]. In this thesis, it is possible to divide the evaluation of the developed work into two groups, evaluation by clinicians and data scientists. This thesis focuses on this level of assessment, directed at data scientists.

Human-level evaluation combats the limitations of application-level evaluations [34] by the generation of human experiments with laypeople on simplified applications, representative of the end tasks without jeopardising their essence [23]. This

type of experiment can be applied by asking a person to choose between a set of explanations which is better [72]. Human-level evaluations are critical to simplify some studies and understand which type of data is more suited to explain the prognosis to patients.

Finally, the function level evaluation uses formal interpretability definition as proxies to analyse the quality of the explanations, without the involvement of human experiments [72]. It is very popular, as it does not require necessary approvals and additional cost and time related to the other type of evaluations, conditions that often are untenable for ML researchers [23]. Due to all these difficulties, the prediction developed models [81] only evaluated the results with proxies intrinsically. The authors considered that a model with fewer features and simpler classifiers was interpretable. Therefore, the generated framework would be relatively interpretable, depending on the number of features used in the classification.

It is fundamental to validate the explainability with human experiments to ensure that a model is indeed explainable or interpretable. It is important to note that there is still a lot to do before these models are ready to be applied in the real world, and the future tasks are challenging, but the path to do so is being set.

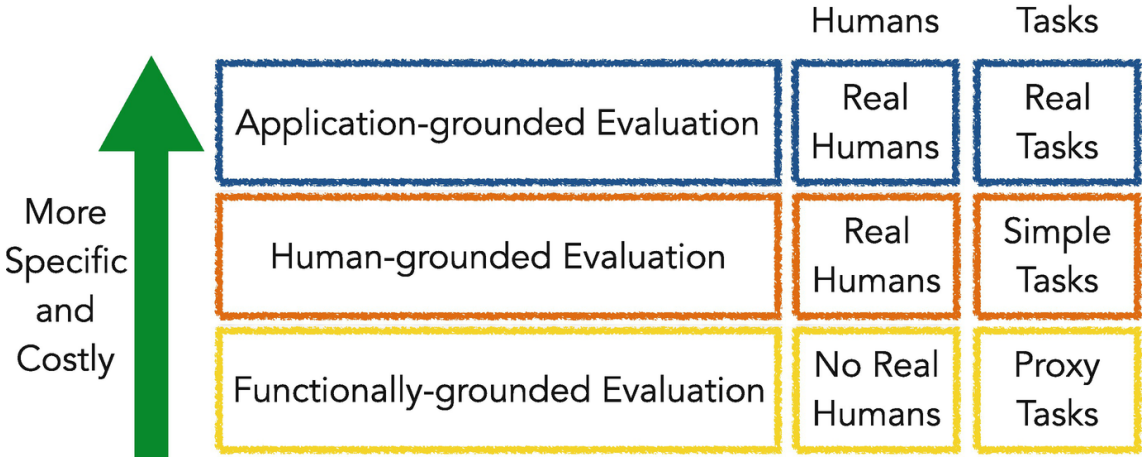

Figure 2.5: Evaluation of Explanations [23].

2.3.3 Explainability methods

As referred before, it is possible to distinguish the methods that produce explanations by their ability to be used in any model or be specific to an application.

With the creation of methods independent of the models, defined as model-agnostic explanation methods, it is possible to achieve high levels of flexibility, since there is no limitation on used techniques. Additionally, it can facilitate model's results comparisons, as the used explanation methods can be the same. The restriction

of applying only intrinsic interpretable models can come with the cost of significant loss of performance compared to black-box models. Generally, as a part of the model-agnostic type of methods, example-based explanations use specific samples of the data set to characterise the behaviour of the models, or to represent the data set distribution without the study of every observation [72].

Although it is possible to use model-agnostic explanations to comprehend neural networks, these models benefit from explanations specific for neural network interpretation. Special tools are necessary to discover the concepts of their many hidden layers, that could not be uncovered by universal models [72]. This type of methods can gradient interpret the different layers of the models, which is a considerable advantage compared to procedures that only consider their final results.

2.3.3.1 Model-Agnostic methods

These methods are based on the generation of feature summaries, by visualisation, their importance degree, or highlighting important interactions between them [72]. Here are some examples of such techniques:

Partial Dependence Plot (PDP)

It shows through graphic plots the marginal influence of one or two independent features in the prediction, given by a ML model, i.e, with the output $g(x)$, the partial dependence of the X_S domain (the features studied) is given by [98]:

$$g_S(x_S) = E_{X_C}[g(x_S,X_C)] = \int g(x_S,X_C)dP(x_C) \tag{2.6}$$

Where X_C is the rest of the features used in the ML model and g_S is the prospect of g on the marginal distribution of X_C.

For example, it is possible to understand how a model predicts the number of bicycles rented ($g(x)$) in a specific season by analysing the PDP of the features humidity, temperature and wind speed at Figure 2.6.

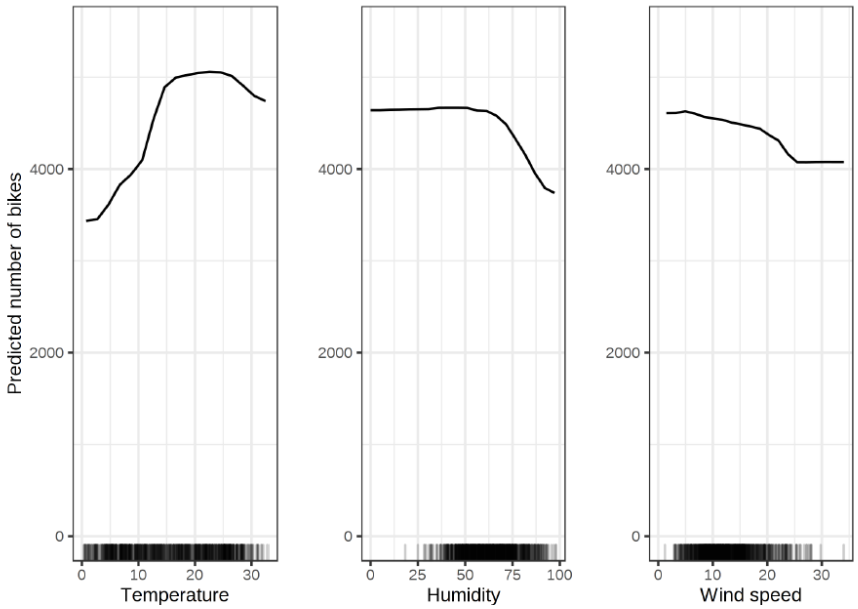

Figure 2.6: PDP for the prediction count model of bicycle renting of the most significant features [72].

It is possible to observe from the plots that the hotter the day is, the more rented bikes, until it is too hot, which leads to a decrease in rented bicycles [72].

This causal interpretation method is straightforward to implement and intuitive. Still, it is limited in the number of features studied, as it is only possible to analyse two features simultaneously. Another problem is that only the average marginal effects are shown, which might hide disparate influences. This leads to the possibility to see results that do not represent the real nature of the problem [72].

The Accumulated Local Effects (ALE) plot is a similar and faster alternative to PDP that demonstrates the features' influence on the model's prediction on average [98], using intervals of data points. It is not affected by bias, but its implementation is more intricate. It is necessary to smooth the plots for a more readable analysis, by decreasing the number of intervals used, which may result in loss of the actual influence of the variables. It also can not be directly linked with Individual Conditional Expectation (ICE) (ICE) plots, i.e. it is not possible to analyse these results in the different samples [72].

Individual Conditional Expectation (ICE) plots

These plots differ from PDP since they show the dependence for each distinct instance, not the overall average. It resolves the problem of heterogeneous effects related to PDP. Still, its interpretation is only evident in one feature at a time. It might be challenging to distinguish relevant characteristics due to the possibility to

overcrowd the plots with instances [98].

Given the bicycle renting example, the corresponding ICE plots are represented in Figure 2.7, demonstrating a similar result as in the PDP.

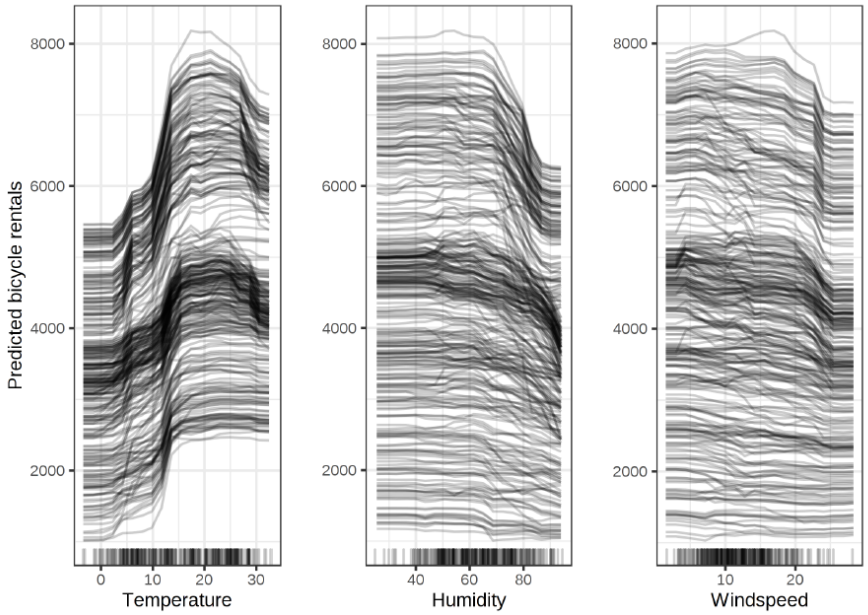

Figure 2.7: ICE for the prediction count model of bicycle renting of the most significant features [72].

Feature Interaction

With the intuit to understand how the relations between features influence the predictions, this method evaluates to what degree two components interact with each other using Friedman's H statistics [31], as well as the interaction of one feature with the rest of the data variables. It recognises all interactions, no matter their form, which is an adequate method to implement before the creation of the PDPs of the relevant discovered interactions. However, these metrics are computationally expensive, and there is no definition of the threshold to distinguish what interactions are high enough to be considered relevant [72].

Permutation Feature Importance

This method consists of evaluating the importance of a feature by permuting it and analysing the prediction's error increase. It gives a global understanding of the model's behaviour considering all interactions between variables, but it is sensible to unrealistic instances. Thus it can be biased [72].

Global Surrogate

Global surrogate method is used to approximate the predictions of black-box models with simpler models that are interpretable. These models are applied after the black-box models' training phase, without the need of any internals' information from them. It uses their prediction as the target to evaluate the ability of approximation of the surrogate model easily, being a very flexible method [28]. It is important to note that this technique does not have information about the actual outcome and, therefore, the resultant interpretability is only related to the model itself.

Another form of model surrogates is local. Local Interpretable Model-Agnostic Explanations (LIME) trains models for specific prediction explanations. This method is very promising, since it is easy to use and produces short and simple explanations. However, its implementation has some problems, as it is a method still in development with room for improvement [72].

Scoped Rules (Anchors)

With reinforcement learning algorithms, this technique explains single predictions by discovering decision rules that link the forecast, considering that the change of other features does not alter the prognosis. Similar to LIME, this method uses perturbations to find IF-THEN rules that anchor the instance to its outcome, with the consequent creation of local explanations. It is a method easily understandable, since it is based on rules, albeit it needs a lot of configuration time, and the method's coverage may vary due to exceeding specification, generally associated with the necessity of discretisation [72].

Shapley Values

A game theory-based method, the Shapley values analysis is grounded with the notion that each feature is a player in a game where the prediction is the payout for an individual instance. For each feature, the Shapley Values method evaluates the model with all possibilities of feature coalitions, with and without the studied variable, in the analysed instance. The objective is to distribute the payout fairly through the features [28]. Returning to the bicycle renting example, Figure 2.8 shows the Shapley values for a predicted value having in mind each feature. In this instance, the most significant contribution to the predictions was humidity and weather, with the number of bicycles below average.

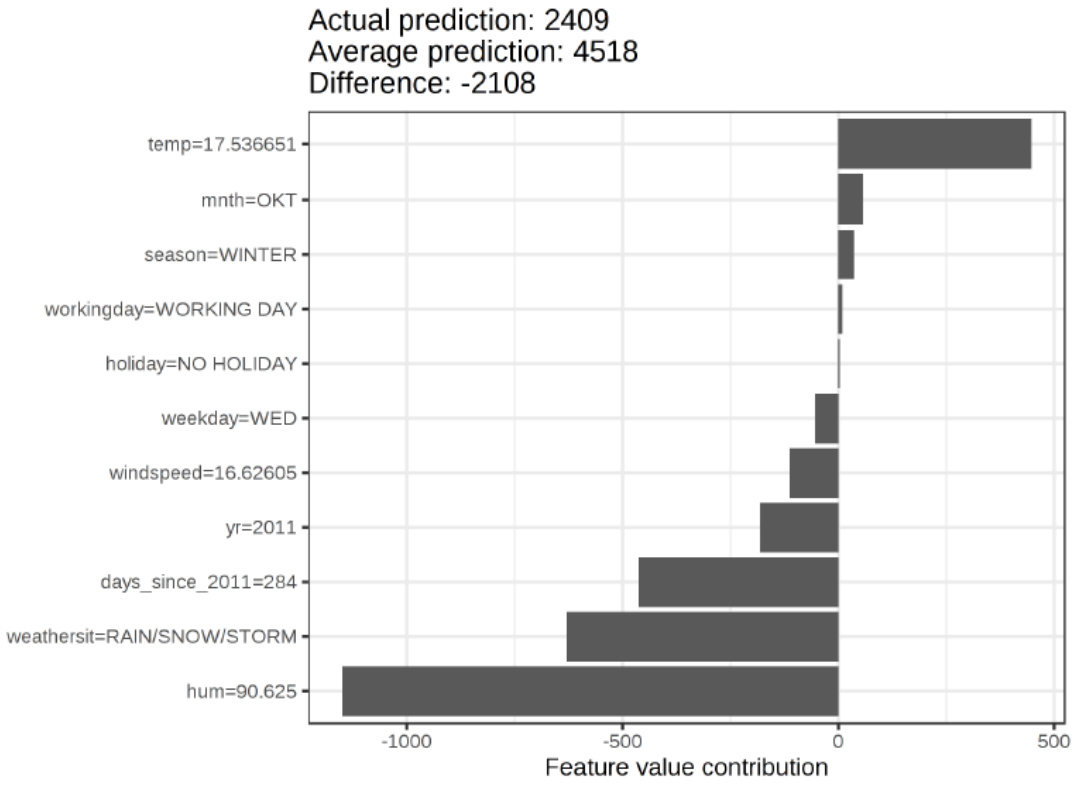

Figure 2.8: Shapley values regarding an instance from the prediction model of the daily rented bicycles number [72].

This method is very time consuming and may include instances that are not realistic when there is a correlation between the variables [72].

2.3.3.2 Example-based methods

As the name suggests, example-based methods differ from the model-agnostic methods because they produce explanations based on significant data instances and not the features characteristics. These methods are great tools to help explain structured data, but to use them in tabular data sets with several features might be difficult and result in a lack of meaningful information due to the absence of human-comprehensibility [72].

A well-known technique associated with example-based interpretability is the already mentioned KNN, a model whose foundation is to compare the similarities between neighbour points to make a prediction. Other methods that can be example-based are the LIME and the Shapley values as they produce explanations to specific observations.

Counterfactual Explanations

This method uses statements in a conditional form that demonstrate how a particular prediction would change if specific variables had different values for that instance [94], i.e., to have hypothetical scenarios that contradict the observation. For example, if someone wants a loan but the request is rejected, with counterfactual examples, this person could understand why it was not approved with explanations like "if the annual income was 5000 euros higher, the loan would be accepted" [72].

This technique is relatively simple to implement, and its explanations are apparent for a human to understand. It is not restricted only for ML systems[72]. However, the possible counterfactual statements are endless for a ML model. Thus it is important to produce explanations that only explore the smallest or the most applicable changes necessary to alter an outcome [94].

Similarly to counterfactual examples, it is possible to use feature alterations, adversarial examples, to confuse the model and consequently evaluate the systems' vulnerability to outliers and attacks [72].

Prototypes and Criticisms

Although insufficient, using examples representing the data behaviour is a great asset to understand the data distribution. Having in mind the data points that do not fit into the prototypical examples, defined as criticisms, is fundamental to understand the complex distribution of data. In other words, *"data points from regions that are not well explained by the prototypes are selected as criticisms"*, [72], as observed in the Figure 2.9.

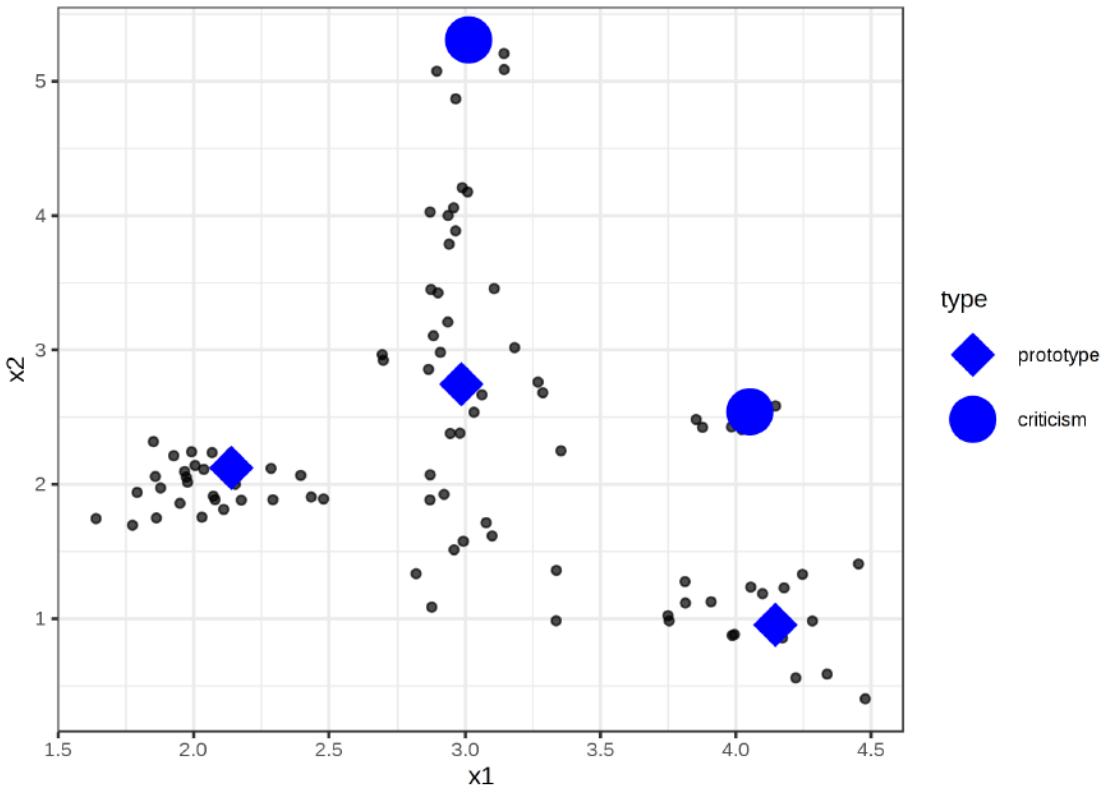

Figure 2.9: Prototypes and criticisms for a data set with two dimensions [72].

An example of methods that find prototypes and criticisms is the Maximum Mean Discrepancy (MMD)-critic [48], a technique that, with MMD statistic, measures the points' similarity and finds prototypes that maximise it. In addition, it also finds the criticisms with a regularised witness function score. This method is easily understandable and can study any form of data. But it does not consider the existence of irrelevant features [72] and the distinction of criticisms and prototypes is generally only based on the desired number of prototypes [40].

Influential Instances

This method aims to identify samples that significantly influence the model's predictions, when removed from the training data, as represented in Figure 2.10 [51]. A popular approach to identify these instances is to remove them and then train the model. This method is defined as deletion diagnostics, but it is costly in terms of computational time, and storage [51]. It is possible to use influential functions instead to overcome these limitations, by increasing the loss weight of an instance using robust statistics without the need to retrain the model [53]. However, this method needs knowledge about the loss gradient, which restricts the application to specific ML models [72].

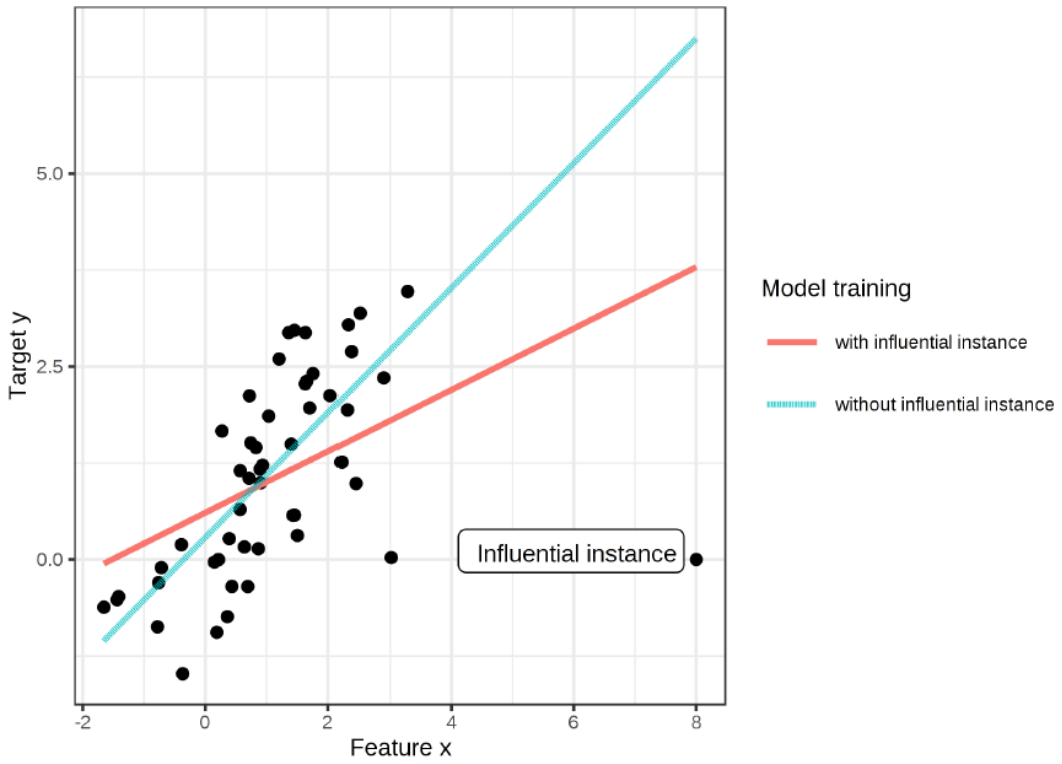

Figure 2.10: Linear model with one feature with and without an influential instance [72].

2.3.3.3 Neural network interpretation

With the popularisation of neural networks in ML comes the need to develop methods that explain their behaviour global and clearly to human users. This task is very challenging, due to the several layers and weights used in these networks that provide high complexity associated with their logic mechanism.

Therefore, most techniques are model specific methods that can discover the information of the hidden layers, by learning created features in each unit or extracting the concepts that each layer learned from the data.

Feature visualisation is a method that shows the new learned features by discovering the data that maximises the activation of a unit. It demonstrated that the first layers learned more straightforward characteristics in images, and increased the complexity along the network. This method can be linked with network dissection, a method that labels units from the neural network with human concepts, to show what that unit learned in an understandable manner [72]. Another method beneficial to deal with images is pixel attribution, a technique that evaluates the relevance of each pixel in the classification of an image.

It is also possible to try to explain a neural network model by simpler models,

with the resource of model distillation or the study of the relevance that each feature has on the prediction [7].

Considering all the complexity of a neural network system, it is impossible to fully interpret the associated internal mechanisms, which is perhaps absurd to affirm that these methods can turn these models entirely explainable. However, these techniques have a fundamental role in keeping users and scientists a bit less out of the dark on comprehending systems too complex for human understanding capacity. They are therefore critical in the development of such systems now and in the future.

2.3.4 Explainability Methods remarks

Since the created framework [81] does not apply neural networks for the prognosis of the disease, methods specifically focused on developing explanations for deep learning will not be used.

From the previous work, some features appeared to have a significant influence on the outcome but, because the dynamics of MS is very complex, it is fundamental to understand the degree of features interactions and how these interactions operate. It would be interesting to implement the PDP methods as the explanations created show visually the relationship between each feature (or pairs of features) with the predicted outcome, which makes it possible to compare directly with the already known clinical information of the dynamics of the disease. This assures the clinicians if the model is learning the correct logic of the features' evolution considering the different courses of MS.

Shapley values are known to provide grounded explanations with a solid logic base behind them, with the advantage of not assuming linear behaviour, an approximation not proven to represent correctly the real-world although frequently used by other methods. Its implementation might unravel important information about each feature's contribution to a result of an instance, possibly increasing the knowledge already acquired by the analysis of the general predictive power of the most features.

Explanations similar to those obtained by counterfactual examples have a high level of human-comprehension, and are typically used as a learning technique in the real-world by everyone. Thus, its application to produce and find causal explanations for the models' predictions is also very appealing.

2.4 Grounded theory (GT)

Evaluating explanations given by models is challenging, and it may be difficult to do so quantitatively. An explanation is an exchange of beliefs. Different people may have different priorities when it comes to an explanation. So, more than just understanding which are the most used explanations, it is more significant to understand people's reasoning to choose them and what they think is important. Thus the developed explanations in this thesis will be presented to data scientists, followed by open question interviews. These interviews must be analysed formally and rigorously. As such, this analysis will be executed by resorting to a very common tool used in major scientific research, the grounded theory (GT).

GT is an inductive research approach that creates theories based on collected data, without a preconceived hypothesis formed beforehand. This approach is frequently used to study qualitative, and sometimes quantitative information, grounded in the collection and analysis of the data systematically. With GT, the experience of the people in the survey guides the research, and the results are a reflection of the found patterns. This concept avoids the research to be led by the investigator's assumptions a priori, but rather by an impartial view of the social phenomena. It is a technique to generate a theory, not to test an existent hypothesis [29].

Since the information is generally acquired by interviews or in observational fields, the data sampling does not focus on population's representation but on collecting emergent considerations about the problem in study, until it achieves theoretical saturation. Theoretical saturation describes the point where there are no new relevant insights obtained in the data collection [47]. Large amounts of interviews and data are not crucial in GT, but the sufficient quantity to execute this saturation [39].

With the ongoing data acquisition, the emerging issues and incidents are noted and constantly compared, where the similarities and differences with the remaining examples are found [47].

The next step is data coding, i.e., combining significant incidents and issues according to their similarities. These incidents may be the topics that most frequently occur or new perspectives that may not be obvious to most but, either way, interesting points of view [39]. With the combinations, categories are formed that are constantly refined, through constant comparison that redefines and reorganises the categories, according to the ongoing data acquisition that, over time, it gets increasingly focused and selective [47]. The categories are linked, and notes, or memos, are written to identify patterns in each theme and between them. This concept

is beneficial to organise and fit the data, and express the emerging formed links. Memos' purpose is to help shape the writing of the final theory, as they transform the information to a conceptual level and help to rework ideas until the final product is achieved. The generation of a theory consists of the organisation and combination of memos and the theoretical draft into a thoroughly connected and attainable hypothesis [29]. A scheme of the approach of GT is presented in Figure 2.11.

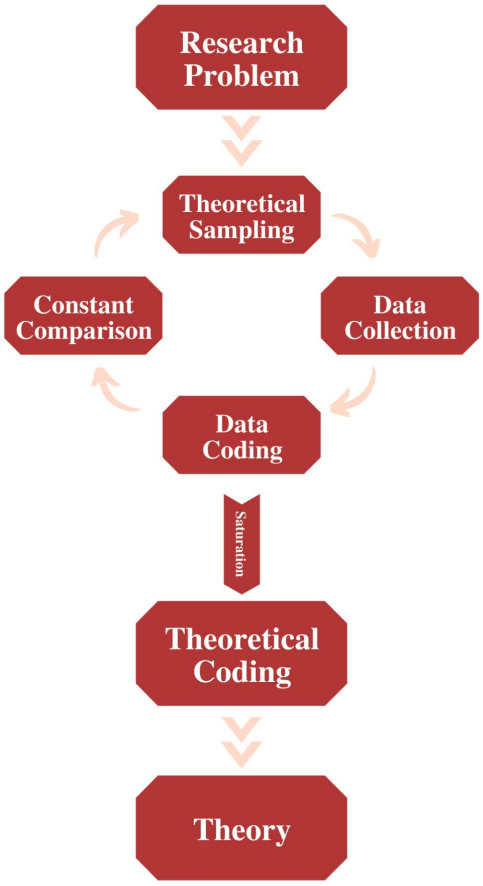

Figure 2.11: GT methodology scheme.

Succinctly, the GT is a methodology that, resourcing to iterative cycles of acquisition of data and the associated comparative analysis, generates explanations and hypothesis, grounded in practical background, not preexisting assumptions [47].

3

State of the art

This chapter provides an explanation synopsis of the current state of the art related to explainability of Multiple Sclerosis (MS) progression models. This thesis focuses on prediction Machine learning (ML) models to offer insights about the progression of MS. Thus, firstly a synopsis is presented that summarizes the developed prediction models over the years, regarding MS progression.

The summary shows all the similarities and discrepancies in the literature and the already achieved in this area of study. The project lies in the production of explanations about the work of Pinto et al. [81]. Therefore this study is then highlighted. An overview of the projects that use explainability in models related to MS is also performed. Lastly, it is shown a summary of studies about explainability in other prognosis problems in healthcare, due to the lack of literature about the creation of human-comprehensible explanations in MS progression models.

3.1 Prediction of MS progression

Some studies have already been developed that use ML to offer a prognosis of the different MS courses. Table 3.1 represents a summary created by Seccia et al. [89] of studies that use clinical data in this area, as this type of data has been characterised as relevant for long term prognosis. Besides the comparison of the different results of performance, this summary also contains information acquired about the most predictive features in each model.

Table 3.1: Summary of studies that predict MS progression with ML extracted from Seccia et al. [89].

Author year	Problem in study	Type of data; subjects	Classifiers	Performance	Most relevant variables in the best model
Bejaro et al. 2011 [5]	EDSS change>1 and EDSS range after 2 years and relapse ocurrence	MRI, MEP and clinical data; 71 + 96	Naive Bayes, decision tree logistic regression and neural network	EDSS change>1: accuracy=75% sensitivity=82% specificity=52% AUC=74% EDSS range: accuracy=80% sensitivity=92% specificity=61% AUC=76% Relapses: accuracy=67% sensitivity=53% specificity=77% AUC=65%	EDSS and MEPs
Wottschel et al. 2015 [96]	Progression of CIS to MS in 1 or 3 years	MRI and clinical data; 74	SVM	In 1 year: sensitivity=77% specificity=66% In 3 years: sensitivity=60% specificity=66%	In 1 year: lesion load, type of presentation and gender In 3 years: age, EDSS at baseline, lesion attributes (count, average distance from center of the brain, average proton density and smallest horizontal distance from the vertical axis)
Yoo et al. 2017 [97]	Progression of CIS to MS in 2 years	MRI and clinical data; 140	Logistic regression, random forest and CNN	accuracy=75.5% sensitivity=78.7% specificity=70.4% AUC=74.6%	Not studied
Zhao et al. 2017 [99]	EDSS change $\geq$ 1.5 at 5 years	MRI and clinical data; maximum of 1693	Logistic regression and SVM	accuracy=67% sensitivity=81% specificity=59%	Progressive: EDSS change, pyramidal function and its change at 1 year of follow-up, disease activity, active disease at baseline, T2 lesion volume Non-progressive: EDSS and disease activity at 0,6 and 12 months, brain parenchymal fraction, ethnicity, race and family history

Table 3.1: Summary of studies that predict MS progression with ML extracted from Seccia et al. [89].

Author year	Problem in study	Type of data; subjects	Classifiers	Performance	Most relevant variables in the best model
Law et al. 2019 [57]	EDSS change ≥ 1 in 2 years for patients with SP MS	MRI and clinical data; 485	Logistic regression, SVM, decision tree, radom forest and decision tree with AdaBoost	sensitivity=59% specificity=61% PPv=32.1% NPv=82.8%	EDSS, timed 25-foot walk and 9-hole peg test
Seccia et al. 2020 [88]	Progression from RR to SP at 0.5 to 2 years	MRI and clinical data; maximum of 1515	SVM, random forest, decision tree with AdaBoost, KNN and CNN	Random forest: accuracy=86.2% sensitivity=84.1% specificity=86.2% PPv=8.9% NN: accuracy=98% sensitivity=67.3% specificity=98.5% PPv=42.7%	Not analysed
Brichetto et al. 2020 [12]	Progression from RR to SP within 4 months	Clinical data and patient described effects; 810	Logistic regression, SVM, KNN and other linear classsifiers	accuracy=82.6%	Lack of information
Zhao et al. 2020 [100]	EDSS change ≥ 1.5 at 5 years of follow-up	MRI and clinical data; 724+400	Logistic regression, SVM, random forest and boosting methods	XGBoost: accuracy=71% sensitivity=79% specificity=69% AUC=78%	EDSS, pyramidal and cerebellar functions,MRI lesions, ambulatory index and disease course (RR, SP and PP)
Pinto et al. 2020 [81]	Progression from RR to SP at 5 years, EDSS¿3 at the 6th or 10th year	clinical data maximum of 187	KNN, SVM decision tree, logistic regression	SP development: sensitivity=76% specificity=77% AUC=86% EDSS at the 6th year: sensitivity=84% specificity=81% AUC=89% EDSS at the 10th year: sensitivity=77% specificity=79% AUC=85%	SP development: Age at onset, EDSS, FS scores (sensory, cerebellar, brainstem and mental), CNS in relapses (brainstem, pyramidal tract and neuropsychological) EDSS¿ 3: EDSS, FS scores and CNS affected functions in relapses

As demonstrated by Table 3.1, the progression of MS problem was tackled from different approaches. Some studies concentrated their efforts to predict the evolution from Clinically Isolated Syndrome (CIS) to MS [96, 97], while others developed a

prognosis of the long-term course of the disease. The difference between problems is clear. While some focus on MS diagnosis, with the prediction of cases that develop MS from CIS, others concentrate their efforts to analyse the progression of the disease, with the prediction of long-term effects through the Expanded disability status scale (EDSS) or the Secondary Progressive (SP) development. This thesis only focuses on the latter.

The analysis of the progression of the disease is made either with the prediction of the SP course from Relapse-remitting (RR) patients [12, 81, 88], or with the evaluation of the disease severity, regarding the values of the EDSS or its changes [5, 57, 81, 99, 100]. Although the EDSS values are commonly studied to evaluate the disease's severity in literature, there is a lack of coherence in terms of the identification of the threshold that distinguishes severe cases. It shows that there is still a lot of deliberation in creating the most accurate definitions, as previously mentioned in 2.1.4. There may be a lot of reasons that explain such inconsistencies, being one the availability of data and the distribution of classes, since it dramatically differs depending on the used threshold in the problem.

The used classification models can vary from one of the simplest classifiers in ML, like K-nearest neighbors (KNN), to one of the most complex and opaque methods, like neural networks. Some studies opt for models easier to explain and understand, whereas others prefer classifiers that offer better performance despite their lack of interpretability.

It is also possible to note that some researchers gave some insights about the logic of the created models, and a level of explainability, with the offer of information about the most relevant features in the predictions [5, 57, 81, 96, 99, 100]. Information about the EDSS is often considered important in the MS prognosis, classified as relevant in every study with this type of data. The functional systems (FS) behaviour, such as the pyramidal, cerebellar and mental functions, is also viewed as relevant in the prediction of the progression of MS [57, 81, 99, 100] as well as characteristics about brain lesions [96, 99, 100]. Although there are some similarities in the analysis of the most relevant features, the results show inconsistency in selecting predictive features. Such inconsistency is also visible in the study from Zhao et al. [99], where some variables were predictive when maximised a positive outcome but not predictive when the focus is a non-progressive result [89]. Despite not being possible to select the universally important features, due to the demonstrated heterogeneity in these studies, this type of analysis is extremely important. The presented information gives necessary insight about the models, even if insufficient to apply them in the real-world. However, such results start the path to comprehend

MS models and to increase the knowledge of the disease's progression overall.

3.1.1 Framework from Pinto et al.

All the created methods are applied to the models developed by Pinto et al. [81] framework. It is important to particularly highlight and explain this study as it is directly associated with the developed thesis. As mentioned before, Pinto et al. [81] developed three frameworks, one to predict a SP course in patients that appear to have RR MS, one to predict the severity of the disease by identifying benign and severe cases in the 6th year of follow-up and the last to predict the severity of the disease in the 10th year after the diagnosis.

This framework was created with k-fold Cross Validation (CV) with a k=10, where this process was repeated 10 times. Therefore with 100 runs, 100 different models were developed in order to explore the whole data and give some security about the obtained performance. To select the most relevant features, firstly the Pearson's linear correlation coefficient was applied to select the best 100 features and then the Area Under the Curve (AUC) to select the best 50. Lastly, the Least Absolute Shrinkage and Selection Operator (LASSO) method was implemented to select the optimal predictive set of features. Posteriorly, some classifiers were used in the predictions, where the linear Support Vector Machine (SVM) was considered the one with the best results. The used pipeline in this study is represented in Figure 3.1.

The data used to create these models consists in static and dynamic data, about patients with RR and SP courses provided by the Neurology Department of Centro Hospitalar e Universitário de Coimbra (CHUC). Patients with Primary Progressive (PP) MS were excluded from these studies, since this course presents distinct manifestations that make it easily identifiable.

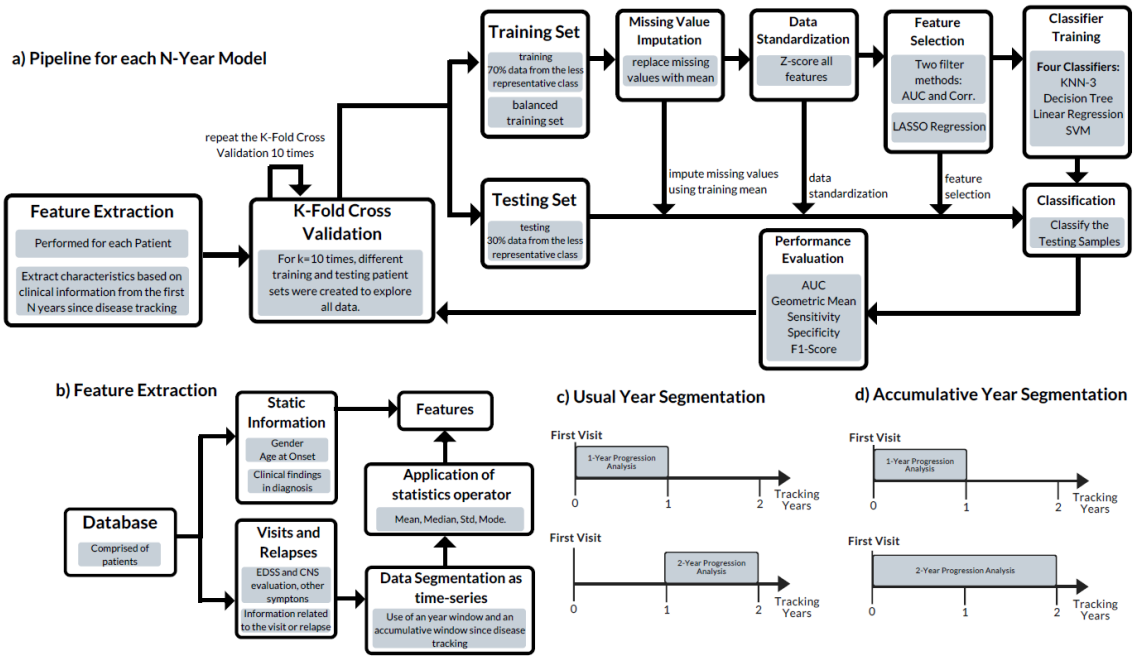

Figure 3.1: ML pipeline from Pinto et al. 2020 [81].

Database description

After a MS diagnosis, a patient is medically monitored through routine consultations every 3 to 6 months, depending on medical opinion. In these appointments, the patient's neurological status is analysed and registered. During unscheduled visits due to relapses, corticosteroids (methylprednisolone) are administered as treatment, and information about which areas of the central nervous system are affected and the severity of the relapse is documented. These data were temporal segmented and afterwards transformed into features. Each resultant dynamic feature represents the information from the visits of a specific year (e.g. the EDSS mean of the first year) or the accumulated information of several years, up until a specific year (e.g. the mean of the visits from the first and second year). Their used values are the statistics mean, median, mode and standard deviation of the considered temporal windows (years).

The static data used in the Pinto et al. studies [81] is acquired at baseline when the diagnosis is obtained being the used information the following characteristics:

- **Gender**;
- **Age of Onset:** calculated using the date of birth and date of diagnosis;
- **Date of birth**;
- **Date of diagnosis**;
- **Supratentorial:** boolean field (yes/no). It specifies if there are initial mani-

festations of MS related to the supratentorial region;

- **Optic Pathways:** boolean (yes/no). It specifies if there are initial manifestations of MS related to optic Pathways;
- **Brainstem-Cerebellum:** boolean (yes/no). It specifies if there are initial manifestations of MS related to the brainstem and/or cerebellum;
- **Spinal Cord:** boolean (yes/no). It specifies if there are initial manifestations of MS related to the spinal cord;
- **Clinical findings:** boolean (yes/no). It specifies if there are initial clinical evidence of manifestations of MS;
- **Magnetic resonance imaging (MRI)**: boolean (yes/no). It indicates if there are MS manifestations visualized in MRI scans (lesions) at baseline;
- **Evoked Potentials:** boolean (yes/no). It indicates if there are MS manifestations visualized in the evoked potentials test at baseline;
- **Cerebrospinal Fluid (CSF):** boolean (yes/no). It indicates if there are MS manifestations visualized in the lumbar puncture exam at baseline.

The dynamic data can be divided into visits and relapses. The used information in the developed models from the database and their missing data ratio is presented in the following Table 3.2.

Table 3.2: Database information concerning visits and relapses, the dynamic information, that was used in the models [81].

Characteristic	Description	Missing data ratio
Visits		**0.00**
Visit Date		0.00
Routine	boolean (yes/no): routine visit;	0.00
Score Pyramidal	numeric (0-6): Pyramidal FS score;	0.20
Score Cerebellar	numeric (0-5): Cerebellar FS score;	0.20
Score BrainStem	numeric (0-5): Brain Stem FS score;	0.20
Score Sensory	numeric (0-6): Sensory FS score;	0.20
Score Bowel & Bladder	numeric (0-6): Bowel & Bladder FS score;	0.20
Score Visual	numeric (0-6): Visual FS score;	0.20
Score Mental	numeric (0-5): Mental FS score;	0.20
Score Ambulation	numeric (0-12): Ambulation FS score;	0.20
Cerebellar Weakness	boolean (yes/no): MS manifestations of cerebellar weakness	0.00
Visual Symptoms	boolean (yes/no): visual symptoms;	0.00
gdAtaxia	boolean (yes/no): manifestations of gait disturbances related to ataxia;	0.00
dysaesthesiae	boolean (yes/no): manifestations of dysaesthesiae;	0.00
ataxiaLowerExtrem	boolean (yes/no): manifestations of ataxia in the lower extremities;	0.00

Table 3.2: Database information concerning visits and relapses, the dynamic information, that was used in the models [81].

Characteristic	Description	Missing data ratio
paresthesiae	boolean (yes/no): manifestations of paresthesiae;	0.00
cognitionPb	boolean (yes/no): manifestations of perturbances in cognition;	0.00
gdParesis	boolean (yes/no):manifestations of gait disturbances related to paresis;	0.00
gdSpasticity	boolean (yes/no): manifestations of gait disturbances related to spasticity;	0.00
mwUpperExtrem	boolean (yes/no): manifestations of muscular weakness in the upper extremities;	0.00
micturitionPb	boolean (yes/no): manifestations of perturbances in micturition;	0.00
fatigue	boolean (yes/no): manifestations of fatigue;	0.00
mwLowerExtrem	boolean (yes/no): manifestations of muscular weakness in the lower extremities;	0.00
moodPb	boolean (yes/no): manifestations of mood perturbances;	0.00
EDSS	numerical (0-10): the EDSS value;	0.00
Relapses		**0.04**
Relapse Date		0.00
Impact ADL Functions	boolean (yes/no): impact on activities of daily life;	0.00
Recovery	boolean (yes/no): the patient recovered from the relapse;	0.83
Severity	numerical (0-2): relapse severity;	0.79
CNS Pyramidal Tract	boolean (yes/no): MS manifestations related to the Pyramidal tract;	0.00
CNS Brain Stem	boolean (yes/no): MS manifestations related to the Brain Stem;	0.00
CNS Bowell & Bladder	boolean (yes/no): MS manifestations related to Bowel and Bladder;	0.00
CNS Neuropsycho Functions	boolean (yes/no): MS manifestations related to Neuropsycho functions;	0.00
CNS Cerebellum	boolean (yes/no): MS manifestations related to the Cerebellum;	0.00
CNS Visual Functions	boolean (yes/no): MS manifestations related to Visual Functions;	0.00
CNS Sensory Functions	boolean (yes/no): MS manifestations related to Sensory Functions;	0.00
Hospital	boolean (yes/no): the relapse required hospitalization;	0.00
Ambulatory	boolean (yes/no): the relapse affected ambulatory capacity;	0.00

3.1.1.1 Features' Predictive Power

The analysis of the data set variables' predictive power by Pinto et al. [81] was executed by counting how many times a specific characteristic, e.g. the *EDSS* or the *ScorePyramidal*, was selected in the feature selection as part of the classification model input in the 100 runs. For example, if a characteristic has the predictive power of 0.67, it was selected as part of the input 67 times in 100 models. Moreover, this information is linked with the type of relationship that each characteristic has with the target, positive (+) if it relates with a benign outcome or negative (−) if it promotes a severe case of MS. Pearson's correlation between features and the target classification output was calculated to study such relations. If the correlation is negative, the feature promotes a mild case, while if the correlation is positive, it is associated with leaning the result to a malignant outcome.

The predictive power results show that the EDSS and the scores of the pyramidal, mental, cerebellar, and bowel and bladder systems have the highest predictive power, which are the only variables with a constant relevance over the studied years, as represented in Figure 3.2. Although Age and Gender appear to have a significant impact in the 1-year framework, that influence does not appear in the other models.

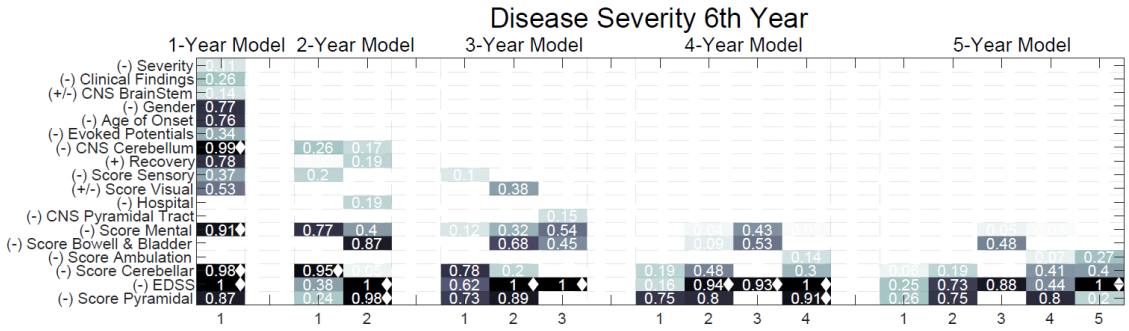

Figure 3.2: Predictive power of each characteristic. The values represent the recurrence of the characteristics in the 100 runs, being the predictive power superior to 0.90 represented by diamonds. The signs represent the influence that the variables have in the classifications: positive (+) if it promotes a good prognosis and negative (−) if it promotes a severe outcome.

Most characteristics negatively influence the outcome, i. e. they promote a severe case of the disease, except in the characteristic Recovery, as expected. However, the characteristics CNS BrainStem and Score Visual do not show a clear relation with the predictions. The authors explain the indicated results by affirming that patients who persistently display these symptoms occasionally have a higher probability of having mild disease effects. In contrast, the patients that suffer from these symptoms irregularly are more likely to be associated with a severe case of MS [81].

In general, the results are in accordance with the literature and what is already known clinically. However, there are still some discrepancies between studies regarding what variables are the most relevant to the MS prognosis.

This shows not only the evolution in the predictive power of the characteristics but also the type of influence that they have in the outcome (positive or negative). This analysis is prevalent among these studies, since it provides fundamental insight into the models. Nevertheless, the method does not consider feature's interactions that are frequently present in complex problems. Moreover, it is not possible to directly compare the recurrence of a characteristic with its impact on the prediction as a characteristic may only be selected as input, e.g. five times in 100 runs, but may significantly influence the classifications of those specific models. Moreover, the amount of analysed patients causes doubts about the conclusions of this study, lacking further tests in other databases.

3.2 Explainability and MS

Recently, a few applications of explainability in MS-related problems have emerged that show the rising need to study such concepts in this area. However, the encountered research on this topic is not aimed to investigate the disease's progression yet, but to explain diagnosis models instead. Therefore, although the studies presented are not directly compared to this thesis, they also help to comprehend further MS-related models.

Aiming to create an understandable model that diagnoses the disease, Eitel et al. [26] created an explainable framework with 3D convolutional neural networks (CNNs). The authors implemented Layer-wise Relevance Propagation (LRP), a method that creates heatmaps from the holdout sets that show the relevance of every voxel in an individual decision [10] for each subject. Dealing with Fluid-attenuated Inversion Recovery (FLAIR) image sequences, the authors used LRP to validate the results of the model by comparing the most relevant features with the already clinically used markers in MRI images of individual predictions. This comparison showed that indeed the identified features were consistent with the clinical knowledge, able to distinguish features only discovered in MRI advanced techniques. The created explanations proved that the model was learning what was supposed to for those samples. However, since the used data is relatively small, it is necessary to do further analysis with more extensive and diverse data sets, to assure that the results are robust and able to generalise to other samples. The used method does not consider voxel interactions nor has information about the inherent mechanisms of

the neural networks. However, it showed that it is capable of learning expected MS markers from FLAIR image sequences by explaining the results in a very clear and intuitive way with heatmaps. This study is, therefore, very important, as this type of research helps to decrease the scepticism that surrounds neural network models in the diagnosis of MS.

The LRP was also used by Creagh et al. [20] in a transfer learning Deep Convolutional Neural Network (DCNN) framework. This framework uses collected ambulatory time-series data from wearable smartphone sensor data to distinguish healthy individuals, patients with mild MS and patients with moderate MS. This analysis made it possible to find relevant ambulatory patterns for diagnosis, such as the gait power. Through the results provided by this method, it was also possible to observe a clear distinction between healthy patients, patients with mild MS and moderated MS in the gait domain. It visually demonstrated the distinct characteristics of the signals for the three different classifications very intuitively. This work used two opened-source data sets, also having problems regarding low-subject data set that the authors helped combat with the implementation of transfer learning. Despite the relevancy of the LRP explanations, the evaluation was only visual, influenced by clinical hypothesis. Thus it is necessary to conduct and analyse these studies in a more controlled manner.

With demographic and MRI information, as well as disease covariates, Reinhold et al. [85] developed a structured causal model based on counterfactual images and variational autoencoders to help diagnose and make a prognosis of MS cases. With a limited data set as well, the study focused on the analysis of counterfactual images regarding questions like 'what happens when the lesion volume is equal to 0 mL?' or 'if the EDSS of this patient was 4, how would the MRI be?'. With exciting results in the training test, the developed model created weak counterfactual explanations in the test and validation sets, which shows its inability of generalisation. Additionally, the counterfactual explanations were not validated significantly, since it is challenging to evaluate hypothetical outcomes caused by imaginary transformations to a specific variable. Despite its flaws, this type of work could possibly be used in the future to assess MS dynamics, as the created explanations might give insights into the evolution of the MS courses.

Even though these studies [20, 26, 85] also focus on MS, their efforts are concentrated mainly in the creation of explainability in diagnosis models, which deviates from this thesis objective. Additionally, there are no similarities in the input data, as this project uses longitudinal clinical data collected from hospital admissions without image information. However, all of them have a fundamental role not only to

help ML models to gain trust from both the medical and computer science experts but also to expand the knowledge about this very complex disease that torments so many.

3.3 Explainability in healthcare models

Since this study is not comparable with other research about ML models that predict MS progression, the next best thing is to find research that applies explainability methods in healthcare, that may present an analogous concept to this thesis. Therefore the research further focused on problems that concentrate on applying explainability in black-box frameworks related to medical issues with characteristics somewhat comparable with MS or the type of problem studied. Therefore, the focus of this research lies in the prognosis of a health problem with ML models, considering data from different time-period moments. This study consisted of the search of articles linked to the words 'machine learning' and 'interpretability' or 'explainability', and 'neurodegerative disease', 'autoimmune disease', 'prognosis', 'risk' or 'healthcare'. Table 3.3 summarises the results of this research.

With this research, it is possible to note the prominence of the use of SHapley Additive exPlanations (SHAP) in black-box models [1, 3, 4, 9, 27, 45, 56, 63, 73, 80, 92]. Methods that are based in the Shapley values are often characterized by the creation of the only explanations that guaranty a fair distribution of effects of each feature [72], a very appealing advantage that may justify its popular implementation.

The Local Interpretable Model-Agnostic Explanations (LIME) are also frequently applied in these problems, since they offer similar information to the Shapley values but with less computational costs. It is important to note that some SHAP methods connect the data of Shapley values and LIME without being computationally expensive. All the advantages related to Shapley values, together with the fact that they offer prediction models and are easy to implement in decisions trees, make SHAP very popular [72].

There is a predominance of more complex prediction classifiers, like neural networks and Extreme Gradient Boosting (XGBoost), representing their significant popularity associated with higher performance results than simpler models.

It is possible to observe the range of different problems and medical areas that include explainability to the discussion. From all-causes mortality risk problems to neurodegenerative diseases diagnosis, understanding black-box models is essential, since patients health is at risk if those models fail.

It should be noted that there is no direct comparison of the explanations given

Table 3.3: Studies associated with explainability in healthcare.

Author, year	Problem in study	Data Type	Recurrence of data acquisition	Classifier	Explainability methods
Antoniadi et al., 2021 [3]	Quality of life prediction for ALS patients	Patient and caregiver 's interviews information and clinical data	3 time-instances with 4 to 6 months intervals	XGBoost	SHAP
El-Sappagh et al., 2021 [27]	Early diagnosis of Alzheimer's disease; Progression from MCI to Alzheimer within 3 years	MRI, PET and clinical data	Baseline	Random Forest	SHAP + 22 explainers based on decision trees and fuzzy rule-based systems
Bloch et al., 2021 [9]	Progression from MCI to Alzheimer without time of conversion defined	MRI and demographic data	Baseline and follow-up	XGBoost and Random Forest	SHAP
Maggesh et al., 2020 [65]	Parkinson Diagnosis	SPECT scan images	Initial screening of each patient	CNN	LIME
Peng et al., 2021 [65]	Mortality risk for hepatitis patients	Clinical data	Baseline and follow-up	Random Forest	SHAP, LIME and PDP
Moncada-Torres et al., 2021 [73]	Prediction of breast cancer survival	Clinical data	Baseline and follow-up	Random Forest, SVM and XGBoost	SHAP
Prentzas et al., 2019 [82]	Stroke prediction	Clinical data	Every 6+ months after recruitment	Random Forest	Decision rules from Georgias framework
Lv et al., 2021 [63]	All-cause mortality; All-cause readmission for patients with heart failure	EHR	All information from the first hospitalization due to heart failure	Random Forest, linear regression, SVM, ANN and XGBoost	SHAP
Cho et al., 2019 [17]	Post-stroke hospital discharge	Clinical data	Information of primary stroke diagnosis	Logistic regression, Random Forest, AdaBoost and MLP	LIME
Athanasiou et al., 2020 [4]	Cardiovascular disease risk in Type 2 Diabetes patients	Clinical data	5 years of follow up	XGBoost	SHAP tree
Jiang et al., 2020 [45]	In-hospital mortality risk in sepsis survivors	Clinical data	first day of readmission data	light gradient-boosting tree	SHAP and PDP
Epifano et al., 2020 [30]	Mortality prediction in septic and all-comers patients	Test results	first-day admission	DNN	Influence functions
Thorsen-Meyer t al., 2020 [92]	90 day mortality prediction	Longitudinal, static and time-series data	Baseline and follow-up	LSTM	SHAP
Lauritsen et al., 2020 [56]	Acute critical illness prediction	EHR	24h admission records	TCN	SHAP
Lundberg et al, 2018 [62]	Prediction of hypoxaemia during surgery	Times-series, dynamic and static data	pre and during surgery data	Gradient Boosting	PDP, Shapley values and averaged feature importance
Agius et al., 2020 [1]	Risk of infection prediction in chronic lymphocytic leukemia patients	Clinical data	Baseline and follow-up	Random Forest, elastic networks, logistic regression and XGBoost	SHAP

by different methods in most of these studies. Thus, it is difficult to evaluate if they are the best explanations for the addressed problems or if there are better alternatives yet to analyse in the future. Since the explainability and interpretability concepts are still in development, most of these studies are just the beginning of the investigations. This newness can be demonstrated by observing the year that the mentioned studies were conducted, where the majority of studies were published in

2019, 2020, or in the present year.

Although there is already some research about explainability in medical problems, there is still a long way to go and a lot more to develop. Not only in MS but also healthcare problems in general, as it is already known that, in some areas of study, the information provided by ML black-box models is not enough.

3.4 Final observations

The inclusion of information about the relevance of features is currently a popular analysis in the available studies about MS prediction models, as well as in other medical issues. However, such research does not consider feature interactions that may drastically influence the conclusions about the model's intrinsic logic. It may give a false sense of simplicity about the dynamics of the models and the disease itself.

The need to guarantee that the prediction models operate with causal thinking is also fundamental, but not completely achieved only by analysing the most predictive features. It is imperative to assure that a model is learning correct reasons to classify a sample and not random links between features and target.

With the increase of information about explainability methods, it is believed that the production of clear explanations can combat the aforementioned limitations. Explainability is already a focal point of study in some healthcare problems. However, it is not a popular approach yet, particularly in MS progression models. The necessity of studies such as this is evident.

These studies commonly create explanations by implementing one or two methods without the analysis if the obtained insight is sufficient to turn a model clearly explainable. The importance of the comparison of multiple explainability methods in the same problem must be reported, as some methods might be more suited than others or might complement other types of explanations.

Additionally, most available studies lack an evaluation of the produced explanations, a common problem, due to the complexity of this topic. The indicated limitation needs addressing, regardless of its difficulty, to provide full validation to all the work already achieved and to put ML models a step closer to being a resource in clinical environments.

4

Methodology

The main goal of this thesis is to understand if the developed work can be applied in a clinical context and what type of produced explanations can influence the most to achieve such accomplishment.

The models developed by Pinto et al. [81] focused on three different predictions, the disease's severity at the sixth and tenth years of follow-up, and the Secondary Progressive (SP) development. Since the available information increases over the years, in each problem, n-year models were created, each using accumulative information up until the respective year. The analysed data consisted of clinical annotations from the first to the fifth year of follow-up, depending on the n-year model generated. The 1-year model studied data from the first year of follow-up, while the 5-year model used accumulative information from the first to the fifth year of follow-up.

The number of different models could cause considerable entropy if all data was studied. Therefore only one model was selected as the focus of this thesis. Firstly the problem regarding the prediction of the disease's severity in the sixth year was chosen. The higher performance of these models was one of the reasons for this decision. Additionally, in the previous work, it was possible to observe that the prediction of the Multiple Sclerosis (MS) severity had more discriminant features than the SP progression. Due to its complexity, this case is intriguing to analyse the true potential of the explanation methods. Considering the different n-year models, the authors assumed that the model about information up until the second year of follow-up presented the best trade-off between the value of performance and the time of the prediction, as represented in Figure 4.1. It is imperative to give a correct prognosis the quickest as possible, since a later prediction and, consequently, a late therapy administration might be ineffective in inhibiting the disease's progression.

With that in mind, the taken approach started with the implementation of different explainability methods that produced explanations about the model that predicts the disease's severity in the sixth year of follow-up, using the clinical infor-

mation of the first two years after diagnosis.

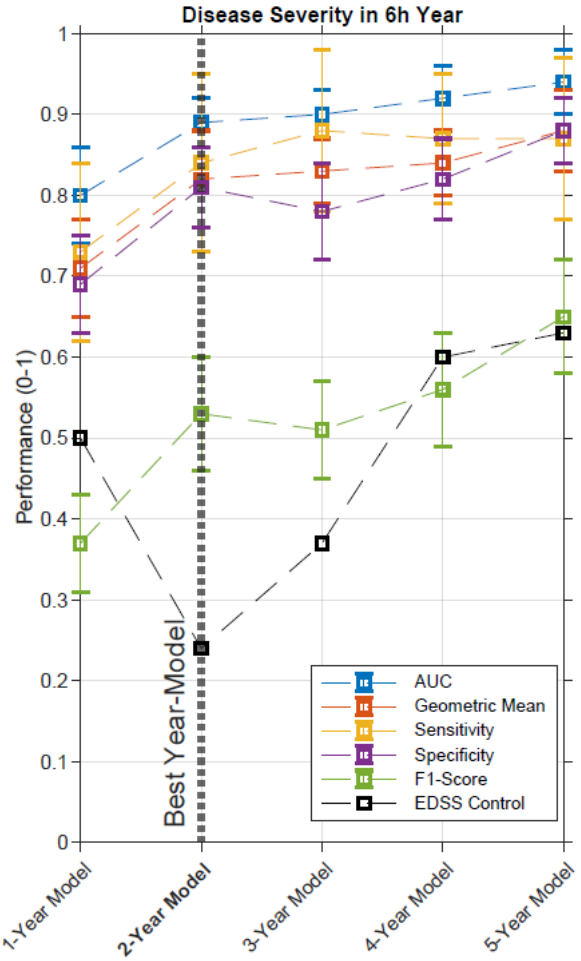

Figure 4.1: SVM classifier performance in every year-framework.

The produced explanations were then compared with each other and with what is already clinically known about the MS progression and its characteristics. Then the results were presented to a group of people that daily work with Machine learning (ML), with different levels of knowledge, to be evaluated. This evaluation focused on understanding if the developed work can guarantee trust and safety if used in the real-world. The produced explanations were presented, followed by an interview of open questioning to know their opinions about the different explanations and understand the reasons that support such responses.

4.1 Model agnostic explainability methods

With all the information about the created models from the Pinto et al. framework [81], several explainability methods were implemented. Such methods can be divided into three subgroups, depending on the type of explanations they generate. Some methods explain the framework's logic, give universal insights and possible inconsistencies about all of the 100 originated models, and methods that only present the global behaviour of specific models. Additionally, methods that produce local explanations about individual predictions of specific samples were also developed and analysed.

The implementation of these methods was performed in MATLAB® version R2021a with the Statistics and Machine Learning Toolbox.

4.1.1 Framework-related methods

These methods give general explanations observed in every model and offer some analysis of the differences between them. The logic of a model directly relates to the input data of the classifiers. In a real-world environment, it is improbable to know a priori the characteristics of the models' input. Thus, it is not possible to conclusively know what is the best model, in practice. Therefore, it is important to understand if, in general, the models are consistently learning the same relations in different sets of data, or if there is a lack of coherence between them.

Three methods were developed that give information about all models, the analysis of the recurrence of features, the calculation of the permutation feature importance, and the analysis of the partial dependence plots (PDPs). Furthermore, it is essential to mention the study of the predictive power by Pinto et al. [81].

Although this analysis was not developed in this thesis, it also considers and evaluates these results. The information given by the indicated method is a form of insight about the created framework and the disease's dynamic itself. Therefore, it is valuable information for the comparison of the different forms of understanding.

Similar to the feature predictive power, the first method executed in this thesis was the analysis of the recurrence of the features obtained in their selection. The goal was to conclude if an entire input set of features stands out in the 100 runs, i.e. if the full selected combination of features is significantly repeated. Additionally, this method aims to possibly show which partial combinations of features are the most recurrent, to give insight into the models' granularity. Firstly, this procedure counted how many times a single feature, and combinations from two to 10 features,

were part of the input in the 100 models. The limit was set to 10 because it was the threshold in Least Absolute Shrinkage and Selection Operator (LASSO), the last technique used in feature selection. Then, the best results were analysed and presented in the form of a dendrogram. Thus, the links were analysed starting from the smallest combination (pairs), adding the best following combination (trios) that contains the previous variables and so on until it is impossible to distinguish the most recurrent group of features.

The following developed method was the permutation feature importance, a method that measures each variable's importance in the classification ability of generalisation. This method was applied by permuting the values of the test data set of one specific feature and calculating the loss of performance in the new classification results by calculating the difference between the original G-mean of the model and the G-mean of the permuted one. The permutation was created by altering the samples' values order of the studied feature and predicting the altered data set again. The permutation order was randomly assigned, which was repeated 200 times for each feature in each model to stabilise the results for each feature in different runs. This analysis was performed in each feature and combinations of two variables. The function *predict* was used with the new permuted data set but with the original model to predict the new altered instances. The same method was implemented with the train data set, to evaluate the feature importance in the models' learning ability. As the obtained values were similar to the results with the test data set, this work presents only the results of the test data set.

Lastly, the PDPs of the most recurrent features from the previous analysis were produced, as well as the most recurrent pair of variables. Therefore, in each model that contained the analysed variable, its PDP was calculated and registered with the use of the function *plotPartialDependence*. Subsequently, it was calculated the mean value of the obtained PDPs from every model and presented a plot with all the PDPs for each model and their mean value. Although Individual Conditional Expectation (ICE) explanations were considered and developed, the results were discarded from this thesis, due to the lack of increase of information when compared with the PDPs explanations, as the results showed that all the effects were equal in the different points.

4.1.2 Model-specific global explanations: Linear regression

The only method that explains model-specific mechanisms in this project is the classifier linear regression. The function *fitlm* created the linear regression models, in which the classification labels of the data test were acquired by the use of the

function *predict*. As there is no highlighted unique recurrent model in the 100 runs, the selection of the to be analysed models consisted of identifying the runs with the best and worst values of G-mean and one model with the most similar G-mean to their mean value. The regression models can be represented by a function with linear relations between the features and the target, where the learning task focus on finding what the most appropriate estimate of coefficients for each variable that best fit the data is:

$$y = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + ... + \alpha_n x_n + \varepsilon \tag{4.1}$$

To distinguish each class, the used criterion was if the function's value was equal or higher than 0.5, the analysed point belonged to class 1. In contrast, the prediction indicated a benign outcome if the function's value was lower than 0.5. With the inherent information resultant from the function *fitlm*, the estimated coefficients (α) of each feature ($x_i i$ = 1,...,n) from the selected models were then documented in a bar plot. This study was also performed in every model, where the total coefficients per variable and their mean were collected. Since these results had a significant variability, the analysis of every model was also excluded from the posterior comparison and evaluation.

4.1.3 Sample specific explanations

Some local methods were implemented to create specific explanations. To produce them, the first procedure was to find what Support Vector Machine (SVM) models had the best, worst and average value of G-mean, being the methods only related to them. With these models, the next move was to find representative data points to which the created explanations were applied. The test data set was divided into points correctly classified as benign (class 0), points correctly identified as malignant (class 1), and the misclassifications to select the indicated samples in each model.

To find a group of points that represent all the data distribution, in each set of samples, a dendrogram of points' similarity was created using the functions *pdist, linkage* and *dendrogram* using the euclidian distance to compare them. With the observation of the dendrograms, 3 to 4 clusters were manually created, since the goal was to analyse 3 to 4 points from each type of points (points belonging to class 0, 1 and misclassifications). Due to the non-systematic distribution of samples in each group, particularly in the observation from class 0, it was not possible to use a clustering technique with a relative homogeneous distribution of points per cluster.

Thus the division was executed manually delimited subjectively. This selection was only created as it was impossible to study every sample in this thesis.

In each cluster created, the point with the highest classification score in module was selected to be posteriorly studied. Figure 4.2 presents an example of this division. The classification score is directly linked with a confidence level in each classification, i. e. the higher of score is in module, the more distance the point has with the decision hyperplane. Therefore, the greater the classifier's certainty to identify the sample as a part of the determined class.

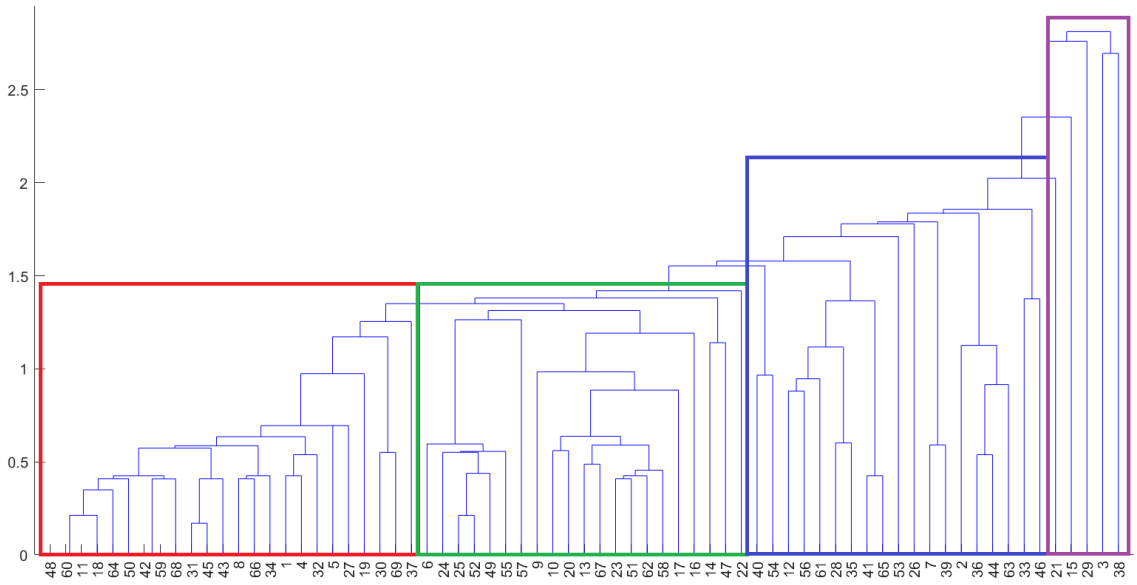

Figure 4.2: Obtained dendrogram about the similarities of points that were correctly classified as benign cases (class 0) in the model with the best g-mean. The manually selected clusters are delimited by the coloured rectangles, being each point represented by each tick in the x-axis.

With all the points selected, three different methods were developed.

The first method studied was the Local Interpretable Model-Agnostic Explanations (LIME). These method's explanations were created by using the function *lime* with the parameter 'DataLocality' defined as 'local'. The number of neighbours necessary to develop LIME was defined with a *rule of thumb*, equal to the round value of the square root of the complete test data set. The indicated choice was made because the optimisation of this parameter is complex, and there was a lack of knowledge to make a more rigorous decision. This function generates synthetic points regarding the predictor data and uses them to create a local intrinsic interpretable model. In this case, it was linear regressions. Then, the function *fit* was used to fit the local simple linear model to the studied point being its results visualised when used the function *plot*.

The Shapley values were also calculated for the selected observations using the function *shapley, fit* and *plot* in the respective order. Only the predefined parameters were used.

The final approach was to develop counterfactual explanations in the form of 'If the feature X had the value X1, then the outcome would be Z, not Y'. The goal is to know the minor change in the features' values that can alter the prediction. With this in mind, firstly, each variable is independently altered, according to its domain. The step used to vary the values of the features systematically also depends on each feature. The domain and step used in the variations are presented in Table 4.1. After analysing all the possible changes in each feature, only the top three minor alterations of values were selected as explanations. This method does not study changes of combinations of features.

Table 4.1: Values' domain and selected intervals for each characteristic

Variables' identifying words	**Domain**	**Interval between values**
EDSS (mode)	0 - 10	0.50
EDSS (mean, median, std)	0 - 10	0.20
Scores Pyramidal, Visual, Sensory and Bowel & Bladder	0 - 6	0.10
Scores Mental, Brainstem and Cerebellar	0 - 5	0.10
Score Ambulation	0 - 11	0.20
Supratentorial, Optic, Spinal, BrainStemCerebellum, clinical findings, evoked potencials, CSF, MRI, and gender	0 - 1	1.00
Age	0 - 100	1.00
Other	0 -1	0.05

4.2 Evaluation of the developed work

After implementing so many methods, the principal question is 'how to evaluate this information?'. With this in mind, the initial goal was to present the data to specialists, both in the MS-related clinical field and to the computer science environment. However, it became clear that the explanations need to be different

for each context since the concerns of the data scientists differ from the ones of a clinician. The explanations presented to data scientists need to probably be more complex in terms of an algorithmic perspective than those shown to clinicians. In addition to the challenge to contact clinicians that directly work with MS, in the available time, it is essential first to assure the models' robustness and confidence in an algorithmic perspective. Therefore, the evaluation was firstly performed at a data science level before being analysed in a clinical context.

This evaluation consisted of a presentation for computer scientists of how the prediction framework was created and the explanations produced. After the presentation, the scientists were interviewed to analyse all the developed work qualitatively. Ten ML scientists were contacted and accepted to be part of this analysis, where eight of them work on the laboratory responsible for this project and two are external investigators. No participant had prior knowledge about what was developed in this thesis.

The presentation consisted of explaining how the prediction framework was created, the performance results of the generated models, and all the explanations produced by implementing the mentioned methods. This presentation was made individually to each participant. Afterwards, each scientist was asked to answer some questions about all the created work. Some results were discarded from this presentation as it had a limited time per person. It would be impossible to fully explain all the data without being too exhausting to the listeners. Therefore only the most relevant aspects of each method and explanations were presented. The interviews were all recorded to posteriorly being analysed. The interview was based on the following questions:

- Regarding the ML context, are the prediction models trustworthy?
- Is all the developed work fit to be applied in a clinical environment?
- What can be improved regarding the prediction framework?
- What explanations are more suited to support the work's decision making?
- Is there any limitations linked with the produced explanations?
- Do different types of explanations give different types of trust? What are they?
- Are there any unfilled gaps in all this work?
- What could be improved and developed to combat those limitations?

After all these questions, it was asked if there was any missing observation that the participant wanted to expose. These statements were also taken into consideration. While listening to the recordings, all the data was transcribed into text

data. Through an iterative analysis, based on Grounded Theory (GT), the emergent topics in the interviews were found and documented, up until the saturation of the data, i.e. when no new information emerges from the interviews. If the data did not saturate with ten interviews, more of them would be performed to achieve this goal. All the emergent topics were afterwards organised in a diagram to explain the obtained results.

5

Results

This chapter shows the produced explanations by the different methods and the results acquired in the interviews. The explanations are divided into three groups, according to the type of given information. Therefore, it is first presented general explanations about all the 100 models, afterwards global explanations about a specific model and, lastly, the methods that generate explanations about a single prediction. The name of the features represents the type of information associated with them, where the number related to a *y* represents the analysed year by the variable, and the *acc* shows that the feature is associated with accumulative information of several years. For example, the feature *EDSS mode 2y acc* is the mode value of the EDSS documented in the visits from the first and second year of follow-up.

5.1 Framework-related explanations

The analysis of the behaviour of all the models started previously with the study of the predictive power by Pinto et al. [81]. The predictive power is described as the recurrence of each characteristic in the 100 runs, associated with the type of influence that the variables have in the prediction (positive, if they promote for a mild case, or negative if they promote for a severe course of Multiple Sclerosis (MS)). The results are presented in Figure 3.2.

The analysis of the recurrence of the features' combinations is represented in Figure 5.1. In the scheme, the numbers represent the number of times a variable or combination was selected in the 100 runs as input of the classifier. The intermittent lines represent the features associated with accumulative information. The saturation is directly related to the links' recurrence, e.g. a combination with 67 selections of input has a saturation equal to 67% and the thickness of each line, which is directly associated with the year of follow-up that the feature is related to. The links on the left side of the scheme represent features with the same studied characteristic.

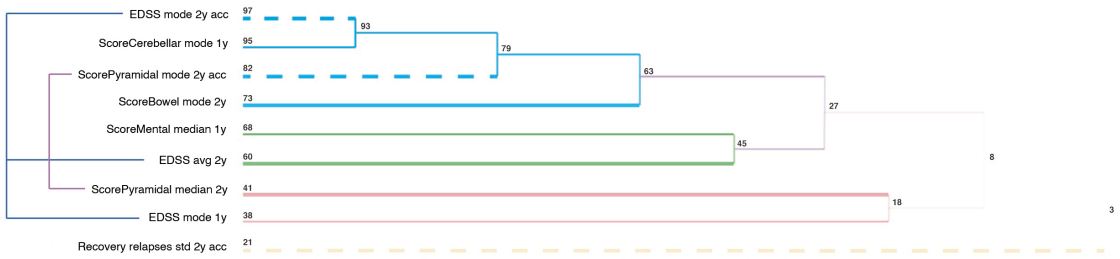

Figure 5.1: Dendrogram-like scheme that shows the recurrence of the variables in the 100 models.

Similarly to the predictive power, the results of the recurrence of variables and their combinations show that the Expanded disability status scale (EDSS) and the Scores of some Functional System (FS) (pyramidal, cerebellar, mental and bowel and bladder) have the highest results. It is relevant to highlight the blue and green clusters due to their prominent recurrence. The feature *EDSS mode 2y acc* is selected in almost all the models as well as the *Score Cerebellar mode 1y*.

The results do not significantly differ from the predictive power as the analysis is based on the same logic, the number of times a variable is selected to be a part of the input in the classification. However, since this method has in regard combinations between features, it enables to observe possible variables' interactions, and therefore it is an asset. The analysis of the models' granularity provides a path to an interesting study that may be worth investigating more thoroughly. However, the indicated explanations lack insight about the relations between variables and target, if positive or negative. This data can easily compare what a model is learning and what is observed in the real-world.

In Figure 5.2, the permutation feature importance shows the combinations with the highest results, where the feature *EDSS mode 2y acc* is related with most interactions. It is also the feature with the highest importance value when analysing individual features. The combinations with the highest loss of performance were the pairs *EDSS mode 2y acc - ScoreVisual median 2y acc* and *EDSS mode 2y acc - ScoreAmbulation avg 2y*. However, these combinations are only present in 1 and 2 models, respectively.

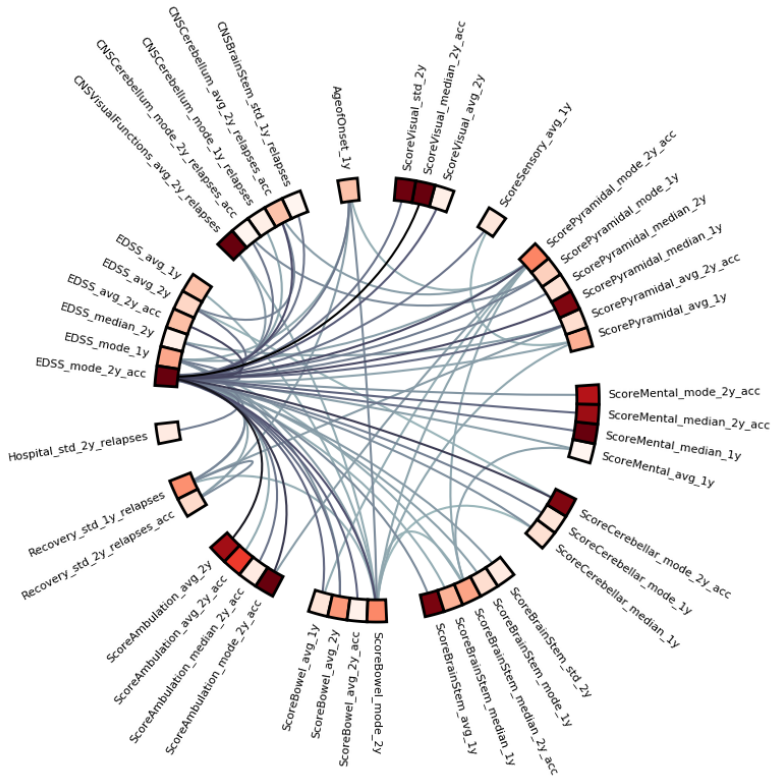

Figure 5.2: Permutation feature importance. Only the interactions with a value higher than the mean value and the standard deviation are demonstrated. Each interaction is represented by the links between features and each coloured rectangle the importance of the correspondent feature. The darker the colour, the higher is the importance value. The individual values vary from -0.039 to 0.100. The values from the links vary from 0.070 to 0.234.

It is important to notice the predominance of the characteristics EDSS and scores of some FS e.g. the scores of the pyramidal system. Although some features individually do not significantly impact the predictions, when combined with other features, they greatly influence the forecast, associated with a high loss of performance. It is possible to verify this by observing, e.g. the interaction *CNSCerebellum mode 2y relapses acc - ScorePyramidal mode 1y*.

This technique offers an effortless way to evaluate the features' importance and combinations. It is easy to understand that, if the values' alteration of a variable triggers an accentuated loss of performance, its impact on the classification is significant. All the adaptability associated with the method is also a favourable point, since it is possible to analyse individual features and interactions between two or more variables. Regardless, it is important to note that, due to all the randomness

linked with each permutation, there is the possibility to generate unrealistic samples that influence the results. This randomness also creates a high variability of results between all the runs, which causes some uncertainty when trying to retrieve conclusions. Like the analysis of the models' granularity, these explanations do not have information about the variables' type of influence in the predictions, i. e. if they promote malignant or benign courses of the disease.

By observing Figure 5.3, it is possible to say that all partial dependence plots (PDPs) from the different models have the same appearance. The influence of each feature is a positive linear relation with the outcome, i.e. they always promote a bad prognosis. Although there is some variance in the slope of the segments, it is not very prominent since the behaviour of the features is very similar in every run.

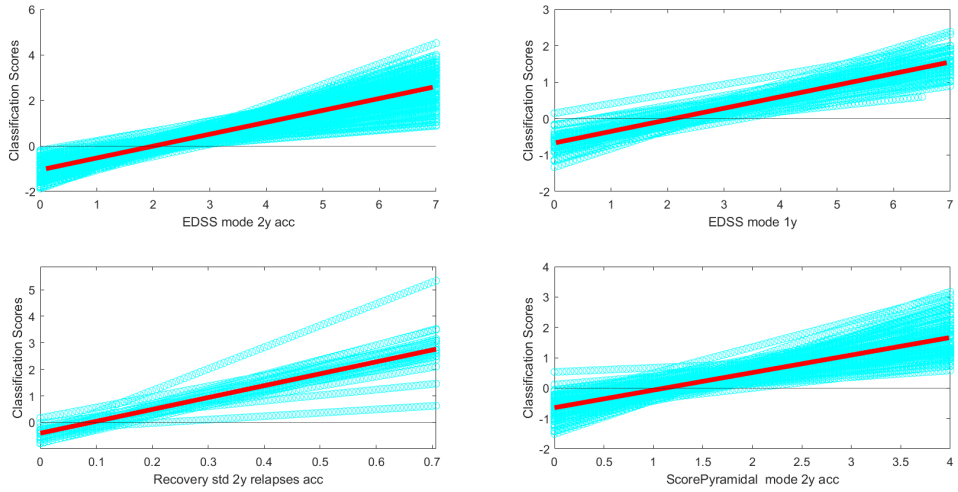

Figure 5.3: PDP of the features *ScorePyramidal mode 2y acc, Recovery std 2y relapses acc, EDSS mode 1y* and *EDSS mode 2y acc*. The blue outlines represent the feature's PDP of each model and the red outline the mean PDP value of every model that contains the analysed feature.

In the graphs of the Figure 5.4, the mean PDPs are identical to the ones shown in Figure 5.3. However, these features have some discrepancies between models as some present a negative influence on the outcome. The logic of those models when learning the relation between the studied variable is sometimes different, showing here some inconsistencies. Additionally, the slope of the segments vary greatly compared with the features analysed in Figure 5.3.

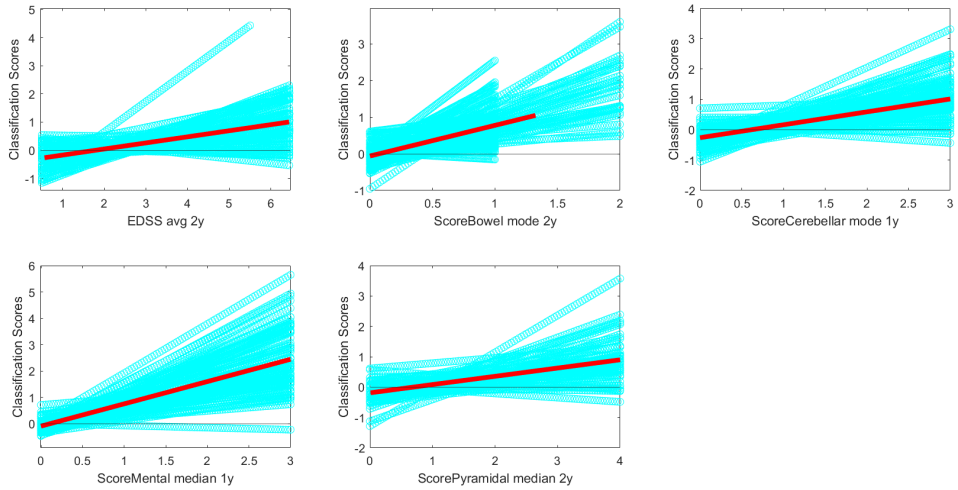

Figure 5.4: PDP of the features *ScorePyramidal avg 2y, ScoreMental median 1y, ScoreCerebellar mode 1y, ScoreBowel mode 2y* and *EDSS avg 2y*. The blue outlines represent the feature's PDP of each model and the red outline the mean PDP value of every model that contains the analysed feature.

The PDPs of the Figure 5.5 show that their mean value has a positive linear influence in the prediction, where the highest the values of the features, the higher the probability of a malignant prediction.

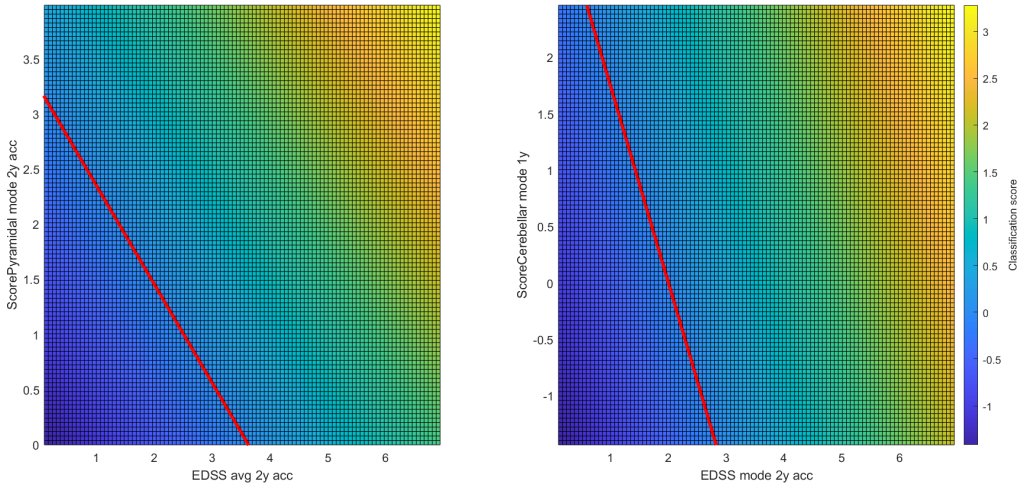

Figure 5.5: PDP of the interactions *ScoreCerebellar mode 1y - EDSS mode 2y acc* and *ScorePyramidal mode 2y acc - EDSS mode 2y acc*. Only the PDP mean value is presented, where the red outline is the representation of the classification score that equals to 0.

With the PDPs, it is possible to verify the influence of each feature (and pairs of features) in the outcome. Consequently, it is also possible to assess if the behaviour

is similar in every model or if there are some incongruities. It also provides a visual idea of the level of effect that each feature has in the outcome, in this case, with the slope of the linear relations. In some features, a small change has a significant impact on the outcome, while in others, some alterations are not relevant to the model's classification.

With no limitations associated with the study of the relation between features and target, this study is a valuable safety tool since it is possible to observe if the variables act like expected, considering the established clinical findings or if the models are not learning as expected. Although, it is important to indicate some restraints regarding the PDPs as it does not take into account interaction between features higher than two variables and does not give clear insights about the variables' weight in the predictions.

5.2 Model-specific explanations

Figure 5.6 demonstrates the estimated coefficients of the linear regression models.

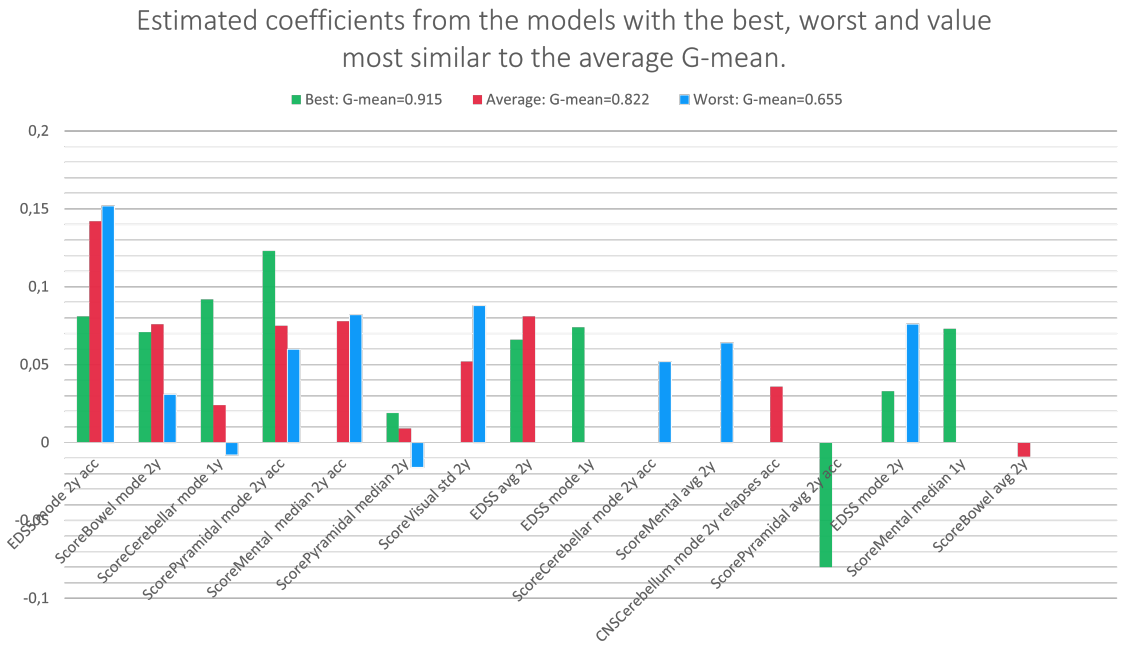

Figure 5.6: Estimated coefficients from the models with the best, worst and value most similar to the average G-mean. A positive coefficient promotes a severe case of MS while a negative coefficient promotes a benign outcome.

Considering the model with the best performance, the feature with the highest coefficient is the *ScorePyramidal mode 2y acc*. However, the feature *ScorePyrami-*

dal avg 2y acc has a strong negative impact on the predictions. In the average performance model, there are also two features with a negative influence but with shallow values (*ScoreCerebellar mode 1y* and *ScorePyramidal median 2y*), not having a significant impact on the outcome. The same scenario is verified with the feature *ScoreBowel avg 2y*. The behaviour of the features *EDSS mode 2y acc* and *ScorePyramidal mode 2y acc* is relatively similar in every model with high positive values of coefficients. However, there are features with distinct weights in different models like the feature *ScoreCerebellar mode 1y*. The values of this feature vary from a high positive coefficient to a low negative impact, being very inconsistent.

As mentioned before, the linear regression classifiers are intrinsically interpretable, i.e. the logic of the models is understandable in such a degree that it is possible to know the respective outcome very clearly with the input. However, these classification models only consider linear feature relations with the target and do not consider interactions between variables. This may jeopardise their performance when compared with more complex models such as support vector machines (SVMs), as it generally does. However, this discrepancy is not visible in the studied problem, as it is represented in Table 5.1.

Table 5.1: Performance of models with best, worst and average performance associated with the linear regression and the linear SVM classifiers.

	Best	**Average**	**Worst**
Linear regression	**G-mean = 91.5%**	**G-mean = 82.2%**	**G-mean = 65.5%**
	Sensitivity = 100.0%	Sensitivity = 81.8%	Sensistivity = 54.6%
	Specificity = 83.6%	Specificity = 82.5%	Specificity = 78.6%
	AUC = 89.0%	**AUC = 86.3%**	**AUC = 84.5%**
	F-score = 62.9%	F-score = 53.0%	F-score = 35.3%
Linear SVM	**G-mean = 92.9%**	**G-mean = 82.2%**	**G-mean = 68.1%**
	Sensitivity = 100.0%	Sensitivity = 81.8%	Sensitivity = 54.6%
	Specificity = 86.2%	Specificity = 82.5%	Specificity = 85.0%
	AUC = 94.0%	**AUC = 89.9%**	**AUC = 84.4%**
	F-score = 66.7%	F-score = 53.9%	F-score = 41.4%

Another limitation of these explanations is associated with the simultaneous analysis of multiple models. Although it is easy to interpret a single model, substantial entropy is linked with all the distinct results of coefficients in the models making the interpretation task challenging. This phenomenon is slightly hinted in Figure 5.6.

5.3 Individual prediction explanations

Three different models (best, worst and average) were considered in these explanations. In each model, three or four samples were selected per group of classifications (class 1 and class 0, and misclassifications in the counterfactual explanations). Due to the high amount of studied samples and its inherent redundancy of information, that do not increase the insights on this subject of study, only a few samples' explanations are presented here. The selection of these samples was performed to represent as well as possible the data set distribution.

The estimated coefficients by Local Interpretable Model-Agnostic Explanations (LIME), in the results of the Figures 5.7, 5.8, 5.9, 5.10, and 5.11, are negative if the variables promote to a severe MS and positive if they influence the prognosis to a mild form of the disease.

Feature	**Observation (Classification score=1.690)**
Hospital std 2y relapses	0.000
EDSS mode 2y acc	1.000
EDSS mode 2y	1.000
ScorePyramidal avg 2y acc	1.636
ScorePyramidal median 2y	1.594
ScoreCerebellar mode 1y	1.214
CNSCerebellum mode 1y relapses	0.133
ScorePyramidal mode 2y acc	1.448
ScoreMental median 1y	0.286
EDSS avg 2y	1.000

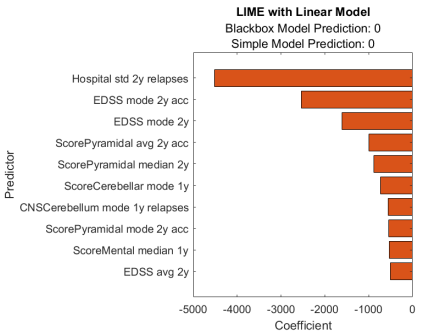

Figure 5.7: Estimated coefficients by the LIME model for the first selected sample belonging to the class 0 (benign case) of the best model.

Feature	**Observation (Classification score=2.372)**
Hospital std 2y relapses	0.049
EDSS mode 2y	1.000
ScoreMental median 1y	0.000
EDSS avg 2y	1.250
CNSCerebellum mode 1y relapses	0.000
ScoreCerebellar mode 1y	2.000
ScorePyramidal median 2y	1.000
EDSS mode 2y acc	1.000
ScorePyramidal avg 2y acc	0.750
ScorePyramidal mode 2y acc	0.000

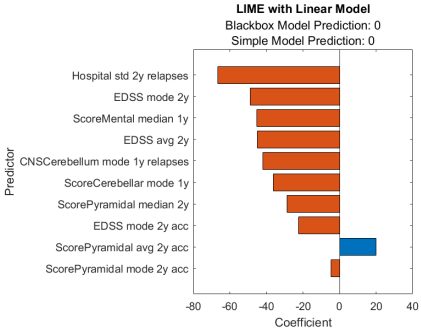

Figure 5.8: Estimated coefficients by the LIME model for the second selected sample belonging to the class 0 of the best model.

Feature	Observation (Classification score=4.136)
Hospital std 2y relapses	0.049
ScoreMental median 1y	0.000
CNSCerebellum mode 1y relapses	0.000
ScoreCerebellar mode 1y	0.000
ScorePyramidal mode 2y acc	0.000
ScorePyramidal median 2y	0.000
EDSS avg 2y	0.000
ScorePyramidal avg 2y acc	0.000
EDSS mode 2y	0.000
EDSS mode 2y acc	0.000

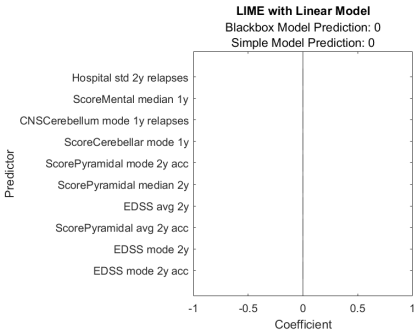

Figure 5.9: Estimated coefficients by the LIME model for the third selected sample belonging to the class 0 of the best model.

Feature	Observation (Classification score=3.496)
Hospital std 2y relapses	0.049
ScorePyramidal mode 2y acc	0.000
EDSS mode 2y	0.000
ScoreMental median 1y	0.000
EDSS avg 2y	0.000
ScorePyramidal median 2y	1.594
ScorePyramidal avg 2y acc	1.500
EDSS mode 2y acc	0.000
CNSCerebellum mode 1y relapses	0.000
ScoreCerebellar mode 1y	2.000

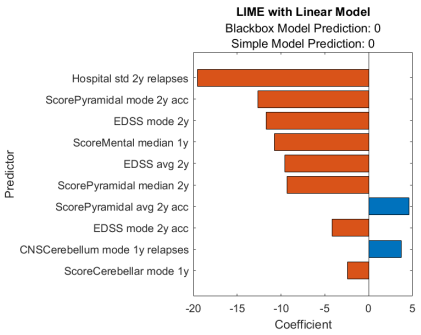

Figure 5.10: Estimated coefficients by the LIME model for the fourth selected sample belonging to the class 0 of the best model.

Firstly it is important to note that the LIME was unable to estimate the prediction coefficients of every feature of the sample in the Figure 5.9.

Some features' values express a great variance in their classification impact, as in some cases, a feature is associated with a high weight in the prediction. In contrast, in other samples, its influence is relatively low, as demonstrated when analysing the results of the feature *EDSS avg 2y*. The variable *Hospital std 2y relapses* has significantly high coefficients in every explanations with constant values throughout, not only the observations of the Figures 5.7, 5.8, and 5.10, but also in the malignant cases. The *ScorePyramidal avg 2y acc* promotes a benign course in the predictions, in the Figures 5.8 and 5.10, which is also present in the results of the linear regression classifier in the best model. Additionally, the feature *CNSCerebellum mode 1y relapses* also has a positive coefficient in prediction of the Figure 5.10.

However, both of these features do not have a significant impact on the predictions.

The results of the Figure 5.11 demonstrate an incorrect classification made by the LIME model. Thus the explanations of this prediction are not plausible as every explanation is directly linked with the simpler LIME model.

Both incorrect predictions and the lack of estimated coefficients appeared in different samples. These cases were not a one time case scenario.

Feature	**Observation (Classification score=0.500)**
EDSS mode 2y acc	1.500
EDSS mode 1y	1.500
Recovery std 2y relapses	0.707
ScoreCerebellar mode 1y	0.000
ScorePyramidal avg 2y acc	1.000
ScoreMental median 1y	0.000
ScoreSensory avg 1y	1.500
EDSS mode 2y	1.500
ScoreBowel mode 2y	0.000
ScorePyramidal median 2y	1.000

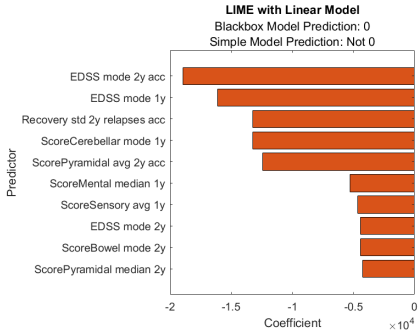

Figure 5.11: Estimated coefficients by the LIME model for a selected sample belonging to the class 0 of the worst model.

The fact that the produced explanations are a simpler and interpretable local model is very helpful since it explains the complete logic of a prediction without jeopardising the performance of the original model, the prediction model. Additionally, this method is relatively versatile since it is possible to restrict the number of considered predictors, which allows to create simpler explanations depending on the problem. Regardless, the lack of knowledge to optimise all the related parameters makes the LIME implementation challenging, possibly where the absence of a rigorous assessment of their values is one of the reasons for the anomalies mentioned before. However, this method still in development may be improved in the future.

The Figures 5.12, 5.13, and 5.14 show the results of the calculus of the Shapley values related to three predictions of malignant cases. For this set of cases, a positive Shapley value promotes a severe scenario, whereas a negative result promotes a mild course of the disease.

Feature	Observation (Classification score=1.117)
EDSS mode 2y acc	4.000
ScoreMental median 1y	0.000
ScorePyramidal mode 2y acc	2.000
ScorePyramidal avg 2y acc	2.000
CNSCerebellum mode 1y relapses	0.000
ScoreCerebellar mode 1y	1.000
EDSS avg 2y	2.614
ScorePyramidal median 2y	1.594
EDSS mode 2y	2.446
Hospital std 2t relapses	0.049

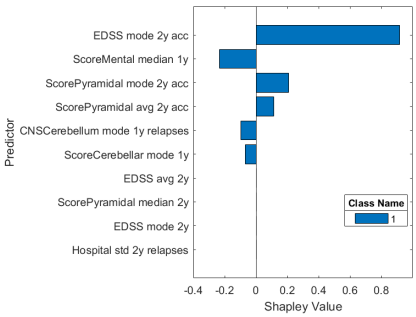

Figure 5.12: Shapley values of the first selected sample belonging to the class 1 of the best model.

Feature	Observation (Classification score=2.579)
EDSS mode 2y acc	6.500
ScorePyramidal mode 2y acc	1.448
CNSCerebellum mode 1y relapses	0.133
EDSS avg 2y	2.614
ScorePyramidal median 2y	1.594
EDSS mode 2y	2.446
ScoreMental median 1y	0.286
ScorePyramidal avg 2y acc	1.636
ScoreCerebellar mode 1y	1.214
Hospital std 2y relapses	0.049

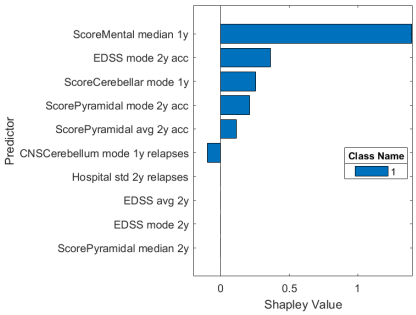

Figure 5.13: Shapley values of the second selected sample belonging to the class 1 of the best model.

Feature	Observation (Classification score=2.512)
ScoreMental median 1y	2.000
EDSS mode 2y acc	3.000
ScoreCerebellar mode 1y	2.000
ScorePyramidal mode 2y acc	2.000
ScorePyramidal avg 2y acc	2.000
CNSCerebellum mode 1y relapses	0.000
Hospital std 2y relapses	0.049
EDSS avg 2y	2.614
EDSS mode 2y	2.446
ScorePyramidal median 2y	1.594

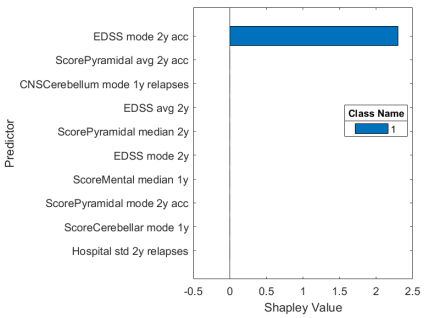

Figure 5.14: Shapley values of the third selected sample belonging to the class 1 of the best model.

In the Figures 5.12 and 5.13, the *EDSS mode 2y acc* was considered the feature with the highest weight in the predictions. It is important to mention that its values in these studied samples are already higher than 3, the threshold in the sixth year. In the third observation, this feature is the second with the highest Shapley value, where the value of the *ScoreMental median 1y* was considered by these explanations as the most relevant in the classification.

It is possible to observe that the Shapley values greatly depend on the value of each feature of the studied sample where each point explanation is very distinct from the others due to this direct link between the attributed weight and each variable's value.

These explanations assume any features' interactions and any relation between them and the outcome, where the results directly measure the weight of the features' values in the studied prediction. Since it is a direct weight of each variable, the explanations are easily understandable and clinically verifiable. They have indirect information about the relation between features and the classification result, i.e. if it is positive or negative. However, even with this information, it is not possible to anticipate what a change in a feature's value may cause, which may not be achievable to create a simple prediction model with the produced explanations.

The counterfactual explanations of samples that belong to severe cases of MS are represented in the Tables 5.2 and 5.3.

Table 5.2: Original data point and the respective counterfactual explanations of the first sample from the class 1 of the best model.

Feature	**Observation (Classification score=1.117)**
EDSS mode 2y acc	4.000
ScoreMental median 1y	0.000
ScorePyramidal mode 2y acc	2.000
ScorePyramidal avg 2y acc	2.000
CNSCerebellum mode 1y relapses	0.000
ScoreCerebellar mode 1y	1.000
EDSS avg 2y	2.614
ScorePyramidal median 2y	1.594
EDSS mode 2y	2.446
Hospital std 2y relapses	0.049

Real outcome: class 1 - Predicted outcome: class 1

1. IF *EDSS mode 2y acc* WAS 1.5 THEN OUTCOME WOULD BE 0.

Table 5.3: Original data point and the respective counterfactual explanations of the second sample from the class 1 of the best model.

Feature	**Observation (Classification score=2.579)**
EDSS mode 2y acc	6.500
ScorePyramidal mode 2y acc	1.448
CNSCerebellum mode 1y relapses	0.133
EDSS avg 2y	2.614
ScorePyramidal median 2y	1.594
EDSS mode 2y	2.446
ScoreMental median 1y	0.286
ScorePyramidal avg 2y acc	1.636
ScoreCerebellar mode 1y	1.214
Hospital std 2y relapses	0.049

Real outcome: class 1 - Predicted outcome: class 1

1. IF *EDSS mode 2y acc* WAS 1.5 THEN OUTCOME WOULD BE 0.

Only a change in the value of the feature *EDSS mode 2y acc* generates an alteration of the outcome in the samples represented in the Tables 5.2 and 5.3. It is possible to assume that, because the values of this feature are already so high, that any value decrease in the other variables would not have a significant impact on the predictions since these alterations are compensated by the elevated values of the *EDSS mode 2y acc*.

Having in mind the samples of the Tables 5.4, 5.5, 5.6, and 5.7, their variables' values are already so low that only drastic changes in the EDSS or the hospital visits due to relapses can cause an outcome transformation. It is also possible to observe that the higher the classification score of the sample, the more drastic the necessary changes need to be to transform the prediction, as expected.

Table 5.4: Original data point and the respective counterfactual explanations of the first sample from the class 0 of the best model.

Feature	**Observation (Classification score=1.690)**
Hospital std 2y relapses	0.000
EDSS mode 2y acc	1.000
EDSS mode 2y	1.000
ScorePyramidal avg 2y acc	1.636
ScorePyramidal median 2y	1.594
CNSCerebellum mode 1y relapses	0.133
ScoreCerebellar mode 1y	1.214
ScorePyramidal mode 2y acc	1.448
ScoreMental median 1y	0.286
EDSS avg 2y	1.000

Real outcome: class 1 - Predicted outcome: class 1

1. IF *Hospital std 2y relapses* WAS 0.2 THEN THE OUTCOME WOULD BE 1
2. IF *EDSS mode 2y acc* WAS 4.5 THEN THE OUTCOME WOULD BE 1
3. IF *EDSS mode 2y* WAS 5.5 THEN THE OUTCOME WOULD BE 1

Table 5.5: Original data point and the respective counterfactual explanations of the second sample from the class 0 of the best model.

Feature	**Observation (Classification score=2.372)**
EDSS mode 2y acc	1.000
EDSS mode 2y	1.000
ScoreCerebellar mode 1y	2.000
ScorePyramidal avg 2y acc	0.750
ScoreMental median 1y	0.000
ScorePyramidal mode 2y acc	0.000
EDSS avg 2y	1.250
CNSCerebellum mode 1y relapses	0.000
ScorePyramidal median 2y	1.000
Hospital std 2y relapses	0.049

Real outcome: class 1 - Predicted outcome: class 1

1. IF *Hospital std 2y relapses* WAS 0.3 THEN THE OUTCOME WOULD BE 1
2. IF *EDSS mode 2y acc* WAS 5.5 THEN THE OUTCOME WOULD BE 1
3. IF *EDSS mode 2y* WAS 7.5 THEN THE OUTCOME WOULD BE 1

Table 5.6: Original data point and the respective counterfactual explanations of the third sample from the class 0 of the best model.

Feature	**Observation (Classification score=3.496)**
EDSS mode 2y acc	0.000
EDSS mode 2y	0.000
ScorePyramidal mode 2y acc	0.000
ScoreCerebellar mode 1y	2.000
ScoreMental median 1y	0.000
EDSS avg 2y	0.000
CNSCerebellum mode 1y relapses	0.000
ScorePyramidal acg 2y acc	1.500
Hospital std 2y relapses	0.049
ScorePyramidal median 2y	1.594

Real outcome: class 1 - Predicted outcome: class 1

1. IF *Hospital std 2y relapses* WAS 0.4 THEN THE OUTCOME WOULD BE 1
2. IF *EDSS mode 2y acc* WAS 6.5 THEN THE OUTCOME WOULD BE 1
3. IF *EDSS mode 2y* WAS 9.5 THEN THE OUTCOME WOULD BE 1

Table 5.7: Original data point and the respective counterfactual explanations of the fourth sample from the class 0 of the best model.

Feature	**Observation (Classification score=4.136)**
EDSS mode 2y acc	0.000
EDSS mode 2y	0.000
ScorePyramidal mode 2y acc	0.000
ScorePyramidal avg 2y acc	0.000
ScoreCerebellar mode 1y	0.000
ScoreMental median 1y	0.000
EDSS avg 2y	0.000
ScorePyramidal median 2y	0.000
CNSCerebellum mode 1y relapses	0.000
Hospital std 2y relapses	0.049

Real outcome: class 1 - Predicted outcome: class 1

1. IF *Hospital std 2y relapses* WAS 0.45 THEN THE OUTCOME WOULD BE 1
2. IF *EDSS mode 2y acc* WAS 7.5 THEN THE OUTCOME WOULD BE 1

The counterfactual explanations are straightforward to comprehend since they appear as a rule, a concept frequently used in learning in the real world. Because they show what is the smallest change to be made to alter a prediction, these explanations might give a new perspective to the specialists in more complex cases. The fact that the alterations do not consider correlation between features, there is

the possibility to create unrealistic data points that influence the interpretation of the explanations. Still, they may be a good form of explaining the prognosis to a patient due to their already mentioned simplicity. Despite this method also offering indirect information if a feature promotes a benign case or a more severe MS course, there is no insight about the impact that each variable has on the classification, lacking therefore essential data about the models' logic.

5.4 Qualitative evaluation

When asked if the developed work was ready to be used in a medical context, the response frequently was affirmative, but it was generally associated with some doubt.

The participants felt overwhelmed by the number of different explanations presented, and they found it difficult to completely understand every method. However, the increase of explanations provides a more complete analysis. With so many distinct explanations, it is easier to assess their reliability, by analysing their coherence. Furthermore, different methods tackle different perspectives that complement each other, giving, as already mentioned, the idea of a complete study. Therefore, the explanations are not fully comparable, in which it exists a hierarchy/ granularity between them. Some help to verify the models' robustness, and others are more focused on explaining the decision making.

The lack of full data familiarity and understanding is related to the presentations' limited available time, as some explanations needed more time and effort. There was a preference for visual explanations as they are easier to apprehend. It was recommended to study further the type of relations between features and target and the feature interactions. The results of the interviews are represented in a diagram in the Figures 5.15 and 5.16.

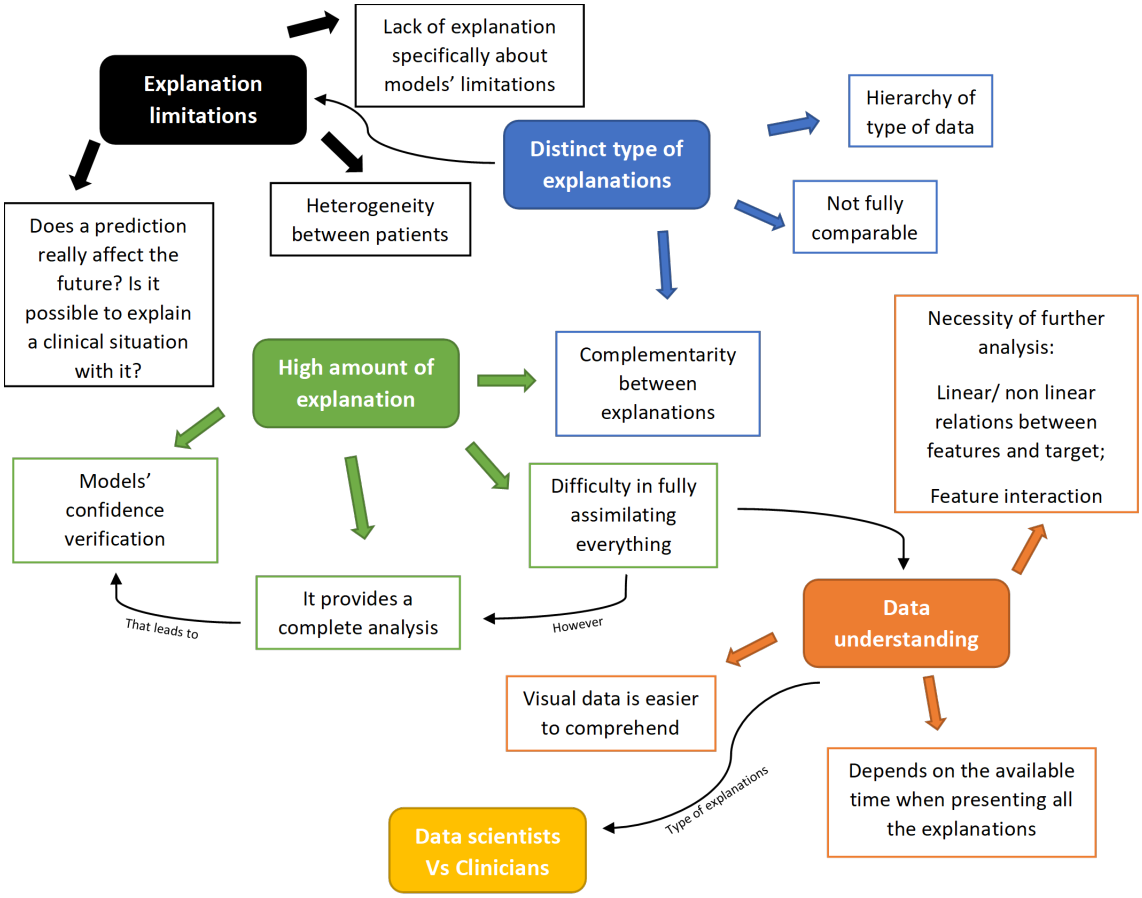

Figure 5.15: Scheme about the significant data collected in the interviews, grouped by categories (part 1).

The responses about the most valuable explanations were diverse. Shapley values were a prevalent choice and were characterised as very intuitive. It is often compared to LIME, a method that was defined by needing more time to be understood. It had contradictory opinions as some participants found its explanations helpful, and others did not think that the advantages of this method were sufficient to justify all its limitations. In this comparison between LIME and Shapley values, frequently, the Shapley values were considered better.

The discrepancies in opinions appeared when discussing the counterfactual explanations as well, since some trust the produced explanations and others do not. The indicated distrust is highly associated with the production of unrealistic data. Nevertheless, the counterfactual examples are very easy to assimilate, and although not considered visual information, they may be very informative to clinicians.

The PDPs were considered interesting to evaluate the features' behaviour and understand that some runs do not operate as the norm. One interviewee said that this method provides excessive information that may not be necessary.

Regarding the permutation feature importance, it was observed that it was useful to analyse feature interactions, although it also created unrealistic data points. The predictive power was often considered a useful analysis as it showed the relevance of feature and their relation with the outcome providing a lot of information in advance. Lastly, the linear regression was important due to its simplicity and the associated performance in this problem, being possibly justifiable to use this classifier in the problem in detriment of SVM models. However, most participants affirmed that they trust the SVM models as their results are always better than the linear regression.

As highlighted in black in the Figure 5.15, it was noted that the study lacked explanations that demonstrated meticulously the models' limitations and failures, which is beneficial information. Additionally, the explanations are directly linked with the input data of the prediction framework. Since the disease's characteristics and effects are highly heterogeneous, these explanations may not suit all cases associated with this disease.

Regarding *Data scientists vs Clinician* category of the scheme in Figure 5.16, according to the interviewees' opinions, to a data scientist, the presented explanations should be technical (from an algorithmic point of view), and it needs to be complete and detailed. However, the explanations should be more 'fluid', less technical and more conceptual for a clinician. It is also important to show a more selective analysis as it might be too overwhelming and unnecessary to present so much information to a medical specialist. This leads to an explanation refinement. With the suggestion to only show some explanations, preferably the most visual and simple information, the participants also proposed to explain the most challenging prognosis. The focus would be the cases that a clinician could not give a clear prediction in the first years of follow-up, but the models correctly identified the outcome.

In addition, it would be important to show the prediction results in probabilities or confidence level metrics in detriment of models only saying if the sample belongs to class 1 or class 0. The use of calibration curves was also suggested.

Since there is no information about the influence that a prediction has in a clinical situation, it was recommended to create explanations with variables that can be controlled by clinicians and patients, such as medication or diet. These explanations would provide a study of what changes would be the most suitable to inhibit the MS progression of a specific patient, giving important additional information to the explanations of a prediction. All these suggestions are represented

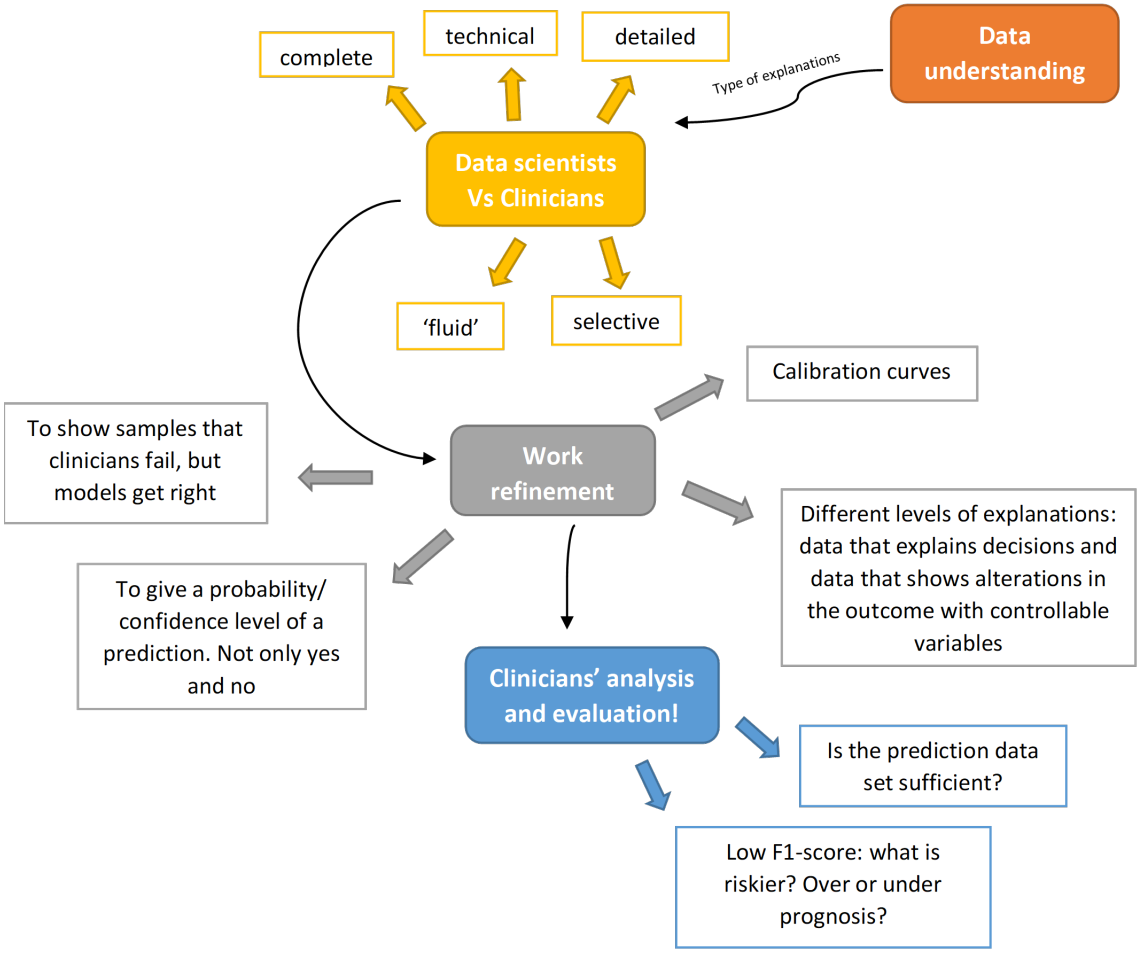

Figure 5.16: Scheme about the significant data collected in the interviews, grouped by categories (part 2).

in the Figure 5.16 as *Work refinement*.

Finally, the most popular insight is the fact that it is indispensable the validation of the results by MS specialists. Some questions only have definitive answers with this communication, e.g. 'should the studies focus on under or over-prognosis?' i.e. what is the best compromise between having mild cases classified as malignant and severe cases not correctly predicted? And 'is the used data set enough to validate all the developed work?'. The next step is, therefore very clear. The principal efforts should be to improve the presented data and contact clinicians to enable this work to move forward and achieve its ending objective.

6

Discussion

This chapter presents a discussion about the used approach in the experimental procedure and the analysis of the results. This chapter is divided by the data about the produced explanations and the qualitative evaluation performed by the interviews.

6.1 Explainability methods

In this master thesis, several explainability methods were tested. The analysis of several explanations provided the possibility to verify the consistencies between the results, which promotes a higher sense of trust in the prediction models and the explanations. It also provided a constant comparison between methods. This comparison is essential to understand which explanations are more suited to assure the safety and reliability of the prediction framework.

Distinct types of explanations were created, from general insights about all developed models to information about single predictions. This range of explanations is essential, as it targets different requirements and concerns inherent to the studied topic.

To the best of knowledge, this may be the first explainability analysis on Multiple Sclerosis (MS) disease progression. However, the use of such approach may be generalised in similar studies, since explainability is increasing popularity among data scientists. Nonetheless, there are some limitations associated with its application in this thesis.

6.1.1 Implementation of the methods

Framework-related explanations

Considering that some of the methods summarise the behaviour of all the models, it is important to highlight the approach used in the permutation feature impor-

tance. This analysis provides very interesting insights about features' interaction, since it enables to examine the influence of different sets of variables. However, no information verifies if the results of each combination of features are the effect of a feature interaction or only the sum of each variable value. This analysis would be beneficial to understand the models' logic with clarity, since it may find hidden behaviours related to the combination of different variables. In addition, this method was related to significant variation in each run. To have conclusive results, it would be interesting to further analyse the data, beyond the calculation of the mean and standard deviation of the values.

Model-specific explanations

Regarding the model-specific explanations, due to the impracticability related with the study of all models separately, only the models with the best, worst, and value most similar to the average G-mean were selected to study model-specific explanations. This decision was made as no complete combination of features were significantly recurrent in the 100 runs. Since it is not possible to select a definitive prediction model yet, this analysis promoted a broader idea of the type of explanations associated with the framework, not just single models. Nevertheless, The results significantly change from model to model, which does not permit to take global conclusions about the behaviour of the disease and predictions.

Sample-specific explanations

Moreover, in the explanations about specific predictions, clusters were created manually to select the samples studied with the local methods, since, with automatic techniques, the distribution of data points was not homogeneous, i.e. the clusters were not formed with a similar number of samples. Some clusters had almost all the observations, while the others contained only one or two observations. However, this decision is associated with the subjectivity of the researcher. The goal of this selection was to provide a set of examples that represent the data distribution. However, since each sample is different, the respective explanations are distinct from each other. Therefore it is not possible to completely compare the results of the several samples nonetheless.

Additionally, by selecting the samples that present the highest classification scores, it was possible to choose the predictions with the highest certainty in the classification. The easiness in the identification of the correct class is also probably verified in a clinical context, where the confirmation of Machine learning (ML) in these cases may not be necessary. However, by presenting such examples, it is

possible to clearly compare them with the insights of the clinicians and, consequently, validate the explanations and used methods.

The implementation of Local Interpretable Model-Agnostic Explanations (LIME) also needs improvement. The parameters were not optimised, but instead selected with a rule of thumb. The lack of optimisation possibly poorly influenced the results. It would be possible to generate more suitable explanations with a complete studied optimisation.

6.1.2 Produced explanations

In general, the produced explanations highlighted the influence of the features related with the EDSS and the scores of some Functional System (FS), with emphasis on the pyramidal, mental, and cerebellar systems. These features were often associated with the highest prediction weights, where the feature *EDSS mode 2y acc* was particularly predictive in most explanations. Since the disease is normally monitored with the analysis of the neurological deterioration, and the MS effects are measured by the EDSS and the scores of the FS, it is easy to understand why those features have the most impact in the classification.

With some exceptions, the influence of the variables is following theoretical knowledge and clinical observations, e.g. the higher the EDSS, the higher the probability of a severe case of MS. However, the exceptions must be taken into account, as they might significantly impact the predictions and show essential limitations of the models. It is, therefore, fundamental to identify those exceptions and further analyse them, since performance metrics are incapable of representing such anomalies. These problems can be demonstrated in the Partial Dependence Plot (PDP) explanations. Some models were associated with variables' effects contrary to the norm. Similar observations were present in specific explanations, such as the analysis of the estimated coefficients of the linear regression and the LIME models. A feature that often showed some inconsistencies was the *ScorePyramidal avg 2y acc*, since it promoted for a mild case of MS in some studied models.

Results constraints

It is important to note that this work was developed with a small data set compared to the literature [12, 57, 88, 99, 100], that lacked demographic diversity and had a high amount of missing values. Since the data set only contained caucasian people, mainly from the central zone of Portugal, the results are only associated with a particular group of patients. Due to this limitation, it is not possible to conclude

that these findings cover diverse patient cases.

The low value of the F1-score of the prediction models also can not be ignored. Although the other metrics have values very satisfactory, the F1-score transmits that the models are classifying a high amount of benign cases as severe MS. This might lead to overmedication if these models were used as a prognosis tool without any medical confirmation. Models are not intended to make the decisions for the clinicians, but rather be a resource tool to help them in complex cases where the prognosis is challenging and unclear. The low values of the F1-score are probably caused by the highly imbalanced data set, which is very common in prognostic healthcare problems.

With so many methods applied, the entropy caused by the number of different explanations can be exaggerated. This diversity also showed that some variables have different levels of influence from method to method. For example, although the feature *ScoreCerebellar mode 1y* was one of the most relevant variables in the recurrence analysis, it had a relatively low impact on the permutation feature importance explanations. Since the methods use distinct mechanisms and logic to calculate the features' effect in the predictions, it is expected that not all the produced information is always completely congruent. Nevertheless, it is impossible without an expert's help to know which method is adequate due to a lack of certainties or dogmas about the disease's dynamics. This study only considers the opinions of the interviewees to analyse the explanations, since, without a formal and rigorous analysis, it is not possible to assume which explanation is the best approach to support the prediction models.

6.2 Qualitative evaluations

To qualitatively evaluate the produced explanations, interviews were performed with data scientists. This approach was based on the grounded theory. Not doing a purely quantitative analysis of the results made it possible to analyse interesting details and ideas, even if only stated by one interviewee. This type of analysis allowed to have new perspectives about the developed work that otherwise would not be apparent, as the participants worked in different fields of ML and thus had different experiences and knowledge to share.

6.2.1 Interviews development

Starting with an evaluation by data scientists was very important, due to the confidence given to the project. Since data scientists have the technical knowledge to understand what was applied in this project, it is possible to rigorously evaluate all the approaches and results. This is fundamental to improve and validate all the work surrounding this thesis, before the presentation of the results to MS specialists. The validation of the results by clinicians is essential to completely evaluate the developed work, but it is currently missing at this stage. In the future, the goal is to reach MS-related clinicians that are available to perform this analysis.

In the interviews, the limited time of the presentations (around 17 minutes) partially hindered the assimilation of the results. Most of the interviewees did not know all the information, as only one participant was already aware of explainability and explainability methods. However, with a more detailed presentation, the interviewees' attention would lose focus. A possible solution would be to perform more than one presentation to promote a complete comprehension of all the data. Although, the first impression with a lack of complete knowledge about this subject might also be considered an asset, since clinicians frequently do not have insights about such topics. Therefore, it provides beforehand the necessary changes to make in the explanations to better understand the results when presented in a clinical context.

It is also important to note that the interviews can be greatly influenced by how one understands and presents the explanations, during the presentation of the results.

6.2.2 Explainability methods evaluation

The opinions of the interviewees allowed us to gather some conclusions. A comparison between LIME and the Shapley values was frequently observed, generally with a preference for the Shapley values. The inherent logic of LIME is to create a new model, that only has in consideration local characteristics of the data set. Thus, more time is needed to understand its mechanisms due to the increased complexity of this method compared to others. This causes doubts that compromise trust in these explanations, as it is difficult to trust something that is not entirely understood. The same can also be stated about the Shapley values, since calculating them might be challenging to comprehend without much experience in the field. However, when presenting all the methods, the complexity associated with the Shapley values was suppressed, possibly a partial reason for its popularity. This method has in mind

every type of interaction, and relations between features and outcome, which is the essential cause for its preference. It guarantees that all the possible learned relations between the data and model are considered, which, therefore, offers confidence that the explanations represent the logic behind each prediction accurately. Nevertheless, this needs to be confirmed by clinicians.

The counterfactual explanations also had contrasting responses. There were opinions that the explanations did not offer much trust and did not increase the knowledge about the predictions, since they only provided superficial insights. However, it was also stated that they would be a great asset to explain the clinicians' decisions due to their simplicity. Additionally, the use of features that could not be directly changed was also a pointed limitation, since the biggest advantage of counterfactual explanations is to show what would be possible to change for a more favourable outcome, for example. The use of features that can be regulated, such as medication and diet in this scenario, would create interesting, and perhaps more useful explanations. They would provide concrete insights about what changes could the patients do to alter their prognosis, even if the influence of those variables was not dramatic. This idea of controlled variables can also be applied to the other explanations. It is a very appealing concept, as it allows a patient to intervene in the decision-making.

The analysis of the recurrence of variables had a very consistent approval by the participants. In feature selection, it is studied and selected the best set of variables, that have the most power to distinguish the different classes. Thus, this type of analysis already provides a significant amount of information about the data dynamics, as it shows the granularity of the 100 models. Therefore it offers the features that have a substantial role in the prognosis due to their high selection as input. If they are constantly selected as input, no matter the train data set differences in each run, the conviction that those features are significant in the disease's prognosis is increasingly assured. This study is, therefore, a standard procedure in similar problems [89]. The predictive power by Pinto et al. [81] was a frequent choice by the interviewees as a good method to support the prediction models. It provided a continuous analysis over the years, as it studied from 1 to 5-year models, which was constantly highlighted. This analysis offers insight into the influence of characteristics in different stages of the disease and, therefore, additional information about the MS dynamics that were lacking in the other methods.

The permutation feature importance was also considered an interesting approach to find the most predictive feature. It was also often chosen as a good resource to help in the analysis of the developed models. With several runs related

to each feature, a limitation of this method was all the dispersed values of performance loss. The performance loss not only depends on the model itself but also on the permutation in each run. This makes it difficult to conclude which features and interactions of variables are the most important in the predictions. The analysis executed about feature interactions was a great asset, since it offered valuable data that is missing from the other methods. Feature interactions are a frequent topic of concern in these studies. They can provide essential links with the target that are nonexistent in the respective features individually. Since a group can be significantly more valuable than the sum of the members, the lack of information about feature interactions can lead to false conclusions about the real impact of the variables in the prognosis.

Some participants mentioned that the application of the Support Vector Machine (SVM) classifier might not be necessary, as the regression models did not present a significant loss of performance. However, it was frequently stated that that they trusted the SVM. The reason to chose the SVM models in detriment of the linear regression is focused on the fact that the SVM models always presented the highest performance in all the n-year models of all the studied problems by Pinto et al. [81]. Since the SVM is linear, the complexity does not increase exponentially in comparison with the linear regression. Still, even if not very significant, the performance increase has high advantages in the real-world since it represents that more patients have a correct prognosis.

Lastly, the PDPs had dividing opinions. While some interviewees stated their information was very helpful, since it was not a quantity of importance or impact, but visual data about the influence of features in the outcome, which complements the other type of explanations, one stated that it might be too much information to interpret.

6.2.3 Global improvements

The simplicity of the explanations is constantly put into consideration. Simpler explanations are easier to understand by all, and therefore they are a more reliable option to be presented in a clinical context. From more visual explanations to rules-based descriptions, although sometimes simpler explanations are not as complete as other options, if the robustness of the work is assured, they probably are sufficient to guaranty safety and trust in a clinical environment.

Moreover, it is essential to consider the available time that each person has to comprehend the different explanations. Generally, people have limited time. Consequently, they are more inclined to straightforward and simple data. Therefore,

the presentation of the information is a crucial component for the acceptance of the developed work. To data scientists, the explanations should be the most technical and detailed as possible to describe completely all the inherent logic associated with the predictions. By doing so, it is assured that no risks are taken with the application of this classification tool as a resource. Similar concerns exist in the clinical environment, but this type of presentation would be overpowering. Clinicians do not study ML in their field. Thus they do not have basic knowledge about all the mechanisms used to create classification models. Therefore, the data should be presented in a more practical manner, without unnecessary redundancy that can confuse the information's receptor.

Some aspects could be improved and further developed considering the opinions of the interviewees. The project is missing explanations about all the limitations of the prediction models. It would be beneficial to present all those issues and discrepancies into one explanation to simplify its interpretation.

Additionally, more than to only claim if a sample is from class 1 or 0, it would be interesting to provide a level of certainty about the predictions. A probability of an observation belonging to a severe case of MS offers the right of uncertainty to the models. A sample with 90% certainty to belong to a malignant course is very different from having a probability of 55%. This data is very relevant to interpret the prediction results. Another suggestion was to generate calibration curves. Calibration curves are plots where the x-axis is the prediction probability, and the y-axis are the actual values. These plots are very helpful to know how close the classification models are to reality. It is important to understand if a model with a good performance adequately represents the reality or if it is over-forecasting or under-forecasting.

The idea to particularly present samples that do not have a conclusive prognosis by the clinicians but are correctly classified by the ML models is very appealing. Those are the cases where this type of tool is necessary. To help, especially, the decision-making of these cases is the primary goal of studies similar to this. For simple cases, the clinicians do not need the confirmation of other resources due to the already confidence in the prognosis. However, the analysis of simple cases is also fundamental, as those are the cases that can demonstrate if the models are reliable and safe.

The problem about the selection of the explanation lies in the fact that the methods are not fully comparable, since their approaches and results are distinct and focused on different topics. Thus, this selection can only be executed considering the qualitative evaluation acquired by the interviews and the opinion of other specialists.

The validation by clinicians is critical, since, without it, it is impossible to achieve the end goal, which is to apply this type of work in a medical context. It is fundamental to adapt the information to clinicians and thoroughly analyse their input on the matter.

6.3 Refined model

With the data collected and studied from the interviews, the path for this work is evident. The validation from clinicians is crucial as the next step. However, the developed work must be refined. The information presented to clinicians must be straightforward and clear. A selection of the most suited explanations is necessary. This selection of explanations was performed with in mind all the suggestions and opinions of the interviewees, as there is no literature about which explainability method is more suited to help prediction models in their decision-making.

It is fundamental to have two types of explanations, explanations that represent the models' global behaviour and explanations that show the reason for an individual prediction. This complementary relationship provides information about different issues that need to be addressed. Global explanations can offer information about the global dynamic of the models, and possibly the dynamic of the disease. In contrast, local explanations about an observation provide the reasons for a decision specific to a patient. It can show new perspectives and reassurance in particular predictions. Due to the heterogeneity of characteristics of the MS patients, local explanations also show with clarity the differences in decisions between different classifications in other samples, which is highly beneficial.

6.3.1 Global explanations

The predictive power provides general but essential information about the relation between the data and the outcome. In addition, it shows the evolution of the impact of the features and, therefore, a perspective that no other explanation can offer. Although a simple approach, this analysis is essential, as it presents the features that are constantly selected in feature selection. The selection is based on the capacity of the variables to distinguish both classes. Thus, the characteristics that have the highest values in the explanation are the ones with the highest power to identify the different cases of MS.

To complement the information of the predictive power, the permutation feature importance is very valuable. These explanations provide information about feature

interactions in an apparent and intuitive manner. In addition to the methods' adaptability to show information about combinations of more than two features, the simplicity associated with the logic behind the importance definition is a great advantage to this method.

Since general information and interactions between features are already covered, the last selected global explanations are the PDPs. Although some interviewees considered that these explanations were excessive and that may bring difficulties, the PDPs information is beneficial to understand the logic of the developed models. Since the PDPs demonstrate the average influence of the features in the predictions, it is possible to understand if their behaviour is identical to what is clinically expected or if there are anomalies that jeopardise the produced work. They are a very valuable safety resource that provides significant trust to users and specialists.

For an easy comprehension of these explanations, the results of the different methods would be combined and grouped to each feature, and created straightforward explanations to the detriment of a presentation of the different methods. It would enable more 'fluid' explanations that focus more on the results, not on the technical aspects of the techniques.

6.3.2 Specific predictions explanations

To cover the necessity of explanations specific to individual prediction, the Shapley values were the preferable choice. As these explanations have in mind every type of relation between features and the predictions, and consider feature interactions, they offer significant advantages in comparison with the other techniques. The direct link between the features values and the classification provides a complete and simple explanation about the impact of the different variables in the outcome, delivering a clear cause for the prognosis of the analysed patient.

The examples shown in this method would be samples that, with initial data from the first year, did not show signs of the actual course of the disease progression (e.g. a patient appeared to present a severe course of the disease, however over the years, the patient only had a mild disability), but were correctly classified by the models.

The ideal scenario would be to have access to samples that clinicians incorrectly predicted. Still, that information would only be possible to have if MS specialists provided their predictions to this study. The most similar concept is the simplification to assume that a patient with the Expanded disability status scale (EDSS) higher than three in the second year of follow-up will continue to worsen and con-

sequently belong to a severe case of MS. This criteria was used to compare the different classifiers in the Figure 4.1 from the chapter 4, where it is defined as *EDSS control*. The samples of the explanations represented in the Figures 6.1 and 6.2 were selected with a similar definition. These patients present the average value of the EDSS in the first year of follow-up equal to three, but are examples of benign cases. One may assume that, with such a high EDSS value in the first year of follow up, the tendency is that the condition of the patient will only worsen over time. This may lead a clinician to believe that the patient has a severe case of MS. However it might not be accurate, as represented by the examples of the Figures 6.1 and 6.2. In these examples, a positive Shapley value demonstrates that that value of the feature promotes a benign course of the disease. In contrast, a negative Shapley value means that the value of a feature influences the prognosis to a severe case of MS.

Feature	**Observation**
ScoreBowel avg 2y	0.000
ScoreCerebellar mode 1y	0.000
ScorePyramidal median 2y	1.000
EDSS mode 1y	3.500
ScoreMental avg 2y acc	0.600
ScorePyramidal mode 2y acc	1.000
ScoreAmbulation median 2y acc	0.000
ScorePyramidal mode 1y	1.000
EDSS avg 2y	2.750
EDSS mode 2y acc	2.500

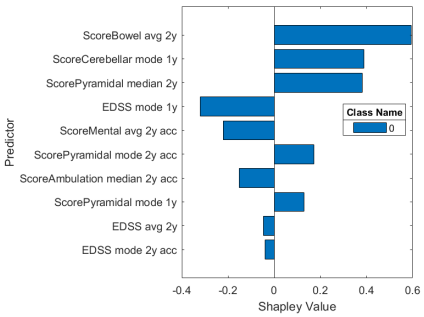

Figure 6.1: Shapley values of the first example of a benign case that presented an average EDSS equal to three on the first year of follow-up.

Feature	**Observation**
ScoreBowel avg 2y	0.000
ScorePyramidal median 2y	1.000
EDSS avg 2y	1.500
ScoreMental avg 2y acc	0.714
EDSS mode 2y acc	1.500
EDSS mode 1y	2.000
ScorePyramidal mode 2y acc	1.000
ScoreAmbulation median 2y acc	0.000
ScorePyramidal mode 1y	2.000
ScoreCerebellar mode 1y	1.000

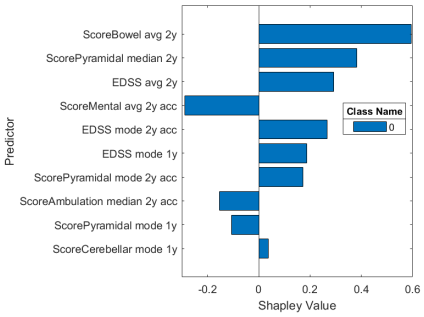

Figure 6.2: Shapley values of the second example of a benign case that presented an average EDSS equal to three on the first year of follow-up..

The Shapley values of these examples show that while the EDSS values in the Figure 6.1 promote for a benign course, in the Figure 6.2 they have a negative

influence on the prediction. With the analysis of those features, it is possible to verify that the values of the EDSS are significantly higher in Figure 6.1 than of the sample in the Figure 6.2. This shows that low values of the EDSS promote a mild disability in the long term while high values promote a malignant course, as expected. These observations show that the EDSS may not be the most relevant predictor in some cases, which demonstrates the complexity of this disease and consequent challenge of prognosis associated with such heterogeneous characteristics.

6.3.3 Work development

It is important to highlight the potential of the counterfactual explanations as they are so intuitive, simple, and provide a new perspective to the analysis of the predictions. However, these explanations would only be significantly beneficial when used with controllable variables, since they explain what needs to change to alter the outcome. Apart from the patients' medication, data that can be externally controlled is not available in the database of this project. Therefore, although it is a good idea that most certainly would increase the value of the developed work, this is not currently executable.

Lastly, the addition of information about the models' certainty in a prediction would help understand the level of confidence that a model has in a patient's prognosis. This provides insights to the clinicians that allow them to better understand the models, adjust the prediction results and explanations offered more adequately, and explain concretely to patients the challenges associated with their prognosis.

A model with these four types of explanations would not only show general feature insights, but also information about feature interactions and a comparison between their behavior in the models and what is clinically observed. It also covers local explanations, as the norm represents not all samples, and it is imperative to understand the different characteristics of the analysed samples. This analysis would offer a broad study of the developed work without unnecessary redundancy. However, this is not the final model. This work will possibly be refined and improved with the validation of the clinicians, as it is an approach based on communication between both fields, MS specialists and data scientists. This communication is essential not only in this problem but in other healthcare problems that are associated with ML models.

7

Conclusion

The overall goal of this study was to understand if the developed work by Pinto et al. [81] has the ability to be applied in a clinical environment, and which explanations are more suited to achieve that objective.

Although it is not possible to confidently state that the models are ready to be applied in practice yet, the accomplished work in this master thesis provided the first step to achieve it.

Several explainability methods were implemented that showed various characteristics about the logic of the prediction framework, as well as the causes of individual predictions. This analysis provided a better comprehension of the decision-making of the models. The feature *EDSS mode 2y acc* showed consistently relevance through most explanations as well as some scores of the FS, namely, the pyramidal, cerebellar, and mental systems.

The qualitative evaluation from the data scientists demonstrated that the explanations that support the prediction models need to be simple, straightforward, and concise to clinicians, but more technical and detailed to validate the results from an algorithmic perspective. This evaluation was essential to understand which improvements are necessary to perform before the presentation of the developed work to Multiple Sclerosis (MS) specialists. The use of the Grounded Theory (GT) to analyse the interviews was extremely important to find and organize all the information without preconceived assumptions. This allowed to discover the main concerns and ideas about the project and consequently a refined model that will be validated in the future. The selected explanations for such model were the predictive power, by Pinto et al. [81], the permutation feature importance, and the partial dependence plots (PDPs) to explain the global behavior of the framework and data relations, and the Shapley values to explain individual predictions.

Due to the low number of patients considered in this study and consequent lack of a diverse and representative data set, it is not possible to affirm that this work completely covers the different cases of MS, which is a big limitation.

Additionally, definitive conclusions can not be made without the validation from clinicians, since it is only possible to know if the explanations actually represent adequately the reality through the MS specialists input.

The future work will focus on that validation, to improve and develop the work, until the models are safe and reliable to help the prognosis of MS. It would also be interesting to add a measure of the models' confidence of each classification. The comparison of these results with different data sets may also be necessary to validate the findings of this study.

Bibliography

[1] R. Agius, C. Brieghel, M. A. Andersen, A. T. Pearson, B. Ledergerber, A. Cozzi-Lepri, Y. Louzoun, C. L. Andersen, J. Bergstedt, J. H. von Stemann, M. Jørgensen, M. H. E. Tang, M. Fontes, J. Bahlo, C. D. Herling, M. Hallek, J. Lundgren, C. R. MacPherson, J. Larsen, and C. U. Niemann. Machine learning can identify newly diagnosed patients with CLL at high risk of infection. *Nature Communications*, 11(1), 2020.

[2] L. S. Aiken, S. G. West, S. C. Pitts, A. N. Baraldi, and I. C. Wurpts. *Multiple Linear Regression*, chapter 18. American Cancer Society, 2012.

[3] A. M. Antoniadi, M. Galvin, M. Heverin, O. Hardiman, and C. Mooney. Development of an explainable clinical decision support system for the prediction of patient quality of life in amyotrophic lateral sclerosis. *Proceedings of the ACM Symposium on Applied Computing*, pages 594–602, 2021.

[4] M. Athanasiou, K. Sfrintzeri, K. Zarkogianni, A. C. Thanopoulou, and K. S. Nikita. An explainable XGBoost-based approach towards assessing the risk of cardiovascular disease in patients with Type 2 Diabetes Mellitus. *Proceedings - IEEE 20th International Conference on Bioinformatics and Bioengineering, BIBE 2020*, pages 859–864, 2020.

[5] B. Bejarano, M. Bianco, D. Gonzalez-Moron, J. Sepulcre, J. Goñi, J. Arcocha, O. Soto, U. D. Carro, G. Comi, L. Leocani, and P. Villoslada. Computational classifiers for predicting the short-term course of Multiple sclerosis. *BMC Neurology*, 11, 2011.

[6] M. Bekkar, H. K. Djemaa, and T. A. Alitouche. Evaluation Measures for Models Assessment over Imbalanced Data Sets. *Journal of Information Engineering and Applications*, 3(10):27–38, 2013.

[7] V. Belle and I. Papantonis. Principles and practice of explainable machine learning. *arXiv preprint arXiv:2009.11698*, 2020.

[8] R. Bergamaschi. Can we predict the evolution of an unpredictable disease like multiple sclerosis? *European Journal of Neurology*, 20:995–996, 2013.

[9] L. Bloch, C. M. Friedrich, and D. Neuroimaging. Data analysis with Shapley values for automatic subject selection in Alzheimer ' s disease data sets using interpretable machine learning. pages 1–32, 2018.

[10] M. Böhle, F. Eitel, M. Weygandt, and K. Ritter. Layer-wise relevance propagation for explaining deep neural network decisions in MRI-based Alzheimer's disease classification. *Frontiers in Aging Neuroscience*, 10(JUL), 2019.

[11] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, and M. Lang. Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics and Data Analysis*, 143:106839, 2020.

[12] G. Brichetto, M. Monti Bragadin, S. Fiorini, M. A. Battaglia, G. Konrad, M. Ponzio, L. Pedullà, A. Verri, A. Barla, and A. Tacchino. The hidden information in patient-reported outcomes and clinician-assessed outcomes: multiple sclerosis as a proof of concept of a machine learning approach. *Neurological Sciences*, 41(2):459–462, 2020.

[13] K. Buzzard, S. Broadley, and H. Butzkueven. What do effective treatments for multiple sclerosis tell us about the molecular mechanisms involved in pathogenesis? *International journal of molecular sciences*, 13:12665–709, 12 2012.

[14] C. Caravagn, A. Jaouën, S. Desplat-Jégo, K. K. Fenrich, E. Bergot, H. Luche, P. Grenot, G. Rougon, M. Malissen, and F. Debarbieux. Diversity of innate immune cell subsets across spatial and temporal scales in an EAE mouse model. *Sci Rep*, 8(5146), 2018.

[15] C. Caravagna. What Is Multiple Sclerosis? *Front. Young Minds*, 7(7), 2019.

[16] G. Chandrashekar and F. Sahin. A survey on feature selection methods. *Computers and Electrical Engineering*, 40(1):16–28, 2014.

[17] J. Cho, A. Alharin, Z. Hu, N. Fell, and M. Sartipi. Predicting Post-stroke Hospital Discharge Disposition Using Interpretable Machine Learning Approaches. *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, pages 4817–4822, 2019.

[18] C. Confavreux and S. Vukusic. Chapter 15 - The clinical course of multiple sclerosis. In D. S. Goodin, editor, *Multiple Sclerosis and Related Disorders*, volume 122 of *Handbook of Clinical Neurology*, pages 343–369. Elsevier, 2014.

[19] J. Correale, M. C. Ysrraelit, and M. P. Fiol. Benign multiple sclerosis: Does it exist? *Curr Neurol Neurosci Rep*, 12:601–609, 2012.

[20] A. P. Creagh, F. Lipsmeier, M. Lindemann, and M. De Vos. Interpretable Deep Learning for the Remote Characterisation of Ambulation in Multiple Sclerosis using Smartphones. 2021.

[21] J. M. De Sa. *Pattern recognition: concepts, methods and applications*. Springer Science & Business Media, 2012.

[22] A. R. T. Donders, G. J. van der Heijden, T. Stijnen, and K. G. Moons. Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59(10):1087–1091, 2006.

[23] F. Doshi-Velez and B. Kim. Towards A Rigorous Science of Interpretable Machine Learning. (Ml):1–13, 2017.

[24] E. Dove. The eu general data protection regulation: Implications for international scientific research in the digital era. *The Journal of Law, Medicine Ethics*, 46:1013–1030, 12 2018.

[25] G. C. Ebers. Environmental factors and multiple sclerosis. *Lancet Neurol*, 7:268–277, 2008.

[26] F. Eitel, E. Soehler, J. Bellmann-Strobl, A. U. Brandt, K. Ruprecht, R. M. Giess, J. Kuchling, S. Asseyer, M. Weygandt, J. D. Haynes, M. Scheel, F. Paul, and K. Ritter. Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. *NeuroImage: Clinical*, 24(August):102003, 2019.

[27] S. El-Sappagh, J. M. Alonso, S. M. Islam, A. M. Sultan, and K. S. Kwak. A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease. *Scientific Reports*, 11(1):1–26, 2021.

[28] R. Elshawi, M. H. Al-Mallah, and S. Sakr. On the interpretability of machine learning-based model for predicting hypertension. *BMC Medical Informatics and Decision Making*, 19(1), 2019.

[29] H. Engward. Understanding grounded theory. *Nursing standard (Royal College of Nursing (Great Britain) : 1987)*, 28(7):37–41, 2013.

[30] J. R. Epifano, R. P. Ramachandran, S. Patel, and G. Rasool. Towards an

explainable mortality prediction model. *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, 2020-September, 2020.

[31] J. H. Friedman and B. E. Popescu. Predictive learning via rule ensembles. *Annals of Applied Statistics*, 2(3):916–954, 2008.

[32] M. Gaspari, G. Roveda, C. Scandellari, and S. Stecchi. An expert system for the evaluation of EDSS in multiple sclerosis. *Artificial Intelligence in Medicine*, 25(2):187–210, 2002.

[33] L. Gautheron, A. Habrard, E. Morvant, and M. Sebban. Metric learning from imbalanced data. *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, 2019-November(9):923–930, 2019.

[34] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An overview of interpretability of machine learning. *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*, pages 80–89, 2019.

[35] L. H. Gilpin, C. Testart, N. Fruchter, and J. Adebayo. Explaining explanations to society. *arXiv*, (Nips), 2019.

[36] G. Giovannoni, H. Butzkueven, S. Dhib-Jalbut, J. Hobart, G. Kobelt, G. Pepper, M. P. Sormani, C. Thalheim, A. Traboulsee, and T. Vollmer. Brain health: time matters in multiple sclerosis. *Multiple Sclerosis and Related Disorders 9 (2016) S5–S48*, 9:5–48, 2016.

[37] M. M. Goldenberg. Multiple Sclerosis Review. *Pharm. Ther*, 37(3), 2012.

[38] B. Goodman and S. Flaxman. European union regulations on algorithmic decision making and a "right to explanation". *AI Magazine*, 38(3):50–57, 2017.

[39] D. H. Grossoehme. Overview of Qualitative Research. *Journal of Health Care Chaplaincy*, 20(3):109–122, 2014.

[40] K. S. Gurumoorthy, A. Dhurandhar, G. Cecchi, and C. Aggarwal. Efficient data representation by selecting prototypes with importance weights. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2019-November(Icdm):260–269, 2019.

[41] J. D. Haines, M. Inglese, and P. Casaccia. Axonal Damage in Multiple Sclerosis. *Mount Sinai Journal Of Medicine*, 78:231–243, 2011.

[42] H. F. Harbo, R. Gold, and M. Tintoré. Sex and gender issues in multiple sclerosis. *Ther Adv Neurol Disord*, 6(4):237–248, 2013.

[43] H.-P. Hartung, J. Graf, O. Aktas, J. Mares, and M. H. Barnett. Diagnosis of multiple sclerosis: revisions of the McDonald criteria 2017 – continuity and change. *Current Opinion in Neurology*, 32(3), 2019.

[44] S. L. Hauser and B. A. C. Cree. Treatment of Multiple Sclerosis: A Review. *The American Journal of Medicine*, 133(12):1380–1390.e2, dec 2020.

[45] Z. Jiang, L. Bo, Z. Xu, Y. Song, J. Wang, P. Wen, X. Wan, T. Yang, X. Deng, and J. Bian. An explainable machine learning algorithm for risk factor analysis of in-hospital mortality in sepsis survivors with ICU readmission. *Computer Methods and Programs in Biomedicine*, 204:106040, 2021.

[46] M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[47] T. J. T. Kennedy and L. A. Lingard. Making sense of grounded theory in medical education. *Medical Education*, 40(2):101–108, 2006.

[48] B. Kim, R. Khanna, and O. Koyejo. Examples are not enough, learn to criticize! Criticism for interpretability. *Advances in Neural Information Processing Systems*, (Nips):2288–2296, 2016.

[49] I. Kister, T. E. Bacon, E. Chamot, A. R. Salter, G. R. Cutter, J. T. Kalina, and J. Herbert. Natural History of Multiple Sclerosis Symptoms. *International Journal of MS Care*, 15:146–156, 2013.

[50] S. Klineova and F. D. Lublin. Clinical Course of Multiple Sclerosis. *Cold Spring Harbor perspectives in medicine*, 8(9):a028928, sep 2018.

[51] S. Kobayashi, S. Yokoi, J. Suzuki, and K. Inui. Efficient estimation of influence of a training instance. *arXiv*, (1):41–47, 2020.

[52] N. Koch-Henriksen and P. S. Sørensen. The changing demographic pattern of multiple sclerosis epidemiology. *Lancet Neurol*, 9:520–532, 2010.

[53] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. *34th International Conference on Machine Learning, ICML 2017*, 4:2976–2987, 2017.

[54] K. Korjus, M. N. Hebart, and R. Vicente. An efficient data partitioning

to improve classification performance while keeping parameters interpretable. *PLoS ONE*, 11(8):1–16, 2016.

[55] J. F. Kurtzke. Rating neurologic impairment in multiple sclerosis. *Neurology*, 33(11):1444, 1983.

[56] S. M. Lauritsen, M. Kristensen, M. V. Olsen, M. S. Larsen, K. M. Lauritsen, M. J. Jørgensen, J. Lange, and B. Thiesson. Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nature Communications*, 11(1):1–11, 2020.

[57] M. T. Law, A. L. Traboulsee, D. K. Li, R. L. Carruthers, M. S. Freedman, S. H. Kolind, and R. Tam. Machine learning in secondary progressive multiple sclerosis: an improved predictive model for short-term disability progression. *Multiple Sclerosis Journal - Experimental, Translational and Clinical*, 5(4), 2019.

[58] E. Leray, T. Moreau, A. Fromont, and G. Edan. Epidemiology of multiple sclerosis. *revue neurologique*, 172:3–13, 2016.

[59] B. Lo Sasso, L. Agnello, G. Bivona, C. Bellia, and M. Ciaccio. Cerebrospinal Fluid Analysis in Multiple Sclerosis Diagnosis: An Update. *Medicina (Kaunas, Lithuania)*, 55(6):245, jun 2019.

[60] F. D. Lublin. New Multiple Sclerosis Phenotypic Classification. *European Neurology*, 72(suppl 1)(Suppl. 1):1–5, 2014.

[61] F. D. Lublin, S. C. Reingold, J. A. Cohen, G. R. Cutter, P. S. Sørensen, A. J. Thompson, J. S. Wolinsky, L. J. Balcer, B. Banwell, F. Barkhof, B. Bebo Jr, P. A. Calabresi, M. Clanet, G. Comi, R. J. Fox, M. S. Freedman, A. D. Goodman, M. Inglese, L. Kappos, B. C. Kieseier, J. A. Lincoln, C. Lubetzki, A. E. Miller, X. Montalban, P. W. O'Connor, J. Petkau, C. Pozzilli, R. A. Rudick, M. P. Sormani, O. Stüve, E. Waubant, and C. H. Polman. Defining the clinical course of multiple sclerosis: the 2013 revisions. *Neurology*, 83(3):278–286, jul 2014.

[62] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K. W. Low, S. F. Newman, J. Kim, and S. I. Lee. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10):749–760, 2018.

[63] H. Lv, X. Yang, B. Wang, S. Wang, X. Du, Q. Tan, Z. Hao, Y. Liu, J. Yan, and Y. Xia. Machine learning-driven models to predict prognostic outcomes

in patients hospitalized with heart failure using electronic health records: Retrospective study. *Journal of Medical Internet Research*, 23(4):1–17, 2021.

[64] H. M and S. M.N. A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining Knowledge Management Process*, 5(2):01–11, 2015.

[65] P. R. Magesh, R. D. Myloth, and R. J. Tom. An Explainable Machine Learning Model for Early Detection of Parkinson's Disease using LIME on DaTSCAN Imagery. *Computers in Biology and Medicine*, 126(October):104041, 2020.

[66] R. A. Marrie. Environmental risk factors in multiple sclerosis aetiologys. *Lancet Neurol*, 3:709–718, 2004.

[67] The Multiple Sclerosis International Federation. Atlas of ms. `https://www.atlasofms.org/map/global/epidemiology/number-of-people-with-ms`, accessed 14.02.2021.

[68] The Multiple Sclerosis International Federation. *Atlas of MS, 3rd Edition*. September 2020.

[69] D. Miller, F. Barkhof, X. Montalban, A. Thompson, and M. Filippi. Clinically isolated syndromes suggestive of multiple sclerosis, part I: natural history, pathogenesis, diagnosis, and prognosis. *The Lancet Neurology*, 4(5):281–288, 2005.

[70] D. H. Miller and S. M. Leary. Primary-progressive multiple sclerosis. *The Lancet Neurology*, 6(10):903–912, oct 2007.

[71] S. Min, B. Lee, and S. Yoon. Deep learning in bioinformatics. *Briefings in bioinformatics*, 18(5):851–869, 2017.

[72] C. Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*, chapter 2. 2019. `https://christophm.github.io/interpretable-ml-book/`.

[73] A. Moncada-Torres, M. C. van Maaren, M. P. Hendriks, S. Siesling, and G. Geleijnse. Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival. *Scientific Reports*, 11(1):1–13, 2021.

[74] I. Muhammad and Z. Yan. Supervised Machine Learning Approaches: a Survey. *ICTACT Journal on Soft Computing*, 05(03):946–952, 2015.

[75] K. Y. Ngiam and I. W. Khor. Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, 20(5):e262–e273, 2019.

[76] B. Nourbakhsh and E. M. Mowry. Multiple Sclerosis Risk Factors and Pathogenesis. *Continuum (Minneap Minn)*, 3:596–610, 2019.

[77] A. Ochoa-Morales, T. Hernández-Mojica, F. Paz-Rodríguez, A. Jara-Prado, Z. T.-D. L. Santos, M. Sánchez-Guzmán, J. Guerrero-Camacho, T. Corona-Vázquez, J. Flores, A. Camacho-Molina, V. Rivas-Alonso, and D. D.-O. de Montellano. Quality of life in patients with multiple sclerosis and its association with depressive symptoms and physical disability. *Multiple Sclerosis and Related Disorders*, 36:101386, 2019.

[78] P. A. Patrician. Multiple imputation for missing data. *Research in Nursing and Health*, 25(1):76–84, 2002.

[79] A. Paulino. Área abaixo da curva roc. `https://medium.com/ensina-ai/%C3%A1rea-abaixo-da-curva-roc-15d2ae93a577`, accessed 23.03.2021.

[80] J. Peng, K. Zou, M. Zhou, Y. Teng, X. Zhu, F. Zhang, and J. Xu. An Explainable Artificial Intelligence Framework for the Deterioration Risk Prediction of Hepatitis Patients. *Journal of Medical Systems*, 45(5), 2021.

[81] M. F. Pinto, H. Oliveira, S. Batista, L. Cruz, M. Pinto, I. Correia, P. Martins, and C. Teixeira. Prediction of disease progression and outcomes in multiple sclerosis with machine learning. *Scientific Reports*, 10(1):21038, 2020.

[82] N. Prentzas, A. Nicolaides, E. Kyriacou, A. Kakas, and C. Pattichis. Integrating machine learning with symbolic reasoning to build an explainable ai model for stroke prediction. *Proceedings - 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering, BIBE 2019*, pages 817–821, 2019.

[83] C. A. Ramezan, T. A. Warner, and A. E. Maxwell. Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. *Remote Sensing*, 11(2), 2019.

[84] C. Reid Turner, A. Fuggetta, L. Lavazza, and A. L. Wolf. A conceptual basis for feature engineering. *Journal of Systems and Software*, 49(1):3–15, 1999.

[85] J. C. Reinhold, A. Carass, and J. L. Prince. A Structural Causal Model for MR Images of Multiple Sclerosis. pages 1–15, 2021.

[86] K. Rejdak, S. Jackson, , and G. Giovannoni. *British Medical Bulletin*, 95:79–104.

[87] A. Scalfari, A. Neuhaus, M. Daumer, P. A. Muraro, and G. C. Ebers. Onset of secondary progressive phase and long-term evolution of multiple sclerosis. *J Neurol Neurosurg Psychiatry*, 85:67–75, 2013.

[88] R. Seccia, D. Gammelli, F. Dominici, S. Romano, A. C. Landi, M. Salvetti, A. Tacchella, A. Zaccaria, A. Crisanti, F. Grassi, and L. Palagi. Considering patient clinical history impacts performance of machine learning models in predicting course of multiple sclerosis. *PLoS ONE*, 15(3):1–18, 2020.

[89] R. Seccia, S. Romano, M. Salvetti, A. Crisanti, L. Palagi, and F. Grassi. Machine learning use for prognostic purposes in multiple sclerosis. *Life*, 11(2):1–18, 2021.

[90] P. C. Sen, M. Hajra, and M. Ghosh. *Emerging Technology in Modelling and Graphics*, volume 937. Springer Singapore, 2020.

[91] A. Thompson, B. Banwell, F. Barkhof, W. Carroll, T. Coetzee, G. Comi, J. Correale, F. Fazekas, M. Filippi, M. Freedman, K. Fujihara, S. Galetta, H.-P. Hartung, L. Kappos, F. Lublin, R. Marrie, A. Miller, D. Miller, X. Montalban, and J. Cohen. Diagnosis of multiple sclerosis: 2017 revisions of the mcdonald criteria. *The Lancet Neurology*, 17, 12 2017.

[92] H. C. Thorsen-Meyer, A. B. Nielsen, A. P. Nielsen, B. S. Kaas-Hansen, P. Toft, J. Schierbeck, T. Strøm, P. J. Chmura, M. Heimann, L. Dybdahl, L. Spangsege, P. Hulsen, K. Belling, S. Brunak, and A. Perner. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. *The Lancet Digital Health*, 2(4):e179–e191, 2020.

[93] S. Vukusic and C. Confavreux. Primary and secondary progressive multiple sclerosis. *Journal of the Neurological Sciences*, 206(2):153–155, feb 2003.

[94] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *arXiv*, pages 1–52, 2017.

[95] S. Wang, J. Tang, H. Liu, and E. Lansing. Encyclopedia of Machine Learning and Data Mining. *Encyclopedia of Machine Learning and Data Mining*, pages 1–9, 2016.

[96] V. Wottschel, D. C. Alexander, P. P. Kwok, D. T. Chard, M. L. Stromillo, N. De Stefano, A. J. Thompson, D. H. Miller, and O. Ciccarelli. Predicting

outcome in clinically isolated syndrome using machine learning. *NeuroImage: Clinical*, 7:281–287, 2015.

[97] Y. Yoo, L. Y. Tang, D. K. Li, L. Metz, S. Kolind, A. L. Traboulsee, and R. C. Tam. Deep learning of brain lesion patterns and user-defined clinical and MRI features for predicting conversion to multiple sclerosis from clinically isolated syndrome. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization*, 7(3):250–259, 2019.

[98] Q. Zhao and T. Hastie. Causal Interpretations of Black-Box Models. *Journal of Business & Economic Statistics*, 39(1):272–281, 2021.

[99] Y. Zhao, B. C. Healy, D. Rotstein, C. R. Guttmann, R. Bakshi, H. L. Weiner, C. E. Brodley, and T. Chitnis. Exploration of machine learning techniques in predicting multiple sclerosis disease course. *PLoS ONE*, 12(4):1–13, 2017.

[100] Y. Zhao, T. Wang, R. Bove, B. Cree, R. Henry, H. Lokhande, M. Polgar-Turcsanyi, M. Anderson, R. Bakshi, H. L. Weiner, and T. Chitnis. Ensemble learning predicts multiple sclerosis disease course in the SUMMIT study. *npj Digital Medicine*, 3(1):1–8, 2020.