UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

Temporal unsupervised learning models to study ALS progression

Afonso Manuel Tito Lopes Seguro

Mestrado em Engenharia Informática

Dissertação orientada por:
Profª. Doutora Helena Isabel Aidos Lopes Tomás
Profª. Doutora Sara Alexandra Cordeiro Madeira

2023

Acknowledgements

I would like to express my immense gratitude to my parents, Luis and Cesarina, for the dedication and unconditional support they have provided me at every stage of my life. I am grateful for the constant presence of my girlfriend, Catarina, who has always been by my side, encouraging me to never give up. I also thank my friends, whose friendship and support have been invaluable throughout my journey. I cannot fail to mention the importance of my teachers, who played a key role in my professional training. In particular, I am immensely grateful to Professor Helena Aidos, whose assistance was invaluable to the success of this project. I would like to extend my thanks to all my fellow students, who have shared every step of this journey with me. Finally, it is with gratitude that I acknowledge the Faculty of Sciences for hosting me during these two years and the LASIGE Research Unit and the Foundation for Science and Technology for their indispensable support.

I have lived with the prospect of an early death for 49 years. I'm not afraid of death, but I'm in no hurry to die, there are so many things I want to do first.

-Stephen Hawking

Abstract

Amyotrophic Lateral Sclerosis (ALS) is a rapidly progressing chronic disease that affects motor neurons, leading to progressive disability and eventually paralysis. Due to its complexity and heterogeneity, the search for effective treatments that can slow down the progression of ALS and improve the quality of life of patients has been a constant challenge in the medical field[54]. Thus, it is important to automatically identify the groups of patients with similar progressions, to improve the prediction of medical procedures. This work is divided into three parts: stratification, prognosis of the type of progression, and prediction of the need for medical procedures. In stratification, groups of patients with similar progressions are found to help new patients predict their needs, resulting in five groups: lower limbs, upper limbs, bulbar, diffuse (with a strong respiratory component), and advanced progressions. Supervised machine-learning models were built to predict the type of progression, using data collected directly at the first consultation, with some classifiers showing an accuracy of over 80%. However, it was difficult to predict patients with diffuse progression, and data balancing techniques were applied which, although they performed slightly worse overall, showed an improvement in this specific group of patients. The need for medical procedures varies according to the type of progression. It was calculated which procedures each cluster tends to need and studied whether creating specific classifiers for each of them would perform better in predicting these procedures when compared to a general classifier. Predictions were made in windows of 90, 180, and 365 days. Comparing the specialized classifiers with the general ones, inconclusive results were obtained at 90 days, but at 365 days, the specialized models showed better results in some procedures, such as predicting non-invasive ventilation and the need for a communication aid device. This project is a step towards a better understanding of ALS in order to contribute to the development of more personalized therapeutic strategies in the future.

Keywords: Amyotrophic Lateral Sclerosis, Machine Learning, Unsupervised Temporal Stratification, ALS Medical Procedures

Resumo Alargado

A Esclerose Lateral Amiotrófica (ELA) é uma doença crónica de progressão rápida que afeta os neurónios motores, levando a uma incapacidade progressiva e, eventualmente, à paralisia. Devido à sua complexidade e heterogeneidade, a procura por tratamentos eficazes que possam retardar a progressão de ELA e melhorar a qualidade de vida dos pacientes tem sido um desafio constante no campo médico[57]. Assim, é importante identificar automaticamente os grupos de doentes com evoluções semelhantes, para melhorar a previsão de procedimentos médicos e ajudar os clinicos a melhor gerirem a doença. Este trabalho está dividido em três partes: estratificação, prognóstico do tipo de progressão e previsão da necessidade de procedimentos médicos. Cada uma destas etapas é fundamental para fornecer informações valiosas que possam orientar os cuidados personalizados aos pacientes com ELA. A estratificação é o primeiro passo abordado neste estudo. Consiste na criação de grupos de pacientes com progressões semelhantes, permitindo que novos pacientes possam antecipar as suas necessidades através da comparação com outros indivíduos do mesmo grupo. Para tal, são exploradas duas abordagens: estratificação univariada e estratificação multivariada. A estratificação univariada leva em consideração apenas o declínio geral do paciente ao longo do tempo. Esta abordagem classifica as progressões como rápidas, lentas ou médias, fornecendo uma visão geral da velocidade de deterioração. No entanto, embora seja útil para calcular o tempo médio de vida, tem uma importância limitada na previsão das necessidades individuais dos pacientes. Por outro lado, a estratificação multivariada considera as diferentes regiões do corpo afetadas pela progressão da ELA. Neste estudo, foram encontrados cinco grupos distintos para a estratificação multivariada, estes grupos distinguem os pacientes com base na apresentação inicial da doença em quatro categorias principais: membros inferiores, membros superiores, bulbar e difusa. O grupo de membros inferiores abrange pacientes cujos primeiros sintomas da ELA se manifestam principalmente nos membros inferiores, resultando em fraqueza muscular e dificuldade de locomoção. Já o grupo de membros superiores inclui aqueles com progressão inicial nos braços e mãos, levando a dificuldades em atividades como agarrar objetos e escrever. A progressão bulbar está relacionada ao comprometimento inicial dos músculos responsáveis pela fala, deglutição e expressão facial. Estes pacientes podem enfrentar dificuldades na articulação de palavras, na deglutição de alimentos e no controlo dos movimentos faciais. A categoria difusa é caracterizada por um forte componente respiratório, mas sem uma localização específica de início, estes pacientes podem experimentar dificuldades respiratórias significativas, embora não haja uma região do corpo claramente afetada inicialmente. Além desses grupos, há

uma categoria adicional destinada a pacientes que já estão em estágios avançados da doença, este grupo abrange casos em que a progressão atingiu um estágio crítico, afetando várias regiões do corpo de forma acentuada.

Para realizar a estratificação, foi adotada a arquitetura LSTM AutoEncoder com a aplicação de KMeans no espaço latente. Uma vez estabelecidos os grupos de estratificação, torna-se crucial fornecer aos clinicos informações prognósticas sobre o tipo de evolução que os seus pacientes podem enfrentar. Para isso, a aprendizagem automática supervisionada é usada, aproveitando todas as informações recolhidas durante a primeira consulta. Com base nesses dados, é possível prever o tipo de evolução que um novo paciente pode vivenciar, focando especialmente nos grupos estabelecidos pela estratificação multivariada. Os resultados obtidos foram positivos, com alguns classificadores a obterem precisões acima de 80% de precisão. No entanto, existe dificuldade na previsão de pacientes com progressões difusas, pois estes representam uma minoria em relação aos outros grupos estabelecidos. Para lidar com este desafio, foram aplicadas técnicas de balanceamento de dados, o que resultou numa melhoria na classificação desses casos específicos. No entanto, é importante observar que, em termos gerais, o desempenho dos classificadores pode ter sido ligeiramente afetado por esta abordagem.

Além da previsão da progressão da doença, outro aspeto fundamental é a identificação das necessidades médicas específicas de cada grupo. Isto ocorre porque as necessidades de procedimentos médicos podem variar de acordo com o tipo de progressão apresentado pelo paciente. Por exemplo, um paciente com paralisia nos membros inferiores provavelmente precisará de uma cadeira de rodas, enquanto um paciente com progressão bulbar pode necessitar mais de um dispositivo auxiliar de comunicação. Com base nesta premissa, foi realizado um estudo para determinar quais procedimentos médicos são mais frequentemente necessários em cada grupo. A partir desses resultados, surge a possibilidade de desenvolver classificadores especializados para cada grupo, visando uma previsão mais precisa das necessidades médicas ao longo do tempo. Ao comparar os classificadores especializados com os classificadores gerais, é possível observar algumas tendências interessantes. No curto prazo, ou seja, num período de 90 dias, os resultados obtidos são inconclusivos devido à dificuldade de previsão nesse intervalo de tempo. No entanto, num período de 365 dias, os resultados são muito mais significativos. Nessa perspetiva, é possível constatar que os modelos especializados tendem a superar a previsão realizada por classificadores que não fazem a distinção entre os diferentes tipos de progressão. Este fenómeno pode ser observado em três casos específicos: previsão da necessidade de ventilação não invasiva para pacientes com progressões difusas, previsão da necessidade de um dispositivo auxiliar de comunicação para pacientes com progressões bulbares e previsão da necessidade de apoio de um cuidador para pacientes em estágios avançados da doença. Em resumo, este projeto constitui um passo significativo para uma melhor compreensão da ELA, visando contribuir para o desenvolvimento de estratégias terapêuticas mais personalizadas.

Palavras-chave: Esclerose Lateral Amiotrófica, Aprendizagem Automatica, Estratificação Temporal não-supervisionada, Procedimentos médicos ELA

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Amyotrophic lateral sclerosis (ALS), often referred to as Lou Gehrig's disease, is a neurodegenerative disease, whose average life span from the onset of symptoms is 3 to 5 years, but in rare cases can reach several decades [71]. The prevalence of the disease is 3 to 5 cases per 100,000 individuals, and in elderly patients, the risk increases, reaching 1 in 300, around the age of 85 [49]. Due to the heterogeneity of the disease, it is rather difficult to assign a therapy that shows real signs of delay [26, 36]. One of the remarkable aspects of ALS research is the collaboration between various scientific disciplines and the global community. Clinicians, geneticists, neuroscientists and bioengineers, among others, have come together to share knowledge and ideas in the search for effective treatments. Research on ALS primarily centers around patient longevity[25], and for pertinent clinical features that could help prognosis prediction[16].
Recently, there has been a growing emphasis in the ALS field on categorizing patients. This categorization involves either organizing patients based on their rate of disease progression[75] or using a collection of predictive characteristics[23]. Early ALS patient classification systems were based on clinical presentation and for diagnosing ALS, but had limited capacity to predict disease prognosis or suggest underlying disease mechanisms[60, 7, 10]. Recent attempts towards ALS patient classification focused on predicting clinical outcomes but were often limited by small sample sizes and sparse clinical information[23, 68, 20, 48]. With the collection of demographic data, medical procedures, and subsequent progression scores, this work aims to group patients according to disease progression and to understand which factors may influence the speed of their development. In addition, by using machine learning to provide as much information as possible about the disease and the patient's needs, it is possible to offer patients a better quality of life while they are affected by the disease, allowing clinicians to make more informed decisions.

1.1 Problem formulation and contributions

Amyotrophic lateral sclerosis is a disease that is difficult to predict due to its heterogeneity, so this project tries to understand it better using machine learning techniques, to find common traits that allow clinicians to better understand this disease, and to better apply the necessary care depending on the type of evolution.

To this end, certain questions formulate the problem to be solved, such as:

- Is it possible to stratify the different types of progression? And identify them at the first consultation?
- Is it possible to stratify and identify different stages of the disease?
- Which variables most influence the different types of progression?
- Does stratification help reduce heterogeneity, and improve prognosis in disease?

The aim is to create groups and relationships, using machine learning and data processing, that demonstrate similar characteristics and to understand whether these have an impact on better predicting the need for medical procedures.
The expected contributions of this work are twofold: scientific, where the stratification of patients can be the basis for the development of further work, in order to help develop the state of the art in the identification and prognosis of ALS. On a clinical level, the development of ALS prognostic techniques will help clinicians to make more informed decisions, and patients to achieve a better quality of life by mastering the disease as much as possible.

1.2 Document structure

The document is organized by the following chapters:

- Chapter 2 - Background, which explores the basics of amyotrophic lateral sclerosis and supervised learning for a better understanding of this work.
- Chapter 3 - Related work, explored articles that address the topic, such as, articles that have used the data that was used in this project, articles that reveal the state of the art of unsupervised temporal learning and articles that talk about the stratification of sclerosis and similar diseases.
- Chapter 4 - Data, which is divided into 3 sections, how they are connoted, their origin and what treatment they took for the different phases of the project.
- Chapter 5 - Methodology, which explains how the methods were developed to achieve the results in stratification and prognosis, both in terms of the type of progression and the medical procedures.
- Chapter 6 - Stratification, explains how both univariate and multivariate stratification were performed.
- Chapter 7 - Prognosis of different types of progressions, where supervised learning is used to predict at the first visit, what type of progression the patient may have using the initial symptoms.

- Chapter 8 - Predicting medical procedures in different types of progression, which looks at which medical procedures are most common in each type of progression and tries to predict when the patient might need them.
- Chapter 9 - Conclusion, presents future work and the conclusion.

Chapter 2

Background

2.1 Amyotrophic Lateral Sclerosis

Amyotrophic lateral sclerosis (ALS) is a progressive disease that affects the nervous system. People with this disease tend to gradually lose strength in their muscles, which leads to limited movement and an inability to be independent. The average survival time from the onset of symptoms is 3 to 5 years, but in particular cases, it can reach up to several decades [71]. The prevalence of the disease is 3 to 5 cases per 100,000 individuals, and in elderly patients, the risk increases reaching up to 1 in 300, with about 85 years[49]. Due to the heterogeneity of the disease, it is quite difficult to assign a therapy to show true signs in its retardation [26, 36]. The main cause of death from ALS, is respiratory failure, and patients are more vulnerable to respiratory infections and poor blood oxygenation. Symptoms can be revealed in 5 forms, according to the Portuguese Amyotrophic Lateral Sclerosis Association (APELA):

- Medullar - First symptoms involve muscles in the arms or legs.
- Bulbar - When there is difficulty articulating words, chewing or swallowing.
- Respiratory - The patient shows shortness of breath on physical exertion or at rest.
- Axial - Muscular weakness is felt in the muscles of the neck or back.
- Diffuse - When it is difficult to localize the first symptoms, they may be spread throughout the body.

2.1.1 Classification of ALS

John Ravits et al[59], divided sclerosis into 3 classifications, clinical phenotypes, molecular neuropathology and genetics.

Clinical phenotypes

Phenotypes refer to the observable or measurable characteristics of an organism, resulting from the interaction between its genes and the environment. These characteristics can include physical

aspects such as eye color, hair type or height, as well as behavioral traits or cognitive abilities. John Ravits et al[59] and Leslie I Grad et al [28], divide sclerosis phenotypes into two different types, based on the neuronal level involved, or based on the somatic region. When viewed at the neuronal level, there are three possible types of evolution:

- Typical - Typical ALS, also known as ”classic”, shows signs in both Upper Motor Neurons (UMN) and Lower Motor Neurons (LMN) and is usually fatal within four years of the onset of symptoms.
- PLS - Primary Lateral Sclerosis (PLS) refers to a syndrome that predominantly involves the degeneration of upper motor neurons (UMN). It is still unclear whether this phenotype is a distinct disorder or a variant of ALS. In most patients with PLS, symptoms start in the legs and ascend in a relatively symmetrical fashion to the arms and bulbar muscles.
- PMA - Progressive Muscular Atrophy (PMA) refers to a syndrome with predominant lower motor neuron involvement. Unlike typical ALS, the onset of PMA can occur in any region of the body, has a higher occurrence in males and usually occurs later in life.

Other phenotypic designations for ALS may be by the region of the body affected at disease onset. Bulbar ALS affects the muscles of speech, chewing and swallowing, with involvement predominantly in the LMN, whereas the pseudobulbar variant indicates involvement predominantly in the UMN, both forms having a similar progression. Limb manifestation ALS is considered the typical primary form and may have regional variants such as upper limb (arms) or lower limb (legs) weakness. Mill’s variant is a rare form of ALS with a progressive hemiplegic pattern, which resembles primary lateral sclerosis.

Molecular neuropathology

Molecular neuropathology is a field of research that focuses on the molecular analysis of neurological diseases. This area is concerned with studying how molecular alterations, such as genetic mutations, abnormal gene expression and epigenetic changes, can lead to diseases of the nervous system.
In 1988, by Leigh et al[40] and Lowe et al [44], it was discovered that motor neurons of the ALS have ubiquitin deposits in the cytoplasm. These deposits are mainly composed of the protein TDP-43, which becomes abnormal and insoluble. This finding led to the belief that ALS is a TDP-43 proteinopathy. These changes are also observed in some cases of frontotemporal dementia. The presence of ubiquitinated TDP-43 is a common feature in most cases of ALS, regardless of the clinical phenotype. The distribution of TDP-43 pathology is broad in the brain, not limited to motor regions.

Genetics

Genetics is the study of genes, heredity and genetic variation in living things. Genes are units of information in DNA (deoxyribonucleic acid) that determine the physical and functional character-

istics of organisms, such as eye color, blood type, height, susceptibility to disease and other traits. Grad et al [27] says that about 5% to 10% of ALS cases are genetically transmitted. This increases to up to 15% - 20% when known genes are tested in patients initially thought to have the sporadic disease. Different genes have been identified in approximately 60% - 70% of families with familial ALS. These genes are associated with diverse clinical features such as lower motor neuron syndromes, upper limb onset, slower progression and possible associations with frontotemporal dementia. In addition, there are specific mutations that can lead to faster or more indolent forms of the disease.

2.2 Machine Learning

Machine learning, is a branch of artificial intelligence that focuses on the development of algorithms and models capable of learning and making decisions based on data, without being explicitly programmed to perform specific tasks.

2.2.1 Data imbalance

In some parts of this work, there was a need to apply data balancing techniques, after comparing some of the most popular ones, the chosen one was SMOTE-ENN. Developed by Batista et al[4], it combines the ability of SMOTE to generate synthetic examples for the minority class and the ability of ENN that removes observations from both classes when they have a different class in relation to the observation and the nearest neighbor of the majority class. In summary the algorithm can be divided into the following steps:
SMOTE:

1. Randomly choose data from the minority class.
2. Calculate the distance between the random data and its k nearest neighbors.
3. Multiply the difference by a random number between 0 and 1 and add the result as a synthetic sample to the minority class.
4. Repeat steps 2 and 3 until the desired proportion of the minority class is reached.

ENN:

1. Determine the value of K, the number of nearest neighbors.
2. Find the K-nearest neighbor of the observation among the other observations in the dataset and return the majority class of this nearest neighbor.
3. If the class of the observation and the majority class of the K-nearest neighbor are different, exclude the observation and its K-nearest neighbor from the dataset.
4. Repeat steps 2 and 3 until the desired proportion of each class is reached.

2.2.2 Unsupervised machine learning

Unsupervised machine learning is an approach in which algorithms are trained to find patterns or structures in input data without the presence of labels or external supervision. Using algorithms that rely on similarities and differences between instances, unsupervised learning seeks to divide data into subsets, or groups, that have common characteristics. These groups can provide valuable information about the structure of the data and the relationships between the different variables[21].

A. Most Typical Models

There are several types of clustering algorithms that can be use to stratify patients, KMeans and Hierarchical Clustering are two of the best known.

Hierarchical Clustering - Clustering methods create groupings based on a hierarchy and can be divided into two categories: divisive and agglomerative, the latter being more commonly used. Agglomerative hierarchical clustering methods work by interactively combining the two closest objects or clusters until all data are grouped in the same cluster. On the other hand, divisive methods follow an opposite approach. At the beginning of the agglomerative Hierarchical Clustering (HC) process, the number of clusters is equal to the number of instances. Then, a proximity matrix is generated to record the distances between each pair of data points. The two closest points are combined into a cluster and the proximity matrix is recalculated by replacing the two points with the centroid of the newly formed cluster and computing its distance from the remaining instances. This process is repeated until there is a single cluster containing all the instances or until the stop conditions (a predefined number of clusters) are met.[21].

KMeans - The KMeans method is a widely used and easily implemented stratification algorithm. It requires the number of clusters to be defined in advance, and uses an iterative approach to assign data instances to each cluster. The process starts with the random selection of k centroids, which represent the center of each cluster. Then, each data instance is compared to the centroids, and assigned to the cluster whose centroid is closest. After each iteration, the centroids are updated with the average of all instances within each cluster. Then, the whole process is repeated until there are no more changes in the clusters or the maximum number of iterations is reached. Although a popular method, KMeans has some limitations. For example, it is necessary to determine the number of clusters in advance, and the number of iterations can be computationally intensive depending on the number of instances and clusters[21].

B. Cluster Validation

Cluster validation can be subdivided into 3 types according to Therese Ullmann et al[72]:

- Internal - Measures intra-cluster and inter-cluster dispersion and is important to ensure that the groupings obtained are useful and consistent with the data.

- External - Compares the clusters assigned by the learning algorithm with the actual clusters, allowing the effectiveness of the clusters to be assessed.
- Stability - Assesses the robustness and stability of clusters, helping to ensure that clusters represent real patterns in the data and are not influenced by random or sampling variations.

NbClust is a package in the R programming language that validates the number of clusters. It contains more than 30 indices, which by majority vote, that is, the number of clusters most chosen from the total of all the indices, determines the best number of clusters to use in the given problem.[14]. Those indices are:

- CH index (Calinski and Harabasz 1974)
- Duda index (Duda and Hart 1973)
- Pseudot2 index (Duda and Hart 1973)
- Cindex (Hubert and Levin 1976)
- Gamma index (Baker and Hubert 1975)
- Beale index (Beale 1969)
- CCC index (Sarle 1983)
- Ptbiserial index (Milligan 1980, 1981)
- Gplus index (Rohlf 1974; Milligan 1981)
- DB index (Davies and Bouldin 1979)
- Frey index (Frey and Van Groenewoud 1972)
- Hartigan index (Hartigan 1975)
- Tau index (Rohlf 1974; Milligan 1981)
- Ratkowsky index (Ratkowsky and Lance 1978)
- Scott index (Scott and Symons 1971)
- Marriot index (Marriot 1971)
- Ball index (Ball and Hall 1965)
- Trcovw index (Milligan and Cooper 1985)
- Tracew index (Milligan and Cooper 1985)
- Friedman index (Friedman and Rubin 1967)

- McClain index (McClain and Rao 1975)
- Rubin index (Friedman and Rubin 1967)
- KL index (Krzanowski and Lai 1988)
- Silhouette index (Rousseeuw 1987)
- Gap index (Tibshirani et al. 2001)
- Dindex (Lebart et al. 2000)
- Dunn index (Dunn 1974)
- Hubert statistic (Hubert and Arabie 1985)
- SDindex (Halkidi et al. 2000)
- SDbw index (Halkidi and Vazirgiannis 2001)

Reval is a tool that helps determine the best clustering solution on a dataset without having prior knowledge of the expected clusters. It uses a stability-based relative cluster validation method that transforms a clustering algorithm into a supervised classification problem, that is, by transforming the labels assigned to the data by the unsupervised algorithm into target labels and using a supervised learning algorithm to predict these same labels, it is possible to identify the best number of clusters by analyzing the error generated by the classification, and comparing it with the different N numbers of clusters. This allows selecting the number of clusters with the lowest expected misclassification error, which is a measure of stability. In summary, Reval is a useful tool to validate the robustness and stability of clusters in a dataset.

2.2.3 Supervised machine learning

Supervised learning is a branch of machine learning that involves using a set of labeled data to train a model. In this type of learning, the goal is to make the model learn to correctly map the inputs (input data) to the correct outputs (target data).

A. Most Typical Models

Decision Tree The basic idea of the decision tree is to divide the dataset into smaller subsets based on the characteristics of the data, until each subset contains only one possible outcome or class. To create a decision tree, the algorithm analyzes the characteristics of the input data and decides which is the best way to divide it into smaller subsets. This division is done based on a division criterion, such as entropy or the gini index, which measures the homogeneity of the resulting subsets[65].

Random Forest Random Forest is a machine learning model based on decision trees. It is an ensemble learning technique that combines multiple decision trees into a single model, with each tree created from a random subset of the training data and a random subset of the features of the data. The idea behind the random forest is that the combination of several decision trees reduces the risk of overfitting compared to a single decision tree. During the test phase, each individual tree of the random forest produces a prediction and the final prediction is given by combining the individual predictions of each tree, usually by voting[6].

Support Vector Machine (SVM) The idea behind SVM is to find a hyperplane that separates the different classes in the feature space of the input data, so that the classes are well defined and maximize the margin between them[34]. The hyperplane is chosen such that the distance between the points of each class and the hyperplane is maximized. These points are called support vectors. Kernel functions are used to map the input data to a higher-dimensional feature space where the classes are more likely to be linearly separable. The most common kernel functions are linear, polynomial and RBF (Radial Basis Function)[22].

Logistic Regression The idea behind logistic regression is to find a function that models the relationship between the features of the input data and the probability of belonging to one of the possible classes[33].

$$z = \beta_0 + \beta_1 x_1 + \beta_i x_i$$

In this logistic regression equation, z is the dependent variable(output labels) and x_i is the independent variable(input labels). In this model, the beta(β_i) parameter, or coefficient, is commonly estimated using maximum likelihood estimation. This method tests different values of beta through several iterations to optimize the best fit. All these iterations produce the "log likelihood" function, and logistic regression seeks to maximize this function to find the best parameter estimate. Once the optimal coefficient (or coefficients if there is more than one independent variable) is found, the conditional probabilities for each observation can be calculated, logged and summed to produce a predicted probability[32].

$$Sigmoid(z) = \frac{1}{1 + e^{-z}}$$

In binary classification, the sigmoid function is used, a probability less than 0.5 predicts 0, while a probability greater than 0.5 predicts 1, transforming a regression algorithm into classification.

K Nearest Neighbor (KNN) KNN is based on the idea of classifying a new data point based on its nearest neighbors in the training data. In the case of classification, KNN considers the K training points closest to the point to be classified and assigns the most common class among these points to the new point. In the case of regression, KNN considers the K training points closest to the point to be predicted and estimates the output for the new point based on the average of these points[43]. The distance between the points can be calculated using various metrics, such as the

Euclidean distance or the Manhattan distance. The value of K is a parameter that must be adjusted for each dataset.

Naive Bayes Naive Bayes assumes that the characteristics (or attributes) of the data are independent of each other, that is, that the presence or absence of one characteristic does not affect the presence or absence of another characteristic[61]. Naive Bayes calculates the probability that a given data point belongs to a specific class, based on the probabilities of each feature for that class. In other words, the algorithm estimates the probability that a given data point belongs to a given class, given the characteristics (or attributes) of that data point[61].

Neural Networks Neural networks are a type of machine learning algorithm that are inspired by the functioning of the human brain. They consist of layers of interconnected neurons that process information to perform classification, regression or other data processing tasks. Each neuron receives inputs from other neurons and applies an activation function to the weighted sum of these inputs to produce an output. The output of the neuron is then passed to other neurons in the next layer and the process is repeated up to the output layer, which produces the final[24] response. Neural networks are particularly useful for tasks involving complex data such as images or audio, such as analyzing medical images or developing drugs. One example, convolutional neural networks are a variant of neural networks that are especially good for computer vision tasks such as image recognition.

B. Cross-validation

Cross-validation (CV) is a popular method for model selection and parameter estimation in supervised learning. It works by splitting the data multiple times in order to estimate the performance of each classifier. A subset of the data, called the training set, is used to train the classifiers, while the rest, the test set, is used to evaluate the performance. There are many ways to split the data, but the most popular are: Leave-One-Out (LOO), Leave-P-Out (LPO) and K-Fold. The first two approaches are considered exhaustive splitters and the last one is a partial splitter. In LOO, each instance is successively removed from the sample and used for model validation, while in LPO, each possible subset of p instances is left out for validation. In CV K-Fold, the data is divided into k subsets, in each iteration, one subset is retained for testing, while the other subsets are used for training purposes, this iteratively alternating the test set each time, so that all K subsets of the data can be tested.

C. Validation metrics

Classification model validation metrics are used to assess the quality of a model's performance on classification tasks. These metrics allow measuring how well the model is able to correctly classify instances into different classes by comparing predictions with the actual values of the classes.

		Predicted classification	
		PC1	PC2
Actual classification	AC1	True Positive(TP)	False Negative(FN)
	AC2	False Positive(FP)	True Negative(TN)

Table 2.1: Confusion Matrix

Confusion Matrix Confusion matrices are an important tool for evaluating the performance of a classification model. They are used to compare the predictions of the model with the true values of the dataset and provide information about the accuracy of the model in each class[21].

A confusion matrix 2.1 is a table that summarizes the predictions of a classification model. It has two dimensions: one for the model predictions and one for the true values. The table cells contain the number of instances correctly and incorrectly classified by the model in each class. An ideal confusion matrix would have only values on the main diagonal (representing correct predictions), while all other cells would be zero.

From a confusion matrix, it is possible to calculate various performance metrics such as accuracy, precision, recall and F1. These metrics are useful for assessing the quality of the classification model and identifying areas that need improvement.

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

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

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

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN}$$

Receiver Operating Characteristic (ROC) are a visual tool used to evaluate the performance of classification models. They represent the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) at different classification thresholds. The Area Under the Curve (AUC) is a numerical metric derived from the ROC curve that provides a consolidated measure of the performance of a classification model.

Chapter 3

Related work

This chapter will be subdivided into three sections, namely, unsupervised temporal learning addressing methods that study the clustering of temporal sequences, stratification of ALS, focusing on stratification methods implemented in the disease, and prognostics in ALS, exploring the creation of models that seek to make specific predictions for ALS.

3.1 Unsupervised Temporal Learning

Liao[41] with Rani and Geeta[58], subdivide unsupervised temporal learning algorithms into three groups, raw-data-based, feature-based and model-based.

3.1.1 Raw-data-based models

These models act directly on the data, without extracting features, variables or parameters, their major modifications are the change of distance/similarity metrics for the comparison of different time series. S.Chandrakala[12], Aurangzeb Khan[35] and S.R.Nanda[51], all used Euclidean distances as a metric, although each had different objectives, and different algorithms. S.Chandrakala and C. Chandra Sekhar[12], proposed a density-based method for clustering multivariable time series, thus presenting Kernel DBSCAN, with heuristic methods in order to find out the best initial parameters. Aurangzeb Khan[35], used a hybrid clustering algorithm, to discover frequent patterns in the stock market, the algorithm consists of using KMeans to subdivide stocks into three movements, dead, slow and fast, and then applied the algorithm ”Most Frequent Pattern (MFP)”, to find the most frequent patterns in each type of group. S.R.Nanda[51], also applied clustering to the stock market, namely for portfolio management, where she used KMeans with the return of several stocks at different time instants to create a portfolio with minimized risks. Another distance metric, used by some researchers, is Dynamic Time Warping (DTW)[52], which consists of creating a matrix, containing all the distances of each point of two series, and then connecting from the end to the beginning, using the smallest distance for each point, making it possible to compare two undone series. T.W.Liao[42], used this metric with K-Means, fuzzy c-means and genetic clustering to group different time series, of different sizes, related to battle simulations, assigning various states during the battle. Vit Niennattrakul and Chotirat Ann Ratanamahatana[52],

in turn, applied KMeans to multimedia data, comparing Euclidean distance to DTW, and showing that the Euclidean Distance is much more general, and that KMeans with DTW, do not converge in certain cases. John Paparrizos and Luis Gravano[54], present a new method of raw-based clustering, but the distance metric consists of cross-relationships, that is, it verifies the similarity between two time series. In the article presented[54], they demonstrate the effectiveness of this algorithm, called KShapes, where it is compared with different state-of-the-art algorithms, managing to overcome all in terms of accuracy. Junjing Yang et al[78], used this algorithm to study energy consumption, where it counts the consumption, by hours, of several buildings, and uses KShapes to create groups with identical consumption, in order to improve the prediction of supervised learning algorithms. There are also other algorithms, such as KernelKMeans[19], which applies a kernel function, identical to SVM, to be able to stratify on linearly non-separable data, and other distance metrics, such as the ”goodness-of-fit”[45].

3.1.2 Feature-based models

Feature-based algorithms apply data reduction techniques, or attribute extraction, and use them as data for stratification. Principal Component Analysis (PCA)[46], is one of the most well-known feature extraction algorithms, it consists of translating a dataset into non-linearly separable variables. For the stratification of time series, it has been used in some articles such as Shaw and King[63], they used this method with two hierarchical clustering algorithms (ward and single linkage) to group oscillations in wind turbines. An identical but rather distant method is Independent Component Analysis (ICA)[67], while PCA compresses, ICA separates. Chonghui GUO et al[31] used this method, splitting the components, and using them with a modified version of KMeans to stratify the stock market. Jian Xin Wu[76], on the other hand, used ICA, not to stratify, but to forecast, in a way, similar to Chonghui, used the components stratified by ICA as input data to SVR, building a more robust forecasting model. Not only component reduction methods, but the use of transforms can also be a type of methods to help in the stratification of time series, in particular, the Wavelet transform, which is identical to the Fourier transform, but allows extracting information as a function of time, not just frequency. M. Vlachos[74], uses Wavelet with KMeans in a different way, because he applies KMeans, and updates the centroids, for all levels of the coefficients resulting from the transform, iteratively, as opposed to, for example, Geert Verdoolaege[73], who applied this transform to time series of blood oxygen levels, then applied KMeans to the distributions of the coefficients.

The feature-based algorithms with DeepLearning are almost all inspired by autoencoders, but with the use of recurrent networks, which have a greater capacity for processing sequences[66]. Thus, it is possible to compress them into a specific number of nodes, some authors assign the size of the latent space to the number of clusters [62, 70], the cluster being assigned to the activated node. For example, Neda Tavakoli et al[70] reports a case study where financial data consisting of time series, are clustered using a novel procedure divided into two phases, the first consists of creating labels from known features and transforming the data from unsupervised to supervised learning,

the second consists of using an autoencoder to cluster the series and return hidden features so that they help predict the assigned labels. Others use this method to compress sequences using only the encoder and then apply regular clustering algorithms, such as KMeans or hierarchical clustering, in the latent space. Zicong Zhang et al[79], for example, used this process to map disease progressions.

3.1.3 Model-based models

This type of algorithm considers that each type of series (clusters) is generated by a certain model or distribution, and the stratification is performed when the models that characterize the individual series are identical. The model ARMA is used by some authors[77, 3], this, is composed by the junction of two simpler models, the auto-regressive and the moving average. Xiong and Yeung[77], is an example, where the authors assumed that several time series were generated from k models ARMA, each model corresponds to a cluster, an expectation-maximization algorithm was used to mix coefficients, such as the parameters of the component models, in order to maximize the log-likehood between the series and the appropriate clusters, also finding the best number of clusters. Hidden Markov models (HMM), are other types of models capable of representing stationary series, Oates et al[53], presented a hybrid model of clustering with HMM to automatically detect the number of models (HMM) generators (clusters), this, begins by differentiating the various series using hierarchical clustering, and then uses them as training data for markov models (each cluster corresponds to a model), all series that lack classification are then assigned to the appropriate clusters, being selected by the largest log-likehoods.

3.2 Stratification in ALS

The stratification of pregressions is a problem with different solutions, several authors have already written different ways to stratify, with the most different techniques on more than different datasets. Divya Ramamoorthy et al[57] created the MOGP, which is built on a mixture of Gaussian processes to cluster different progressions of ALS, this algorithm was tested on a dataset with more than 3500 patients, focusing on the ALSFRS_R score in patients with at least 1 year of follow-up. Johann de Jong et al[17], on the other hand, used DeepLearning for the stratification of progressive diseases, namely recurrent variational autoencoders, called VaDER, with a focus on Alzheimer's and Parkinson's diseases. Harold H G Tan et al[69], used a somewhat different method to stratify subtypes of ALS, because it did not use the progressions that connote evolution, but rather, magnetic resonance imaging of different areas of the brain. From 488 patients, a probabilistic clustering algorithm was used, which determined that there were three subtypes of ALS, pure motor, frontotemporal and cingulate-parietal-temporal, each with distinct clinical profiles. Robert Kueffner et al[37], develop a study which contains more than 30 teams, developing machine learning and stratification algorithms, based on a dataset containing more than 10000 patients. All participants had to follow a baseline, starting with data pre-processing, then stratification, and prediction of life span in each cluster. Focusing more on stratification, this study revealed

several algorithms to cluster the various progressions of ALS, testing from 2 to more than 100 the number of clusters, but integrating the various methods applied, a consensus was revealed, namely in 4 groups, slow progressions, rapid progressions, late stage of the disease and early stage.
The prediction of the need for Non Invasive Ventilation (NIV) is part of many of the articles presented with these data. Sofia Pires et al[56], is no exception, being the focus of the article, the prediction of NIV, subdividing the patients according to the rate of progression. This rate is calculated by subtracting 48 (maximum possible score) from the initial ALSFRS_R value of the first visit, and dividing this total by the number of months between the onset of symptoms and the first visit. Then, the distribution of patients is calculated, with the 25% with higher and lower ratios being called fast and low progressions, and the 50% around the average, the average progressions. Marta Gromicho et al[30] used exactly the same clustering method, but for a different purpose, using Bayesian networks to establish relationships between some variables and with the different types of progressions.

3.3 Prognosis in ALS

The prognosis of ALS is useful for patients, families and doctors, so the development of new ALS prognostic techniques is the focus of study for many authors. M.A. del Aguila et al[18], is one of them. Interviewing a population of 180 subjects with ALS from Washington, he sought information on potential prognostic factors. Older age, female gender and any bulbar characteristics at the onset of the disease are key prognostic factors that tend to reduce the average life span. Machine learning also plays a key role in the development of new prognostic methods, Michael S. Bereman et al [5], used proteins taken from the blood of several patients to predict ALS, identifying a set of proteins that account for 49% of the variation in the ALSFRS score. Vincent Grollemund et al[29] used a dataset of more than 5000 patients to build a model to estimate the prognosis of survival at 1 year, using UMAP to map numerous variables in 2D, and subdivided the projection into 3 zones, each relative to the survival rate. Jason Ackrivo et al[1], attempted to predict the onset of respiratory failure in ALS, with a dataset of 765 patients, using variables such as the patient's age, the interval between the onset of symptoms and diagnosis, or the value of the ALSFRS_R score. The main outcome assessed was respiratory failure up to 6 months, defined by the start of non-invasive ventilation, tracheostomy or death. External validation was also applied, testing on a dataset of more than 7000 patients, obtaining results which, although worse, were very positive. Other projects have already been carried out with the Lisbon dataset, many of them referring to quantifying the importance of static variables in the speed of the disease.Tiago Leão et al[39], use Bayesian networks in the prediction of subscores, both before patients need of NIV, and after, so it is possible to understand which variables are most important at various times throughout the disease. Another study with a similar purpose was Marcel Muller et al[50], who used LSTM networks to predict the patient's respiratory evolution (subscore R), with the addition of the SHAP tool, to identify which variables have the greatest influence on the prediction. Not so much focused on the variables, other articles address more the issue of predicting the need for NIV, for example,

Diogo Soares et al [64], created a prediction model, using data mining methods, biclustering for static data and triclustering for temporal data, thus predicting, within 90 days, whether the patient would need NIV or not. With the same intention, but with a different approach, Telma Pereira et al[55], used conformal learning, that is, a framework built on classifiers, which when applied to regression problems, results in an interval that estimates the possible outcome, the size of which varies according to the degree of certainty of the models, so it is possible to predict the need for NIV, and what time interval they may need (3, 9 or 12 months).

Chapter 4

ALS Data

This chapter talks about the data, how it is transformed into numerical values to be used in machine learning algorithms, where it comes from, how it is divided, and what treatment it has received.

4.1 Description

The data used in this project are divided into 2 sets, both containing clinical data from respiratory and neurophysiological tests, and demographic factors.
The first set comes from the Hospital de Santa Maria, Lisbon, where all observations come from 1995 to 2022, from about 1560 patients, with 8066 tests and 51 characteristics, each consultation occurs on average every 3 months, and each patient has on average 5.17 clinical observations.
The second set, comes from two centers ALS, located in Torino and Novara, for a total of 2156 patients, where all clinical observations occur between 1995 and 2020, and each patient has on average 7.22 clinical observations, both the average time between each consultation and the number of characteristics is the same as the first set.
The data was preprocessed according to André Carreiro[9], where a hierarchical clustering-based approach is used to group all tests by dates, forcing all tests with a relatively close date to be in the same group. There are some restrictions in the algorithm, two observations of the same test cannot be in the same group, and all observations have to have the same NIV state, although the non-evasive ventilation(NIV), is not relatively important for this project, the creation of these snapshots allowed an easier handling of the temporal sequences by the tests performed.
Both datasets are divided into two parts, the static data, which contains data that does not change over time, mostly made by initial assessments, medical procedures and demographic information. And the temporal data, which contain the scores that connote different aspects of the patient's evolution. The table 4.2 gives a better view of the data.

4.2 ALSFRS-R Scale

The connotation of ALS progression can be performed using a functional rating scale (ALSFRS_R). This measures 12 aspects of motor function, ranging from the ability to swallow to the

ability to climb stairs and breathe, each function is scored from 0 (no ability) to 4 (normal), with a maximum score of 48[11]. This test, when taken over a period of time, allows to connote the progression of the disease in numerical values, giving the possibility to better study the evolution.

Question	About	SubScores	Score
1	Speech	ALSFRSb	ALSFRS_R
2	Salivation		
3	Swalling		
4	Handwritting	ALSFRSsUL	
5	Cutting food		
6	Dressing and hygiene		
7	Turning in bed	ALSFRSsLL	
8	Walking		
9	Climbing stairs		
10	Dyspnea(difficulty breathing)	R	
11	Orthopnea(shortness of breath while lying down)		
12	Breathing insufficiency		

Table 4.1: Question topics for determining ALSFRS_R score

The table 4.1 gives a better view of how the ALSFRS_R score is computed, showing the topic of each question, the group where it belongs (subscore), and how the sum of all subscores gives the ALSFRS_R score. The same table is inspired in [11].

Currently, 12 questions are asked, and scores are calculated from them. However, older versions of this form contained only 10 questions, and the last question (old10), related to breathing, was replaced by three others. It is normal in some data, especially in older patients, to still have clinical observations based on 10 questions.

4.3 Data Preprocessing

The different phases of the project required different treatments, the following subsections present how the data were treated in each situation.

4.3.1 Stratification

The stratification, both univariate and multivariate, had an identical treatment, the only variation being in the choice of which scores to use, since the univariate only uses ALSFRS_R, while the multivariate uses its subscores (ALSFRSb,ALSFRSsUL,ALSFRSsLL,R). As in either case, multivariate or univariate, it is necessary to remove all static variables, and only use the convenient scores, as the objective is to stratify taking into account only the first year of follow-up, and as each consultation is performed on average every 3 months, only the first 5 consultations were con-

Name	Type	Temporal/Static	SubGroup
REF	Categorical	Static	Identification
Birth_year	Numerical	Static	Onset Evaluation
Age_onset	Numerical	Static	Onset Evaluation
Gender	Categorical	Static	Demographics
Ethnicity	Categorical	Static	Demographics
NIV	Categorical	Static	Medical Procedure
Date_NIV	Date/Categorical	Static	Date
Tracheostomy	Categorical	Static	Medical Procedure
PEG	Categorical	Static	Medical Procedure
Date_PEG	Categorical	Static	Date
UMNvsLMN	Categorical	Static	Onset Evaluation
Onset	Categorical	Static	Onset Evaluation
Limb_O	Categorical	Static	Onset Evaluation
Limbs_Impairment	Categorical	Static	Onset Evaluation
Limbs_Side	Categorical	Static	Onset Evaluation
Height (m)	Numerical	Static	Onset Evaluation
Weight	Numerical	Static	Onset Evaluation
Weightloss_10%	Categorical	Static	Onset Evaluation
ALS_familiar_history	Categorical	Static	Medical and Family History
Ever_smoked	Categorical	Static	Onset Evaluation
Blood_hypertension	Categorical	Static	Medical and Family History
Diabetes	Categorical	Static	Medical and Family History
Dyslipidemia	Categorical	Static	Medical and Family History
Thyroid	Categorical	Static	Medical and Family History
Autoimmune	Categorical	Static	Medical and Family History
Stroke	Categorical	Static	Medical and Family History
Cardiac_disease	Categorical	Static	Medical and Family History
Primary_cancer	Categorical	Static	Medical and Family History
SOD1 Mutation	Categorical	Static	Genetic
C9orf72	Categorical	Static	Genetic
TARDBP mutation	Categorical	Static	Genetic
FUS mutation	Categorical	Static	Genetic
DateOf1stSymptoms	Date/Categorical	Static	Date
DateOfDiagnosis	Date/Categorical	Static	Date
Date_Critical	Date/Categorical	Static	Date
ALSFRS_*/P*	Numerical	Temporal	Functional Scores
ALSFRSb	Numerical	Temporal	Functional Scores
ALSFRSsUL	Numerical	Temporal	Functional Scores
ALSFRSsLL	Numerical	Temporal	Functional Scores
R	Numerical	Temporal	Functional Scores
ALSFRS_R	Numerical	Temporal	Functional Scores
%FVC	Numerical	Temporal	Respiratory Tests
firstDate	Date/Categorical	Temporal	Date
lastDate	Date/Categorical	Temporal	Date
medianDate	Date/Categorical	Temporal	Date

Table 4.2: Data type, classification and subgroup

sidered, the first consultation being consultation 0. All consultations above 5 were discarded. In summary, all data processing can be summarized in the following steps:

- Removal of static attributes and ALSFRS_R score or subscores, depending on the type of stratification.
- Removal of patients with less than 5 visits (1 year).
- Removal of patients with NaN in some of the scores.
- Removal of visits from the 5th onwards, forcing the remaining patients to contain only the first 5 visits.
- Removal of attributes related to dates.
- Data normalization (StandardScaler).

4.3.2 Prognosis of different types of progressions

As the main idea is to inform the patient directly in the first consultation, what type of ALS progression he may have, it was necessary to collect as much information as possible available from the first consultation, for this, it was decided to use all static data with the exception of medical procedures, such as NIV, Percutaneous Endoscopic Gastrostomy (PEG) and tracheostomy, since these are only indicated at a more advanced stage of a disease, the variable weightloss$\geq$10%, was also removed, for the same reasons. However, while some variables were removed, others were added, because although they are not static data, the addition of variables such as the ALSFRS_R score from the first consultation is information that contributes to a better prediction of the type of progression that the patient may develop, and as mentioned, as this information is assigned directly at the first consultation, it can be used for prognosis. We also added a variable called progression_rate, taken from the articles [56, 30], to understand how the patient has evolved from the date of the first symptoms to the date of the first visit, and it is given by:

$$ProgressionRate = \frac{48 - ALSFRSR_{1stVisit}}{\Delta t_{1stSymptoms;1stVisit}}$$

This can be described by the equation above, subtracting the ALSFRS_R score from the maximum value and dividing by the time, in months, between the first symptoms and the first consultation. It was also applied to all the subscores, with the difference that instead of 48 as the maximum value, 12 was used, this being the highest possible value for each subscore.
Summarizing the whole process, the following list shows in order all the steps performed.

- Removal of all patients who were not previously stratified
- Separation of static data and subscores referring to the first visit
- Addition of progression_rates

- Removal of dates, NIV, Tracheostomy, PEG and weightloss$\geq$10%
- Removal of variables with more than 50% NaN
- Univariate imputation, mean for numerical variables and mode for categorical ones
- Normalization of data (StandardScaler).

4.3.3 Forecasting medical procedures

For the prognosis from the multivariate stratified groups, the snapshots created by André Carreiro[9] were used, those are divided into 5 datasets, each one representing a medical procedures:

- C1 - Non-invasive ventilation(NIV)
- C2 - Communication aids
- C3 - Percutaneous endoscopic gastrostomy(PEG)
- C4 - Need for a caregiver
- C5 - Wheelchair

A snapshot is related to a consultation, and the aim is to predict when patients will need these procedures (C1, C2, C3, C4, C5) in a time scale (90, 180 or 365 days), so there was a need to treat snapshots. Basically, it is almost identical to the previous subsection (4.3.2), however, the formula for calculating progression_rates has been changed, since the value of the first consultation does not matter, but rather, of each consultation individually, also dividing by the distance in months from the date of the first symptoms to the date of the consultation.
All steps can be represented as follows:

- Removal of patients who were not used in the stratification
- Removal of variables with more than 50% NaN
- Addition of 5 new attributes, namely a variation of the progression_rates, where instead of using the date of prognosis and the subscore value of the first visit, the date of the given visit and the subscore value of that same visit are used.
- Addition of BMI (body mass index) and removal of height and weight
- Univariate imputation, mean for numerical variables and mode for categorical ones.
- Normalization of data

Due to the imbalance of the data, it was decided to apply SMOTE-ENN to the snapshots.

Chapter 5

Methodology

5.1 Proposed stratification models

The process of selecting the best model to cluster ALS sequences is identical for both univariate sequences (only taking into account the ALSFRS_R score) and multivariate sequences (taking into account the subscores ALSFRSb, ALSFRSsUL, ALSFRSsLL and R). Four component extraction algorithms (AutoEncoder, PCA, ICA and Wavelet transform) were tested, where KMeans or Hierarchical clustering with different parameters was then applied to the extracted components, this for both univariate and multivariate. To check which algorithm was best, the comparison process was divided into two steps: The first, looking for which is the best number of clusters in each algorithm present, this using internal validation(NbClust). The second, using the stability validation (Reval), checks all the algorithms used with their respective best number of clusters, which will present the lowest normalized stability.
Figure 5.1 below represents the entire process of finding the best model.

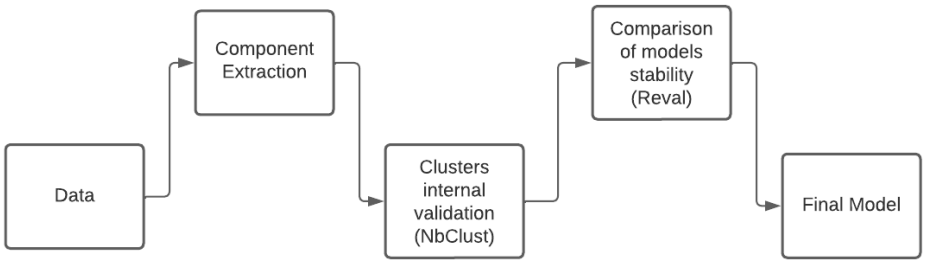

Figure 5.1: Process for finding the best model

5.1.1 Multivariate stratification

Graph 5.2 represents the number of clusters, determined by Nbclust (white numbers inside the graph), and the calculation of their normalized stability, for the multivariate sequences.

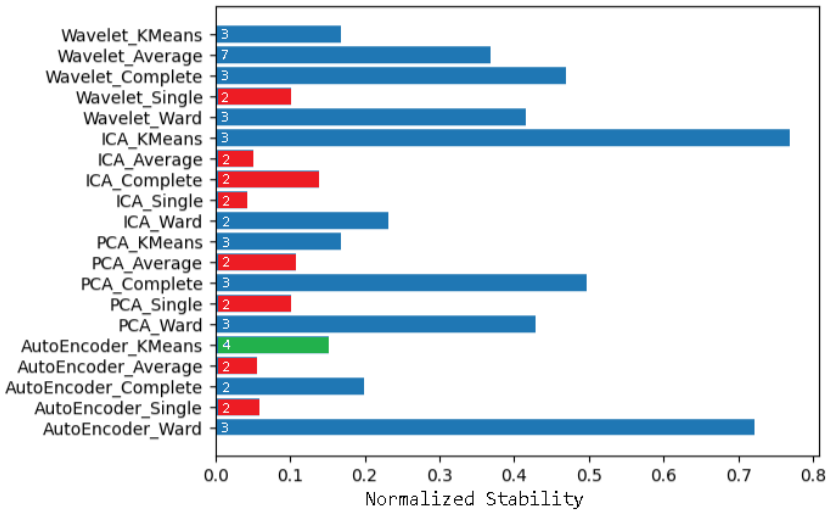

Figure 5.2: Number of clusters (white numbers) and normalized stability of various algorithms for multivariate sequences

All the graphs in red are candidates for choosing the best clustering algorithm, however, they show discrepancies between the various clusters in terms of the number of patients, for example, when using the hierarchical clustering algorithm with single linkage, one of the clusters only contains one patient, while the opposite one has 1200, so, and removing all the algorithms that do not show reasonable ratios (graphs in red), the algorithm that obtained the best performance was Autoencoder with KMeans (graph in green).

5.1.2 Univariate stratification

As mentioned above, univariate stratification consists of creating clusters using only the ALS-FRS_R score.

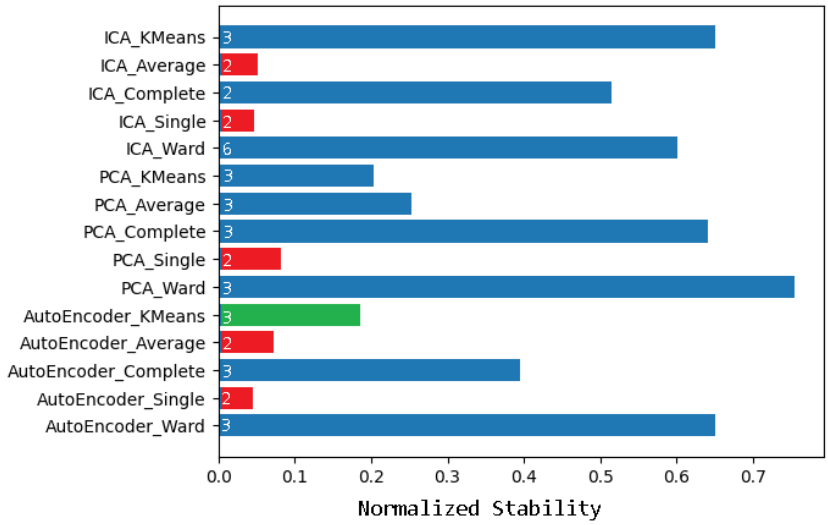

Figure 5.3: Number of clusters (white numbers) and normalized stability of various algorithms for univariate sequences

Figure 5.3, shows all the algorithms, the number of clusters with the most votes according to NBClust (numbers in white) and the normalized stability. Since what is sought are the lowest values, the graphs in red, are the first to be selected as possible candidates, however, the relationship between the number of patients per cluster is quite bad, as they present large discrepancies, the smallest being 1259 to 4. Thus, the first algorithm that presents a low normalized stability, and a good relationship between the number of patients per cluster is AutoEncoder with KMeans (green graph), the same as multivariate stratification.

5.1.3 LSTM AutoEncoder with KMeans

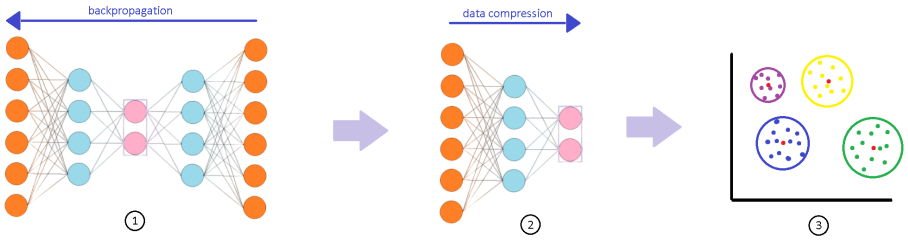

Figure 5.4: AutoEncoder sketch with KMeans

The LSTM AutoEncoder algorithm with KMeans, was the one that obtained the best performance in both multivariate and univariate, it consists of three simple steps.

1. The first step is to train the AutoEncoder with the data, so that the output is as identical as possible to the input provided, in this way, the latent space is forced to compress the information into a smaller number of nodes than the size of the original information.

2. The second step consists in removing the encoder, and using it to compress the previously trained information.

3. The third step, with the compressed information, uses the KMeans algorithm and identify the clusters.

The figure above 5.4, is a graphical sketch of how these three steps are elaborated.
The LSTM network chosen is made up of 7 layers, which are interspersed between LSTM and batchnormalization cells, in the 4th layer, where the latent space is located, it is made up of 20 perceptual neurons (Dense), this network uses Adam as the optimizer, and ”mean_squared_error” as the loss function, since the aim is to match the output result to the input.

5.2 Prognosis of different types of progressions

The idea is to predict, from static data, the type of progression that a patient may develop, to indicate to clinicians, at a patient’s first visit, what this may imply for their future, allowing them to give clinicians a more refined view of what to expect from the disease.

The available data were divided into 80% for training, where stratified 10-fold cross-validation was applied, due to the imbalance of the data, and 20% for testing.

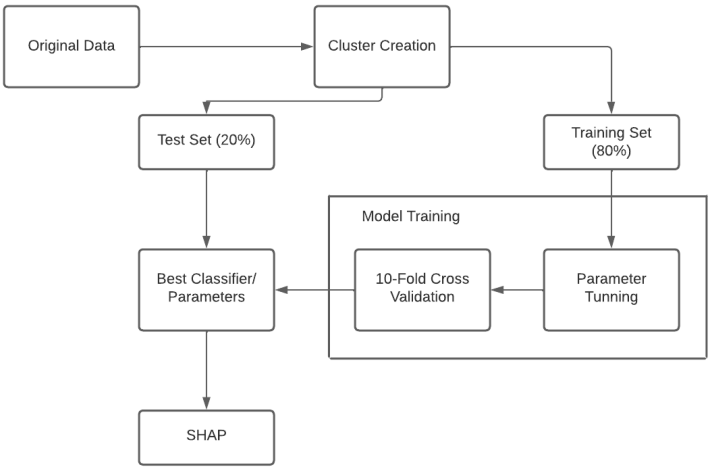

Figure 5.5: Diagram of the training and testing of models to differentiate types of progressions

Figure 5.5 is an outline of the entire process used to make the prognosis for a new patient. The first step was to define the clusters, which can be reconstructed in subsection 5.1.3, where the process of creating the clusters is explained. After being labeled according to the type of progression, the data is divided into 80% training and 20% testing. The training data is used to compare different machine learning models, where hyperparameter tuning was applied to find the best parameters for each model, and 10-fold cross-validation was applied to identify the best supervised learning model.
Table 5.1 displays the models that were utilized, and the optimum hyperparameter obtained for each model.

Models	Hyperparameter
RandomForestClassifier	criterion: gini; n_estimators: 120
DecisionTreeClassifier	criterion: gini; splitter: best
SVC	C: 1; kernel: linear
LogisticRegression	C: 1; penalty: none; solver: saga
KNeighborsClassifier	algorithm: auto; n_neighbors: 20; weights: distance
GaussianNB	var_smoothing: 0.0187

Table 5.1: Models and their best parameters for classifying different types of progression

Finally, the SHAP is applied in order to find out which variables most influence the prognosis of different types of progression.

5.3 Forecasting medical procedures in different types of progressions

The objective of predicting certain medical procedures according to the type of progression is to understand whether creating classifiers specialized only in one type of evolution improves the prediction of certain medical procedures, when compared to general classifiers that do not differentiate the types of disease progressions.
The data of the specialized models were divided into 80%/20%, and the 80% of the general model, is the junction of all 80% of the specialized, from the figure 5.6 it is possible to see the outline of the construction of the training data of the general model, from the various clusters.
The main reason why the 80%/20% division of the general models is carried out in this way is because of the representativeness of some clusters, since some of these have very few patients (e.g. Diffuse). When dividing the general data into training/testing, it can happen that the clusters with low representativeness are not fairly represented in either training or testing, so with the division shown in figure 5.6 it is possible to guarantee the representation of all the clusters, both in training and testing.

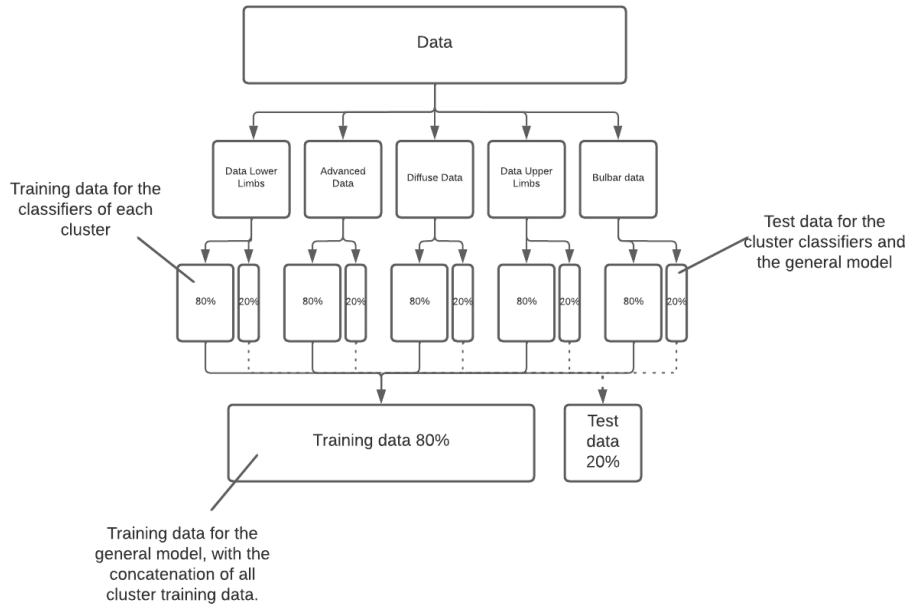

Figure 5.6: Diagram of how data is split

Only one supervised learning model was used for training, RandomForests, in order to better understand the influence of clusters, since the aim is to compare whether specialized models trained on each type of progression(cluster) are better than general models trained on general data, not to find the best possible model for such a dataset. To do this, hyperparameterisation was applied to find the best parameters, and stratified cross-validation divided into 10 parts was used, as mentioned in subsection 4.3.3, SMOTE-ENN was applied, due to the imbalance of the data, both in the general and specialised sets. This whole process was repeated 5 times.

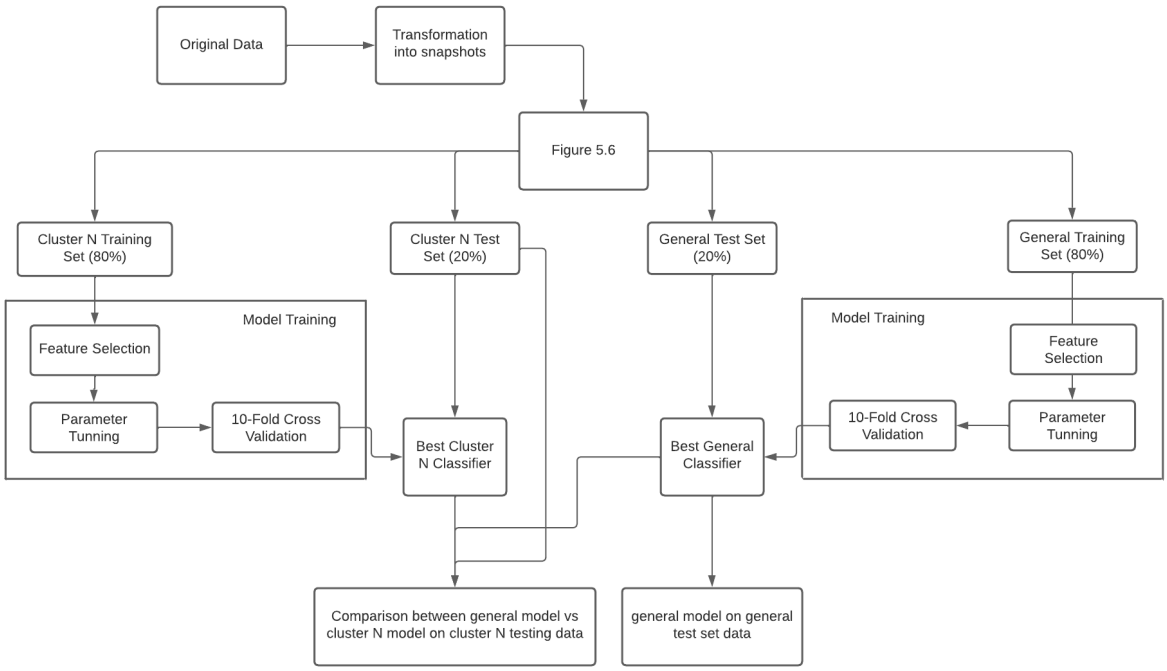

Figure 5.7: Process for training and testing medical procedure prediction models

Figure 5.7 shows how the models are trained and tested, although it does not refer to any specific cluster(N), this diagram shows how the specialized models are compared against the general ones. The basic idea is to remove two pieces of information at the end:

1. A comparison between the test data of a specialized model, against the general one.
2. To understand if any specialized model can perform better on its test data than the general models on the general test data.

From the diagram, it can be seen how the data is divided, and how the general models and specialized models are trained. Feature selection and hyperparameterization were applied to find out which variables and parameters best optimize the models, and 10-Fold cross-validation was applied to see how good the model is in general, before comparisons with other data. Next, a comparison is made between the general model and the specialised model, comparing these on the specific test data for a given cluster, to see if there is any improvement in creating these cluster-specific models.

Chapter 6

Stratification Results

6.1 Multivariate stratification

The multivariate stratification is identical to the univariate one, except that not only the ALSFRS_R score is taken into account, but also its sub-scores (ALSFRSb, ALSFRSsUL, ALSFRSsLL, R). This section analyzes the characteristics of the resulting clusters, identifying the greatest disparities between them and what each one represents.

6.1.1 Analysis and reformulation

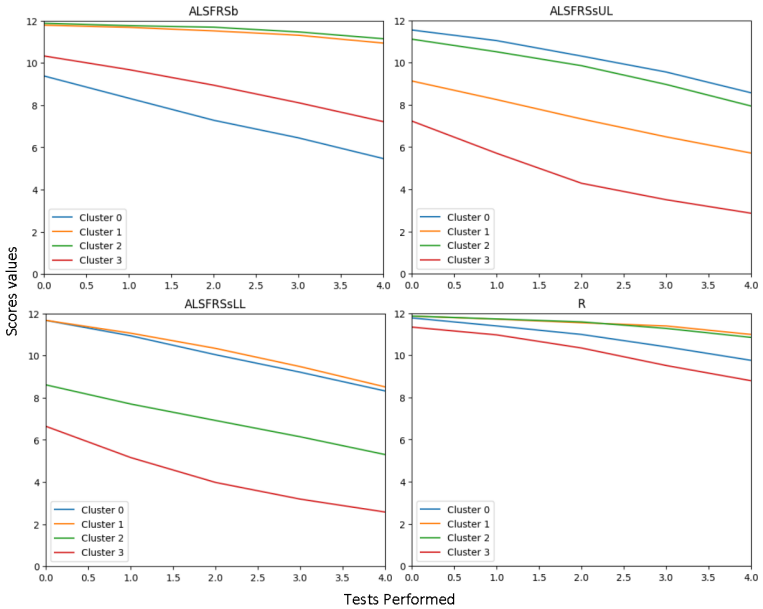

Figure 6.1: Mean multivariate progressions

Figure 6.1 represents the average progressions of the clusters, by inspection, for each subscore, it is possible to see what each group represents, but when viewed as a whole, it is difficult to see the difference between each cluster. For this reason, it was decided to apply a treatment, transforming a line graph into a bar graph, using their averages, first between each cluster, and then between

the subscore, which allows better inspection and different conclusions to be drawn. The following figure 6.2 shows how this treatment is carried out.

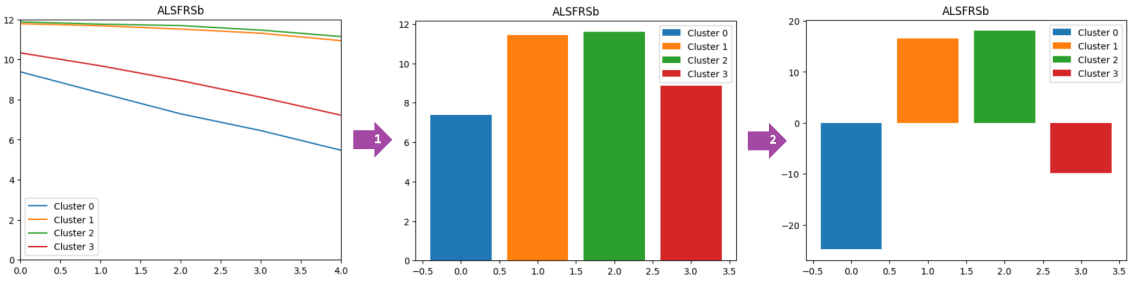

Figure 6.2: Example of the transformation from a subscore to a bar chart

Figure 6.2, shows the example of the transformation of a subscore, namely the ALSFRSb, to a bar graph of easier understanding, this process is divided into 2 steps:
The first, consists in calculating the average of each progression, that is, an average of all the values of the subscore in question over the year of which the patient was followed, this for each group.
The second step is to divide each cluster by the average of all the clusters, calculating the percentage difference.
This makes it much more noticeable, in the total set of all subscores, what each cluster represents. Figure 6.3, shows the transformation in all the subscores, and by inspection, it is possible to directly remove 4 different types of progression.

- Cluster 0 contains only two subscores that stand out, namely ALSFRSb, and R,with the great discrepancy being in the subscore ALSFRSb, attributed this cluster to patients with a mostly bulbar evolution.

- Patients in cluster 1, present only a discrepancy in the ALSFRSUL subscore, consisting of patients containing progressions with an initial presentation in the upper limbs.

- Patients in cluster 2 have the same behavior as patients in cluster 1, but in a different subscore (ALSFRSsLL), so this cluster presents patients with progression whose first symptoms were in the lower limbs.

- In cluster 3, the patients who are part of it, present low values, when compared to the average, in all subscores, revealing that they are already in an advanced stage of the disease, calling this cluster as advanced.

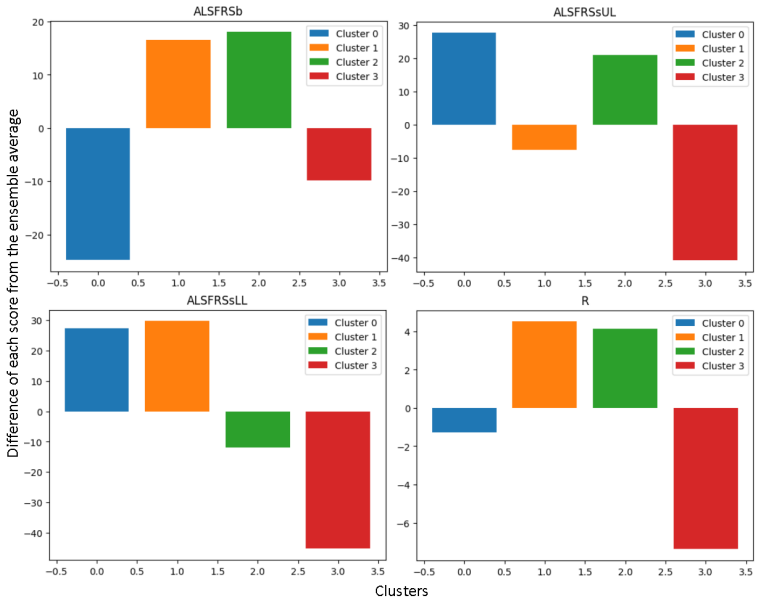

Figure 6.3: Clusters(4) after treatment in each subscore

However, clinically, and according to some studies, not all types of ALS progression are represented in these 4 clusters, as there are people whose disease has a diffuse initial revelation, this means, they present with various symptoms without it being possible to understand or categorize the specific type of progression (lower limbs, upper limbs or bulbar). For this reason, it was decided to create a 5th cluster, which contains patients who show various signs of different types of progression.

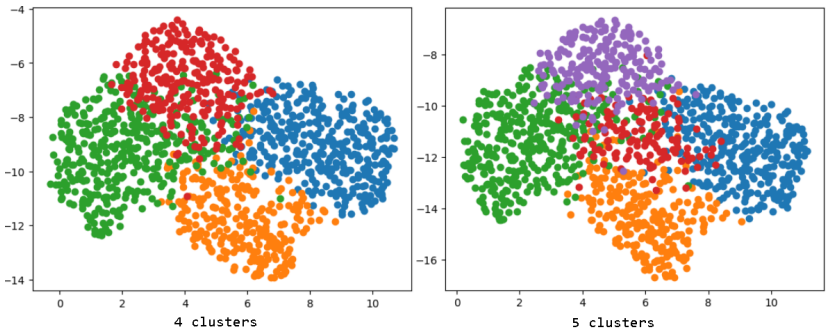

Figure 6.4: Graphical comparison between 4 and 5 clusters, applying dimensionality reduction with UMAP

It was compared the representation in 2 dimensions between several dimensionality reduction algorithms (T-SNE, UMAP, PACMAP), to understand where the addition of a 5th cluster comes from. Of the three representations applied, the one that best presents itself is the UMAP (figure 6.4), where the origin of the 5th cluster is easily understood, this one, consists of the union of the 4 original clusters.

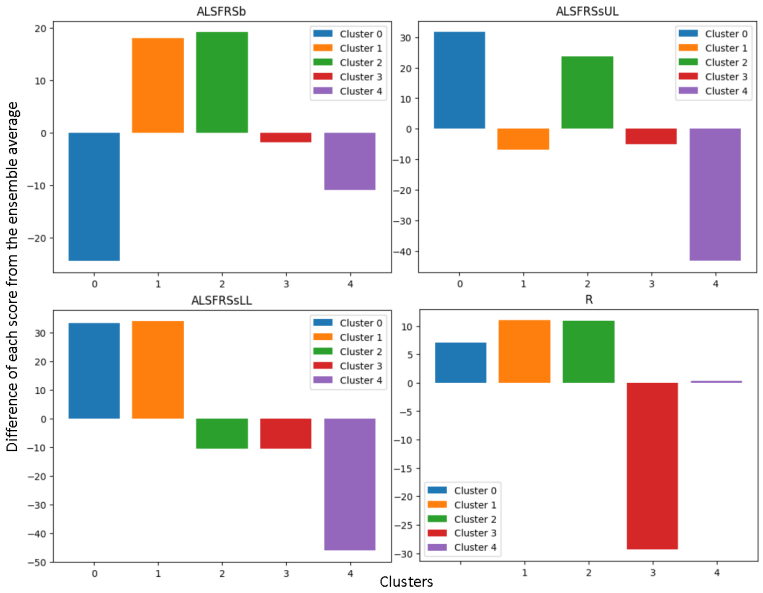

Figure 6.5: Clusters(5) after treatment in each subscore

By inspecting the graphs, figure 6.5, it is easy to see that the initial clusters, namely, the bulbar, upper and lower limbs are still represented (clusters 0, 1 and 2), however, a new cluster was formed (3), which consists of a strong respiratory component, but with low values in all other subscores (ALSFRSb, ALSFRSsUL and ALSFRSsLL), thus, as represented in figure 6.4, it is possible to verify the influence of all the other clusters, differentiating the respiratory component. The advanced cluster(4), also contributed to the construction of this one(3), being still strongly influenced by the decay of all its subscores, except the respiratory one.

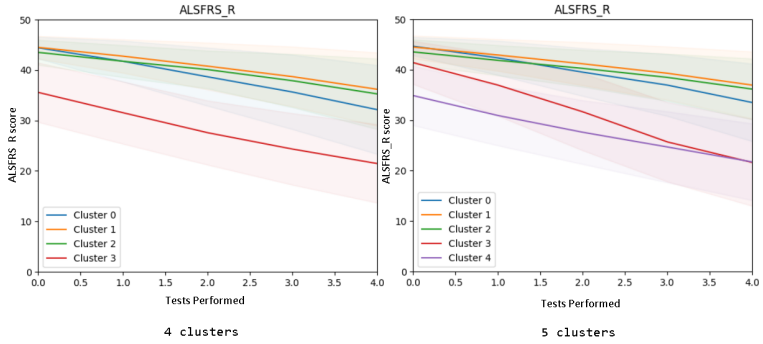

Figure 6.6: ALSFRS_R scores of 4 and 5 clusters

From figure 6.6, which represents the ALSFRS_R score of all these clusters, making it easier to understand at what stage of the disease each cluster is, it is possible to determine that clusters 0,1 and 2 are in identical stages, even with different initial symptoms, while cluster 4 (advanced), is much more advanced, with a much lower average, and patients in cluster 3, despite starting at an initial stage identical to clusters 0,1 and 2, end up having a much faster deterioration, ending with almost the same value as the advanced cluster (4).

6.2 Univariate Stratification

The univariate stratification consists in the creation of clusters, only using the ALSFRS_R score. This section analyzes the groups of results from the univariate aggregation and looks at average life expectancy.

6.2.1 Analysis

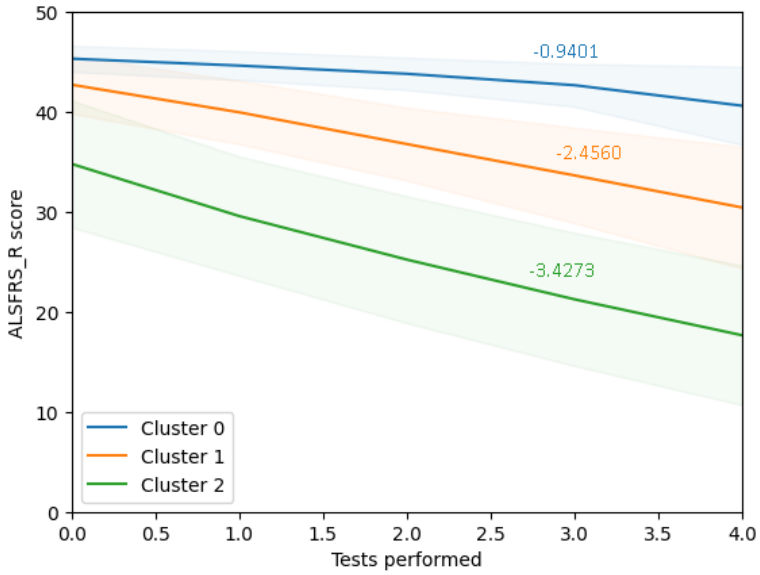

Figure 6.7: Averages

Figure 6.7 shows the average progression of the ALSFRS_R score with the standard deviation, making the evolution of each group noticeable when analyzed together. From the figure 6.7, it is possible to get a basic idea of how the patient's progression will be, depending on whether it is a fast or slow deterioration. Using the slopes is a way to quantify how fast the disease is evolving, the lower the slope, the faster the progression. When inspecting the graph, it is easy to see that cluster 2 is a fast progressor, by losing more than 10 points in ALSFRS_R over the year. On the other hand, cluster 0 is the slowest progressor, since it remained almost constant, without losing almost any ALSFRS_R point. Cluster 1 is a medium progressor, not decaying as fast as cluster 2, but not as slow as cluster 1.

Cluster	Life Expectation	
	Mean	Std
0	14.61	7.49
1	10.41	5.50
2	8.53	7.49

Table 6.1: Average number of tests by each cluster

Cluster	Life Expectation	
	Years	Months
0	4	6
1	3	2
2	2	6

Table 6.2: Average life expectation in years and months by each cluster

A metric for comparing evolutions is the average life expectation, in the table 6.2 by the number of years and months, or in the table 6.1 by the number of tests, there is a certain disparity between the life span of each cluster. This calculation is performed based on the number of patient tests, allowing to have a perception of the average life expectancy of each group. However, there is a weakness in this approach as some patients are still early in their disease and therefore have not had many tests, which may reduce the average. Still, it is possible to compare life expectancy between each group of patients. As can be seen, the life span of each group can be related to the slope of the curve. The higher the value of the slope, the more tests and years of disease the group has, which means that patients with slower progression tend to live longer, while patients with faster progression tend to live less, especially when looking at group 2.

Chapter 7

Prognosis of different types of progressions

The idea of predicting the clusters described in section 5.2 from static data is to indicate directly to clinicians, at the first consultation with their patients, what kind of evolution can be expected, what it might entail in the future, allowing to give to the clinician a more refined view of what to expect from the disease, providing better management and helping to decide which treatments are best to apply. The SHAP tool was also applied to understand which static variables had the most impact on the prediction of clusters, also allowing to understand which factors tend to influence each type of disease progression.

7.1 Classification of groups

Several supervised learning models were compared, GridSearchCV was applied in the search for the best parameters and then cross-validation, in the following table 7.1, it is possible to see the metrics of each model, in order to be able to compare them.

	Precision	Recall	F1
RandomForestClassifier	0.80±0.03	0.80±0.03	0.79±0.03
DecisionTreeClassifier	0.73±0.03	0.73±0.03	0.73±0.03
SVC	0.81±0.03	0.82±0.02	0.81±0.03
LogisticRegression	0.82±0.03	0.82±0.03	0.81±0.03
KNeighborsClassifier	0.74±0.04	0.72±0.03	0.69±0.03
GaussianNB	0.74±0.03	0.75±0.03	0.74±0.03
	Accuracy	AUC	MCC
RandomForestClassifier	0.80±0.03	0.94±0.01	0.75±0.03
DecisionTreeClassifier	0.73±0.03	0.83±0.02	0.65±0.04
SVC	0.82±0.02	0.95±0.01	0.77±0.03
LogisticRegression	0.82±0.03	0.95±0.01	0.77±0.03
KNeighborsClassifier	0.72±0.03	0.89±0.02	0.65±0.03
GaussianNB	0.75±0.03	0.89±0.02	0.68±0.04

Table 7.1: Averages of various model metrics for cluster prediction

By inspection, the model that obtains the best performance, being a technical tie between logistic regression, is the SVM, despite obtaining almost all the identical metrics, these obtain a better performance when looking at the AUC and MCC. With the application of the hyperparameterization, the SVM model applied to the test data, contains the value of C equal to 1, and applies a linear kernel. The following shows the value of the metrics in the portion of data reserved for testing, it is possible to see that this model has a reasonably good classification, without suffering from overfitting.

- Precision - 0.81
- Recall - 0.82
- F1 - 0.81
- Accuracy - 0.81
- Roc_Auc - 0.94
- Matthews_CorrCoef - 0.77

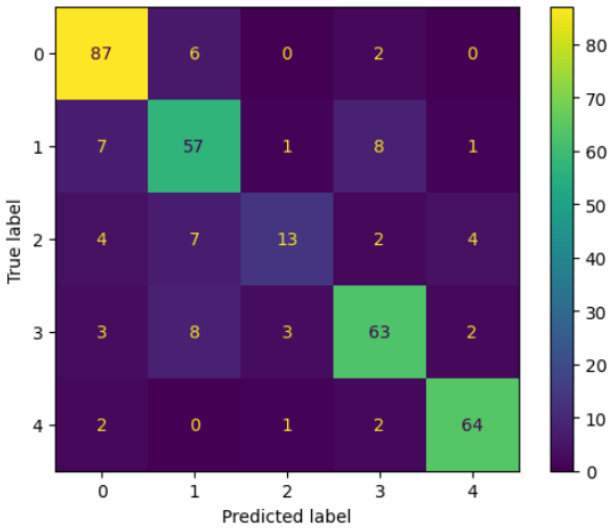

Figure 7.1: Confusion matrix of test data

But from the figure 7.1, the confusion matrix, cluster 2, called diffuse, it is possible to realize that its classification obtains a worse performance than all the others.
One of the reasons for the worse performance in the classification of the diffuse cluster is the amount of data, being the dataset imbalanced, the little acquisition of cluster 2 compared to all the others, can influence the bad categorization of this, so it was tried to apply SMOTE-ENN in the training phase, so that the performance of the classifier can be corrected and improved.

In this case, the best performing model was the random forest, taking into account the table 7.2, although it is a technical tie with the SVM, these obtain better performance. The best hyperparameters were entropy, to measure homogeneity, and 150 estimators. The matrix 7.2 is the

	Precision	Recall	F1
RandomForestClassifier	0.78±0.03	0.77±0.03	0.77±0.03
DecisionTreeClassifier	0.76±0.03	0.73±0.04	0.74±0.04
SVC	0.78±0.04	0.77±0.04	0.76±0.04
LogisticRegression	0.77±0.04	0.76±0.04	0.76±0.04
KNeighborsClassifier	0.70±0.04	0.68±0.04	0.66±0.04
GaussianNB	0.69±0.04	0.70±0.03	0.67±0.04
	Accuracy	AUC	MCC
RandomForestClassifier	0.77±0.03	0.93±0.02	0.71±0.04
DecisionTreeClassifier	0.73±0.04	0.84±0.02	0.67±0.05
SVC	0.77±0.04	0.92±0.02	0.71±0.05
LogisticRegression	0.76±0.04	0.76±0.02	0.70±0.05
KNeighborsClassifier	0.68±0.04	0.87±0.02	0.60±0.05
GaussianNB	0.70±0.03	0.86±0.03	0.62±0.04

Table 7.2: Averages of various model metrics for cluster prediction with SMOTE

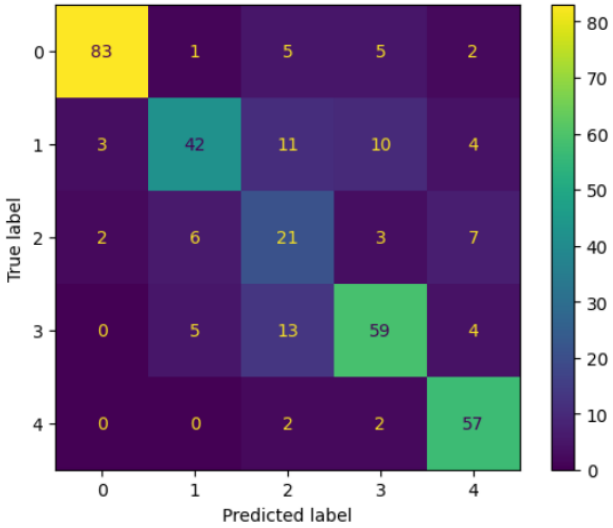

Figure 7.2: Confusion matrix of test data(SMOTE-ENN)

result of applying it to the 20% test cases. Despite the increase of well classified cases in the Diffuse(2) cluster, the improvement is not very visible, because when dividing the correct cases over the number of total cases in the cluster, that is, measuring the precision focused on the diffuse cluster(2), there is a difference of 10 percentage points, with the classification with SMOTE-ENN of 53.8% and without 43.3%, so, despite an improvement in this cluster, in general, the application of SMOTE-ENN obtains worse performance, this when inspecting the other metrics, as can also be seen from the following results.

- Precision - 0.77
- Recall - 0.75
- F1 - 0.76

- Accuracy - 0.75
- Roc_Auc - 0.92
- Matthews_CorrCoef - 0.69

7.2 SHAP

The SHAP tells the importance of each variable for the classification, so in order to facilitate the distribution of the categorical variables, onehotencoder was used. Figure 7.3 shows the 20 variables that have more importance in the correct classification, the figure 1 in the appendix, is described with all the variables. Analyzing the figure, it is possible to see that from the first consultation, certain information already characterizes what type of evolution the patient may have, for example, the first three, Limb_O(onset)_ul(upper limbs), Limb_O_ll(lowerlimbs) or Onset_bulbar, can characterize respectively the cluster of the lower limbs(1), upper limbs(4) and bulbar(5), others, such as the first score values (ALSFRSsLL, ALSFRSsUL, R and ALSFRSb) also have a great weight in the classification, and it can be concluded that from the first consultation, with the evaluation of the scores, and the Onset, it is possible to understand what type of progression the patient will have, without the need for other less related variables, such as age, hypertension, height and weight, and even progression_rate.

Figure 7.4, shows the distribution of each variable, being the blue referred to lower values (0 in binary variables), and the red the highest (1 in binary variables), when these are to the right, it means that they fit more in the cluster with higher quotation (upper limbs(4), bulbar(5)) and to the left, clusters with lower quotation (lower limbs(1) and advanced(2)). Thus, by inspection, it is easy to see that certain variables have an obvious distribution to what they represent, for example, Limb_O_ul and Onset_bulbar have distributions with red values in the highest clusters, and Limb_O_ll the opposite. This confirms how it is possible to identify the type of progression the patient will have just by looking at the subscores and Onsets taken at the first visit.

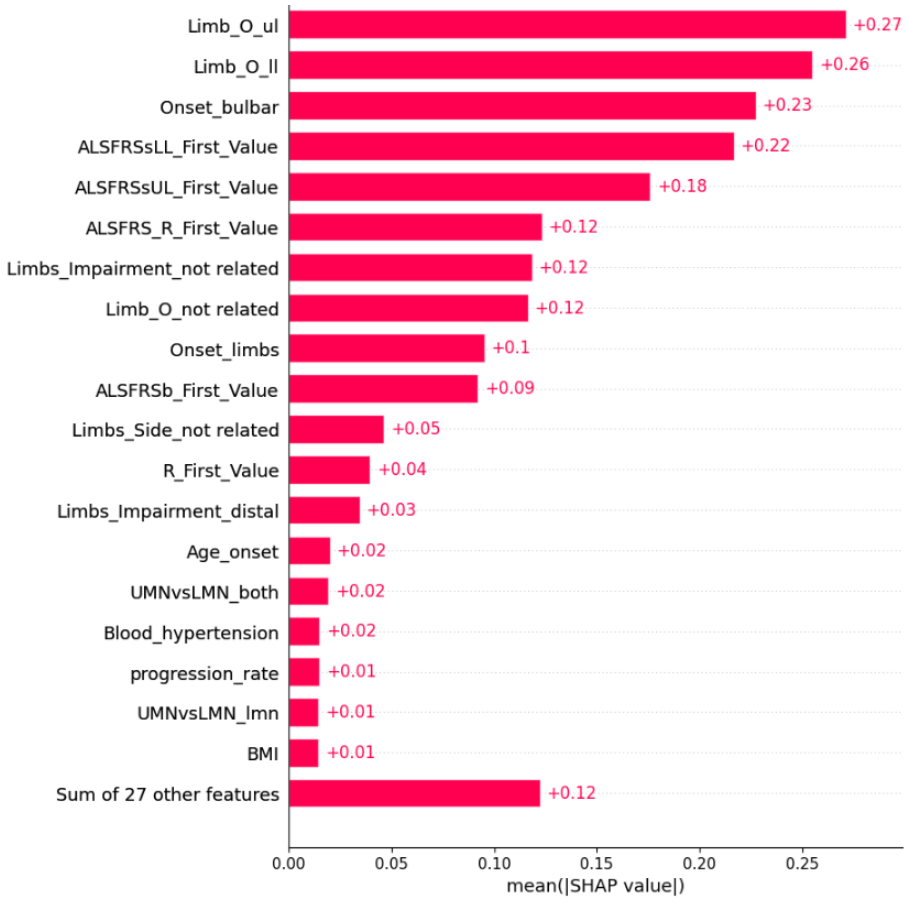

Figure 7.3: Importance of each variable (SHAP)

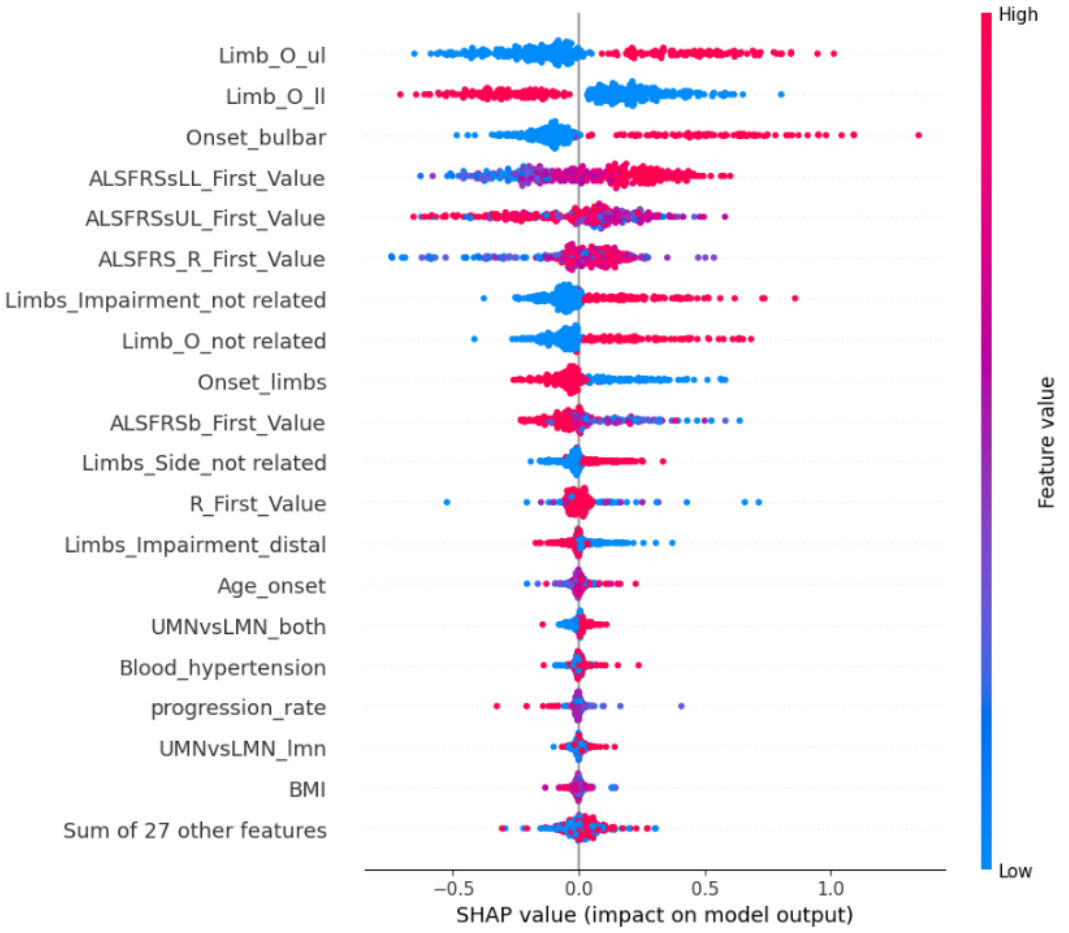

Figure 7.4: Distribution of each variable (SHAP)

Chapter 8

Forecasting medical procedures in different types of progressions

The aim of this chapter is to understand whether the groups created with the approach from section 5.1.1(Multivariate stratification), differentiating the types of progression, help to predict the need for a certain medical procedure. The basic idea is to detect which medical procedures have the highest incidence in each type of progression (group), and to see if there is any improvement in predicting these medical procedures by creating specialized classifiers for each group, instead of a general classifier that doesn't distinguish between each type of progression (group).

8.1 Most common medical procedures for each type of progression

In order not to encourage errors, the clusters were recalculated and the results saved so that all models can be compared with the same data. Thus, the new reformulation was assigned as:

- Cluster 0 - Lower Limbs
- Cluster 1 - Advanced
- Cluster 2 - Diffuse
- Cluster 3 - Upper Limbs
- Cluster 4 - Bulbar

Already mentioned in section 4.3.3, the snapshots are subdivided into five datasets as well, each referring to a medical procedure, containing the need for a certain procedure at 90, 180 and 365 days.

- C1 - Non-invasive ventilation (NIV)
- C2 - Auxiliary communication device
- C3 - Percutaneous endoscopic gastrostomy(PEG)
- C4 - Caregiver

- C5 - Wheelchair

When grouping the snapshots by their respective clusters, and dividing those containing the requirements of certain procedures by the total number of the cluster, it is possible to visualize the following tables 8.1, 8.2, 8.3.

90 days	LowerLimbs	Advanced	Diffuse	UpperLimbs	Bulbar
C1	1.94%	4.29%	26.48%	3.07%	4.41%
C2	0.70%	4.12%	2.29%	1.07%	12.79%
C3	0.40%	2.56%	2.61%	1.09%	5.38%
C4	5.56%	36.17%	18.55%	13.08%	7.73%
C5	3.78%	9.04%	5.55%	1.99%	1.94%

Table 8.1: Percentage of snapshots requiring a medical procedure, over the total number of snapshots per cluster at 90 days

180 days	LowerLimbs	Advanced	Diffuse	UpperLimbs	Bulbar
C1	6.27%	13.17%	59.50%	8.63%	12.66%
C2	2.30%	12.07%	6.36%	3.07%	31.97%
C3	1.60%	7.89%	7.32%	3.61%	16.38%
C4	14.69%	62.5%	41.40%	30.87%	21.54%
C5	10.39%	24.90%	15.92%	6.10%	6.37%

Table 8.2: Percentage of snapshots requiring a medical procedure, over the total number of snapshots per cluster at 180 days

365 days	LowerLimbs	Advanced	Diffuse	UpperLimbs	Bulbar
C1	15.43%	29.18%	84.42%	19.23%	26.39%
C2	6.22%	24.38%	15.17%	8.26%	59.71%
C3	4.22%	18.00%	15.81%	8.53%	36.97%
C4	31.68%	82.45%	68.10%	54.01%	46.27%
C5	24.04%	49.93%	34.76%	15.16%	15.78%

Table 8.3: Percentage of snapshots requiring a medical procedure, over the total number of snapshots per cluster at 365 days

From these tables, it is possible to realize that there are some discrepancies of some clusters, in certain medical procedures, for example, in C1 (need for non-evasive ventilation), the diffuse cluster, reveals a much greater need, either at 90, 180 or 365 than the other groups, this leads to believe that people who have had initial manifestations of diffuse disease with a strong respiratory component end up needing NIV more recurrently than the other clusters. Another example is that people with a bulbar evolution, in procedure C2 (Auxiliary communication device) and a little in C3 (PEG), obtain a much higher percentage of those events/procedures, when compared to the others, revealing that this type of evolution ends up being more susceptible to need these two procedures. Finally, it is also possible to realize that in general, patients who are already in an

advanced stage of the disease, end up needing, in general, more procedures than the others, which is natural, due to their evolution.

8.2 Forecasting medical procedures

As discussed in subsection 5.3, the aim is to compare whether specialized models, that is, models trained with a certain type of ALS profile, can better predict the need for certain medical procedures than general models that do not differentiate between different types of developments.

To this end, tables 8.4, 8.5, 8.6, 8.7, 8.8, show the means and standard deviations of the results obtained.

C1(AUC)		
90 days	General model	Clusters models
LowerLimbs	0.5345±0.0291	0.5308±0.0265
Advanced	0.6167±0.0439	0.5992±0.0468
Diffuse	0.5594±0.0401	0.524±0.0406
UpperLimbs	0.5649±0.0231	0.5733±0.0216
Bulbar	0.5444±0.0381	0.5473±0.0178
General	0.574±0.0091	
180 days	General model	Clusters models
LowerLimbs	0.7126±0.0199	0.6843±0.0451
Advanced	0.7058±0.0367	0.702±0.0359
Diffuse	0.5865±0.0473	0.5527±0.0987
UpperLimbs	0.7183±0.0381	0.7136±0.0276
Bulbar	0.6886±0.0204	0.7062±0.0218
General	0.7188±0.016	
365 days	General model	Clusters models
LowerLimbs	0.7592±0.0183	0.7695±0.0162
Advanced	0.7645±0.0072	0.7665±0.0112
Diffuse	0.5575±0.1123	0.7365±0.0628
UpperLimbs	0.7832±0.018	0.7927±0.0185
Bulbar	0.7404±0.0214	0.7311±0.031
General	0.7718±0.0052	

Table 8.4: AUC of test data from classifiers per cluster against overall, plus overall test data from C1

C2(AUC)		
90 days	General model	Clusters models
LowerLimbs	0.7035±0.0236	0.7268±0.0525
Advanced	0.7633±0.0428	0.7157±0.0217
Diffuse	0.6933±0.062	0.6853±0.0781
UpperLimbs	0.6654±0.0256	0.626±0.0483
Bulbar	0.6688±0.0315	0.632±0.0103
General	0.7483±0.0132	
180 days	General model	Clusters models
LowerLimbs	0.7804±0.0408	0.8526±0.0436
Advanced	0.8277±0.0219	0.81±0.0171
Diffuse	0.8691±0.0291	0.8116±0.0333
UpperLimbs	0.797±0.0284	0.7727±0.0121
Bulbar	0.6808±0.0216	0.6869±0.0268
General	0.8428±0.0144	
365 days	General model	Clusters models
LowerLimbs	0.8187±0.0301	0.8685±0.0244
Advanced	0.845±0.0117	0.8391±0.0173
Diffuse	0.859±0.0213	0.8563±0.03
UpperLimbs	0.8321±0.0287	0.8697±0.0377
Bulbar	0.6836±0.0173	0.7351±0.0109
General	0.8798±0.01	

Table 8.5: AUC of test data from classifiers per cluster against overall, plus overall test data from C2

C3(AUC)		
90 days	General model	Clusters models
LowerLimbs	0.6906±0.1397	0.553±0.0794
Advanced	0.6394±0.0373	0.6408±0.0794
Diffuse	0.6626±0.0822	0.6201±0.1306
UpperLimbs	0.6906±0.0937	0.6911±0.1162
Bulbar	0.6251±0.0473	0.6043±0.0505
General	0.675±0.0238	
180 days	General model	Clusters models
LowerLimbs	0.747±0.0318	0.7471±0.0278
Advanced	0.732±0.0378	0.7534±0.0368
Diffuse	0.7072±0.0074	0.7355±0.023
UpperLimbs	0.7734±0.0226	0.7805±0.0161
Bulbar	0.7267±0.0181	0.7038±0.0277
General	0.7749±0.0132	
365 days	General model	Clusters models
LowerLimbs	0.8397±0.0306	0.8929±0.0317
Advanced	0.8247±0.0244	0.8516±0.026
Diffuse	0.7644±0.0291	0.783±0.0257
UpperLimbs	0.8453±0.0209	0.8806±0.0204
Bulbar	0.7426±0.007	0.7319±0.0166
General	0.8491±0.0052	

Table 8.6: AUC of test data from classifiers per cluster against overall, plus overall test data from C3

C4(AUC)		
90 days	General model	Clusters models
LowerLimbs	0.7147±0.0238	0.7193±0.0384
Advanced	0.6031±0.0358	0.6286±0.0238
Diffuse	0.6766±0.0607	0.6288±0.0547
UpperLimbs	0.714±0.0188	0.7368±0.0166
Bulbar	0.6949±0.0157	0.7298±0.0347
General	0.7244±0.0135	
180 days	General model	Clusters models
LowerLimbs	0.7339±0.0046	0.7611±0.0102
Advanced	0.6305±0.0422	0.6437±0.0299
Diffuse	0.6343±0.0458	0.6439±0.0219
UpperLimbs	0.7224±0.0163	0.7133±0.0173
Bulbar	0.7382±0.017	0.738±0.0186
General	0.7535±0.0056	
365 days	General model	Clusters models
LowerLimbs	0.7542±0.0187	0.7856±0.013
Advanced	0.6112±0.0938	0.7556±0.0239
Diffuse	0.7003±0.0437	0.756±0.021
UpperLimbs	0.7781±0.0286	0.7473±0.0309
Bulbar	0.739±0.0139	0.7272±0.016
General	0.7772±0.0114	

Table 8.7: AUC of test data from classifiers per cluster against overall, plus overall test data from C4

C5(AUC)		
90 days	General model	Clusters models
LowerLimbs	0.6713±0.0311	0.6953±0.0311
Advanced	0.6426±0.0329	0.6487±0.0296
Diffuse	0.6075±0.056	0.6581±0.0519
UpperLimbs	0.6298±0.0392	0.6548±0.0692
Bulbar	0.6452±0.0605	0.7205±0.0725
General	0.6639±0.0174	
180 days	General model	Clusters models
LowerLimbs	0.7774±0.0282	0.7713±0.0287
Advanced	0.6795±0.0196	0.6783±0.0166
Diffuse	0.6712±0.0284	0.673±0.049
UpperLimbs	0.799±0.0152	0.8049±0.0198
Bulbar	0.7837±0.0398	0.8251±0.0501
General	0.778±0.0104	
365 days	General model	Clusters models
LowerLimbs	0.8342±0.0145	0.8346±0.014
Advanced	0.7138±0.0239	0.7077±0.0283
Diffuse	0.7875±0.033	0.7833±0.0115
UpperLimbs	0.8136±0.0107	0.8248±0.0236
Bulbar	0.7602±0.0082	0.7766±0.0186
General	0.8109±0.0061	

Table 8.8: AUC of test data from classifiers per cluster against overall, plus overall test data from C5

There are two main questions witch can be concluded from this tables:

- In which cases did the specialized models perform better than general models on cluster test data?
- In which cases did the specialized models perform better than general models on general test data?

For both questions, the results calculated for 90 days will not be taken into account, because the standard deviations tend to be quite high, making it impossible to draw any conclusions about the data. Cases referring to the first question only occur 4 times, the cases checked were:

- C1, 365, Diffuse
- C2, 365, Bulbar
- C4, 180, Lower limbs
- C4, 365, Advanced

These 4 cases are the only ones where the standard deviation range does not coincide, although there are cases where there is a large discrepancy of values, such as C4, diffuse at 365 days, the overlap of standard deviations reveals a ”technical tie”. But taking into account only these 4 cases,

and relating to the tables 8.2 and 8.3, it is possible to denote that the clusters that present a greater discrepancy in relation to the need for medical procedures, are exactly the same, except for C4 at 180 days, as those represented above, revealing that general models, when faced with groups of people with greater need for certain medical procedures, tend to perform worse than models specialized in the type of patient evolution. Unfortunately, it was only possible to verify these cases, mostly for medical procedures at 365 days, since the less time of prediction, the more the standard deviation tends to increase, often not being possible to conclude anything.
For the second question, only 3 cases were verified:

- C3, 365, Lower Limbs
- C3, 365, Upper Limbs
- C5, 365, Lower Limbs

In all these cases involving limb issues, notable patterns emerge. For instance, in C5 cases where a wheelchair is needed, specialized models outperform general ones due to the natural tendency for individuals with lower limb problems to require a wheelchair.
However, in C3 cases related to PEG, despite bulbar progression being more common, patients with limb onset actually receive better classifications than the general population.
Two main reasons explain this. First, when examining table 8.3, specifically C3 data for LowerLimbs and UpperLimbs, there are percentages that are too low (4.2% and 8.5%) to draw any meaningful conclusions, as they do not represent the samples adequately. Second, when analyzing the number of patients requiring PEG in the LowerLimbs and UpperLimbs clusters, percentages of 25.07% and 34.89% indicate a significant number of patients. Even though their progression type isn't closely related to the PEG procedure, there is still a considerable portion of patients who require it.

Chapter 9

Conclusion

9.1 Conclusion

The aims of this project are to uncover and explore the ALS disease, focusing mainly on separating and identifying different types of progression in order to reduce heterogeneity, and on predicting medical procedures. With this work, several non-supervised learning algorithms were tested, to stratify different types of progressions of ALS, from an AutoEncoder and Kmeans, it was possible to subdivide 5 different types of progressions, only by 1 year of follow-up, both at the univariable level, taking into account the general decline of the patient, or multivariate, dividing the various types of progression by the region of the body where the disease most manifested. With the stratification of these groups, it was tried to predict at the prognostic level, what type of progression the patient could pursue only taking into account the data of the first appointment, with a degree of F1 and accuracy in the 80% in some supervised learning models, it was also studied which variables most influence the type of progression.
Be able to differentiate the type of progression from the initial symptoms, it was studied which medical procedures tend to appear more according to the given group, comparing if a general machine learning model, that is, trained by the data without any differentiation of clusters, can obtain better performance than models specialized in certain types of progressions. With this, it was possible to see that in overall, general models tend to perform well, but in certain cases, where a certain medical procedure tends to appear with much higher incidence due to a certain type of progression, specialized models perform better, showing that there is a minimal advantage in dividing and studying ALS by different types of evolution.

9.2 Future Work

As future work, there are some topics that could be explored to improve this project. Stratification models only accept a fixed size of time sequences. This presents a huge disadvantage, since a lot of information is lost when truncating patients with more than one year of follow-up, or discarding patients with less. That's why the application of non-supervised machine learning models that can process variable time sequences can be a great help in not wasting so much information.

Even without changing the model, a different approach to processing the input data can also be considered, such as allowing the processing of longer time sequences, and instead of removing patients with little information, filling in the missing information with zeros, or other information that can be discarded.

In the prediction of the need for medical procedures, the comparison between general models and models specialized in a given type of progression ends up with few results, because in many cases, the standard deviation is so high that it is difficult to conclude how good the model is (especially at 90 days). Increasing the number of repetitions of these calculations can help reduce the standard deviations, and thus increase the cases where the specialized models perform better.

In addition, it would be interesting to investigate the influence of genetic factors of the disease, depending on the average life span, or the speed of progression that the patient may present.

Acronyms

ELA Esclerose Lateral Amiotrófica

ALS Amyotrophic Lateral Sclerosis

APELA Portuguese Amyotrophic Lateral Sclerosis Association

UMN Upper Motor Neurons

LMN Lower Motor Neurons

PLS Primary Lateral Sclerosis

PMA Progressive Muscular Atrophy

HC Hierarchical Clustering

SVM Support Vector Machine

SVR Support Vector Regression

KNN K Nearest Neighbor

GBDT Gradient Boosting Decision Trees

CV Cross-validation

ROC Receiver Operating Characteristic

TPR True Positive Rate

FPR False Positive Rate

AUC Area Under the Curve

MCC Matthews correlation coefficient

NIV Non Invasive Ventilation

PEG Percutaneous Endoscopic Gastrostomy

LSTM Long Short-Term Memory

SHAP SHapley Additive exPlanations

MFP Most Frequent Pattern

DTW Dynamic Time Warping

PCA Principal Component Analysis

ICA Independent Component Analysis

ARMA AutoRegressive–Moving-Average

HMM Hidden Markov Model

SMOTE Synthetic Minority Oversampling TEchnique

ENN Edited Nearest Neighbor

LOO Leave-One-Out

LPO Leave-P-Out

UMAP Uniform Manifold Approximation and Projection

Bibliography

[1] Jason Ackrivo, John Hansen-Flaschen, E Paul Wileyto, Richard J Schwab, Lauren Elman, and Steven M Kawut. Development of a prognostic model of respiratory insufficiency or death in amyotrophic lateral sclerosis. *European Respiratory Journal*, 53(4), 2019.

[2] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. 2010.

[3] Roberto Baragona. A simulation study on clustering time series with metaheuristic methods. *Quaderni di Statistica*, 3:1–26, 2001.

[4] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29, 2004.

[5] Michael S Bereman, Joshua Beri, Jeffrey R Enders, and Tara Nash. Machine learning reveals protein signatures in csf and plasma fluids of clinical value for als. *Scientific reports*, 8(1):16334, 2018.

[6] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[7] B Rix Brooks. El escorial world federation of neurology criteria for the diagnosis of amyotrophic lateral sclerosis. subcommittee on motor neuron diseases/amyotrophic lateral sclerosis of the world federation of neurology research group on neuromuscular diseases and the el escorial" clinical limits of amyotrophic lateral sclerosis" workshop contributors. *Journal of the neurological sciences*, 124:96–107, 1994.

[8] Robert H Brown and Ammar Al-Chalabi. Amyotrophic lateral sclerosis. *New England Journal of Medicine*, 377(2):162–172, 2017.

[9] André Carreiro. *An integrative approach for prognostic prediction in neurodegenerative diseases.* PhD thesis, 2016.

[10] Mamede De Carvalho and Michael Swash. Awaji diagnostic algorithm increases sensitivity of el escorial criteria for als diagnosis. *Amyotrophic Lateral Sclerosis*, 10(1):53–57, 2009.

[11] Jesse M Cedarbaum, Nancy Stambler, Errol Malta, Cynthia Fuller, Dana Hilt, Barbara Thurmond, Arline Nakanishi, Bdnf Als Study Group, 1A complete listing of the BDNF

Study Group, et al. The alsfrs-r: a revised als functional rating scale that incorporates assessments of respiratory function. *Journal of the neurological sciences*, 169(1-2):13–21, 1999.

[12] S Chandrakala and C Chandra Sekhar. A density based method for multivariate time series clustering in kernel feature space. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1885–1890. IEEE, 2008.

[13] Pimwadee Chaovalit, Aryya Gangopadhyay, George Karabatis, and Zhiyuan Chen. Discrete wavelet transform-based time series analysis and mining. *ACM Computing Surveys (CSUR)*, 43(2):1–37, 2011.

[14] Malika Charrad, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. Nbclust: an r package for determining the relevant number of clusters in a data set. *Journal of statistical software*, 61:1–36, 2014.

[15] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[16] Huub Creemers, Hepke Grupstra, Frans Nollet, Leonard H van den Berg, and Anita Beelen. Prognostic factors for the course of functional status of patients with als: a systematic review. *Journal of neurology*, 262:1407–1423, 2015.

[17] Johann de Jong, Mohammad Asif Emon, Ping Wu, Reagon Karki, Meemansa Sood, Patrice Godard, Ashar Ahmad, Henri Vrooman, Martin Hofmann-Apitius, and Holger Fröhlich. Deep learning for clustering of multivariate clinical patient trajectories with missing values. *GigaScience*, 8(11):giz134, 2019.

[18] MA Del Aguila, WT Longstreth, V McGuire, TD Koepsell, and G Van Belle. Prognosis in amyotrophic lateral sclerosis: a population-based study. *Neurology*, 60(5):813–819, 2003.

[19] Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 551–556, 2004.

[20] Marwa Elamin, Peter Bede, Anna Montuschi, Niall Pender, Adriano Chio, and Orla Hardiman. Predicting prognosis in amyotrophic lateral sclerosis: a simple algorithm. *Journal of neurology*, 262:1447–1454, 2015.

[21] Peter Flach. *Machine learning: the art and science of algorithms that make sense of data*. Cambridge university press, 2012.

[22] Terrence S Furey, Nello Cristianini, Nigel Duffy, David W Bednarski, Michel Schummer, and David Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.

[23] Jeban Ganesalingam, Daniel Stahl, Lokesh Wijesekera, Clare Galtrey, Christopher E Shaw, P Nigel Leigh, and Ammar Al-Chalabi. Latent cluster analysis of als phenotypes identifies prognostically differing groups. *PloS one*, 4(9):e7107, 2009.

[24] Matt W Gardner and SR Dorling. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636, 1998.

[25] Eleni Georgoulopoulou, Nicola Fini, Marco Vinceti, Marco Monelli, Paolo Vacondio, Giorgia Bianconi, Patrizia Sola, Paolo Nichelli, and Jessica Mandrioli. The impact of clinical factors, riluzole and therapeutic interventions on als survival: a population based study in modena, italy. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 14(5-6):338–345, 2013.

[26] Namita A Goyal, James D Berry, Anthony Windebank, Nathan P Staff, Nicholas J Maragakis, Leonard H van den Berg, Angela Genge, Robert Miller, Robert H Baloh, Ralph Kern, et al. Addressing heterogeneity in amyotrophic lateral sclerosis clinical trials. *Muscle & nerve*, 62(2):156–166, 2020.

[27] Leslie I Grad, Guy A Rouleau, John Ravits, and Neil R Cashman. Clinical spectrum of amyotrophic lateral sclerosis (als). *Cold Spring Harbor perspectives in medicine*, page a024117, 2016.

[28] Leslie I Grad, Guy A Rouleau, John Ravits, and Neil R Cashman. Clinical spectrum of amyotrophic lateral sclerosis (als). *Cold Spring Harbor perspectives in medicine*, 7(8):a024117, 2017.

[29] Vincent Grollemund, Gaétan Le Chat, Marie-Sonia Secchi-Buhour, François Delbot, Jean-François Pradat-Peyre, Peter Bede, and Pierre-François Pradat. Development and validation of a 1-year survival prognosis estimation model for amyotrophic lateral sclerosis using manifold learning algorithm umap. *Scientific reports*, 10(1):13378, 2020.

[30] Marta Gromicho, Tiago Leão, Miguel Oliveira Santos, Susana Pinto, Alexandra M Carvalho, Sara C Madeira, and Mamede De Carvalho. Dynamic bayesian networks for stratification of disease progression in amyotrophic lateral sclerosis. *European Journal of Neurology*, 29(8):2201–2210, 2022.

[31] Chonghui Guo, Hongfeng Jia, and Na Zhang. Time series clustering based on ica for stock data analysis. In *2008 4th international conference on wireless communications, networking and mobile computing*, pages 1–4. IEEE, 2008.

[32] Frank E Harrell et al. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*, volume 608. Springer, 2001.

[33] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[34] Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, et al. A practical guide to support vector classification, 2003.

[35] Aurangzeb Khan, Khairullah Khan, and Baharum B Baharudin. Frequent patterns minning of stock data using hybrid clustering association algorithm. In *2009 International Conference on Information Management and Engineering*, pages 667–671. IEEE, 2009.

[36] Matthew C Kiernan, Steve Vucic, Kevin Talbot, Christopher J McDermott, Orla Hardiman, Jeremy M Shefner, Ammar Al-Chalabi, William Huynh, Merit Cudkowicz, Paul Talman, et al. Improving clinical trial outcomes in amyotrophic lateral sclerosis. *Nature Reviews Neurology*, 17(2):104–118, 2021.

[37] Robert Kueffner, Neta Zach, Maya Bronfeld, Raquel Norel, Nazem Atassi, Venkat Balagurusamy, Barbara Di Camillo, Adriano Chio, Merit Cudkowicz, Donna Dillenberger, et al. Stratification of amyotrophic lateral sclerosis patients: a crowdsourcing approach. *Scientific reports*, 9(1):690, 2019.

[38] Isotta Landi, Veronica Mandelli, and Michael V Lombardo. reval: A python package to determine best clustering solutions with stability-based relative clustering validation. *Patterns*, 2(4), 2021.

[39] Tiago Leão, Sara C Madeira, Marta Gromicho, Mamede de Carvalho, and Alexandra M Carvalho. Learning dynamic bayesian networks from time-dependent and time-independent data: Unraveling disease progression in amyotrophic lateral sclerosis. *Journal of Biomedical Informatics*, 117:103730, 2021.

[40] PN Leigh, BH Anderton, A Dodson, J-M Gallo, M Swash, and DM Power. Ubiquitin deposits in anterior horn cells in motor neurone disease. *Neuroscience letters*, 93(2-3):197–203, 1988.

[41] T Warren Liao. Clustering of time series data—a survey. *Pattern recognition*, 38(11):1857–1874, 2005.

[42] TW Liao, B Bolt, J Forester, E Hailman, C Hansen, RC Kaste, and J O'May. Understanding and projecting the battle state. In *23rd Army Science Conference, Orlando, FL*, volume 25, 2002.

[43] Ting Liu, Andrew Moore, Ke Yang, and Alexander Gray. An investigation of practical approximate nearest neighbor algorithms. *Advances in neural information processing systems*, 17, 2004.

[44] J Lowe, G Lennox, D Jefferson, K Morrell, D McQuire, T Gray, M Landon, FJ Doherty, and RJ Mayer. A filamentous inclusion body within anterior horn neurones in motor neurone disease defined by immunocytochemical localisation of ubiquitin. *Neuroscience letters*, 94(1-2):203–210, 1988.

[45] Duc Thanh Anh Luong and Varun Chandola. A k-means approach to clustering disease progressions. In *2017 IEEE International conference on healthcare informatics (ICHI)*, pages 268–274. IEEE, 2017.

[46] Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993.

[47] Jessica Mandrioli, Sara Biguzzi, Carlo Guidi, Elisabetta Sette, Emilio Terlizzi, Alessandro Ravasio, Mario Casmiro, Fabrizio Salvi, Rocco Liguori, Romana Rizzi, et al. Heterogeneity in alsfrs-r decline and survival: a population-based study in italy. *Neurological Sciences*, 36:2243–2252, 2015.

[48] Benoît Marin, Philippe Couratier, Simona Arcuti, Massimiliano Copetti, Andrea Fontana, Marie Nicol, Marie Raymondeau, Giancarlo Logroscino, and Pierre Marie Preux. Stratification of als patients' survival: a population-based study. *Journal of neurology*, 263:100–111, 2016.

[49] Sarah Martin, Ahmad Al Khleifat, and Ammar Al-Chalabi. What causes amyotrophic lateral sclerosis? *F1000Research*, 6, 2017.

[50] Marcel Müller, Marta Gromicho, Mamede de Carvalho, and Sara C Madeira. Explainable models of disease progression in als: Learning from longitudinal clinical data with recurrent neural networks and deep model explanation. *Computer Methods and Programs in Biomedicine Update*, 1:100018, 2021.

[51] SR Nanda, Biswajit Mahanty, and MK Tiwari. Clustering indian stock market data for portfolio management. *Expert Systems with Applications*, 37(12):8793–8798, 2010.

[52] Vit Niennattrakul and Chotirat Ann Ratanamahatana. On clustering multimedia time series data using k-means and dynamic time warping. In *2007 International Conference on Multimedia and Ubiquitous Engineering (MUE'07)*, pages 733–738. IEEE, 2007.

[53] Tim Oates, Laura Firoiu, and Paul R Cohen. Clustering time series with hidden markov models and dynamic time warping. In *Proceedings of the IJCAI-99 workshop on neural, symbolic and reinforcement learning methods for sequence learning*, volume 17, page 21. Citeseer, 1999.

[54] John Paparrizos and Luis Gravano. k-shape: Efficient and accurate clustering of time series. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, pages 1855–1870, 2015.

[55] Telma Pereira, Sofia Pires, Marta Gromicho, Susana Pinto, Mamede de Carvalho, and Sara C Madeira. Predicting assisted ventilation in amyotrophic lateral sclerosis using a mixture of experts and conformal predictors. *arXiv preprint arXiv:1907.13070*, 2019.

[56] Sofia Pires, Marta Gromicho, Susana Pinto, Mamede Carvalho, and Sara C Madeira. Predicting non-invasive ventilation in als patients using stratified disease progression groups. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 748–757. IEEE, 2018.

[57] Divya Ramamoorthy, Kristen Severson, Soumya Ghosh, Karen Sachs, Answer ALS, Jonathan D Glass, Christina N Fournier, Pooled Resource Open-Access ALS Clinical Trials Consortium, James Berry, Kenney Ng, et al. Identifying patterns of als progression from sparse longitudinal data. *medRxiv*, pages 2021–05, 2021.

[58] Sangeeta Rani and Geeta Sikka. Recent techniques of clustering of time series data: a survey. *International Journal of Computer Applications*, 52(15), 2012.

[59] John Ravits, Stanley Appel, Robert H Baloh, Richard Barohn, Benjamin Rix Brooks, Lauren Elman, Mary Kay Floeter, Christopher Henderson, Catherine Lomen-Hoerth, Jeffrey D Macklis, et al. Deciphering amyotrophic lateral sclerosis: what phenotype, neuropathology and genetics are telling us about pathogenesis. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, 14(sup1):5–18, 2013.

[60] John M Ravits and Albert R La Spada. Als motor phenotype heterogeneity, focality, and spread: deconstructing motor neuron degeneration. *Neurology*, 73(10):805–811, 2009.

[61] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[62] Sunil Saumya and Jyoti Prakash Singh. Spam review detection using lstm autoencoder: an unsupervised approach. *Electronic Commerce Research*, 22(1):113–133, 2022.

[63] CT Shaw and GP King. Using cluster analysis to classify time series. *Physica D: Nonlinear Phenomena*, 58(1-4):288–298, 1992.

[64] Diogo F Soares, Rui Henriques, Marta Gromicho, Mamede de Carvalho, and Sara C Madeira. Learning prognostic models using a mixture of biclustering and triclustering: Predicting the need for non-invasive ventilation in amyotrophic lateral sclerosis. *Journal of Biomedical Informatics*, 134:104172, 2022.

[65] Yan-Yan Song and LU Ying. Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2):130, 2015.

[66] Ralf C Staudemeyer and Eric Rothstein Morris. Understanding lstm–a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586*, 2019.

[67] James V Stone. Independent component analysis: an introduction. *Trends in cognitive sciences*, 6(2):59–64, 2002.

[68] Xiaowei William Su, Zachary Simmons, Ryan Michael Mitchell, Lan Kong, Helen Elizabeth Stephens, and James Robert Connor. Biomarker-based predictive models for prognosis in amyotrophic lateral sclerosis. *JAMA neurology*, 70(12):1505–1511, 2013.

[69] Harold HG Tan, Henk-Jan Westeneng, Abram D Nitert, Kevin van Veenhuijzen, Jil M Meier, Hannelore K van der Burgh, Martine JE van Zandvoort, Michael A van Es, Jan H Veldink, and Leonard H van den Berg. Mri clustering reveals three als subtypes with unique neurodegeneration patterns. *Annals of Neurology*, 92(6):1030–1045, 2022.

[70] Neda Tavakoli, Sima Siami-Namini, Mahdi Adl Khanghah, Fahimeh Mirza Soltani, and Akbar Siami Namin. An autoencoder-based deep learning approach for clustering time series data. *SN Applied Sciences*, 2:1–25, 2020.

[71] Kim Traxinger, Crystal Kelly, Brent A Johnson, Robert H Lyles, and Jonathan D Glass. Prognosis and epidemiology of amyotrophic lateral sclerosis: analysis of a clinic population, 1997–2011. *Neurology: Clinical Practice*, 3(4):313–320, 2013.

[72] Theresa Ullmann, Christian Hennig, and Anne-Laure Boulesteix. Validation of cluster analysis results on validation data: A systematic framework. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3):e1444, 2022.

[73] Geert Verdoolaege and Yves Rosseel. Activation detection in event-related fmri through clustering ofwavelet distributions. In *2010 IEEE International Conference on Image Processing*, pages 4393–4396. IEEE, 2010.

[74] Michail Vlachos, Jessica Lin, Eamonn Keogh, and Dimitrios Gunopulos. A waveletbased anytime algorithm for k-means clustering of time series. In *Workshop on Clustering High Dimensionality Data and Its Applications, at the 3rd SIAM International Conference on Data Mining, San Francisco, CA, USA*, 2003.

[75] Henk-Jan Westeneng, Thomas PA Debray, Anne E Visser, Ruben PA van Eijk, James PK Rooney, Andrea Calvo, Sarah Martin, Christopher J McDermott, Alexander G Thompson, Susana Pinto, et al. Prognosis for patients with amyotrophic lateral sclerosis: development and validation of a personalised prediction model. *The Lancet Neurology*, 17(5):423–433, 2018.

[76] JianXin Wu and JiaoLong Wei. Combining ica with svr for prediction of finance time series. In *2007 IEEE International Conference on Automation and Logistics*, pages 95–100. IEEE, 2007.

[77] Yimin Xiong and Dit-Yan Yeung. Mixtures of arma models for model-based time series clustering. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 717–720. IEEE, 2002.

[78] Junjing Yang, Chao Ning, Chirag Deb, Fan Zhang, David Cheong, Siew Eang Lee, Chandra Sekhar, and Kwok Wai Tham. k-shape clustering algorithm for building energy usage patterns analysis and forecasting model accuracy improvement. *Energy and Buildings*, 146:27–37, 2017.

[79] Zicong Zhang, Changchang Yin, and Ping Zhang. Temporal clustering with external memory network for disease progression modeling. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 956–965. IEEE, 2021.

Appendices

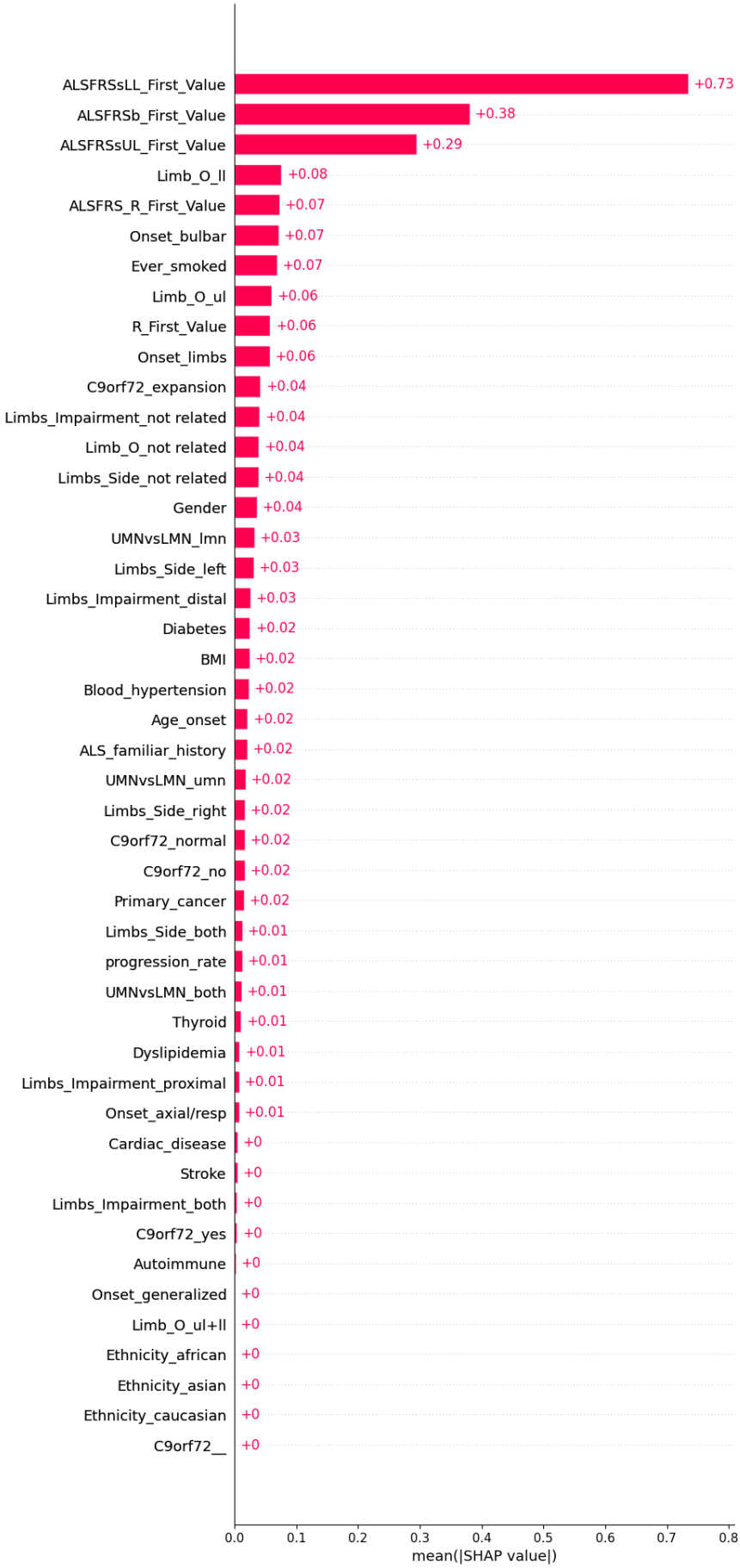

Figure 1: Importance of each variable (SHAP)

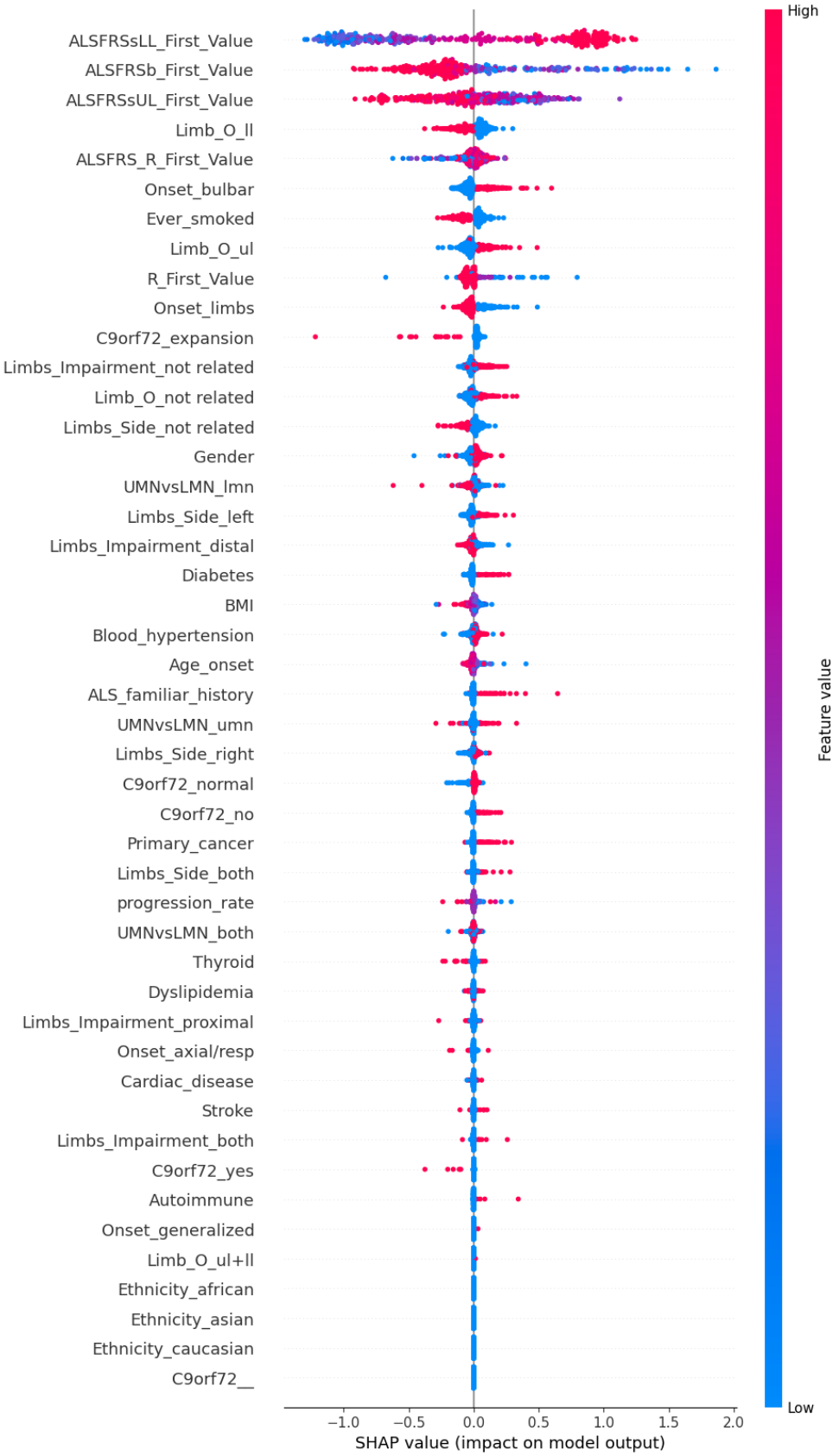

Figure 2: Distribution of each variable (SHAP)