UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE INFORMÁTICA

BICLUSTERING ELECTRONIC HEALTH RECORDS TO UNRAVEL DISEASE PRESENTATION PATTERNS

Joana Sofia Santos de Matos

Mestrado em Ciência de Dados

Dissertação orientada por:

Professora Doutora Sara Alexandra Cordeiro Madeira

2019

For my grandfather and godfather Romeu.
I wish you were here to see this.

"There are only patterns, patterns on top of patterns, patterns that affect other patterns.
Patterns hidden by patterns. Patterns within patterns.
If you watch close, history does nothing but repeat itself.
What we call chaos is just patterns we haven't recognized."

— Chuck Palahniuk, *Survivor*

Acknowledgements

This work would not have been possible without the FCT funding to NEUROCLINOMICS2 (PTDC/EEI-SII/1937/2014) research project, and the LASIGE Research Unit's (UID/CEC/00408/2019) hosting.

First of all, I would like to thank the team involved in the project, from Faculdade de Ciências da Universidade de Lisboa (FCUL) and Instituto de Medicina Molecular (IMM) João Lobo Antunes: my advisor, Prof. Sara Madeira, for her guidance and support; Dr. Mamede de Carvalho for his indispensable knowledge and experience; Manuel Figueiredo and Marta Gromicho for validating and correcting the data; and Sofia Pires for her help in creating the class labelling used in the second Task. Additionally, I would also like to thank Rui Henriques from Instituto Superior Técnico (IST) for his aid in how to use the BicPAMS algorithm for our experiments.

In the past year, I had the privilege of meeting many awesome people here at LASIGE, whom I already consider my friends. I'll dearly miss the companionship we built during our meal times and festive events, so I really hope we won't lose contact!

Last but not least, I want to thank my family. Most of all my parents, Albano and Maria João, for their loving encouragement of my tardy pursuit of a master's degree, and my dear husband-to-be, Pedro, for his endless patience and love during this endeavor.

Resumo

A Esclerose Lateral Amiotrófica (ELA) é uma doença neurodegenerativa heterogénea com padrões de apresentação altamente variáveis. Dada a natureza heterogénea dos doentes com ELA, aquando do diagnóstico os clínicos normalmente estimam a progressão da doença utilizando uma taxa de decaimento funcional, calculada com base na Escala Revista de Avaliação Funcional de ELA (*ALSFRS-R*).

A utilização de modelos de Aprendizagem Automática que consigam lidar com este padrões complexos é necessária para compreender a doença, melhorar os cuidados aos doentes e a sua sobrevivência. Estes modelos devem ser explicáveis para que os clínicos possam tomar decisões informadas.

Desta forma, o nosso objectivo é descobrir padrões de apresentação da doença, para isso propondo uma nova abordagem de Prospecção de Dados: Descoberta de Meta-atributos Discriminativos (DMD), que utiliza uma combinação de *Biclustering*, Classificação baseada em *Biclustering* e Prospecção de Regras de Associação para Classificação. Estes padrões (chamados de Meta-atributos) são compostos por subconjuntos de atributos discriminativos conjuntamente com os seus valores, permitindo assim distinguir e caracterizar subgrupos de doentes com padrões similares de apresentação da doença.

Os Registos de Saúde Electrónicos (RSE) utilizados neste trabalho provêm do conjunto de dados JPND ONWebDUALS (*ONTology-based Web Database for Understanding Amyotrophic Lateral Sclerosis*), composto por questões standardizadas acerca de factores de risco, mutações genéticas, atributos clínicos ou informação de sobrevivência de uma coorte de doentes e controlos seguidos pelo consórcio ENCALS (*European Network to Cure ALS*), que inclui vários países europeus, incluindo Portugal.

Nesta tese a metodologia proposta foi utilizada na parte portuguesa do conjunto de dados ONWebDUALS para encontrar padrões de apresentação da doença que: 1) distinguissem os doentes de ELA dos seus controlos e 2) caracterizassem grupos de doentes de ELA com diferentes taxas de progressão (categorizados em grupos Lentos, Neutros e Rápidos). Nenhum padrão coerente emergiu das experiências efectuadas para a primeira tarefa. Contudo, para a segunda tarefa os padrões encontrados para cada um dos três grupos de progressão foram reconhecidos e validados por clínicos especialistas em ELA, como sendo características relevantes de doentes com progressão Lenta, Neutra e Rápida. Estes resultados sugerem que a nossa abordagem genérica baseada em *Biclustering* tem potencial para identificar padrões de apresentação noutros problemas ou doenças semelhantes.

Palavras Chave: Esclerose Lateral Amiotrófica, *Biclustering* baseado em Prospecção de Padrões, Classificação baseada em *Biclustering*, Prospecção de Regras de Associação para Classificação, *Biclusters* Discriminativos

Abstract

Amyotrophic Lateral Sclerosis (ALS) is a heterogeneous neurodegenerative disease with a high variability of presentation patterns. Given the heterogeneous nature of ALS patients and targeting a better prognosis, clinicians usually estimate disease progression at diagnosis using the rate of decay computed from the Revised ALS Functional Rating Scale (ALSFRS-R).

In this context, the use of Machine Learning models able to unravel the complexity of disease presentation patterns is paramount for disease understanding, targeting improved patient care and longer survival times. Furthermore, explainable models are vital, since clinicians must be able to understand the reasoning behind a given model's result before making a decision that can impact a patient's life.

Therefore we aim at unravelling disease presentation patterns by proposing a new Data Mining approach called Discriminative Meta-features Discovery (DMD), which uses a combination of Biclustering, Biclustering-based Classification and Class Association Rule Mining. These patterns (called Meta-features) are composed of discriminative subsets of features together with their values, allowing to distinguish and characterize subgroups of patients with similar disease presentation patterns.

The Electronic Health Record (EHR) data used in this work comes from the JPND ONWebDUALS (ONTology-based Web Database for Understanding Amyotrophic Lateral Sclerosis) dataset, comprised of standardized questionnaire answers regarding risk factors, genetic mutations, clinical features and survival information from a cohort of patients and controls from ENCALS (European Network to Cure ALS), a consortium of diverse European countries, including Portugal.

In this work the proposed methodology was used on the ONWebDUALS Portuguese EHR data to find disease presentation patterns that: 1) distinguish the ALS patients from their controls and 2) characterize groups of ALS patients with different progression rates (categorized into Slow, Neutral and Fast groups). No clear pattern emerged from the experiments performed for the first task. However, in the second task the patterns found for each of the three progression groups were recognized and validated by ALS expert clinicians, as being relevant characteristics of slow, neutral and fast progressing patients. These results suggest that our generic Biclustering approach is a promising way to unravel disease presentation patterns and could be applied to similar problems and other diseases.

Keywords: Amyotrophic Lateral Sclerosis, Pattern Mining-based Biclustering, Biclustering-based Classification, Class Association Rule Mining, Discriminative Biclusters

Resumo Alargado

A Esclerose Lateral Amiotrófica (ELA) é uma doença neurodegenerativa heterogénea que afecta o Sistema Motor humano, num período de tempo relativamente curto. A doença pode surgir numa determinada região do corpo e com o tempo afectar também outras. Alguns sintomas comuns são: fraqueza nos membros, dificuldades em falar e engolir, insuficiência respiratória e alterações cognitivas/comportamentais. A principal causa de morte é a eventual insuficiência respiratória. O tempo de sobrevivência é altamente variável, dependendo da velocidade da progressão da doença, que aquando do diagnóstico é frequentemente estimada pelos clínicos utilizando uma taxa de decaimento funcional, calculada com base na Escala Revista de Avaliação Funcional de ELA (ALSFRS-R). Esta escala é composta por 12 perguntas cujo valor para cada uma varia entre 0 e 4, com um valor máximo de 48 (quanto maior o valor total, maior a capacidade funcional do doente).

Até ao presente dia, continua a ser necessário descobrir testes de diagnóstico fiáveis ou biomarcadores que ajudem os clínicos a conseguir diagnosticar esta doença de forma rápida e eficaz, dado o alto grau de variabilidade existente nos fenótipos observados, história familiar, genes envolvidos, vias moleculares e factores ambientais que a poderão provocar. Tendo isto em consideração, acredita-se que existem diversos mecanismos que podem causar a neurodegenerescência em doentes com ELA.

Dentro deste contexto, o nosso objectivo é descobrir padrões de apresentação da doença. Estes padrões (que neste trabalho serão chamados de Meta-atributos) são compostos por subconjuntos de atributos discriminativos conjuntamente com os seus valores, permitindo assim distinguir e caracterizar subgrupos de doentes com padrões similares de apresentação da doença. Uma possível forma de identificar este tipo de padrões é através da utilização de modelos de Aprendizagem Automática, que conseguem lidar com a complexidade dos mesmos, permitindo compreender melhor a doença, criar tratamentos mais específicos para os doentes e aumentar o seu tempo de sobrevivência. Idealmente, esses modelos deverão ser explicáveis, uma vez que os clínicos têm de conseguir compreender o raciocínio por detrás de um resultado do modelo antes de tomar qualquer decisão com impacto na vida de um doente.

Os Registos de Saúde Electrónicos (RSE) utilizados neste trabalho provêm do conjunto de dados JPND ONWebDUALS (*ONTology-based Web Database for Understanding Amyotrophic Lateral Sclerosis*), composto por questões standardizadas acerca de factores de risco, mutações genéticas, atributos clínicos ou informação de sobreviência de uma coorte de doentes e controlos seguidos pelo consórcio ENCALS (*European Network to Cure ALS*), que inclui vários países europeus, incluindo Portugal.

Assim, neste trabalho é proposta uma nova abordagem exploratória para a Prospecção de Dados, chamada Descoberta de Meta-atributos Discriminativos (DMD). A base da mesma assenta na utilização

de *Biclustering*, uma técnica de Prospecção de Dados que permite identificar *Biclusters*: observações com padrões coerentes em certos subconjuntos de atributos (e os seus respectivos valores) em conjuntos de dados bidimensionais. Até à data estas técnicas foram aplicadas com sucesso em dados médicos, permitindo descobrir grupos de entidades biológicas significativamente correlacionadas num subconjunto de condições ou pontos no tempo. De forma a encontrar *Biclusters* com sobreposição de forma eficiente utilizou-se uma variante - *Biclustering* baseado em Prospecção de Padrões - que necessita da prévia categorização dos dados, o que pode implicar alguma perda de informação. Os *Biclusters* são considerados discriminativos caso pelo menos 75% das observações incluídas nos mesmos pertençam a uma dada classe, embora esta não seja considerada quando os *Biclusters* estão a ser prospectados. Entre vários conjuntos de *Biclusters* obtidos em experiências diferentes, foram considerados melhores os que tinham maior número de *Biclusters* discriminativos.

Sobre esses resultados são utilizadas técnicas de Classificação baseada em *Biclustering* e Prospecção de Regras de Associação para Classificação, para descobrir quais os padrões mais discriminativos para cada classe considerada. Por um lado a primeira abordagem utiliza uma matriz de identificadores dos sujeitos (doentes ou controlos) $\times$ identificadores dos *Biclusters* discriminativos que indica em que *Biclusters* os sujeitos estavam presentes. Utilizando os identificadores dos *Biclusters* como atributos, conseguimos verificar que conjuntos de atributos (e os seus valores) são considerados mais importantes na classificação e comparar com os atributos individuais. Por outro lado, a segunda abordagem tira partido da teoria de conjuntos para encontrar os subconjuntos de atributos (e os seus valores) que estão mais associados com cada classe. Estas duas abordagens foram utilizadas em paralelo para tirar partido das características explicativas dos modelos utilizados e também para validar a robustez dos resultados obtidos. Várias experiências (incluindo *baselines*) foram incluídas para efeitos comparativos. Uma nota importante é a de que a abordagem DMD foi desenhada para ser genérica, podendo ser implementada de outras formas, utilizando outros algoritmos e até para tipos de dados diferentes, dependendo apenas das capacidades do algoritmo de *Biclustering* escolhido.

Nesta tese a implementação da metodologia DMD foi feita recorrendo a diversos *softwares* e linguagens de programação. O pré-processamento dos dados foi todo efectuado utilizando o software KNIME, incluindo a limpeza, transformação, categorização, adição de classe e selecção de atributos. O *Biclustering* baseado em Prospecção de Padrões foi corrido utilizando o algoritmo BicPAM (que faz parte do *software open-source* BicPAMS), a partir de código Java criado para correr conjuntos de experiências a partir de todas as combinações de parâmetros definidas (*ExperienceSet*). Adicionalmente foi criado código para traduzir os estágios intermédios dos *Biclusters* (de índices de categorias para valores de categorias, e destes últimos para as etiquetas finais, mais legíveis), que ocorreram como consequência de trabalhar com dados categóricos no BicPAMS. A Classificação baseada em *Biclustering* foi feita utilizando a biblioteca *scikit-learn* para a linguagem Python, recorrendo a modelos *Random Forest* para obter métricas de importância dos atributos utilizados na classificação. Finalmente, a Prospecção de Regras de Associação para Classificação foi feita em Python, tirando partido da biblioteca *open-source* SPMF, escrita em Java.

Seguidamente, a metodologia proposta foi utilizada na parte portuguesa do conjunto de dados ON-WebDUALS para encontrar padrões de apresentação de ELA que: 1) distinguissem os doentes de ELA

dos seus controlos e 2) caracterizassem grupos de doentes de ELA com diferentes taxas de progressão (categorizados em grupos de progressão Lenta, Neutra e Rápida, obtidos através de aplicação de um algoritmo de Maximização de Expectativas).

Na primeira tarefa nenhum padrão coerente emergiu das experiências efectuadas, não tendo sido encontrados resultados convergentes entre as técnicas de Classificação baseada em *Biclustering* e Prospecção de Regras de Associação para Classificação. Embora fosse uma tarefa mais complexa que a segunda, a selecção de atributos efectuada e o excesso de atributos vindos dos *Biclusters* discriminativos encontrados poderão ter contribuído para a ausência de resultados conclusivos. Contudo, foi possível verificar que não houve atributos individuais a serem apontados como os mais importantes para a classificação. Desta forma foi possível concluir que para distinguir os doentes dos controlos terão de ser considerados subconjuntos de atributos (e seus valores).

Na segunda tarefa foram encontrados padrões para cada um dos três grupos de progressão (Lenta, Neutra e Rápida), tendo sido reconhecidos e validados por clínicos especialistas em ELA. Os doentes de progressão Lenta foram identificados pelos valores consistentemente máximos de algumas perguntas da escala ALSFRS-R. Já para os doentes de progressão Neutra, a pergunta 5 da escala ALSFRS-R (relacionada com o corte de comida e manejar utensílios) com o valor 3, indicando algum decaimento funcional. Finalmente, os doentes de progressão rápida foram caracterizados pelos valores mais baixos dos atributos *Atraso no diagnóstico* e *Tempo de transição entre a região 1 e 2*.

Estes resultados sugerem que a nossa abordagem genérica baseada em *Biclustering* tem potencial para ser identificar padrões de apresentação noutros problemas ou doenças semelhantes.

Contents

List of Figures

List of Tables

Acronyms

2D	Two-dimensional
AD	Alzheimer Disease
ALS	Amyotrophic Lateral Sclerosis
ALSFRS-R	ALS Functional Rating Scale (Revised)
API	Application Programming Interface
ARFF	Attribute-Relation File Format
ARM	Association Rule Mining
AUC	Area Under Curve
C9orf72	Chromosome 9 open reading frame 72
CAR	Class Association Rule
CI	Confidence Interval
CK	Creatine Kinase
CSV	Comma-Separated Values
CV	Cross-Validation
DM	Data Mining
DMD	Discriminative Meta-features Discovery
DT	Decision Tree
EHR	Electronic Health Record
EM	Expectation-Maximization
EMG	Electromyography
ENCALS	European Network to Cure ALS
ESCO	European Skills/Competences, qualifications and Occupations
fALS	Familial ALS
FIM	Frequent Itemset Mining
FN	False Negatives
FNR	False Negative Rate
FP	False Positives
FPR	False Positive Rate
FS	Feature Selection
FTD	Frontotemporal Dementia

FVC	Forced Vital Capacity
GO	Gene Ontology
GUI	Graphical User Interface
HDL	High Density Lipoprotein
JPND	Joint Programme - Neurodegenerative Disease
LDL	Low Density Lipoprotein
LMN	Lower Motor Neuron
MAR	Missing at Random
MCAR	Missing Completely at Random
MCC	Matthews Correlation Coefficient
ML	Machine Learning
MRI	Magnetic Resonance Imaging
MS	Multiple Sclerosis
NEUROCLINOMICS2	Unravelling Prognostic Markers in NEUROdegenerative diseases through CLINical and OMICS data integration
NIV	Non-Invasive Ventilation
NMAR	Not Missing at Random
NP	Non-deterministic Polynomial (problem)
NSAID	Nonsteroidal anti-inflammatory drug
ONWebDUALS	ONTology-based Web Database for Understanding ALS
PD	Parkinson's Disease
ROC	Receiver Operating Characteristic
SNIP	Sniff Nasal Inspiratory Pressure
TN	True Negatives
TNR	True Negative Rate
TP	True Positives
TPR	True Positive Rate
TSV	Tab-Separated Values
TXT	Text File
UMN	Upper Motor Neuron
XLSX	Office Open XML Workbook (Excel File)

Chapter 1

Introduction

1.1 Context and Motivation

The work described in this thesis was done in the context of the NEUROCLINOMICS2 (Unravelling Prognostic Markers in NEUROdegenerative diseases through CLINical and OMICS data integration, with Ref. PTDC/EEI-SII/1937/2014) project at the LASIGE Research Unit (Ref. UID/CEC/00408/2019). This project's main objectives are to understand Amyotrophic Lateral Sclerosis (ALS) disease progression patterns and predict prognostic markers for personalized medicine. In this thesis we focused on the discovery of disease presentation patterns.

ALS is a heterogeneous neurodegenerative syndrome which affects the human motor system in a relatively short time period. Common symptoms are limb weakness, speaking and swallowing impairment, respiratory insufficiency and cognitive/behavioural changes. Eventually it culminates in respiratory failure, which is appointed as the main cause of death. Survivability is highly variable, depending on the speed of disease progression [12, 32].

To this day, it remains crucial to unravel definite diagnostic tests or biomarkers to help clinicians make a fast and clear diagnosis for this disease, given the high degree of variability in phenotype, family history, genes involved, molecular pathways and environmental factors that might induce it. Thus, it is believed that distinct mechanisms cause the neurodegeneration in ALS patients [32].

In this context, the use of Machine Learning models able to unravel the complexity of disease presentation patterns is paramount for disease understanding, targeting improved patient care and longer survival times. Furthermore, explainable models are vital, since clinicians must be able to understand the reasoning behind a given model's result or prediction before making a decision that can impact a patient's life. [8].

1.2 ONWebDUALS dataset

The JPND **ONWebDUALS** two-dimensional dataset is the result of a collective effort from ENCALS, a consortium composed of partners from European countries. This dataset is the first in Europe to contain answers of standardized patient questionnaires of ALS patients and controls, regarding their demographics, lifestyle, genetic mutations, family occurrences of ALS (or similar neurodegenerative diseases) and more. The major motivations behind the compilation of this dataset were to investigate the interplay between demographics, genetic mutations, clinical features and survival to discover causal relationships linking the patients' specific risk factors and ALS genotype-phenotype [49, 12]. So far, the dataset includes information from Patients and Controls from four European countries (Turkey, Germany, Portugal and Poland) or, more concretely, five different cities (Antalya, Hannover, Jena, Lisbon and Warsaw).

The data considered for each subject (Patient or Control) can be seen as:

- a set of static features (which could not change over time), which includes the information about the subject's demographics, disease severity, co-morbidities, medication, genetic information, habits, trauma/surgery information and occupations;
- a set of temporal features (which could change over time), such as disease progression rate (e.g. ALSFRS-R scale measurements, pulmonary function tests measurements) and clinical laboratory investigations.

Considering what was aforementioned, the general layout of the data can be seen in Figure 1.1:

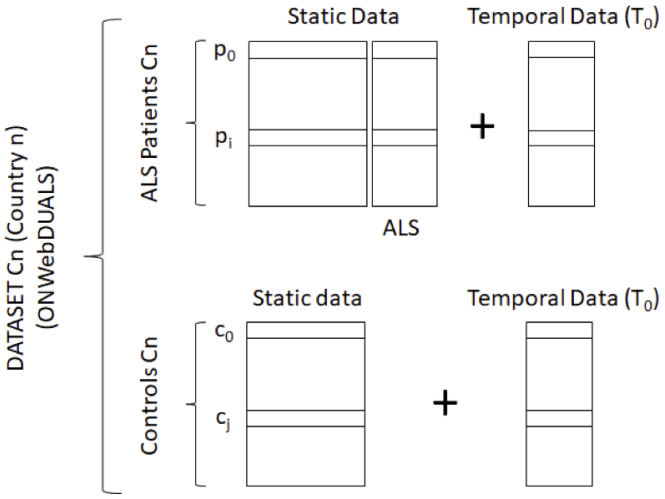


Figure 1.1: Layout of the ONWebDUALS dataset.

However, regarding the temporal features, it is important to state that a single time point (shown as a temporal snapshot T_0 in Figure 1.1) was available in this dataset, thus being the only one considered. Succinctly, it implies that this dataset can be globally treated as static data. Another major consideration

is that disease-related features are specific to the ALS Patients (not being present in the Controls' data). This can also be seen in Table 1.1, where the feature sets from the questionnaires are reported.

Feature Set	Patients	Controls
Demographic Information	Yes	Yes
Disease Features	Yes	No
Clinical Signs	Yes	No
Disease Severity and Progression Rate	Yes	No
Investigations (laboratorial)	Yes	No
Co-morbidities (other diseases)	Yes	Yes
Medication	Yes	Yes
Genetic Information	Yes	Yes
Habits / Trauma and Surgery	Yes	Yes
Occupations	Yes	Yes

Table 1.1: Feature sets from the ONWebDUALS dataset standardized questionnaires.

The analysis in this thesis is the first performed over this data, therefore special considerations had to be had in handling and cleaning the data. An exploratory analysis performed on the original dataset (over 600 features) revealed some issues: some features had irrelevant or erroneous values that needed to be treated or uniformized before discretization and others had duplicate column names. In addition, only Patients and Controls selected in Lisbon (472 Patients and 300 Controls) were considered for further analysis. This was done for several reasons: the Portuguese subjects were followed by one of our expert clinicians, having less missing information per subject and less prone to biases.

The entire data treatment performed is thoroughly detailed in *Section 3.1.1 Data Pre-processing*. The number of effectively considered features varied with the task and respective experiment, which are thoroughly described in *Chapter 4 Discriminative Meta-features Discovery: A Case Study in the Portuguese ONWebDUALS Dataset*.

1.3 Objectives and Contributions

The main objective of this thesis is to unravel disease presentation patterns of ALS patients on Electronic Health Record (EHR) data from the aforementioned ONWebDUALS dataset. These patterns can be found in the form of relevant subsets of features and their values (Meta-features), which caracterize and discriminate class-labeled two-dimensional data.

This main objective encompasses two clinical problems we tried to address as secondary goals. The first goal is to help improve diagnosis of the disease, to start the patients' treatments as early as possible.

And the second goal is to help with the prognosis of ALS patients to help assess disease progression, refine therapeutical trial design and improve patient care.

To accomplish these goals, in this thesis is suggested a new Data Mining approach called Discriminative Meta-features Discovery (DMD). To the best of our knowledge, this approach uniquely combines Data Mining and Machine Learning techniques in order to find class-discriminative patterns in two-dimensional data. At its base is a specialised pattern-based Data Mining technique called Biclustering, which was used to identify any potentially relevant subsets of features (and their respective values) in this complex two-dimensional dataset. Up-to-date this technique has been successfully applied in healthcare data, allowing the discovery of groups of biological entities or individuals meaningfully correlated on a subset of conditions [44].

More concretely, the DMD approach is a workflow composed of several steps. First of all, the data is pre-processed (including discretization and class-labelling). Then Pattern Mining-based Biclustering is applied to the pre-processed data to find discriminative Biclusters and their patterns. Finally, the patterns from the Biclustering phase are further processed in two separate ways (using Biclustering-based Classification and Class Association Rule (CAR) Mining) to find the most discriminative Meta-features.

Finally, in the context of this thesis the DMD approach was applied to the Portuguese portion of the ONWebDUALS dataset, where the aforementioned secondary goals were translated into concrete Tasks with distinct investigation directions:

1. Discover Meta-features which best distinguish the ALS patients from their controls, if any;
2. Discover Meta-features which caracterize ALS patients' progression groups, if any.

1.4 Thesis Outline

This thesis is organized as follows. *Chapter 2 Background and Related Work* contains all the background information, including a thorough disease description, important definitions and related work. *Chapter 3 Discriminative Meta-features Discovery* focuses on outlining the newly suggested DMD approach that was used to complete each one of the defined tasks. *Chapter 4 Discriminative Meta-features Discovery: A Case Study in the Portuguese ONWebDUALS Dataset* reveals the results of the performed experiments. Finally, *Chapter 5 Conclusions and Future Work* disclosures this thesis' main conclusions and suggests possible future work.

Chapter 2

Background and Related Work

2.1 Amyotrophic Lateral Sclerosis

Amyotrophic Lateral Sclerosis (ALS) is an idiopathic and heterogeneous neurodegenerative syndrome, which affects the upper and lower human motor system. The onset area of the body where the neurodegeneration starts may vary, but within weeks or months progressive motor deficits ensue, difficulting proper nutrition and causing cognitive and/or behavioural changes. Eventually it culminates in respiratory failure, appointed as the main cause of death. Survivability is highly variable, with the most common case being around 3-4 years, depending mostly on the origin onset area of the body and speed of disease progression [13, 32].

ALS is usually classified in two main forms: familial (fALS) and sporadic (sALS). Familiar ALS comprises 5-10% of the cases, and it is the best understood form of the disease, since more than 30 genes and loci of major effect involved are, as of date, identified. The most frequent gene mutations occur on C9orf72, SOD1, FUS and TARDBP (which codes for the TDP-43 protein) genes. Sporadic ALS includes the remaining 90% of the patients. Even though it is the most frequent type, so far it has been difficult to concretely identify the underlying causes, since only 15% of the sALS cases can be explained by genetic factors. The only established risk factors are advanced age, male gender and some very specific genetic mutations [13, 34].

Environmental factors have been proposed to explain the high prevalence of ALS in certain populations, like smoking, exposure to pesticides and organic toxins, electromagnetic radiation and high levels of exercise. However, only smoking was proven to have a definite evidence of risk. Studies of relevant sample size are direly needed to test additional factors, such as dietary habits (fat and glutamate rich diets), gut microbiome and psychological stress [13, 34].

ALS incidence is homogeneous across Europe, with an incidence rate of 2.16 per 100000 person years

(95% CI 2.0 to 2.3). Genderwise, for men the incidence rate is higher (3.0 per 100000 person years, 95% CI 2.8 to 3.3) than among women (2.4 per 100000 person years, 95% CI 2.2 to 2.6). These rates tend to increase with the advance of age for both genders, more pronouncedly after 40 years of age, reaching its apex at 70-74 years for men and 65-69 years for women. After that, disease occurrence declines rapidly [42].

On a more global scale, it has been proposed that different genetic etiologies underlying motor neuron degeneration may exist across major ethnic groups. One example of such was found in Japan, where OPTN gene mutations have been appointed as responsible for autosomal recessive ALS in Japanese families [57].

The El Escorial criteria is used to diagnose patients who have a history of progressive muscle weakness that has spread to one or more regions of the body, and that any other disease cannot explain. Its final result is a degree of probability that the patient has ALS: definite, probable, probable (laboratory-supported) and possible [13].

To assess functionality and accurately track progression of patients' disability, the Revised ALS Functional Rating Scale (ALSFRS-R) is used. This scale is composed of 12 questions, where each one can be evaluated in a scale from 0 to 4, totalling to a maximum of 48 points (indicating no dysfunction). This scale takes not only limb and bulbar function in consideration, but also the degree of respiratory disfunction, allowing a good evaluation of patients' quality of function and life [7].

Additionally, it was recently discovered that ALS shares pathobiological features (e.g.: toxic aggregates of TDP-43 protein) with Frontotemporal Dementia (FTD), generating a whole spectrum of disease phenotypes in between. Some patients even tend to suffer from both conditions at once, or have someone in their family which suffers from one of the two [13].

Regarding possible treatments, currently the only widely available drug known to prolong ALS patients' survival is Riluzole, shown to increase life span for approximately 3 months. To alleviate the respiratory symptoms, Non-Invasive Ventilation (NIV) is commonly advised, since it extends survival with an effect size greater than Riluzole (median survival increase of 7 months), especially if started as soon as muscle weakness is detected [13].

In the last decades many technological advances have sped up the discoveries regarding this disease, but it remains crucial to unravel definite diagnostic tests or biomarkers to help clinicians make a fast and clear diagnosis. In addition, this would greatly help to assess disease progression, refine therapeutical trial design and start the patients' treatments as early as possible [32].

In conclusion, taking into consideration the high degree of variability in phenotype, family history, genetic mutations, molecular pathways and possible environmental factors involved, it is believed that different mechanisms cause the neurodegeneration in ALS patients [13].

2.2 Unsupervised Learning

In Machine Learning, **Unsupervised Learning** includes all techniques whose learning process is unsu-

pervised due to the inexistence of a class label for each element of the dataset (no ground truth), such as Clustering, Biclustering and Pattern Mining. These techniques are widely used in Data Mining to group data observations based on their features' values [22].

2.2.1 Clustering

Given a dataset with n observations, $X = \{x_1, \ldots, x_n\}$, the **Clustering** task aims to find subsets of observations (**Clusters**), $\{I_1, \ldots, I_r\}$, where every $I_i \subseteq X$ satisfies certain intra-cluster (within a Cluster) and inter-cluster (between different Clusters) criteria of (dis)similarity over the whole space [28]. A Clustering example is shown in Figure 2.1.

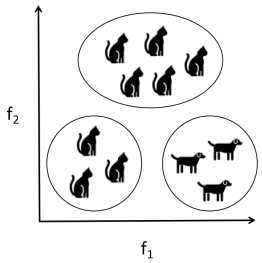


Figure 2.1: Illustrative example of Clustering.

Clustering can help discover previously unknown groupings within a dataset, and currently it is used in a myriad of very different applications (e.g. business intelligence, biology, image pattern recognition, web search and security). From a Data Mining point of view, it can be used as a standalone tool to retrieve new knowledge about and from the data, but it can also be used as a pre-processing step for other algorithms like classifiers. *In extremis*, a Cluster can be also seen as an implicit class: the objects in a Cluster are similar to each other and, at the same time, are different from the objects in other Clusters, which allows for **automatic classification**. Finally, it can also be used for **outlier detection** [22].

2.2.2 Biclustering

A **two-dimensional** (2D) dataset (or matrix) can be defined by n observations (rows) $X = \{x_1, \ldots, x_n\}$, m attributes (columns) $Y = \{y_1, \ldots, y_m\}$, and $n \times m$ elements (values) a_{ij}. Given a real-valued or symbolic matrix A, the **Biclustering** task's objective is to find a set of Biclusters $\mathcal{B} = \{B_1, \ldots, B_q\}$ (a **Biclustering solution**), such that each Bicluster B_i satisfies specific criteria of homogeneity and statistical significance [28]. Figure 2.2 shows an example of a Biclustering solution.

A **Bicluster** $B = (I, J)$ is a subspace given by a subset of rows $I \subseteq X$ which show a coherent pattern observed for a subset of columns $J \subseteq Y$. It is considered **maximal** if and only if there is no other Bicluster $B' = (I', J')$ such that $I \subseteq I'$ and $J \subseteq J'$, while $B \in \{B_1, \ldots, B_q\}$ and $B' \in \{B_1, \ldots, B_q\}$, meaning that it cannot be expanded further in any of the two dimensions [28, 25].

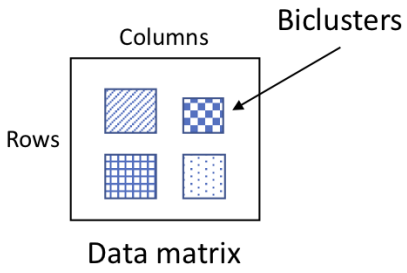


Figure 2.2: Illustrative example of a Biclustering solution.

The **homogeneity** criteria determines the structure, coherence and quality of a Biclustering solution, where it can be said that:

- The **structure** is described by the number, size, shape and position of said Biclusters (Figure 2.6);
- The **coherence** of a Bicluster is defined by the observed correlation of values (**correlation assumption**) and the allowed deviation from expectations (**coherence strength**);
- The **quality** of a Bicluster is defined by the type and amount of tolerated noise (values or symbols that differ from the expected pattern) and missing elements.

Taking these points into consideration, it sounds reasonable that the homogeneity criteria to apply to a given dataset should depend on its **regularities**: the possible domain of the dataset's features - either if they are real-valued, symbolic or non-identically distributed - and their respective distribution [28].

To guide the search for Biclusters, a Biclustering algorithm uses a **merit function**, which evaluates how good a found Bicluster is based on the values of its elements. The chosen merit function is highly correlated with the characteristics of the Biclusters it can obtain, since it defines the type of homogeneity being sought in each one of them. It is important to use a different type of merit function to evaluate the quality of the identified Biclusters, in order to avoid evaluation biases [44, 28]. It is of note that **pattern-based merit functions** exist, allowing to assess the maximality of Biclusters with well-defined patterns composed of a set of symbols from one dimension, repeated over the other. These functions accommodate principles from Biclustering to handle non-constant patterns, sparse data and to minimize the drawbacks of discretization procedures (e.g. loss of information) by alleviating noise [28, 25, 65].

A Bicluster is **statistically significant** if its probability of occurrence deviates from expectations, that is to say that the found pattern has a very low probability of occurring by chance in the given dataset. To derive such a conclusion for a given Bicluster, the **p-value** probability from a statistical hypothesis test against a null data model is usually used [28]. In the same way as homogeneity, the statistical significance criteria to apply to a given dataset should depend on its regularities. Any Biclusters found must be subject to statistical assessments in order to:

1. Measure and minimize the risk of including irrelevant Biclusters in the solution (**false positives**, error of type-I);

2. Guide the Biclustering task without increasing the risk of excluding relevant Biclusters (**false negatives**, error of type-II) [28, 24].

However, even though this criterion is majorly important to guarantee the soundness of a given Biclustering solution, there is no agreed ground truth on how to verify and promote it. Many algorithms are guided by merit functions that enforce homogeneity over statistical significance, but having the former does not imply the existence of the latter, since it is very common for small Biclusters to have good homogeneity levels by chance. Therefore, both criteria need to be combined and considered when choosing between solutions [28, 24].

Besides what was said above for **quality**, it is important to underline that some actions included in the pre-processing (normalization and discretization) and post-processing (merging, filtering, extending and reducing) phases contribute for Bicluster quality adjustment to account for noise [28]. Moreover, some algorithms and frameworks have imputation procedures or dedicated interpretations to deal with missing values [18, 27].

2.2.2.1 Types of Biclusters

Typically each subspace problem to solve has its own specificity, thus **different types of Biclusters** may be needed or considered interesting. The type of found Biclusters depends on the Biclustering algorithm and possibly its parameterization. As defined in [44], these different types can be categorized in four major classes: 1) Biclusters with constant values, 2) Biclusters with constant values on rows or columns, 3) Biclusters with coherent values and 4) Biclusters with coherent evolutions. Figure 2.3 shows some examples:

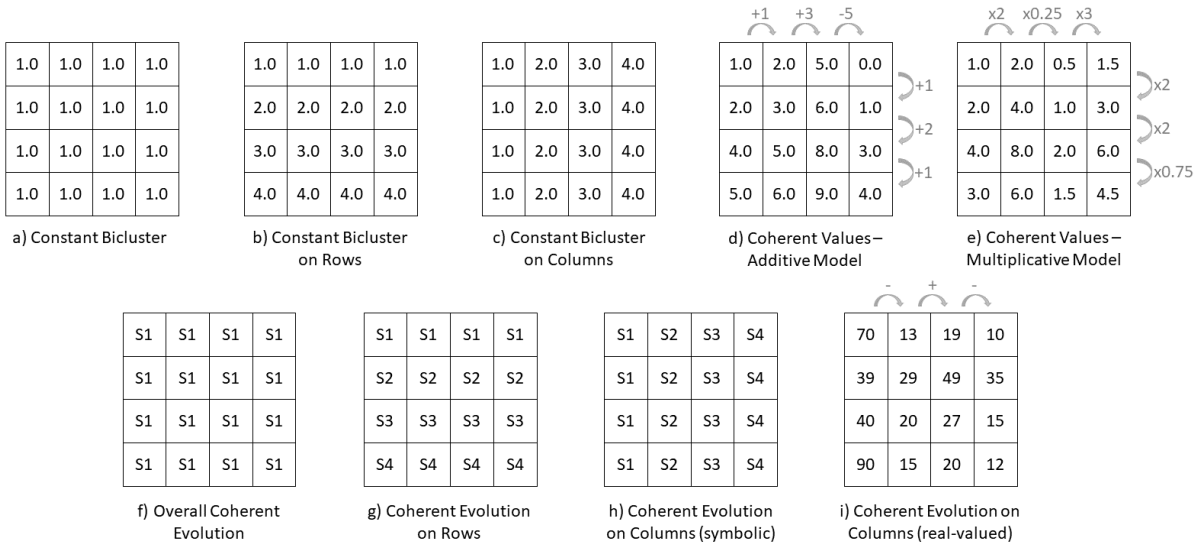


Figure 2.3: Examples of different types of Biclusters (adapted from [44]).

2.2.2.2 Biclustering Models

At the basis of the aforementioned different types of Biclusters stand mathematical models which describe them and define their coherence.

Regarding **Constant Valued Biclusters**, they're considered *perfect* if the Bicluster is a subspace (I, J) where all real-valued elements a_{ij} are equal for all $i \in I$ and all $j \in J$: $a_{ij} = \mu$. An example of this is given in Figure 2.3 a). However, since real data usually has noise, perfect Biclusters are rare to find. This means that a regular Constant Bicluster is better defined by $a_{ij} = \mu + \eta_{ij}$, where η_{ij} is the noise amount associated with the real value μ of a_{ij} [44].

When looking at **Biclusters with Constant Values on Rows or Columns**, we can have two types of models: **Additive** or **Multiplicative**, depending on the relation between the values. These Biclusters are considered *perfect* if the Bicluster is a subspace (I, J) where all elements a_{ij} are obtained using one of the following expressions:

Model Type	Expression
Additive on Rows	$a_{ij} = \mu + \alpha_i$
Multiplicative on Rows	$a_{ij} = \mu \times \alpha_i$
Additive on Columns	$a_{ij} = \mu + \beta_j$
Multiplicative on Columns	$a_{ij} = \mu \times \beta_j$

Table 2.1: Model types and respective expressions for Biclusters with Constant Values.

where α_i is the adjustment value per row $i \in I$ and β_j is the adjustment value per column $j \in J$. Examples for the Additive case can be found in Figure 2.3 b) and c), by adding one unit either in the rows or the columns, respectively. By adding the noise parcel as before we get the regular versions of these Biclusters [44].

If the Additive and Multiplicative models were used to obtain Biclusters on both rows and columns at the same time, **Biclusters with Coherent Values** would be obtained. In the same fashion, for the *perfect* case we have:

Model Type	Expression
Additive	$a_{ij} = \mu + \alpha_i + \beta_j$
Multiplicative	$a_{ij} = \mu \times \alpha_i \times \beta_j$

Table 2.2: Model types and respective expressions for Biclusters with Coherent Values.

Examples of this model can be seen in Figure 2.3 d) and e), with the respective increments in grey for each row/column. The addition of the noise amount can also be applied to this model as well, in order to find regular Biclusters of these types [44].

When dealing with the possibility of **Bicluster Overlapping**, it is fair to consider that the value of an element in the matrix may be composed of several layers, either in an additive or multiplicative way. This is formalized by the **Plaid** model, given by $a_{ij} = \sum_{k=0}^{K} \theta_{ijk}\rho_{ik}\kappa_{jk}$, where K is the number of layers (Biclusters) and the value of θ_{ijk} indicates the contribution of each Bicluster k specified by ρ_{ik} and κ_{jk}, both binary terms which represent the membership of row i and column j in Bicluster k, respectively. It is of note that this notation allows the representation of the previously specified models of Biclusters depending on the definition of θ_{ijk}. For example, if $\theta_{ijk} = \mu_k$, then the Plaid model would identify a set of K Constant Biclusters.

From the Plaid model two more restrictive models can be derived: the **General Additive Model** (GAM) and the **General Multiplicative Model** (GMM) [44]. For the former, as defined above in the Plaid model, every element a_{ij} represents a sum of additive models each representing the contribution of the Bicluster $(I, J)_k$ to the value of a_{ij} in case $i \in I$ and $j \in J$. Examples of this can be seen in the following figure:

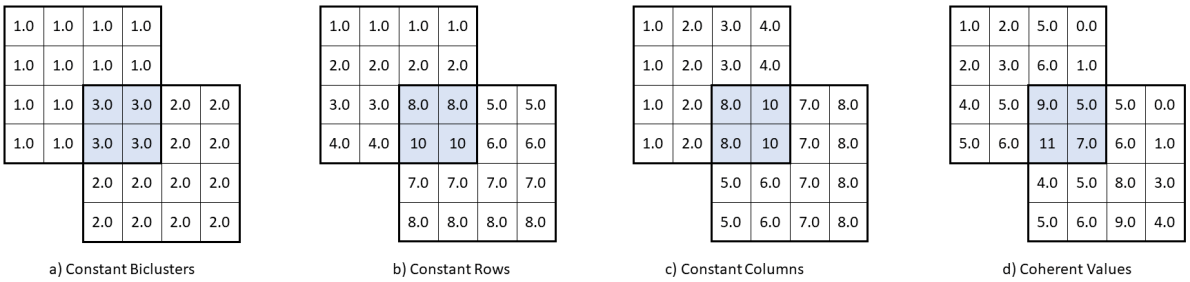


Figure 2.4: Overlapping Biclusters with GAM (adapted from [44]).

For the latter, every element a_{ij} represents a product of contributions of the Bicluster $(I, J)_k$ to the value of a_{ij} in case $i \in I$ and $j \in J$. It means that in this case the value of each element is given by $a_{ij} = \prod_{k=0}^{K} \theta_{ijk}\rho_{ik}\kappa_{jk}$. Equivalently, some examples are provided:

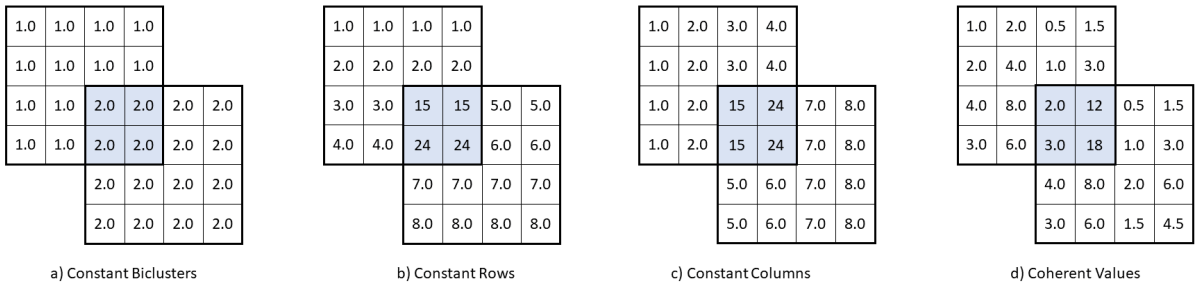


Figure 2.5: Overlapping Biclusters with GMM (adapted from [44]).

When working with real-valued or integer data, it is possible to have sign-changes intertwined in the

patterns. Those **Symmetries** are considered relevant in some areas (e.g. to find activation or repression mechanisms in regulatory processes on Gene Expression data), and can only be captured by Biclustering algorithms that support them [27].

Finally, the last kind of models that needs to be mentioned are the ones that find **Biclusters with Coherent Evolutions**. These models permit to address the problem of finding coherent evolutions (or trends) across the rows and/or columns of the matrix without regarding the exact value of the elements. Additionally, these models are the only ones that can be applied to symbolic/categorical data. Examples of this can be found in Figure 2.3 from f) to h) for categorical data and Figure 2.3 i) for real-valued data [44]. In this work we consider Biclusters with Coherent Evolutions to be the target since the data being processed by the algorithm is purely categorical.

2.2.2.3 Types of Bicluster Structures

Another important aspect is the Biclusters **structure** that a Biclustering algorithm is able to discover. In its simplest form, an algorithm will assume one of two things: that there is only one Bicluster in the data matrix, or that it contains K Biclusters, where K is the number of Biclusters expected to be found. If it is the latter case, according to Madeira et al. [44] several different types of structures within a set of Biclusters can be encountered:

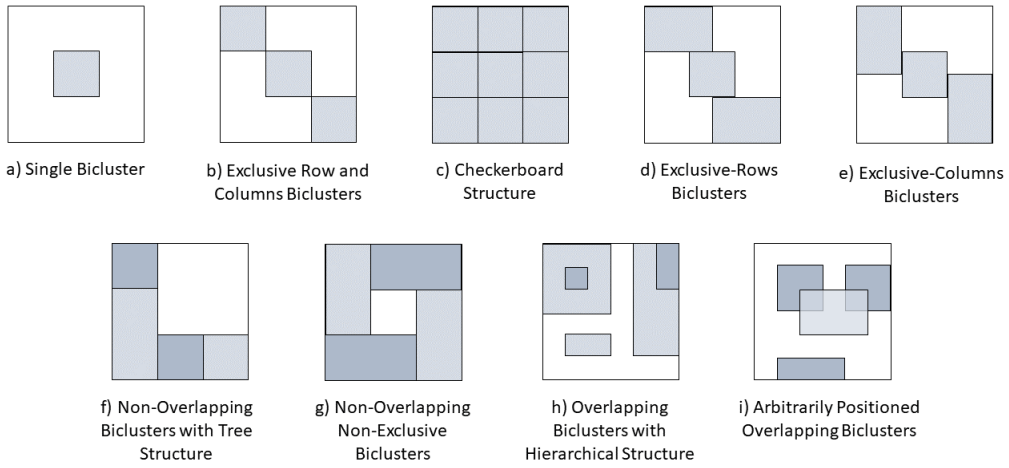


Figure 2.6: Examples of different types of Bicluster structures (adapted from [44]).

Additionally, Figure 2.6 b) to e) presents structures which assume exhaustive Biclusters, in which every row and column in the matrix belongs to at least one Bicluster. However, most of the structures presented above are very restrictive. In real data, it is most probable that some rows and columns do not belong to any Bicluster and that Biclusters can overlap each other, like it is seen in Figure 2.6 i) [44]. Other stopping conditions besides the number of Biclusters to discover are also possible, which will be seen later

on *Chapter 3 Discriminative Meta-features Discovery*. In this work we will use a Biclustering algorithm able to find arbitrarily positioned overlapping Biclusters (Figure 2.6 i)) since structure restrictions do not make sense in our problem.

2.2.2.4 Biclustering Evaluation

To define evaluation measures for Biclustering solutions, we need to distinguish two types of possible datasets:

1. **Synthetic data**, artificially generated data with planted Biclusters, used to test the accuracy of a given algorithm against a true and known solution (**ground truth**, also known as **true** or **hidden** Biclusters);
2. **Real data**, any dataset without a ground truth [28].

Solutions obtained from applying Biclustering algorithms to a given dataset will vary according to the chosen algorithm and its parameterization. Therefore, metrics to evaluate and compare said solutions are important to find the best algorithm, if an algorithm is working well or the best result between experiments. In this thesis we are interested in the last case. Evaluation metrics (or indices) can be divided in three main categories: external, internal and relative.

External metrics determine the similarity between an obtained solution and apriori knowledge, and can be used with synthetic data (**Accuracy-based** views) or with real data plus additional domain information (**Domain Significance** views). In addition, they can be used to compare different algorithms. These are the most abundant type of metric and the ones usually preferred, since they have greater precision and are easier to define, use and implement [60].

Internal metrics contrast the obtained solution with the intrinsic structure of the dataset. These metrics have to be used when no ground truth is available, although they are not as precise as the external ones. However, they are usually adapted from Clustering concepts (e.g. Cluster compactness and separation) which are hard to extend when Bicluster overlapping is allowed. Therefore, when working with real data they tend to be avoided, mostly if there is a possibility of creating a synthetic dataset [60].

Relative metrics are used to compare different configurations of input parameters and solutions, in order to find the optimal set of parameters for a given dataset. The usual procedure is to evaluate and rank each solution obtained from the different user-defined combinations of parameters to select the best. To obtain the best results, all combinations should be investigated. However, these metrics are very complex to formulate and thus almost non-existent, since Biclustering algorithms are very heterogeneous relatively to their input parameters [60].

Finally, if no usable metric exists, a possible alternative is to define a criteria of usefulness depending on the problem to solve and the characteristics of the Biclusters to find the best solution. In this thesis we used such a solution, by choosing the best Biclustering solutions according to the number of discriminative Biclusters in them (as specified later in this document, in Section 2.3.2 Discriminative Biclustering and in Section 3.1.3 Pattern Mining-based Biclustering).

2.2.3 Clustering vs Biclustering

Regular Clustering solutions are limited to grouping objects one dimension at a time, by using all features that describe a given set of objects [28]. This means that they derive a **global model** from the data, which is a restriction when dealing with 2D data spaces with only locally correlated values [28, 44]. This can be clearly seen in Figure 2.7 below:

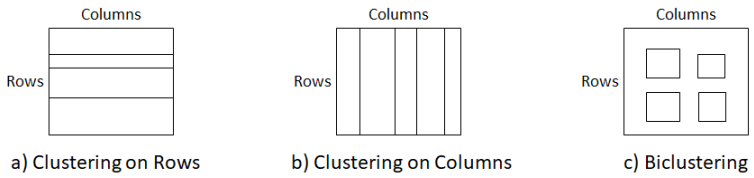


Figure 2.7: Clustering and Biclustering Comparison (adapted from [18]).

Data Mining techniques like Biclustering are then necessary to identify any potentially relevant subspaces in complex 2D datasets, as they can perform local clustering on both dimensions simultaneously (**local model** [44]). Additionally, groups of biological entities or individuals are usually meaningfully correlated only on a subset of conditions/records [28].

However, Biclustering is computationally expensive due to combinatorial optimization. Best case scenario, this task is an **NP** (Non-deterministic Polynomial) problem, meaning it is solvable in polynomial time, but it becomes more complex when searching for non-exclusive and non-exhaustive Biclusters. Due to this, most algorithms follow heuristics or stochastic approaches, by producing sub-optimal solutions or adding constraints to simplify the problem. Other approaches, such as Pattern Mining-based Biclustering, target exhaustive enumeration while using restrictions during the search for efficiency [25].

2.2.4 Pattern Mining

Consider a finite set of items L, and P as an itemset where $P \subseteq L$. One **transaction** t can be defined as a pair (t_{id}, P) with $id \in \mathbb{N}$. A finite set of transactions $\{t_1, \ldots, t_n\}$ then composes an **itemset database** D over L. If another itemset $P' \subseteq P$, it implies that $P' \subseteq (t_{id}, P)$. The **coverage** of an itemset P, ϕ_p, is the set of all transactions in D where P occurs: $\phi_p = \{t \in D \mid P \subseteq t\}$.

From this we can derive the **support** of P in D, sup_P, that can be either **absolute** (the size of ϕ_p, $\mid \phi_p \mid$) or **relative** ($\mid \phi_p \mid / \mid D \mid$). More simply, the support is the frequency (absolute or relative) of the itemset P in the database D [25]. Figure 2.8 below exemplifies the computation of the coverage and support of an itemset P over an itemset database D.

L = {A, B, C, D, E}
P = {A, C}

D:

t_{id}	items
1	{A, B, C, D}
2	{A, B, E}
3	{B, C, D, E}
4	{A, C}

|D| = 4

ϕ_P = {1, 4} $|\phi_P|$ = 2

absolute sup_P = $|\phi_P|$ = 2
relative sup_P = $|\phi_P|/|D|$ = 2/4 = 0.5 = 50%

Figure 2.8: Example of an itemset database D with the coverage and support of an itemset P.

Given an itemset database D and a user-defined minimum support threshold θ, the **Frequent Itemset Mining** (FIM) problem consists on the computation of the set $\{P \mid P \subseteq L, sup_P \geq \theta\}$. Thus, a **Frequent Itemset or Pattern** is an itemset P with $sup_P \geq \theta$. In simpler terms, an itemset is frequent if it appears in a dataset with a frequency equal or greater than the user-specified threshold [25].

A **closed frequent itemset** is a frequent itemset which has no superset with the same support. A **maximal frequent itemset** is a frequent itemset for which all supersets are not considered frequent [25]. The closed frequent itemsets are a subset of the frequent, just like the maximal frequent are a subset of the closed frequent, as shown in Figure 2.9.

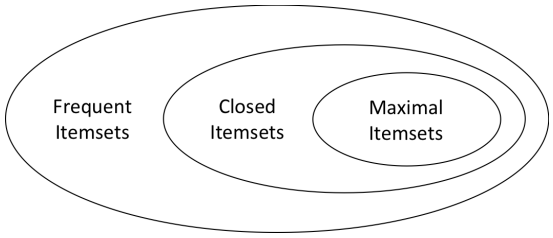


Figure 2.9: Frequent, closed frequent and maximal frequent itemsets.

FIM relies on **monotonic** and **anti-monotonic** properties to prune over combinations of itemsets in order to gain efficiency. Considering two itemsets P and P', where $P' \subseteq P$, and a predicate M, M is monotonic if $M(P) \Rightarrow M(P')$ and anti-monotonic when $\neg M(P') \Rightarrow \neg M(P)$. This means that the support of P is bounded by the support of P', implying that if any subset P' is not frequent, then P is not frequent as well [25].

Three major search strategies can be used to perform FIM:

1. **Apriori-based**, which applies the monotonicity principle (an itemset is candidate if all its subsets are frequent) to iteratively combine $(k-1)$-itemsets to generate new candidate k-itemsets in k scans, until no new candidate groups can be found;
2. **Pattern Growth**, which builds a frequent-pattern tree from an ordered list of frequent items to be later mined, based on prefix paths co-occurring with growing suffix patterns;

3. **Vertical projection**, which compiles the set of transaction ids where each item appears and then grows the itemsets using a depth-first strategy, by intersecting the sets of transaction ids to minimize scanning the database [25].

The first two search strategies listed above consider itemset databases in the **horizontal format**: $\{t_{id}, P\}$, where t_{id} is the transaction id and P is the set of items involved in said transaction, as seen in Figure 2.8. The last one considers the database is in the **vertical format**: $\{item : \{t_{id_1}, \ldots, t_{id_n}\}\}$, where each item is associated with the set of transaction ids from the transactions where it appears [22].

2.2.5 Pattern-Mining based Biclustering

As aforementioned, traditional Biclustering algorithms use flexible merit functions to guide the data space exploration. However, constraints like searching for a fixed number of Biclusters or retaining only non-overlapping structures are put in place, to ease the problem's complexity [25].

Pattern Mining-based approaches require the redefinition of those functions in terms of support and other relevant metrics, but in doing so they allow for a scalable exhaustive space search. This, in turn, produces a flexible structure containing an arbitrarily high number of Biclusters, while still catering for homogeneity and statistical significance criteria. These approaches can be divided in three major steps: mapping, mining and closing. All three steps are relevant in affecting the solution's coherence, structure and quality [25, 28].

The first step - **mapping** - is responsible for the normalization and discretization of the data matrix. Since the data has to be itemized to use Pattern Mining-based solutions, usually this step is mandatory for real-valued data. However, it can be optional if the algorithm is able to discretize the data internally or if the data is already discretized. In this phase it is also performed the handling of outliers, missing values and noisy elements. The second step - **mining** - is the core task, composed by the application of the target pattern miners that will model the type of Biclusters that can be found in the given solution. The last step - **closing** - deals with the post-processing of the mined patterns, to mostly improve the quality of the found solution. This is done through **merging** (affecting structure and dealing with overlapping while still maintaining homogeneity), **extension** (to improve noise tolerance) and **filtering** (to deal with Biclusters numerosity and compactness) [25].

2.2.5.1 Advantages and Disadvantages of Pattern Mining-based Biclustering

Pattern Mining-based approaches' major benefits are:

- Efficient exhaustive searches;
- Dealing with missing and noisy data;
- Ability to use different models to search for specific types of Biclusters;
- Annotating the significance of Biclusters to assess the pattern's relevance;

- Producing non-exhaustive, non-exclusive structures of Biclusters where overlapping is allowed.

Thus, and in accordance with Henriques et al. [25], these approaches are well suited to find shared local patterns within physiological and clinical data. However, performing data discretization always implies some loss of information. In particular, when discretizing real-valued features, the **items-boundary problem** can occur: assigning two elements with similar real-values to two different categories due to their closeness to the interval boundary. To minimize this problem, **multi-item assignments** can be applied: elements with values near a cut-off point of discretization can be assigned to the two categories associated with the closest ranges of values [25, 24].

2.2.5.2 Meta-features

When using Pattern Mining-based Biclustering to look for Constant Biclusters on Rows or Columns (Figure 2.3 b) and c)) and Coherent Evolution on Rows or Columns (Figure 2.3 g) and h)) it is possible to discern a **representative pattern** of each real-valued or categorical Bicluster, respectively. These patterns are composed by subsets of features (and their respective values) which characterize a Bicluster's contents. They can be seen as frequent patterns, representing a distinct data space with greater discriminative power than when considering only their constituent feature-value pairs individually, allowing them to capture additional underlying knowledge in the data [22, 44].

In this thesis, when those patterns can characterize and discriminate class-labelled data they will be called **Meta-features**. For example, assuming features F_i with $i \in \mathbb{N}$ on columns and observations on rows, the pattern for the Bicluster in Figure 2.3 h) is $\{F_1 = S1, F_2 = S2, F_3 = S3, F_4 = S4\}$. By finding the most class-discriminative Meta-features it is possible to unravel disease presentation patterns.

2.3 Supervised Learning

In Machine Learning, **Supervised Learning** includes all techniques whose learning process is considered supervised due to the existence of a label for each element of the dataset (the ground truth that should be predicted), such as Classification (where a categorical class is being predicted) or Regression (where a numeric value is being predicted) [22]. In this thesis the focus is **Classification**.

2.3.1 Classification

Classification can be defined as the process of finding a model that describes and distinguishes data classes or concepts. Such models, called **Classifiers**, are used to predict categorical/symbolic class labels for new observations, and are constructed in two phases:

1. The **Training phase**, where the Classifier is built by learning from the Training data: the model internally defines at least one decision boundary from the data features, as seen in Figure 2.10;

2. The **Test phase**, where the trained Classifier is evaluated with the remainder (and unseen by the model) Test data [22].

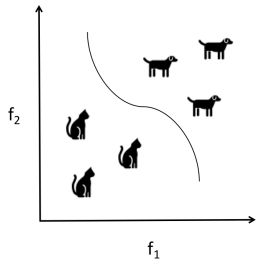


Figure 2.10: Illustrative example of a decision boundary on a Binary Classification problem.

For Classification, the used datasets must be composed of observations/tuples and their associated class labels. A **tuple** X is an n-dimensional feature vector, $X = (x_1, x_2, \ldots, x_n)$, which contain its n feature values [22]. For example, the dataset's data tuples for Figure 2.10 should contain the values for their two features (f_1 and f_2), and each tuple must also have a class label (cat or dog).

Classification problems can be divided according to the number of classes found on the dataset: with only two class labels it is a **Binary Classification** problem (as seen in Figure 2.10); when there are more than two class labels it is considered a **Multiclass Classification** problem. Multiclass Classification problems are more complex than the Binary (they can have several decision boundaries), and many real-world problems are of the former kind [40].

2.3.1.1 Classifier Example - Decision Trees

In this section, a representative Classifier model - Decision Trees - which supports Binary and Multiclass Classification is introduced. **Decision Trees** (DTs) are widely used models in Data Mining and Machine Learning, due to being easy to understand, visualize and interpret [29]. The majority of algorithms that build a DT from the Training data (e.g. ID3 [52], C4.5 [53] and CART [3]) employ a greedy top-down recursive divide-and-conquer strategy: the data is recursively divided into smaller partitions by selecting in each iteration the feature which separates best the data tuples into their respective classes (the **splitting attribute**), by using an **attribute selection measure** (e.g. Information Gain [52], Gain Ratio [53] or Gini Index [3]). The objective of these measures is to obtain, as much as possible, **pure** partitions, where all tuples belong to the same class [22]. However, the Information Gain is biased in favor of categorical features with more levels, with the Gain Ratio being proposed to tackle this issue [10].

While running the algorithm, if after a split all the tuples in a partition belong to the same class, a **leaf** node is created and labeled with the said class. Otherwise, the remaining data is split again on the next feature that divides it best. Finally, if it reaches a point where there are no more features available on which to split any last tuples, the class for the last leaf is decided by majority voting (the most frequent class on those tuples) [22].

On the example seen in Figure 2.11, the splitting attributes are enclosed in rectangles, and the leaf nodes in ellipses. The root split also shows that a DT can have any number of children for each node. The partitioning scenarios depend on the type of feature being split: for categorical features usual cases are splitting by all categories or one-vs-rest, while for real-valued features thresholds are defined (e.g. $A < 5$ and $A \geq 5$). After building the model, in order to classify a new tuple, the DT's nodes are traveled from the root, following the branches according to the new tuple's feature values until reaching a leaf node with a class. By travelling all tree branches a set/list of prediction rules can also be obtained [22].

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

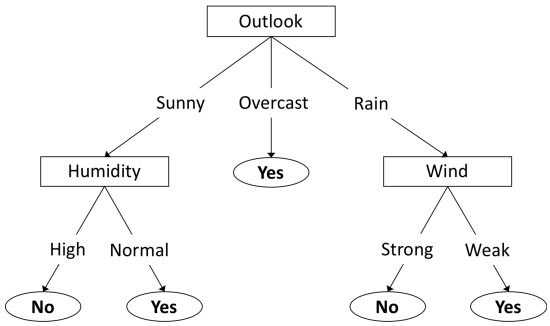


Figure 2.11: Example of a Decision Tree built from the PlayTennis example dataset (adapted from [46]).

DTs are suitable for Binary and Multiclass classification, particularly when the number of data features is not too large, since they tend to overfit [54]. To minimize this, strategies like **pruning** to remove less informative branches can be applied. Moreover, the DTs construction strategy implicitly selects the most important data features, since the features which better split the data space appear higher on the tree (or in the tree at all) [22].

2.3.1.2 Model Selection and Evaluation

Classifier Performance Evaluation Metrics

The construction of a Classification model tends to be an iterative process, hence justifying the need of the Test phase. Without proper validation the model might not be able to correctly predict the class label of new observations, rendering it less useful than it could be. In a Classification problem there are two types of tuples: **Positive tuples**, the tuples from the main class of interest; and **Negative tuples**, the tuples from the remanining classes [22]. Based on this, it is possible to further compute the intermediate components of evaluation measures:

- **True Positives** (TP): number of Positive tuples correctly labeled;

- **True Negatives** (TN): number of Negative tuples correctly labeled;
- **False Positives** (FP): number of Negative tuples incorrectly labeled (**type-I errors**);
- **False Negatives** (FN): number of Positive tuples incorrectly labeled (**type-II errors**) [22].

These four terms are usually condensed in a tabular form, called a **Confusion Matrix** (Table 2.3),

		Predicted class		
		Positive	Negative	Total
Actual class	Positive	TP	FN	P
	Negative	FP	TN	N
	Total	P'	N'	$P + N$

Table 2.3: Confusion Matrix.

where the actual Positive and Negative tuples are identified as P and N, respectively. In similar fashion, P' and N' identify the tuples predicted as Positive and Negative by the model. Taking this into consideration, Table 2.4 characterizes some of the most usual evaluation metrics.

Metric	Formula	Description
Accuracy, Recognition Rate	$\frac{TP+TN}{P+N}$	Percentage of Test set tuples correctly classified
Error Rate, Misclassification Rate	$\frac{FP+FN}{P+N}$	Percentage of Test set tuples incorrectly classified (1 - Accuracy)
Recall, Sensitivity, True Positive Rate (TPR)	$\frac{TP}{P}$	Proportion of Positive tuples that are correctly classified
Specificity, True Negative Rate (TNR)	$\frac{TN}{N}$	Proportion of Negative tuples that are correctly classified
False Negative Rate (FNR), Miss Rate	$\frac{FN}{P}$	Proportion of Positive tuples that are incorrectly classified (1 - Recall)
False Positive Rate (FPR), False Alarm Rate	$\frac{FP}{N}$	Proportion of Negative tuples that are incorrectly classified (1 - Specificity)
Precision	$\frac{TP}{P'}$	Percentage of tuples correctly classified as Positive

Metric	Formula	Description
F-measure, F, F_1, F-score	$\frac{2 \times precision \times recall}{precision + recall}$	Harmonic mean of Precision and Recall
Matthews Correlation Coefficient (MCC)	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$	Correlation coefficient between the observed and predicted classifications

Table 2.4: Evaluation Metrics for Binary Classification (adapted from [22]).

All metrics in Table 2.4 can be extended and applied to the Multiclass problem. The metric most frequently used to evaluate the performance of a Classifier is Accuracy, which poses no problem if the dataset is **Class balanced** (equal proportions for all classes, or at least approximately). However, if a class imbalance exists, Accuracy tends to provide fallacious evaluations due to the **Accuracy Paradox**: for example, if a given class A is dominant in the dataset, composing 99% of the tuples, then predicting that every tuple is of class A will have an Accuracy of 99%. In this case, other measures should also be used to obtain reliable estimates (e.g. Precision, Recall or F-measure) [22].

Receiver Operating Characteristic (ROC) curves can be used to visually compare Classifiers by displaying the trade-off between the rate at which a model can classify correctly Positive tuples (TPR) versus the rate at which it misclassifies Negative tuples (FPR) for different portions of the Test set.

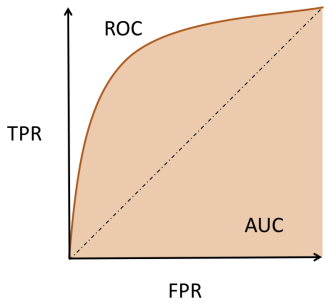


Figure 2.12: Receiver Operating Characteristic curve and Area Under Curve.

As it can be seen in Figure 2.12, any increase in TPR (meaning the classifier is more performant) implicates an increase in FPR (which is not desirable). From this curve it is possible to derive an estimate of the model's Accuracy: the **Area Under Curve** (AUC). In this regard, the diagonal line in Figure 2.12 indicates the point where the model is just performing as badly as random guessing (AUC = 0.5). Thus, the farthest the ROC curve is from this line, the more accurate the model is (AUC = 1) [22].

Generalizations of the ROC curve for Multiclass problems exist, using pairwise comparisons by extending the Binary problem described above [36]. However, to plot them it is necessary to draw one for each class, which in turn may prove confusing in conjunction with some dataset partitioning techniques

(e.g. with k-fold CV the ROC curve for only one class is calculated by finding the mean curve between all folds) [4]. Nonetheless, when the ROC curve is not feasible, the derived AUC value is still a valid metric for model comparison.

Finally, when generically comparing Classifiers, other aspects besides metrics are also relevant:

- **Speed**: the computational costs from generating and using the given Classifier;
- **Robustness**: the Classifier's capability to still make correct predictions when given data with noise or missing values;
- **Scalability**: the ability to construct a Classifier from large amounts of data in an efficient manner (e.g. most models tend to scale poorly with the number of features, taking a lot of time to train);
- **Interpretability**: the amount of insight provided by the Classifier's results; this aspect is subjective and hard to compare, but some Classifiers are more intuitive and easier to interpret than others (e.g. Decision Trees) [22].

Data-related Aspects

Additionally, to avoid misleading estimates due to **overfitting** (where the model learns too well the idiosyncrasies of the training data and is not able to generalize) or **underfitting** (where the model is unable to either learn the training data or to classify new data correctly) of the model to the data it is necessary to have additional considerations:

- the dataset must be **representative** of the reality, meaning that it should include observations from all classes the model is expected to predict;
- the data used to train the model must be **class balanced** as much as possible to avoid classification biases (using over-sampling or under-sampling techniques);
- the dataset should be split with **stratification** (sampling to include tuples in the same class proportions as found in the dataset) to ascertain that the model is trained and evaluated properly for all present classes [14].

Dataset Partition Methods

Different partitioning methods can be used to create the Training and Test sets. The simplest of them is the **Holdout** method, where the data is randomly split into those two sets. Frequently used proportions for the Training and Test sets are, respectively: 2/3 and 1/3, 80% and 20%, or a division in between [22]. It is possible to be used with stratification [4].

Random subsampling is a extension of the holdout method, where it is repeated k times. The overall accuracy estimate is calculated as the average of the accuracies obtained from each iteration. Other possible technique is **Bootstrapping**, where the Training tuples are sampled from the dataset with replacement, having equal chance of being selected again and re-added to the Training set [22].

Lastly, in k-**fold Cross-Validation**, the original dataset is randomly partitioned into k mutually exclusive subsets or "folds", each of approximately equal size. An example with 5 folds can be seen below in Figure 2.13.

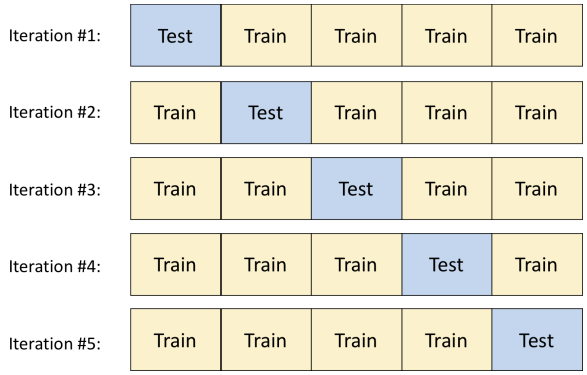


Figure 2.13: Division of an example dataset for each iteration in 5-fold Cross-Validation.

In k-fold Cross-Validation the process of training and testing is performed k times, once per iteration. In iteration i, partition D_i is used as the Test set, and the remaining partitions are collectively used to train the model. The final evaluation metrics are calculated by performing the mean of each metric between all folds. For example, the final accuracy is calculated from the overall number of correct classifications from the k iterations, divided by the total number of tuples in the initial data.

This method distinguishes itself from the Holdout and Random subsampling methods because each sample is used the same number of times for training and once for testing. This method helps to prevent overfitting in the case where Holdout-based methods might have better evaluation scores due to a "lucky" division/sampling of the data [22]. Variants of the Cross-Validation (CV) method also exist, such as:

- **Leave-one-out** (LOO): When k is set to the number of tuples in the original dataset, leaving only one tuple per iteration on the Test set;
- **Stratified** Cross-Validation: the stratification guarantees that the class proportions of the tuples in each fold is approximate to those in the original data;
- **Repeated** Cross-Validation: the partitioning of the dataset in folds is done n times, with the dataset being shuffled before each repetition [4].

Empirically, stratified 10-fold Cross-validation is recommended to estimate accuracy due to its low bias and variance. Repetition can also be added to obtain more robust model evaluation metrics. Nonetheless, the number of folds might be reduced if not enough examples of each class end up being present in each fold to allow a correct training/validation of the model [22].

2.3.1.3 Improving Classification Accuracy

Ensemble Methods

An **Ensemble** method is a Classification algorithm made up of a combination of individual classifiers $(C_1, C_2, \ldots, C_k)$ trained from subsets of the original dataset D $(D_1, D_2, \ldots, D_k)$ that given a new data tuple all classifiers vote to return a class label prediction based on the combination of votes:

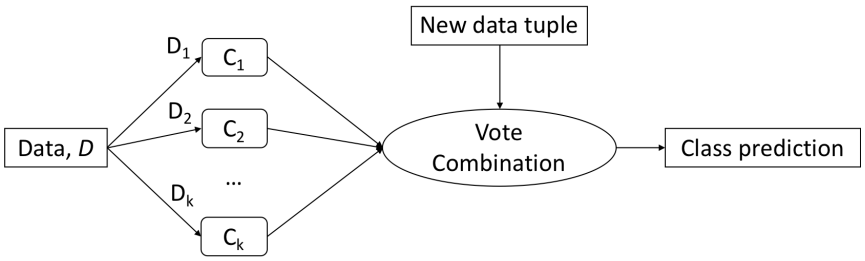


Figure 2.14: Inner workings of an Ensemble method (adapted from [22]).

Ensembles are popular choices since they tend to be more accurate than their component classifiers, mainly when there is significant diversity between the individual components. This happens because different types of classifiers have distinct strengths and weaknesses, and in an ensemble they can compensate for each other: even if the base classifiers make mistakes, the ensemble only misclassifies a tuple if more than half classify it erroneously [22].

A thorough survey on Ensemble methods applied to Classification is available in [58]. Nonetheless, some of those methods are worth mentioning here. One of them is **Boosting**, where weights are assigned to each Training tuple and k classifiers are trained one at a time: after training one classifier C_i, the Training tuples weights are updated so the next classifier, C_{i+1}, will focus on classifying correctly the previously misclassified tuples. In the end, the vote combination is also weighted, with the votes of the most accurate classifiers counting more.

Finally we have **Bagging**, or **bootstrap aggregation**, which works just like what was seen in Figure 2.14, with bootstrapping being used to create the Training datasets for the individual classifiers and majority voting as the vote combination. A known example of Bagging are the **Random Forests** (RF), composed only of Decision Trees generated by using a random subset of features at each node to determine the split. Moreover, they are very robust to errors and outliers, able to compensate for the overfitting individual DTs tend to suffer (as long as the number of trees is large) and are capable of returning internal estimates of Feature Importance [22]. More details on Feature Importance metrics can be found below in *Section 2.3.1.4 Feature Importance*. In this work we use Random Forests to improve both predictability and explainability.

Class Imbalance

Class imbalance exists when a dataset has very different proportions of tuples from different classes, more concretely when the main class of interest (Positive) is rare, either for Binary or Multiclass Classification. This implies that a Classifier will have different error rates per class [22].

Traditional Classification algorithms assume that the cost of a False Positive is the same than of a False Negative. However, depending on the Classifier's main task, this may be a dangerous assumption. For example, in medical diagnosis, although still undesirable, it is preferable to have False Positives than False Negatives, since the latter may imply lack of further clinical investigation, necessary treatments or preventive measures for the patient. Thus, this behaviour has to be kept in check when dealing with class imbalanced data [22].

Several approaches can be used to deal with this matter:

- **Over-sampling** resamples the Positive tuples (the rare class) so that the resulting Training set contains an equal number of Positive and Negative tuples;
- **Under-sampling** randomly eliminates Negative tuples (the majority class) until there are an equal number of Positive and Negative tuples [22].

Ensemble methods can also be combined with sampling to help diminish this issue. For data with two classes, Over-sampling and Under-sampling are effective. However, according to the literature the presented methods are not very adequate to deal with Multiclass imbalance, which is a problem currently being worked upon [22, 59, 71].

2.3.1.4 Feature Importance

An essential part of the biomarker discovery task is to understand how the predictive features have an influence on the variable of interest (the target class), after training a good enough Classification model from the data. Simply put, it is necessary to ascertain the **Features' Importance** in the Classification to gain knowledge about the underlying biological processes. Such a metric provides a score that indicates how useful or valuable each feature was in the model's construction. The more a feature is used to better split the data space, the higher its relative importance will be [1, 43].

However, due to their inner workings, it is easier to perceive the importance of a single feature in some Classifier models than others. For example, Support Vector Machines (SVMs) perform mathematical transformations of the data to be able to separate the classes with an hyperplane, so it might be difficult to interpret which features matter most. Thus, to gain interpretability, simpler (linear) methods tend to be used more often, sometimes at the cost of missing complex dependencies on the data [1].

A good middle ground can be to use tree-based algorithms like Random Forests, which can find interpretable non-linear prediction rules while still getting superior performance. Diverse Feature Importance metrics have been proposed for the RF Classifier, such as:

- the naïve **Variable Importance**, calculated by counting the number of times each variable is selected by all the trees in the forest [62];

- the **Gini Importance**, sometimes simply called Feature Importance [4], is a weighted mean of the individual trees' improvement in the splitting criterion (the Gini Index) produced by each feature [2, 43, 62];
- the **Permutation Importance**, where the values of each feature are permuted to remove its association with the target class; in this way each feature's importance is calculated as the difference in the prediction accuracy before and after the permutation of its values [2, 62, 1, 43].

The first two metrics described above have a known bias towards categorical features with more categories or grouped values. Therefore, depending on the dataset it may be necessary to avoid this bias, in which case the Permutation Importance metric should be used [1].

With Permutation Importance the values of each feature are permuted by using random shuffling and different random seeds, implying a degree of variability in the importance values between sequential runs. The final ranking values are usually obtained by performing the mean of the values obtained for each feature in all the runs. Zero or negative values of this metric imply that the feature is not important for the classification or is even harming the model's performance, respectively. Finally, it is of note that these Importance values do not sum up to one between all features, because they are not normalized. The main usefulness of this metric is to consider the feature's Importance relatively to each other instead of their absolute values [55].

2.3.2 Discriminative Biclustering

When dataset observations are class-labelled, Biclustering can still be applied with additional discriminative criteria in order to distinguish classes and support real-world decisions. This means that the Biclustering task tries to discover class-discriminative Biclusters, where a particular class has a significantly high support in each Bicluster [25].

If the found Biclusters are class-discriminative, their patterns are **discriminative patterns** with important applications in finding biomarkers in medical data [61]. Thus, for the types of Biclustering models specified in *Section 2.2.5.2 Meta-features* those patterns are Meta-features, allowing us to obtain class-discriminative subsets of features (and their respective values).

2.3.2.1 Supervised Measures of Bicluster Validity

When performing a Clustering task over class-labeled data, the class label is not taken into consideration by the algorithm. Although this might seem a redundant thing to do, it is useful to validate the performance of the said Clustering algorithm over the data, since the ground truth is available. Thus, it is possible to use metrics normally used in Classification problems, which will quantify the degree of objects of the same class in a Cluster [63].

First of all, it is necessary to infer the data's class distribution over all Clusters using p_{ij}, the probability that an element of cluster i belonging to the class j. The said probabilities should be calculated for all

clusters and classes, as $p_{ij} = m_{ij}/m_i$, where m_{ij} is the number of elements of the class j in cluster i and m_i is the total number of elements in cluster i [63]. After that, the following metrics can be employed:

- **Entropy**, the extent to which each Cluster i includes elements of one class: $e_i = \sum_{j=1}^{C} p_{ij} \log_2 p_{ij}$, where C is the number of classes; for a set of Clusters the total Entropy can be seen as the sum of the Entropies weighted by the Cluster size: $e = \sum_{i=1}^{K} \frac{m_i}{m} e_i$, where K is the number of Clusters and m the total number of data elements;
- **Purity**, another way to calculate the extent to which each Cluster i includes elements of a single class, given by $purity(i) = \max_j p_{ij}$; for a set of Clusters the total Purity can be seen as $purity = \sum_{i=1}^{K} \frac{m_i}{m} purity(i)$, where K is the number of Clusters;
- **Precision**, the fraction of elements of the class j in Cluster i: $precision(i, j) = p_{ij}$;
- **Recall**, the degree to which all elements of the class j are included in Cluster i: $recall(i, j) = m_{ij}/m_j$, where m_j is the number of elements of class j;
- **F-Measure**, the harmonic mean of Precision and Recall to measure the degree to which a Cluster i includes only elements of a given class j: $F(i, j) = (2 \times precision(i, j) \times recall(i, j))/(precision(i, j) + recall(i, j))$.

All metrics mentioned above range between 0 and 1 except Entropy, which ranges from 0 to $\log_2 C$. Additionally, although Entropy and Purity have the same objective, they measure it differently: when all Cluster elements are from the same class it means the Entropy value is 0, but the Purity value is 1 [63].

Biclustering algorithms are also unsupervised, but their results can further evaluated by using the ground truth provided by class labels. In this context, the metrics above can also be applied to find class-discriminative Biclusters, by evaluating the class labels from the dataset rows present in a Bicluster as a Cluster of one-dimensional points. For example, a threshold on Purity levels can be set to filter out all Biclusters with less than a given Purity percentage. However, Purity alone does not allow to discern which is the most present class in a given Bicluster, so, in order to obtain this knowledge, the Precision for each class has to be considered as well. An example is given in Figure 2.15, where we have two discriminative Biclusters (B_1 and B_3) and one that is not discriminative, B_2, since it has 50% of elements of each class.

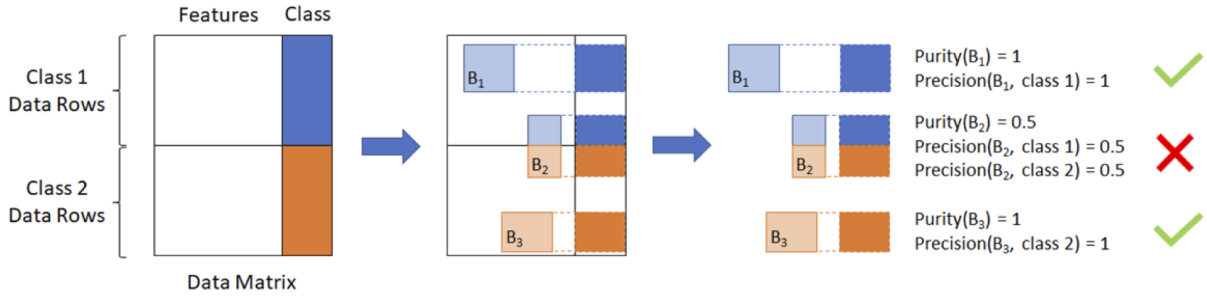
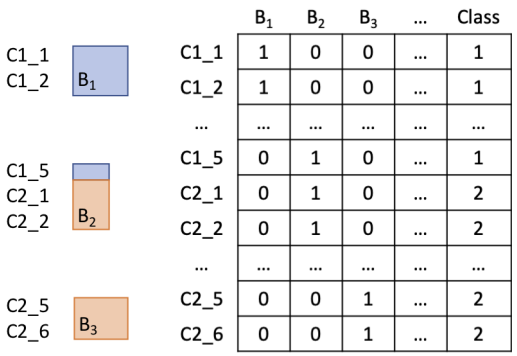


Figure 2.15: Example of class-discriminative Biclusters' mining.

2.3.3 Biclustering-Based Classification

In Biclustering-based Classification, a set of found Biclusters is used as class-discriminative features, taking advantage of the subset of features' space of the pattern from each Bicluster. This is done to improve on regular Classification, which considers only individual feature space [6, 5].

Many variants on how to use this approach are found in the literature. One possible way is to build a matrix where the Biclusters are considered as features (columns) and the remaining dimension (rows) is composed of the observations found in the original data (e.g. Patients or Controls). Regarding the matrix values, if an observation was included in a given Bicluster, that cell is filled with a 1; otherwise, with a 0. In this way it is possible to easily interpret which are the most discriminative subsets of features and objects considered by the Classification model [5]. An example can be seen in the Figure 2.16 (whose Biclusters derive from the example in Figure 2.15).



	B_1	B_2	B_3	...	Class
C1_1	1	0	0	...	1
C1_2	1	0	0	...	1
...	...	...	...	...	...
C1_5	0	1	0	...	1
C2_1	0	1	0	...	2
C2_2	0	1	0	...	2
...	...	...	...	...	...
C2_5	0	0	1	...	2
C2_6	0	0	1	...	2

Figure 2.16: Example of Subject $\times$ Biclusters Matrix for Biclustering-based Classification.

In this example the identifiers of each observation (row) are given by the letter "C" before the class label, followed by an underscore character and an integer number (e.g. the second observation from class 1 is called C1_2). Assuming that only those three Biclusters existed, the cells filled with ellipsis (except for the class column) would be filled with zeros.

2.4 Association Rule Mining

When working with categorical data another useful type of patterns that can be obtained are Association Rules. To follow on this, it might be useful to review what was defined in *Section 2.2.4 Pattern Mining*: in this section we shall consider L as a finite set of items, P as an itemset where $P \subseteq L$ and an itemset database D over L composed by a finite set of transactions (in form of pairs (t_{id}, P) with $id \in \mathbb{N}$) $\{t_1, \ldots, t_n\}$. An **Association Rule** is an implication with the form $P \rightarrow P'$, where $P \subseteq L$, $P' \subseteq L$, and $P \cap P' = \emptyset$. The left side of the rule is called the **antecedent**, while the right side is the **consequent**.

2.4.1 Rule Interest Metrics

A rule can be, at its core, characterized by two metrics of interest: its **Support**, $sup_{P \to P'}$, or the frequency with which both sides of the rule appear in a transaction, given by $sup(P \cup P')$; and the **Confidence**, $conf_{P \to P'}$, which is the strength of the rule, a notion that a transaction containing the items from the antecedent will also contain the ones from the consequent, given by $\frac{sup(P \cup P')}{sup(P)}$ [25].

The first step to find these rules is to perform FIM over a discretized dataset to obtain the frequent itemset database D. Then, considering a user-defined minimum Confidence threshold δ, the **Association Rule Mining** (ARM) algorithm can compute $\{(P, P') \mid P \subseteq L, P' \subseteq L, conf_{P \to P'} \geq \delta\}$, returning the frequent rules whose confidence is equal or greater than the user-specified threshold [25].

The space of all possible rules obtained from a database D can be massive, and will need filtering to find the most interesting ones. This can be done by using measures of interest such as **Lift**, a measure of the correlation/dependency between both sides of the rule, given by $\frac{sup(P \cup P')}{sup(P) \times sup(P')}$. Depending on its value, the Lift of a rule $P \to P'$ has different meanings:

- Lift < 1: P is **negatively correlated** with P', meaning that P' is unlikely to occur in a transaction if P does;
- Lift = 1: P and P' are **not correlated**, meaning their occurrences are independent;
- Lift > 1: P is **positively correlated** with P', meaning that P' is likely to occur in a transaction if P does [22].

2.4.2 Filtering Uninteresting Association Rules

Depending on the itemset database and the thresholds used to limit the Association Rule Mining, the obtained amount of rules may be too great, making their interpretation hard or even unfeasible. Many ARM algorithms tend to only use Support and Confidence on their rule filtering.

However, even with a certain degree of subjectivity when defining what is an uninteresting rule, using only these two metrics is usually not enough to exclude many of them, particularly if the mining task involves low support thresholds or looking for long patterns. Additionally, a high Confidence metric can be sometimes deceiving, since it does not take into consideration the popularity of the items on both sides of the rule, but only that of the antecedent. To compensate for it, the Lift (or other correlation-related metric) should be added to the filtering, if possible.

A feasible approach to limit the number of generated association rules beforehand is to create the item database D from the closed frequent itemsets instead of the frequent itemsets, reducing the number of generated patterns while still completely preserving the information within them [22].

Moreover, one can also remove the rules deemed **redundant** after the mining. Redundant rules are called as such since they do not convey additional information in the presence of other broader rules. More concretely, if a rule B is a super rule of another rule A ($A \subseteq B$) and B has an equal or lower Lift value, then B is redundant [70, 21].

2.4.3 Associative Classification

Associative Classification is a Data Mining approach where Association Rules are generated from frequent patterns and used for Classification, in order to find strong associations between the items in the antecedent (feature-value pairs which compose the patterns) and the class labels [22].

Consider a tuple database T with n features, $F_1, F_2, \ldots, F_n$, a class label feature, F_{class} and where all features are categorical (by default or discretized). An item p_i is a feature-value pair of the form (F_i, v_i), where F_i is a feature which has the value v_i [22]. The rules used in Associative Classification are called **Class Association Rules** (CARs), where the $l \leq n$ items found in the antecedent implicate (and thus, are associated with) the class label C found in the consequent:

$$(F_1 = v_1) \wedge (F_2 = v_2) \wedge \cdots \wedge (F_l = v_l) \Rightarrow (F_{class} = C). \tag{2.1}$$

Then, after finding the CARs from the frequent itemsets/patterns (satisfying both user-defined levels of support and confidence), these can be organized to create a rule-based Classifier [22]. However, if the main purpose is not to create a model to predict the class of new surging tuples but instead to find the most class-discriminative sets of items (the ones most correlated with every possible class label), then the Lift measure of these CARs should also be calculated: the higher the Lift, the more discriminative the rule is for the given class.

2.5 Dimensionality Reduction

Given a dataset with a large number of features (hundreds or more) and a class label, the existence of irrelevant or redundant features regarding that class is highly likely. The increase in dimensionality of the data triggers an effect commonly called the **curse of dimensionality**, where the extra dimensions cause the data to become more and more sparse, effectively disrupting the performance (speed and accuracy-wise) of most Machine Learning and Data Mining algorithms [68]. To prevent such effect from happening, Feature Selection (FS) techniques should be applied to the data before model training.

Three different kinds of FS strategies are available:

- **Filter** methods, where individual features or subsets are evaluated in an independent way of any learning algorithm;
- **Wrapper** methods, which determine the best subset of features (between all combinations) by comparing the obtained evaluation metrics of a learning algorithm using each subset (with k-fold CV);
- **Embedded** methods, that try to combine the interactions with a learning algorithm (akin to wrapper methods) with the efficiency of the filter methods (by not iterating through all possible subsets of features).

Filter methods are more computationally efficient than wrapper methods, especially when the number of dataset features is large. However, filter methods are less accurate than wrapper methods, since the lack of a learning algorithm to guide the selection process can lead to the non-optimality of the selected features for the learning tasks. Embedded methods emerged to combine the best of both [39].

2.5.1 Dimensionality Reduction for Categorical Data

In this thesis, the discretization of all dataset features was necessary in order to apply Pattern Mining-based Biclustering to the data. Additionally, to run it in a timely efficient manner, FS had to be performed to choose the best features for Classification from the dataset. Since Pattern Mining-based Biclustering is able to find subsets of rows which show a coherent pattern observed for a subsets of columns (features) in form of Biclusters, it was decided that features that were not very relevant by themselves should be removed. Thus, on the subsequent sections several filter-based FS techniques applicable for categorical data will be reviewed.

2.5.1.1 Clinical Expert Knowledge

According to Heinze et al. in [23], it is of major importance to use background knowledge to guide FS. Therefore, when working with EHR data interaction with the clinical experts must be highly regarded to:

- discern which features might be (or not) important for a given research question or analysis;
- identify possible **confounding** features, where hidden associations between individual features (that ideally should be independent among them, but not from the class) may exist;
- help pinpoint any other care concerning the data.

2.5.1.2 Missing Values

Missing values can occur in data for many reasons. As such, several types of missing values exist:

- **Missing Completely at Random** (MCAR), when the probability of a missing value is independent of the feature itself and any external influences (e.g. human input error, lost medical sample);
- **Missing at Random** (MAR), when the probability of a missing value is still independent of the feature itself but not of external influences, with the missing values having a predictable pattern (e.g. sensor fail);
- **Not Missing at Random** (NMAR), when the probability of a missing value is dependent of the feature itself (e.g. if a patient with neurodegenerative symptoms was not able to complete a given test, the medical practitioner may not subject him/her to other harder tests) [35].

When dealing with missing values several types of approaches can be used, such as imputation, omission and analysis.

Imputation

A common approach for the treatment of missing values is **data imputation**. Since measures of central tendency (mean/median) cannot be used with categorical data, a new category (e.g. ”unknown”) or the most frequent value (mode) can be used to fill in the gaps. However, value imputation always adds some degree of bias to the data [22].

Omission

As stated before, missing values are common in medical data, being mostly of the NMAR type. If the majority of the data (according to a user-defined threshold) in an observation or feature is missing making imputation unfeasible, another option is to remove them entirely. Although it implies loss of potentially useful information, observations or features with a great majority of missing values may end up being a direct source of noise for models [35].

Analysis

Finally, it is also possible that depending on the task at hand, some algorithms are able to handle the existence of missing data, thus avoiding the necessity of their preemptive treatment [27].

2.5.1.3 Collinear Features

According to Yu et al. [67], in Classification problems the best features are relevant to the class concept but not redundant to any of the other relevant features. Redundant features do not convey any new information due to high correlation (**collinearity**), tending to introduce noise and overfitting in models.

To calculate the correlation between categorical features, the Pearson’s Chi-Square test with a 2×2 contingency table is used. Then the Chi-Square statistic’s value is normalized to a range [0,1] using the Cramér’s V measure of association, where 0 represents no correlation and 1 a strong correlation [56]. Missing values can be treated as another possible feature value [16]. Usually when features have a correlation above a certain high threshold percentage, one of them is removed.

2.5.1.4 Homogeneous Features

When dealing with real-valued data, low variance features are considered poor in information, making them not very useful to discriminate between class-labeled data. For categorical data it is not possible to use the variance due to the features’ types. However, to check for features with very homogeneous values, measures that deal with class distributions in groups (e.g. Entropy seen in *Section 2.3.2.1 Supervised Measures of Bicluster Validity*) can be used [67].

2.5.1.5 Independence of Target Class

According to Jin et al. [30], by calculating the Chi-Square statistic between each feature and the class label, it is possible to discover the features most associated with the class (the ones with the highest Chi-Square statistic values). This means that lower values of the statistic will be present for the features less associated (more independent) with the class label. With this in mind, it is possible to create a FS filter.

2.6 Related Work

The DMD approach used in this thesis takes advantage of well-known techniques of ML and DM. However, the combination of those, to the best of our knowledge, is unique. Thus, in this section, the related work to each of the individual areas the said approach focuses upon will be described.

2.6.1 Biclustering in Healthcare Records

According to Nezhad et al. [47], precision medicine's goal is to develop tailored treatment schemes for different patient subgroups, acknowledging that **for certain diseases some risk factors may be more significant than others for specific subgroups**. Thus, they proposed SUBIC (Supervised Biclustering), a new method which uses convex optimization to detect and prioritize risk factors. Here the use case disease was hypertension, in a dataset composed of people from a vulnerable demographic subgroup (African-American), where the class labels for each patient were determined with help of clinicians.

2.6.2 Pattern Mining-Based Biclustering Algorithms

A thorough description of the state-of-the-art Pattern Mining-based Biclustering algorithms for non-temporal data was done by Henriques et al. [25]. In particular, one of the algorithms in comparison - **BicPAM** - is the result of the integration of dispersed contributions on Pattern Mining-based Biclustering with novelty methods to deal with more flexible expression profiles and varying levels of missing values or noise [25, 27]. BicPAM was deemed superior to the other approaches in several aspects because it can:

- Work with real-valued and categorical data;
- Discover all the maximal Biclusters in an efficient manner, while validating their homogeneity;
- Guarantee competitive computational complexity even when dealing with noisy data and adaptations to the Biclustering solution are needed;
- Handle medium-to-high levels of missing values and noise;
- Unravel biologically relevant solutions [25, 27].

Additionally, BicPAM is available to use freely through the **BicPAMS** (Biclustering based on PAttern Mining Software) tool [26], along with other state-of-the-art Pattern Mining algorithms, in two different utility modes: a Graphical User Interface (GUI) for explorative analysis and an Application Programming Interface (API) mode for further integration with the user's Java code. Supported input file formats are *TXT* and *ARFF* (Attribute-Relation File Format [18]). Other algorithms able to search for Biclusters in categorical data can be found in [48], [65] and [9]. In this thesis we used BicPAM for the reasons stated above.

2.6.3 Previous Uses of the ONWebDUALS data

The ONWebDUALS dataset has been used previously in publications related with the NEUROCLINOMICS2 project and members of the ENCALS consortium have been producing scientific publications based on its data to withdraw conclusions about the disease:

- One such example can be found in [34], where potential environmental factors were investigated in order to find a cause for the great majority of sALS cases (the portion left unexplained by genetic factors);
- Another article regarding the temporal study of anatomical region progression of the disease is being worked upon (not yet published) [20].

However, none of the work done so far with this dataset targeted the goals in this thesis or used Biclustering techniques.

2.6.4 Progression Groups in ALS

In [50] and [51] Pires et al. proposed a Supervised Learning approach to predict, in reasonable time intervals, when a patient would start needing NIV respiratory assistance, using patient stratification and a new approach based on progression groups (relative to the speed of disease progression: Slow, Neutral or Fast). In this thesis, for Task 2, the class for each Patient was obtained by another NEUROCLINOMICS2 project member using an Expectation-Maximization (EM [11]) approach that followed the work proposed in [51]. This approach was used to classify them as belonging to one of the same three progression groups according to the values of their ALSFRS-R scale decay rate in the ONWebDUALS dataset.

Nonetheless, other possible patient stratifications exist, like the one proposed by Kueffner et al. in [33], which combined the efforts of 30 different investigation teams to cluster the patients in four categories: Early Stage, Slow Progressing, Fast Progressing and Late Stage.

2.6.5 Biclustering-Based Classification in Healthcare

Inspiring examples of Biclustering-based Classification are given by Carreiro et al. [5, 6], for time series data. The ability to deal with time points was a characteristic of the used Biclustering algorithm. A similar approach is used in this work for non-temporal data.

In [6], the Biclusters are class-discriminant features used to identify subsets of genes coherently expressed over a subset of contiguous time-points, allowing to pinpoint potentially relevant biological processes related the progression of Multiple Sclerosis patients and their response to the standard treatment with Interferon-β.

Then in [5], following the aforementioned work, a Classification approach based on sets of similar Biclusters (Meta-biclusters) obtained by Hierarchical Clustering, after cutting a dendrogram at a given level is used. After that step, a binary matrix with Meta-biclusters as columns, patients as rows and a final column for the patient class is constructed, to represent which Meta-biclusters contain Biclusters from that patient (1 for Yes, 0 for No).

2.6.6 Associative Classification in Healthcare

Lavrač et al. [37] present a case study using Associative Classification for subgroup discovery. This technique was employed to detect and describe risk groups of Coronary Heart Disease (CHD), where endangered individuals could show only slightly abnormal values of risk factors or combinations of different risk factors.

Chapter 3

Discriminative Meta-features Discovery

In this chapter a new Data Mining methodology called **Discriminative Meta-features Discovery (DMD)** is proposed to find disease presentation patterns in two-dimensional EHR data. Furthermore, its implementation done for this thesis is thoroughly described.

The main objective was to design an exploratory approach that would allow to obtain discriminative, understandable and intuitive descriptions of medical concepts, which in this work have already been introduced as Meta-features. The necessity for **explainable models** in medicine is great since clinicians must be able to understand the reasoning behind a given model result or prediction before making a decision that can impact a patient's life. This type of models are also called "white-box" models, in contrast with "black-box" models (e.g. Artificial Neural Networks) whose internal decisions are very hard to interpret [8].

To accomplish this goal, several Machine Learning and Data Mining techniques were combined. Figure 3.1 shows a simplified workflow of DMD's most important phases. First, Pattern Mining-based Biclustering is run on discretized and class-labelled data to find discriminative data patterns in the form of Biclusters. Afterwards those Biclusters are further used in two distinct branches:

- with Biclustering-based Classification models to understand what features (or subsets of features) are more important for a good classification, by the means of a Feature Importance metric;
- with Class Association Rules in order to find the subsets of features (and respective values) that show greater association with each class label.

These two approaches are used in parallel to take advantage of the explainable characteristics of the underlined models and also to validate the obtained results. Additional flows pointing from the data to both strategies used after Pattern Mining-based Biclustering (dashed arrows) are present to account for baseline tests (with all features and a selected subset of them through FS) defined to understand if the new approaches were being performant.

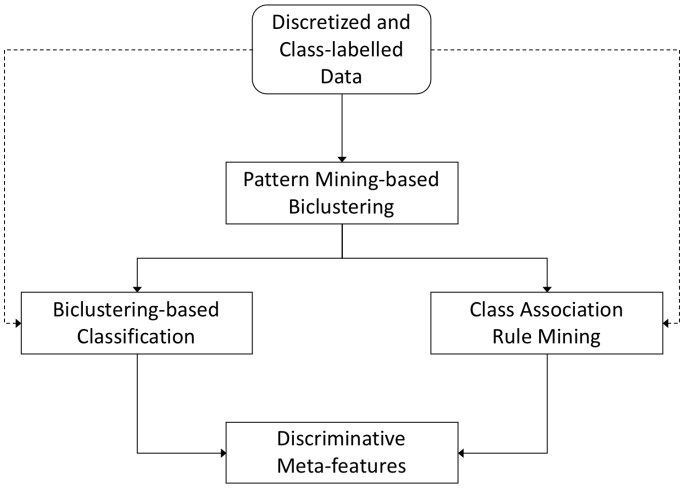


Figure 3.1: Simplified workflow of the proposed DMD approach.

3.1 DMD Implementation

The general definition of each phase seen in the previous section can be implemented in a myriad of different ways and using distinct algorithms. We believe it can be adapted to other types of data (e.g. temporal data) if the chosen Pattern Mining-based Biclustering algorithm is capable of dealing with it.

Figure 3.2 details the workflow of the DMD methodology, including the technologies used to implement each layer and each experiment undertook (highlighted with a distinct alphabet letter) in this thesis. The next subsections further detail the work done in each major phase (light orange blocks on the figure, or light grey if printed in greyscale).

3.1.1 Data Pre-processing

In order to clean, discretize and label the ONWebDUALS dataset for each Task data pre-processing pipelines were created using the **KNIME** tool. KNIME is an open source software written in Java which allows the creation of pipelines that combine the retrieval (from files, databases or the web), exploration, analysis and visualization of scientific data, while also allowing the integration of scripts written in programming languages widely used in Data Science (e.g. R, Python and Java) and abilities from other external tools (e.g. machine learning from WEKA [18]). These pipelines are built by linking nodes with specific tasks to each other, creating an orderly pipeline with flow control, if necessary. Thus, this tool was used due to its integration and easy reproducibility of experiments capabilities [16].

One pipeline per Task was designed to create the datasets as described in *Chapter 1 Introduction*. The Data Pre-processing steps of these pipelines can be found in Figure 3.3.

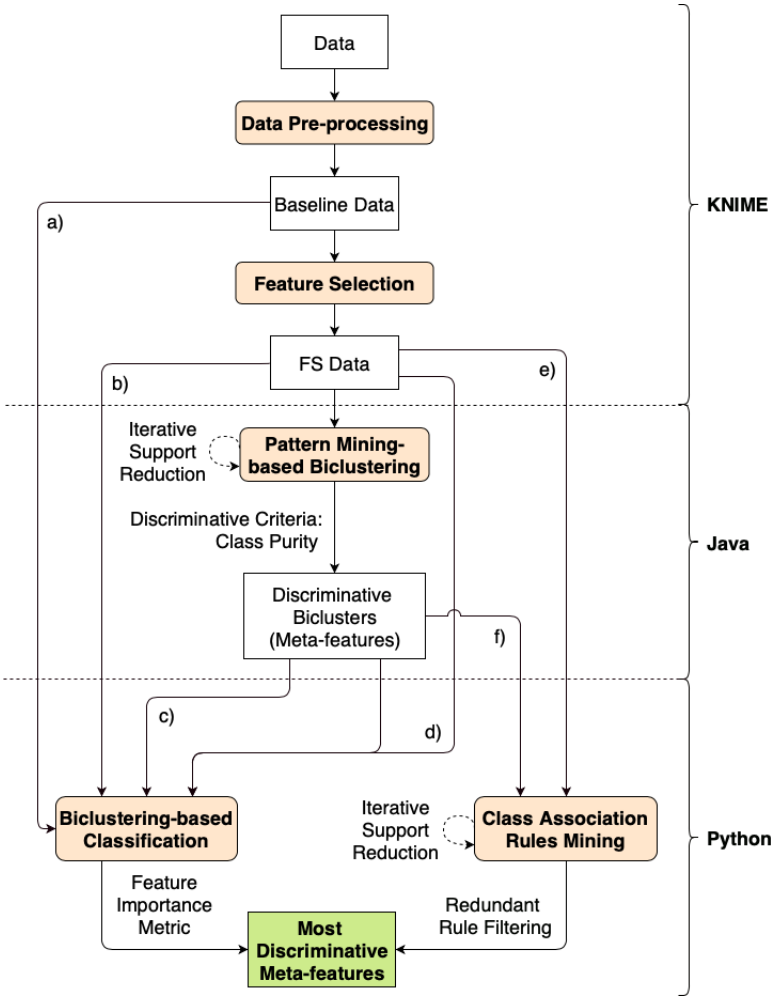


Figure 3.2: Detailed workflow of the proposed DMD approach.

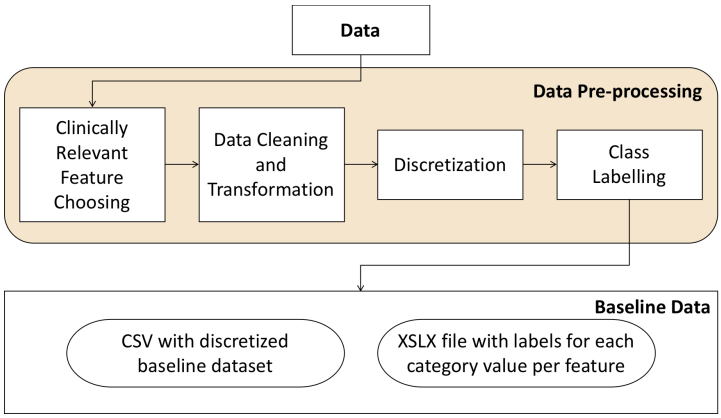


Figure 3.3: Data Pre-processing steps and output.

Clinically Relevant Feature Choosing

All dataset columns were investigated and deemed useful or not for the current investigation with the help of our expert clinicians. It was taken in consideration the clinical usefulness, measure correlation, data type, frequency and clarity of existing values and if it was feasible to discretize the values of the given column. Time-related data (like dates) when relevant were converted to continuous data (e.g. *Date of 1st Symptoms* was converted to *Age (1st Symptoms)*) before interval discretization, since only static data was being considered.

Data Cleaning and Transformation

On the questionnaires made to the patients and controls in each question, three fields were available besides the proper place for a relevant answer: "NA" (Not Applicable), "NR" (Not Relevant) and "NF" (Not Feasible). While the distinction of these answers might be relevant for a clinician, it would generate additional entropy or noise for a Biclustering algorithm. For that reason, those uninformative values were removed and considered as missing values.

Regarding erroneous values, any error that could not be corrected was also converted to missing values. No missing value imputation was deemed necessary because the chosen Biclustering algorithm is able to deal with them while mining the data for Biclusters. For more information on this matter please see *Section 3.1.3 Pattern Mining-based Biclustering*.

Finally, some columns in the dataset file had duplicate names, due to the existence of similar questions on different questionnaire sections that were not properly distinguished. This situation was improved by clarifying those columns' names so their context could be fully understood without the need for additional information.

Discretization

In order to apply Pattern Mining-based Biclustering to the ONWebDUALS dataset it had to be completely discretized. For features that were already categorical, textual integers starting in "1" were attributed as category values, complying with was already done for questions with closed answers (e.g. Yes = "1", No = "2") and to keep things simple [65]. The only exception to this were the European Skills/Competences, qualifications and Occupations (ESCO) codes (with two distinct levels) used to normalize the patients' occupations, which can have category values like "00", "01", etc [15].

Most features with real values were discretized with the help of the clinicians, given their greater domain knowledge. Some of them could be easily discretized in either "Good"/"Bad" values (e.g. *High Density Lipoprotein* (HDL)) due to the existence of threshold information, while others were split in specific time intervals (e.g. *Timing of transition between onset regions*, in blocks of months). Any other real-valued features were discretized according to their quartile distribution (e.g. *tobacco exposure*, in pack-years) using box-plots, including mild and extreme outliers, if any. Specific features regarding ALS screening metrics (e.g. *Sniff Nasal Inspiratory Pressure* (SNIP)) were discretized according to the

intervals defined in their respective literature [41, 31]. The final result of the dataset discretization can be found on *Appendix A Discretized ONWebDUALS Dataset*.

Class Labelling

In both Tasks (defined in *Section 1.3 Objectives and Contributions*) the class feature is called *group*. The added class labels at the end of this Data Pre-processing phase were specific for each Task. For Task 1 the Patients were simply labeled with class "1" and Controls with class "2". For Task 2 the class for each Patient was obtained as aforementioned in *Section 2.6.4 Progression Groups in ALS*. To comply with the labelling in Task 1 the progression groups "Slow", "Neutral" or "Fast" were converted to classes "1", "2" and "3", respectively.

Baseline Data Output

In Figure 3.3 the several outputs of the Data Pre-processing phase are shown:

- A Comma-Separated Values (CSV) file with the discretized baseline dataset, to be used as a Baseline with all features in the Biclustering-based Classification (Figure 3.2, a) Baseline All Features);
- An Excel (XLSX) file, with the category labels for each feature's category values, to translate the final results from the Pattern Mining-based Biclustering and Class Association Rule Mining, whose algorithms employ intermediate encodings.

3.1.2 Feature Selection

This phase was still performed in the KNIME pipeline, right after Data Pre-processing. All the thresholds in this phase were determined empirically. As shown in Figure 3.4, to reduce the dimensionality of the data several FS techniques were used, by removing features:

- With over 70% of missing values;
- With over 70% Correlation with another, in pairwise comparison;
- With very high homogeneity (<10% maximum Entropy for the given features);
- Less associated with the class, according to the Chi-Square statistic.

The Chi-Square step at the end of the FS pipeline allowed to choose a fixed number of features (k = 20), hence picking the k most associated with the class. In case gender and age-related features were not chosen automatically, they were still added to the final feature set due to their clinical relevance.

This phase has several outputs:

- the discretized FS dataset in ARFF format for the Biclustering phase;
- the discretized FS dataset in CSV format to be used as a Baseline with FS in the Biclustering-based Classification (Figure 3.2, b) Baseline FS Features);

- a Tab-Separated Values (TSV) file to map the categorized values of each feature to the indices according to their appearance in the ARFF file.

This last TSV file was needed to translate the mined Biclusters to the original data values, since the used Biclustering algorithm works with the category indices when dealing with nominal data (more details in the next subsection).

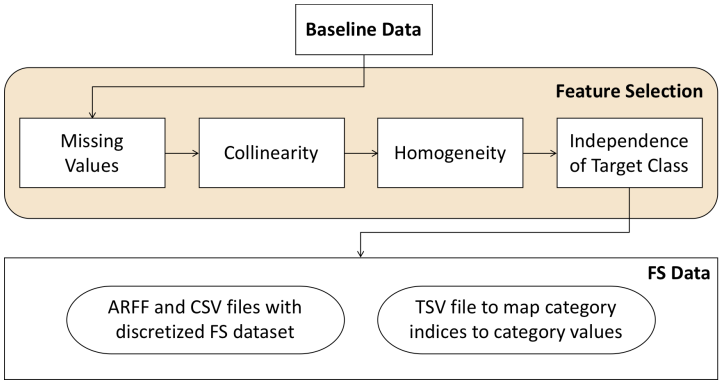


Figure 3.4: Feature Selection steps and output.

3.1.3 Pattern Mining-based Biclustering

The BicPAM algorithm was chosen to carry out the Pattern Mining-based Biclustering due to all the capabilities already seen in *Section 2.6.2 Pattern Mining-Based Biclustering Algorithms*.

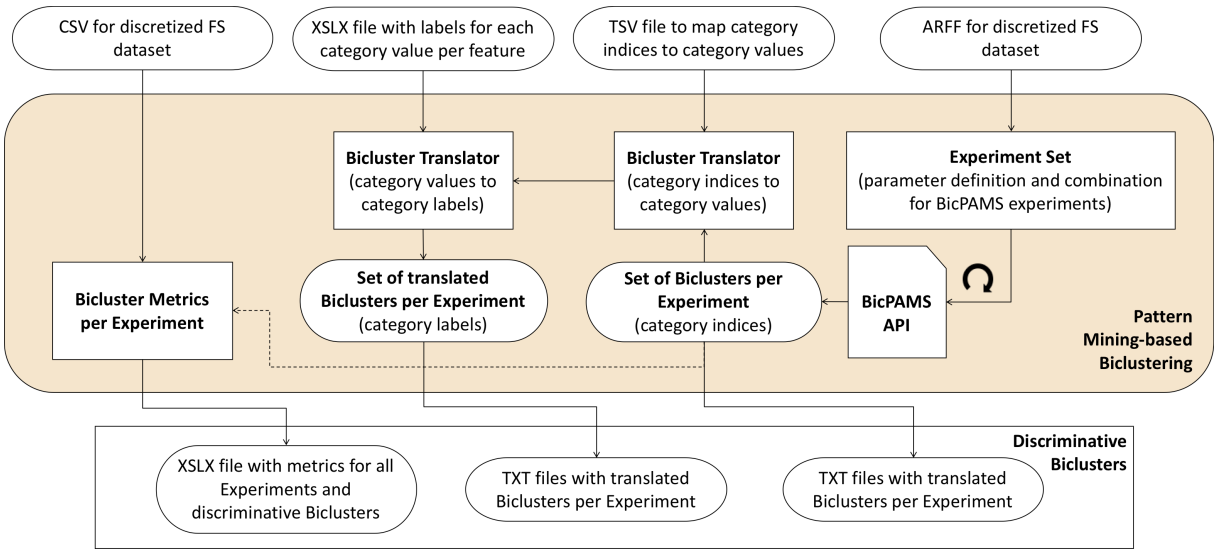


Figure 3.5: Pattern Mining-based Biclustering details.

As depicted in Figure 3.5, most previously computed outputs are used in this phase. The ARFF file with the discretized FS data is passed to Java code acting as a wrapper for BicPAMS API. This code was created to facilitate the definition of batches of experiments (Experiment Sets): for each BicPAM parameter an array of possible value is passed, and an experiment is created for each possible combination of parameter values. In Table 3.1 an explanation of each parameter can be found.

Parameter	**Definition**
Coherency Assumption	Defines the correlation of values within a Bicluster. For Constant models, a pattern is preserved across rows (or columns). Additive, Multiplicative, Order-preserving and Plaid models are also available, considering Symmetries or not.
Coherency Strength	Number of items which determines the allowed deviations from expected values. When working with categorical values, needs to be a value higher than the maximum number of categories in all features.
Quality (%)	Specifies the maximum number of allowed noisy/missing elements. A value of 100% implies that no noise is allowed, meaning the pattern is always the same throughout the Bicluster.
Pattern Representation	To search for maximal Biclusters Closed patterns (default) must be used.
Orientation	Where the pattern should be observed: across rows (default) or columns.
Normalization	Option that allows to normalize data per Row, Column or for the Overall data elements. Can be ignored by selecting the None option.
Discretization	Option that allows to discretize data either by selecting cut-off points of a Gaussian distribution (default) or fixed ranges of values. Can be ignored by selecting the None option.
Noise Handler	Multi-item assignments can be considered to handle deviations on the expected values within a Bicluster caused by noise or discretization issues. Can be ignored by selecting the None option.
Missings Handler	Option to specify the treatment of missing values. The Remove option (default) excludes them from the searches and the Replace option uses WEKA's imputation methods to fill them.
Remove Uninformative Elements	Option to remove uninformative data elements. Normally used with Gene Expression data to ignore genes that are not being expressed/repressed. Can be ignored by selecting the None option.
Stopping Criteria	The stopping criteria of the search algorithm. It has three possible values: 1) minimum number of Biclusters before merging (default), 2) minimum covered area by the discovered Biclusters (% of the elements of the input data matrix), and 3) minimum support threshold (% of overall rows per Bicluster).

Parameter	Definition
Stopping Criteria Value (%)	The value associated with the chosen Stopping Criteria.
Minimum Bicluster Columns	Minimum number of columns (features) that need to be present in each Bicluster.
Pattern Miner	Biclustering algorithm used for the Bicluster search. CharmDiffsets (default), AprioriTID and CharmTID are made available for closed pattern representations.
Scalability	Boolean option that applies assertive FS to guarantee the scalability of the searches in large datasets. Can be ignored by selecting the False option.
Merging Procedure	Closing option to merge Biclusters either in a Heuristic or Exhaustive fashion.
Filtering Procedure	Another closing option applied after merging to guarantee compact Biclustering solutions. One Bicluster is filtered if it has not enough Dissimilar Elements, Dissimilar Rows or Dissimilar Columns against a larger Bicluster (basically if the first is included in the second).
Filtering Procedure Value (%)	Percentage value for the Filtering Procedure parameter. For example, a value of 20% here with a filter for Dissimilar Elements guarantees that Biclusters sharing more than 80% of their elements against a larger Bicluster are removed.

Table 3.1: BicPAMS Parameters' Description

Figure 3.5 also shows that a great amount of intermediate translation work was necessary in order to obtain the final Biclusters in a readable form. This was necessary since BicPAMS internally uses WEKA's *Instance* class to create the dataset object, which deals with categorical data efficiently by working with the indexes (starting in 0) of the category values as they appear in the ARFF file, instead of the values themselves. An example of Translation using the *Age (on date of consultation)* feature can be seen in *Listing 3.1*, where the feature had its categories interpreted internally in the following fashion: the category value 6 was a 0, the 8 was a 1, the 9 was a 2, and so on.

```
...
@ATTRIBUTE 'Age (on date of consultation)' {6, 8, 9, 7, 4, 5, 3, 10}
...
```

Listing 3.1: Example of feature declaration on a ARFF file.

Due to this issue, initially the found Biclusters were almost impossible to read (or even seemed to make no sense). To compensate for it, two Translation phases were implemented after the Bicluster

search: a first to translate category indexes to category values, and then another to translate the category values to their labels. On the first Translation phase the TSV file mentioned in *Section 3.1.2 Feature Selection* would then contain for that feature the reverse translation needed to convert from category indexes to category values. The matching example for the *Age (on date of consultation)* feature can be found on *Listing 3.2*.

...
‘‘Age (on date of consultation)’’ ‘‘0 –> 6’’ ‘‘1 –> 8’’ ‘‘2 –> 9’’ ...
...

Listing 3.2: Example of feature translation from category indexes to category values.

Then, on the second Translation phase, the XLSX file indicated in *Section 3.1.1 Data Pre-processing* is used to convert the feature's category values to their respective labels. The matching example for the *Age (on date of consultation)* feature can be seen in *Table 3.2*.

Feature values	**Feature labels**
1	[0, 10[years
2	[10, 20[years
3	[20, 30[years
4	[30, 40[years
5	[40, 50[years
6	[50, 60[years
7	[60, 70[years
8	[70, 80[years
9	[80, 90[years
10	>= 90 years

Table 3.2: Example of feature translation from category values to category labels.

To guarantee their correctness (for *debug* purposes), their intermediate contents for all experiences were kept in Text files (TXT). Finally, after running all the experiments defined in an Experiment Set, a XLSX file is produced with three worksheets:

- Metrics for each Biclustering solution as a whole (Precision and Entropy) and for its found Biclusters (Entropy, Purity, statistical significance, number of rows, percentage of dataset rows and Precision for all classes);
- Experiment-level metrics to easily generate plots for tendencies (e.g. Average percentage of lines per Bicluster, number of found Biclusters, number of discriminative Biclusters, etc);
- The discriminative Biclusters (their unique id and their class) for each experiment.

In this thesis, every Bicluster returned in a solution (discriminative or not) was statistically significant (p-value < 0.05). Moreover, a Bicluster was only considered discriminative if at least 75% of the subjects in it belonged to the same class (Purity $\geq$ 75%).

To choose the best Biclustering solution between the considered ones it was not possible to use External evaluation metrics due to the lack of ground truth and the shortage of Internal evaluation metrics which support categorical data (only one measure was found, the Goodman and Kruskal τ coefficient, which works by measuring the association of nominal variables [48][45]). Therefore, a criteria of usefulness based on the number of discriminative Biclusters was used to find the best solution: the more discriminative Biclusters a solution had, the better. This is illustrated in Figure 3.2 by the Discriminative Biclusters (Meta-features) block.

3.1.4 Biclustering-based Classification

In this phase Random Forest (RF) classifiers from Scikit-learn [4] Python library were used to determine the most important features (or subsets of features). RFs were chosen for being robust against overfitting, fast to train and able to return Feature Importance metrics. Since our data had features with different numbers of categories, the Permutation Importance metric was used to prevent biases.

Experiments (indicated in Figure 3.2) were performed for:

a) Baseline with all features (Raw);
b) Baseline with a subset of the features (FS);
c) Matrix of subject ids $\times$ discriminative Bicluster ids (Meta-features);
d) Merged data (joining the data from b) + c)).

The number of tree instances used in each RF classifier was 300, and the Permutation Importance values (using the MLxtend library [55]) were calculating by performing the mean out of 10 runs. Additionally, for the experiments involving Biclusters - c) and d) - random sampling was used on the most frequent class(es) to set the same number of discriminative Biclusters for all classes. The random seeds were kept static for experiment reproducibility. Classifier evaluation metrics were calculated using stratified 10-fold Cross Validation. Particularly, the Specificity metric was calculated by using a Python library complementary to Scikit-learn, called Imbalance-learn [38].

3.1.5 Class Association Rule Mining

The Java SPMF library [17] was called from Python scripts to mine regular ARs from frequent closed itemsets. Given that this generated a great amount of rules the following steps were taken to ease interpretation:

1. Filter rules without the class as the single consequent;

2. Split CARs to different files according to class;
3. Remove redundant rules per class.

This was inevitably more time consuming than using a CAR mining algorithm that also calculated the Lift. However, such an algorithm could not be found and to modify an existing one proved outside of the scope of this work. Experiments (indicated in Figure 3.2) were performed for:

e) Baseline with a subset of the features (FS) and their values;
f) Bicluster Features and Values (Meta-features), where for each Bicluster its pattern plus discriminative class were considered as transactions, repeated as many times as the Bicluster's number of rows.

The Baseline e) was calculated with the subset of the features used for experiment b) seen in the previous section, since to find the rules from the dataset with all features used in experiment a) for both Tasks proved too time-consuming.

For the f) experiment, by using the Bicluster patterns as transactions, we try to find the most common patterns associated with the classes they are discriminative for. Since a certain degree of Bicluster overlapping may exist, this technique may find patterns that would not be discovered otherwise. Moreover, for this same experiment, sampling was also applied to consider the same discriminative Biclusters as in *Section 3.1.4 Biclustering-based Classification*.

For each experiment the minimum support was also iteratively lowered until a significant number of CARs for all classes were found. The Minimum Confidence and Lift thresholds used for all experiments were 90% and 1, respectively.

3.1.6 Workflows and Code

The produced workflows and code used in the DMD approach implementation used in this thesis are available in the GitHub repository indicated in *Appendix D GitHub Repository*.

Chapter 4

Discriminative Meta-features Discovery: A Case Study in the Portuguese ONWebDUALS Dataset

In this chapter the proposed DMD approach was applied to perform experiments for both primary tasks. All figures present in this Chapter can also be found in the GitHub repository indicated in *Appendix D GitHub Repository*.

4.1 Task 1 - Discriminative Meta-features between Portuguese ALS Patients and Controls

4.1.1 Data and Settings

As stated in *Chapter 1 Introduction*, this Task's objective is to discover Meta-features which best distinguish the Portuguese ALS Patients from their Controls, if any. After the Data Pre-processing phase, the two Baseline datasets (Figure 3.2 a) Baseline with All Features and b) Baseline FS) had 99 and 20 features, respectively. The class label was provided by the clinicians in the ONWebDUALS dataset. The number of Portuguese Patients and Controls used in this experiment are reported in Table 4.1.

Class Label	Absolute Frequency	Relative Frequency
Patients	472	61%
Controls	300	39%
Total	772	100%

Table 4.1: Descriptive Frequency Analysis per Class of Used Data for Task 1.

Table 4.2 shows the parameterization used in the Pattern Mining-based Biclustering phase in Task 1.

Parameter	Values
Coherency Assumption	Constant Biclusters
Coherency Strength	50 items
Quality (%)	100 (No noise)
Pattern Representation	Closed patterns
Orientation	Pattern on Rows
Normalization	None
Discretization	None
Noise Handler	None
Missings Handler	Remove
Remove Uninformative Elements	None
Stopping Criteria	Minimum Support
Stopping Criteria Value (%)	{60, 55, 50, 45, 40, 35, 30, 25, 20, 15, 10, 5, 2.5, 1}
Minimum Bicluster Columns	3
Pattern Miner	CharmDiffsets
Scalability	False
Merging Procedure	Heuristic
Filtering Procedure	Dissimilar Elements
Filtering Procedure Value (%)	25

Table 4.2: BicPAMS Parameter Values for Task 1.

To distinguish ALS Patients from their Controls we wanted to find the largest possible subsets of similar subjects (Patients or Controls) for given subsets of features. This means that the interesting Biclusters to search for are maximal Biclusters with Coherent Evolution on Columns (Figure 2.3 h)), where the pattern is observed in the Bicluster's rows.

No normalization, discretization or noise handling were deemed necessary, since the data was already discretized beforehand and only Biclusters with the exactly same pattern (Meta-feature) in all their rows were considered interesting (maximum Quality). Regarding missing values, the Biclustering algorithm was instructed not to account for them in the found Bicluster patterns.

The only variable parameter between the experiments was the Stopping Criteria Value (%), meaning that the Minimum Support value used in each experiment was iteratively lowered until discriminative Biclusters with the highest possible number of rows started to be found. This parameter's values started at 60% to account for the largest portion of subjects between all classes (Patient class).

The minimum number of columns was set to at least 3 to avoid discovering Biclusters with a small amount of features. The CharmDiffsets algorithm was chosen to mine the Biclusters, given that it has

a better memory management than its available counterparts (CharmTID and AprioriTID). This is done by using Diffsets: memory structures that help to reduce drastically the amount of memory needed for intermediate results, by keeping only the differences of a candidate pattern from its prefix pattern in the transactions [69].

Heuristic merging was used for computational efficiency reasons. Finally, by using the aforementioned filtering parameters, Biclusters sharing more than 75% of their elements against a larger Bicluster were filtered out.

For the Biclustering-based Classification phase no further parameterization than the one in *Section 3.1.4 Biclustering-based Classification* was necessary. In the same fashion, the used parametrization for the Class Association Rule Mining phase was mostly discussed in *Section 3.1.5 Class Association Rule Mining*, with only the minimum Support percentage being iteratively lowered until a considerable number of rules for each class were found. The exact levels of Support used in each experiment can be found further below in Table 4.7.

4.1.2 Results and Discussion

4.1.2.1 Pattern Mining-based Biclustering

Figure 4.1 shows the evolution of the average number of rows between all Biclusters present in each Biclustering solution (one per Experiment) while the Minimum Support value (Relative Support) is being lowered. For the 60.0 % and 50.0 % Experiments the value is zero because no Biclusters were found (shown in Figure 4.3).

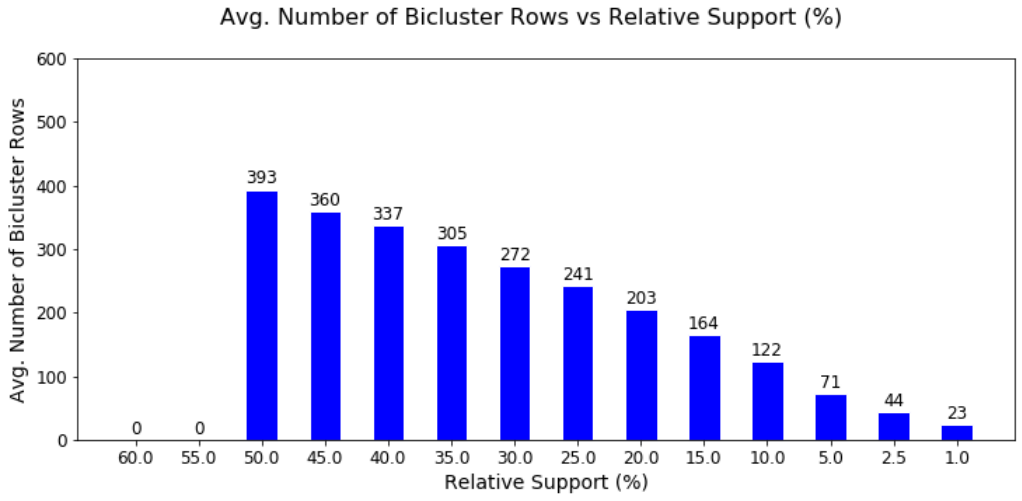


Figure 4.1: Average Number of Bicluster Rows vs Relative Support Percentage [Link].

Figure 4.2 displays the Purity evolution between the different Biclustering solutions while the Min-

imum Support value (Relative Support) is being lowered. As before, for the 60.0 % and 50.0 % Experiments the value is zero because no Biclusters were found. This shows that the Purity is somewhat constant, with a value of approximately 0.6 (60%).

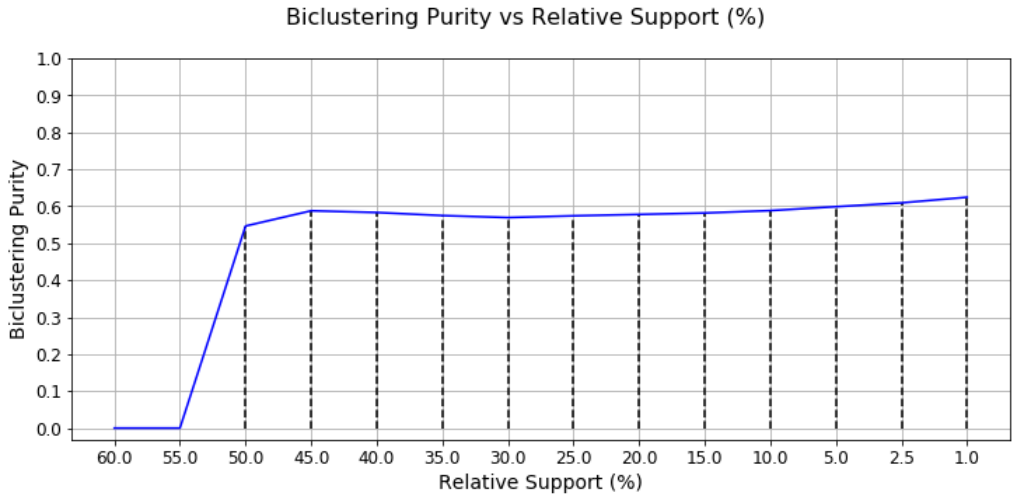


Figure 4.2: Biclustering Solution Purity vs Relative Support Percentage [Link].

As previously indicated, a criteria of usefulness based on the number of discriminative Biclusters was used to find the best solution.

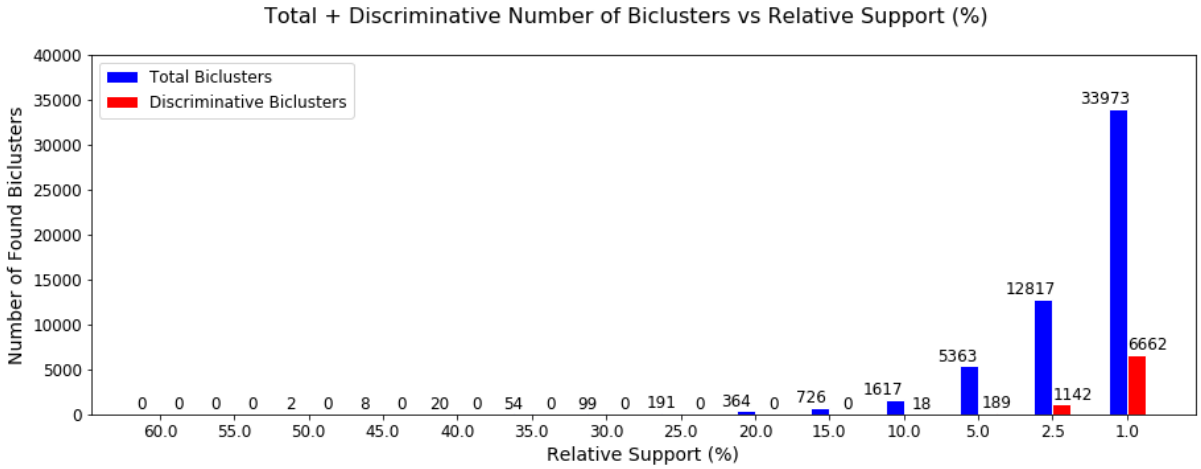


Figure 4.3: Number of Total/Discriminative Biclusters vs Relative Support Percentage [Link].

Figure 4.3 shows the number of total Biclusters (in blue/darkest color) and the number of discriminative Biclusters (in red/lightest color) found between the different Biclustering solutions while the Minimum Support value (Relative Support) is being lowered. We can see that discriminative Biclusters start

to appear at 10.0% Support levels and lower. However, at this point it was still necessary to validate if a considerable amount of Biclusters for each class was discovered.

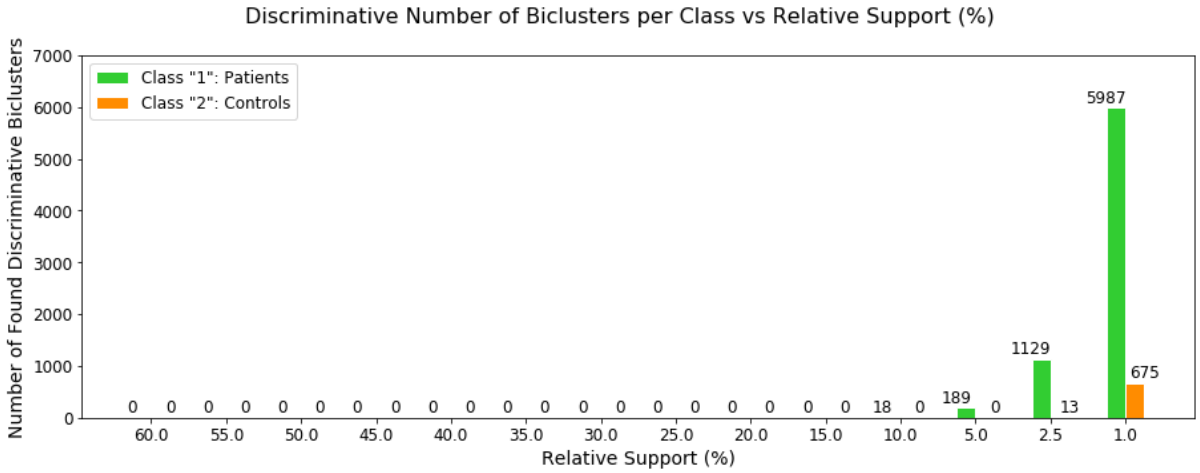


Figure 4.4: Number of Discriminative Biclusters per Class vs Relative Support Percentage [Link].

Thus, as reported in Figure 4.4, the number of discriminative Biclusters per class was determined. Only for the 1.0% Support experiment a significant amount of discriminative Biclusters for each class were present, with this being chosen as the best. We can observe here that for the best experiment many more discriminative Biclusters were found for the Patient class than for the Controls, which conveys the natural heterogeneity of the Patient portion of the dataset. Due to this, random sampling was applied to the most prevalent class to equalize the number of discriminative Biclusters (in this case, 675 for each class). To sum up, Table 4.3 provides a general view of the number of features used in all the Classification experiments for Task 1.

Dataset Version	Number of Features
a) Baseline All Features	99
b) Baseline FS	20
c) Matrix Meta-features	675 * 2 = 1350
d) Merged data	1350 + 20 = 1370

Table 4.3: Descriptive Analysis of Used Data Features for Task 1.

The list of features present in each Baseline dataset can be found in *Appendix B ONWebDUALS dataset features for Task 1*. Additionally, the complete list of discriminative Bicluster patterns for the 1.0% Support experiment can be found in *Appendix D.2.1 Discriminative Bicluster Patterns*.

4.1.2.2 Biclustering-based Classification

The evaluation metrics computed for the four experiments delineated in *Section 3.1.4 Biclustering-based Classification* by using 10-fold CV are present in Table 4.4 below.

Metric	Train	Test
Accuracy	1.00 (+/- 0.00)	0.69 (+/- 0.05)
Precision	1.00 (+/- 0.00)	0.63 (+/- 0.09)
Recall / Sensitivity	1.00 (+/- 0.00)	0.47 (+/- 0.12)
F-measure	1.00 (+/- 0.00)	0.54 (+/- 0.10)
Specificity	1.00 (+/- 0.00)	0.83 (+/- 0.05)
MCC	1.00 (+/- 0.00)	0.32 (+/- 0.13)
AUC	1.00 (+/- 0.00)	0.65 (+/- 0.06)

a) Baseline All Features.

Metric	Train	Test
Accuracy	1.00 (+/- 0.00)	0.66 (+/- 0.05)
Precision	1.00 (+/- 0.00)	0.59 (+/- 0.10)
Recall / Sensitivity	1.00 (+/- 0.00)	0.44 (+/- 0.10)
F-measure	1.00 (+/- 0.00)	0.50 (+/- 0.08)
Specificity	1.00 (+/- 0.00)	0.81 (+/- 0.06)
MCC	1.00 (+/- 0.00)	0.26 (+/- 0.11)
AUC	1.00 (+/- 0.00)	0.62 (+/- 0.05)

b) Baseline FS.

Metric	Train	Test
Accuracy	0.99 (+/- 0.00)	0.74 (+/- 0.06)
Precision	1.00 (+/- 0.00)	0.67 (+/- 0.08)
Recall / Sensitivity	0.99 (+/- 0.00)	0.64 (+/- 0.10)
F-measure	0.99 (+/- 0.00)	0.65 (+/- 0.08)
Specificity	1.00 (+/- 0.00)	0.80 (+/- 0.06)
MCC	0.99 (+/- 0.00)	0.45 (+/- 0.12)
AUC	0.99 (+/- 0.00)	0.72 (+/- 0.06)

c) Matrix Subject ID $\times$ Biclusters.

Metric	Train	Test
Accuracy	1.00 (+/- 0.00)	0.74 (+/- 0.05)
Precision	1.00 (+/- 0.00)	0.74 (+/- 0.11)
Recall / Sensitivity	1.00 (+/- 0.00)	0.53 (+/- 0.09)
F-measure	1.00 (+/- 0.00)	0.61 (+/- 0.09)
Specificity	1.00 (+/- 0.00)	0.88 (+/- 0.06)
MCC	1.00 (+/- 0.00)	0.45 (+/- 0.12)
AUC	1.00 (+/- 0.00)	0.70 (+/- 0.06)

d) Merged Data.

Table 4.4: Evaluation Metrics for RF Classifier - 10-fold CV in Task 1.

Given the results presented above we can see that:

- Every Training evaluation metric in every experience had a value of 1 or very near, which combined with the much lower Test metric values indicates overfitting of the RF model to the data;
- Feature Selection alone (Table 4.4 b)) worsened the classification over the Baseline with all features (Table 4.4 a));
- Using the matrix of subject ids $\times$ discriminative Bicluster ids by itself (Table 4.4 c)) slightly improved over the Baseline with all features (Table 4.4 a));

- Merging the individual feature space with the Meta-features space (Table 4.4 d)) had mixed results: it improves the matrix experiment results on most metrics except for Recall/Sensitivity and its related F-measure.

Nonetheless, the Merged Data experiment (Table 4.4 d)) seems to be the overall best from the four.

Most Important Features

The features considered most important by the RF models for each Classification experiment can be found below, from Figure 4.5 to Figure 4.8, in descending order of mean Permutation Imputation values. When more than 30 features are considered for an experiment (as seen in Table 4.3), only the 30 with higher mean Permutation Imputation values are shown due to space limitations.

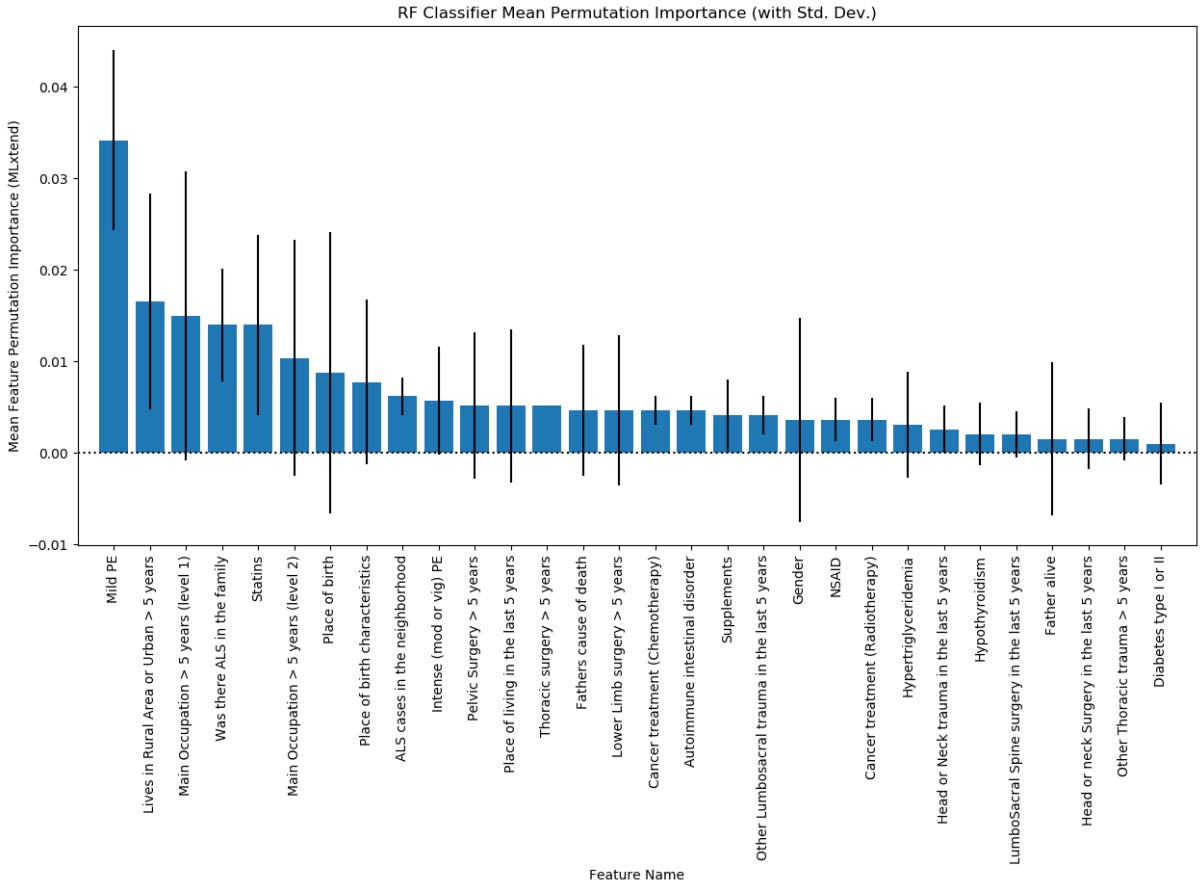


Figure 4.5: Top-30 Most Important Features - a) Baseline All Features for Task 1 [Link].

According to Figure 4.5 for the experiment a) Baseline All Features, the 5 most important features were *Mild PE* (Physical Exercise), *Lived in Rural or Urban Area more than 5 years ago*, *Main Occupation more than 5 years ago (level 1)*, *Was there ALS in the family* and *Statins*. *Mild PE*'s importance is around

the double for the model compared to the other four features. Some studies sought an association between physical exercise and ALS, with conflicting results [13]. It cannot be seen in Figure 4.5, but the features' mean Permutation Imputation values eventually reach zero and even negative values, meaning that the model could benefit from their removal.

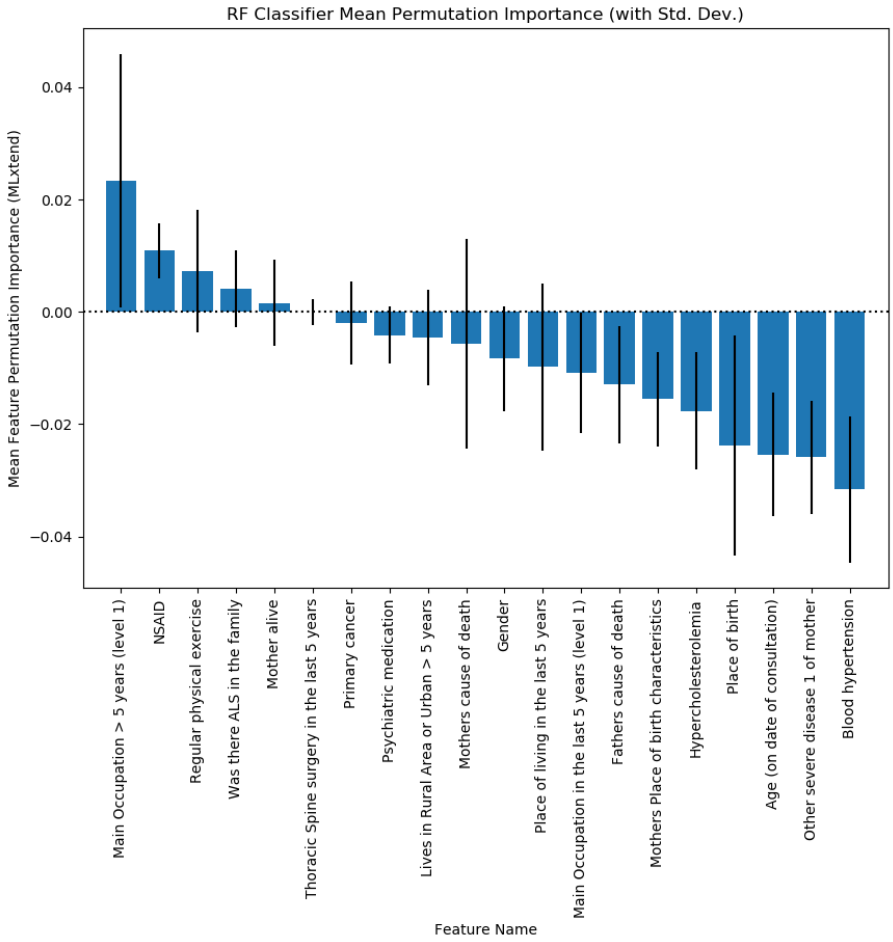


Figure 4.6: Top-30 Most Important Features - b) Baseline FS for Task 1 [Link].

For the experiment b) Baseline FS only 20 features were used in total, thus all of them being visible in Figure 4.6. Here we can see that the 5 most important features changed a bit: *Main Occupation more than 5 years ago (level 1)*, *Nonsteroidal anti-inflammatory drug (NSAID)*, *Regular physical exercise*, *Was there ALS in the family* and *Mother alive*. Additionally, the mean Permutation Importance of the best features are overall lower than in Figure 4.5. Figure 4.6 shows clearly why the Classification results (Table 4.4 b)) were the worse of the lot: most of the remaining features were considered harmful to the model, given their mean Permutation Importance negative values. This may also imply that the FS applied was not the best for the problem at hand.

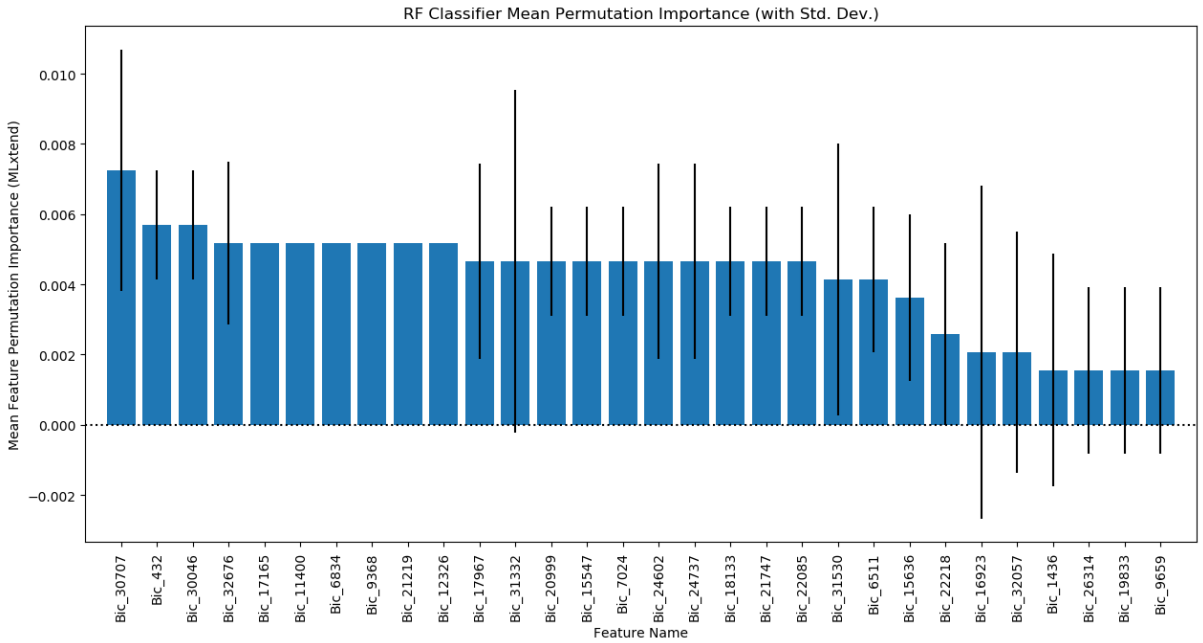


Figure 4.7: Top-30 Most Important Features - c) Matrix Subject ID × Biclusters for Task 1 [Link].

Figure 4.7 contains the mean Permutation Importance values for the top-30 most important features in experiment c) Matrix Subject ID × Biclusters, which considers only Biclusters. In a general fashion, the mean Permutation Importance values are much more uniform and lower than before, probably due to the high number of features/Biclusters used for this experiment (Table 4.3). The patterns of the top-5 Biclusters can be found in Table 4.5.

Bicluster Id	Class	Rows	Pattern
Bic_30707	Controls	9	{*Blood hypertension* = No, *NSAID* = No, *Mother alive* = No, *Regular physical exercise* = No, *Thoracic Spine surgery in the last 5 years* = No, *Main Occupation in the last 5 years (level 1)* = 10 Pensioner / Out of job, *Main Occupation more than 5 years ago (level 1)* = 10 Pensioner / Out of job, *Place of living in the last 5 years* = Large town > 100 000, *Lived in Rural or Urban Area more than 5 years ago* = Urban area}
Bic_432	Patients	24	{*Hypercholesterolemia* = No, *Primary cancer* = No, *NSAID* = No, *Other severe disease 1 of mother* = 11 Diseases of the circulatory system, *Mother alive* = No, *Mothers cause of death* = 11 Diseases of the circulatory system, *Thoracic Spine surgery in the last 5 years* = No, *Lived in Rural or Urban Area more than 5 years ago* = Urban area}

Bicluster Id	Class	Rows	Pattern
Bic_30046	Controls	7	{*Gender* = Male, *Place of birth* = Large town > 100 000, *Mothers Place of birth characteristics* = Urban area, *Hypercholesterolemia* = No, *Primary cancer* = No, *Psychiatric medication* = No, *NSAID* = No, *Was there ALS in the family* = No, *Mother alive* = No, *Regular physical exercise* = No, *Thoracic Spine surgery in the last 5 years* = No, *Place of living in the last 5 years* = Large town > 100 000, *Lived in Rural or Urban Area more than 5 years ago* = Urban area}
Bic_32676	Controls	7	{*Mothers Place of birth characteristics* = Rural area, *Blood hypertension* = No, *NSAID* = No, *Mother alive* = No, *Thoracic Spine surgery in the last 5 years* = No, *Main Occupation in the last 5 years (level 1)* = 10 Pensioner / Out of job, *Main Occupation more than 5 years ago (level 1)* = 10 Pensioner / Out of job, *Place of living in the last 5 years* = Large town > 100 000, *Lived in Rural or Urban Area more than 5 years ago* = Urban area}
Bic_17165	Patients	28	{*Age (on date of consultation)* = [50, 60[years, *Mothers Place of birth characteristics* = Rural area, *Blood hypertension* = No, *Hypercholesterolemia* = No, *NSAID* = No, *Was there ALS in the family* = No, *Thoracic Spine surgery in the last 5 years* = No, *Lived in Rural or Urban Area more than 5 years ago* = Urban area}

Table 4.5: Top-5 Most Important Bicluster Patterns - c) Matrix Subject ID × Biclusters for Task 1.

In this table we can see that the found patterns generally have a large number of features and the Biclusters for the Patient class tend to have a greater number of rows than the ones found for the Controls. However, the appearance of several Biclusters of the Controls class as the most important for the Classification indicates a greater homogeneity in the Controls data comparing with the Patients' data, which was expected. The complete list of Bicluster patterns present in Figure 4.7 can be found in *Appendix D.2.2 Top-30 Most Important Bicluster Patterns - c) Matrix Subject ID × Biclusters*.

Finally, Figure 4.8 below reports the mean Permutation Importance values for the top-30 most important features in experiment d) Merged Data. Only one individual feature appears in this top-30 (*Father's cause of death*), with the rest of the positions being occupied by Biclusters. The patterns of the top-5 Biclusters for this last experiment can be found in Table 4.6. This overall predominance of Biclusters as the most important features clearly indicates that individual features might not be enough to distinguish Patients from Controls, making it necessary to consider subsets of features for that effect. Nonetheless, given the size and variety of the found patterns it was hard to discern a clear subset of features which would allow to discriminate between these two classes. The complete list of Bicluster patterns present in

Figure 4.8 can be found in *Appendix D.2.3 Top-30 Most Important Bicluster Patterns - d) Merged Data*.

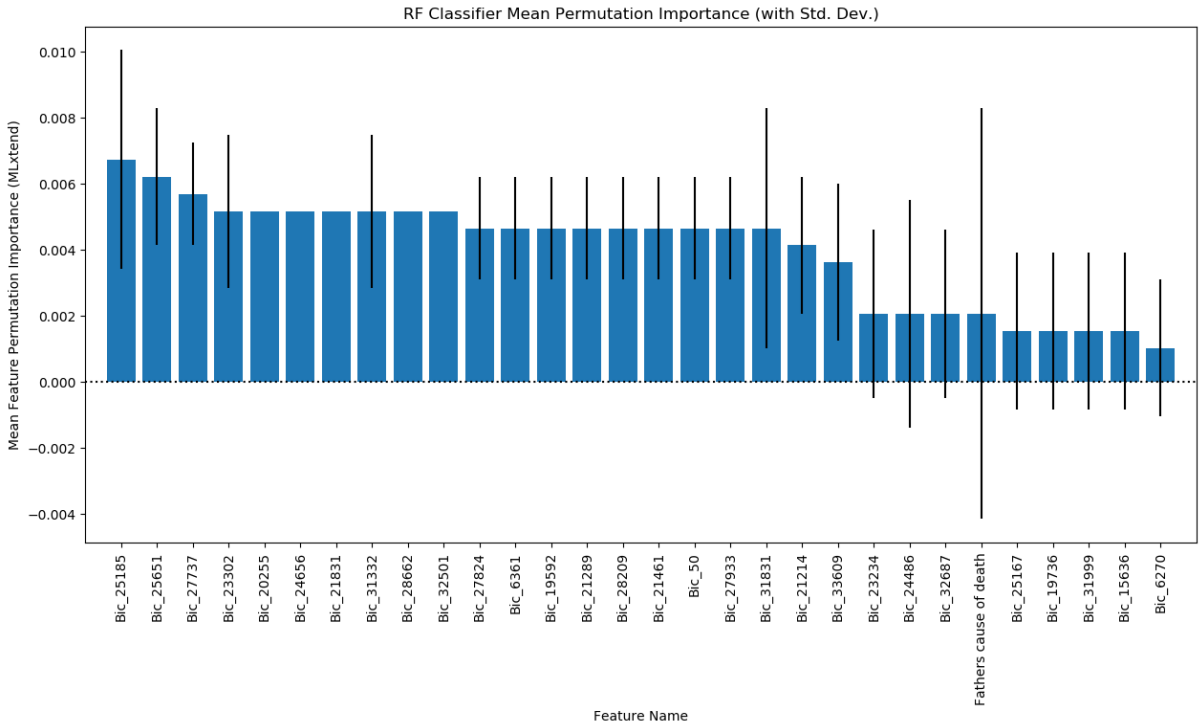


Figure 4.8: Top-30 Most Important Features - d) Merged Data for Task 1 [Link].

Bicluster Id	Class	Rows	Pattern
Bic_25185	Controls	24	{*Gender* = Male, *Blood hypertension* = Yes, *Primary cancer* = No, *Psychiatric medication* = Yes, *Was there ALS in the family* = No, *Mother alive* = No, *Mother's cause of death* = 21 Symptoms, signs or clinical findings, not elsewhere classified, *Thoracic Spine surgery in the last 5 years* = No, *Main Occupation in the last 5 years (level 1)* = 10 Pensioner / Out of job}
Bic_25651	Controls	9	{*Blood hypertension* = Yes, *NSAID* = No, *Was there ALS in the family* = No, *Mother alive* = No, *Thoracic Spine surgery in the last 5 years* = No, *Main Occupation in the last 5 years (level 1)* = 10 Pensioner / Out of job, *Main Occupation more than 5 years ago (level 1)* = 24 Business and administration professionals, *Place of living in the last 5 years* = Large town > 100 000, *Lived in Rural or Urban Area more than 5 years ago* = Urban area}

Bicluster Id	Class	Rows	Pattern
Bic_27737	Controls	9	{*Blood hypertension* = Yes, *Hypercholesterolemia* = Yes, *Was there ALS in the family* = No, *Other severe disease 1 of mother* = 02 Neoplasms, *Thoracic Spine surgery in the last 5 years* = No, *Main Occupation in the last 5 years (level 1)* = 10 Pensioner / Out of job, *Place of living in the last 5 years* = Large town > 100 000, *Lived in Rural/Urban Area more than 5 years ago* = Urban area}
Bic_23302	Controls	10	{*Gender* = Female, *Blood hypertension* = Yes, *Psychiatric medication* = No, *Was there ALS in the family* = No, *Mother alive* = No, *Father's cause of death* = 11 Diseases of the circulatory system, *Regular physical exercise* = No, *Thoracic Spine surgery in the last 5 years* = No, *Main Occupation in the last 5 years (level 1)* = 10 Pensioner / Out of job, *Lived in Rural or Urban Area more than 5 years ago* = Urban area}
Bic_20255	Patients	8	{*Age (on date of consultation)* = [60, 70[years, *Primary cancer* = No, *NSAID* = No, *Was there ALS in the family* = No, *Mother alive* = No, *Father's cause of death* = 12 Diseases of the respiratory system, *Regular physical exercise* = No, *Thoracic Spine surgery in the last 5 years* = No}

Table 4.6: Top-5 Most Important Bicluster Patterns - d) Merged Data for Task 1.

4.1.2.3 Class Association Rule Mining

The experiments with Class Association Rules outlined in *Section 3.1.5 Class Association Rule Mining* found rules to characterize both classes considered in this Task (Patients and Controls). Table 4.7 contains the minimum thresholds of Support used, the size of the transaction database and the number of rules found (before and after filtering the redundant rules):

Experiment	Class	Support (%)	Transactions	Rules	Non-Redundant Rules
e)	Patients	1.25%	772	1903	517
e)	Controls	1.25%	772	44	28
f)	Patients	5%	15350	113	37
f)	Controls	5%	15350	23	8

Table 4.7: Metrics of Class Association Rule Mining experiments for Task 1.

All non-redundant rules for the Controls on both experiments had Lift values above 2, showing high

levels of association between the found patterns and the respective class. However, for the Patients class the Lift values started on 1.4 and tended to be found in larger numbers. This complies with what was seen in the previous section, where the Patient-related data contains more distinct patterns than the Controls. The complete list of non-redundant rules for all experiments and classes can be found in *Appendix D.2.4 Non-Redundant Class Association Rules*.

Given the limited space, the most relevant rules for each Experiment and class were chosen in the following fashion: since the Confidence thresholds were kept very high (minimum of 90%), from the set of rules with the higher Lift values, the ones with the largest Support were selected. Additionally, since the number of transactions between the two experiments (e) and f)) is strikingly different (by two orders of magnitude), the Relative Support (%) is used to show how frequently the given patterns appear instead of the number of rows/transactions. Thus, for Task 1, the said rules are present in Table 4.8.

Experiment	**Class**	**Rule**	**Support (%)**	**Lift**
e)	Patients	*Hypercholesterolemia* = No $\wedge$ *NSAID* = No $\wedge$ *Was there ALS in the family* = Yes $\Rightarrow$ *Class* = Patients	$\approx$ 2.59	$\approx$ 1.64
e)	Controls	*Main Occupation more than 5 years ago (level 1)* = 42 Customer services clerks $\wedge$ *NSAID* = No $\wedge$ Primary cancer = No $\wedge$ *Thoracic Spine surgery in the last 5 years* = No $\wedge$ *Was there ALS in the family* = No $\Rightarrow$ *Class* = Controls	$\approx$ 1.55	$\approx$ 2.57
f)	Patients	*Place of birth* = Village with < 1000 inhabitants $\Rightarrow$ *Class* = Patients	$\approx$ 17.8	$\approx$ 1.7
f)	Controls	*Main Occupation more than 5 years ago (level 1)* = 23 Teaching professionals $\wedge$ *Thoracic Spine surgery in the last 5 years* = No $\Rightarrow$ *Class* = Controls	$\approx$ 6.72	$\approx$ 2.42

Table 4.8: Most Relevant Class Association Rules for Task 1.

From the table above we can see that using ARM techniques leads to smaller patterns than those seen in the previous section, since the common items between the different transactions are found to form the antecedent of each rule. Moreover, the antecedents of the rules found in the f) experiment tend to be smaller than those from e), which might help explain their higher levels of Support.

The rules obtained for the Controls in both experiments showed higher Lift values than for Patients. The most relevant rules for the Patient class are totally different between the two experiments, and do not have many features in common with the patterns obtained in the previous section. However, a thorougher analysis of the large amount of obtained rules might discover some common ground between the results.

4.2 Task 2 - Discriminative Meta-features between Progression Groups on Portuguese ALS Patients

4.2.1 Data and Settings

As stated in *Chapter 1 Introduction*, this Task's objective was to discover Meta-features which caracterize ALS patients' progression groups, if any. These groups are relative to the speed of disease progression: Slow, Neutral or Fast, as aforementioned in *Section 2.6.4 Progression Groups in ALS*. After the Data Pre-processing phase, the two Baseline datasets (Figure 3.2 a) Baseline with All Features and b) Baseline FS) had 198 and 22 features, respectively. The effective numbers of Portuguese Patients used in this experiment are reported in Table 4.9.

Class Label	Absolute Frequency	Relative Frequency
Slow	149	32%
Neutral	190	40%
Fast	133	28%
Total	472	100%

Table 4.9: Descriptive Frequency Analysis per Class of Used Data for Task 2.

Table 4.10 shows the parameterization for the Pattern Mining-based Biclustering phase in Task 2.

Parameter	Values
Coherency Assumption	Constant Biclusters
Coherency Strength	50 items
Quality (%)	100 (No noise)
Pattern Representation	Closed patterns
Orientation	Pattern on Rows
Normalization	None
Discretization	None
Noise Handler	None
Missings Handler	Remove
Remove Uninformative Elements	None
Stopping Criteria	Minimum Support
Stopping Criteria Value (%)	{40, 35, 30, 25, 20, 15, 10, 5, 2.5, 1}
Minimum Bicluster Columns	3
Pattern Miner	CharmDiffsets [69]
Scalability	False

Parameter	Values
Merging Procedure	Heuristic
Filtering Procedure	Dissimilar Elements
Filtering Procedure Value (%)	25

Table 4.10: BicPAMS Parameter Values for Task 2.

The parameterization for the Pattern Mining-based Biclustering phase was kept mostly the same in comparison with Task 1, since to distinguish between the several ALS Patient classes we also want to find maximal Biclusters with Coherent Evolution on Columns (Figure 2.3 h)). The only variable parameter between the experiments was still the Stopping Criteria Value (%), to find the discriminative Biclusters with the highest possible number of rows. For this Task it started at 40% given the largest portion of patients between all classes (Neutral class).

As in Task 1, for the Biclustering-based Classification phase no further parameterization than the one seen in *Section 3.1.4 Biclustering-based Classification* was necessary. Likewise, the used parametrization for the Class Association Rule Mining phase was mostly discussed in *Section 3.1.5 Class Association Rule Mining*, with the Minimum Support being iteratively lowered until a considerable number of rules for each class were found. For this Task the levels of Support used in each experiment can be found in Table 4.16.

4.2.2 Results and Discussion

4.2.2.1 Pattern Mining-based Biclustering

Figure 4.9 shows the average number of rows' evolution between all Biclusters present in each Biclustering solution (one per Experiment) while the Minimum Support value (Relative Support) is lowered.

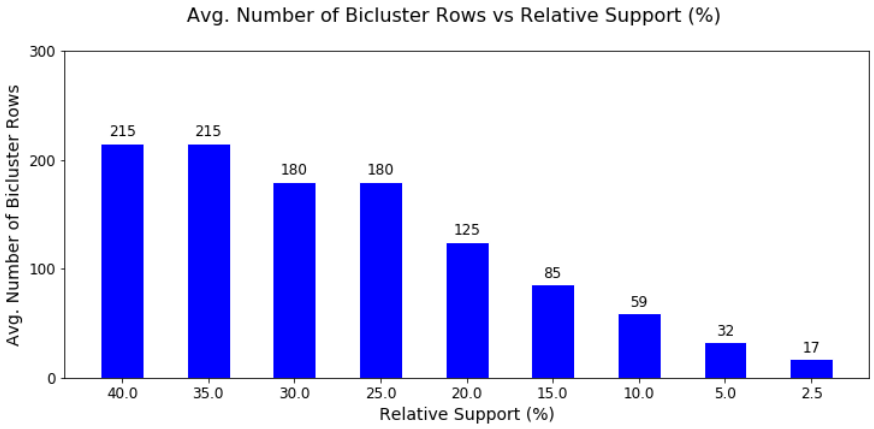


Figure 4.9: Average Number of Bicluster Rows vs Relative Support Percentage [Link].

Some experiments have the exact same average of Bicluster rows due to the discovery of the exact same number of Biclusters with different levels of Relative Support (this is shown in Figure 4.11). Figure 4.10 displays the Purity metric's evolution between the different Biclustering solutions while the Minimum Support value (Relative Support) is being lowered. This shows that the Purity is slowly rising, starting on 0.5 (50%) and towards 0.6 (60%).

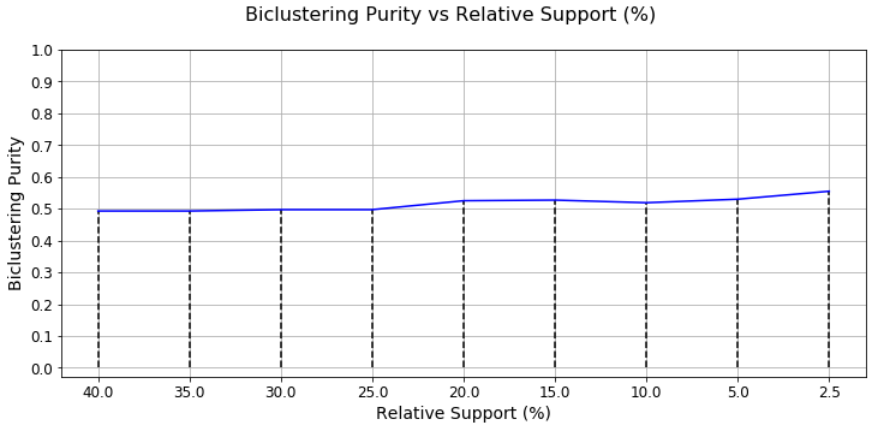


Figure 4.10: Biclustering Solution Purity vs Relative Support Percentage [Link].

The same criteria of usefulness based on the number of discriminative Biclusters used in Task 1 was employed to find the best Biclustering solution. Figure 4.11 shows the number of total Biclusters (in blue/darkest color) and the number of discriminative Biclusters (in red/lightest color) found between the different Biclustering solutions while the Minimum Support value (Relative Support) is being lowered. For this Task discriminative Biclusters also start to appear at 10.0% Support levels and lower. However, a considerable amount of Bicluster for each class was still needed.

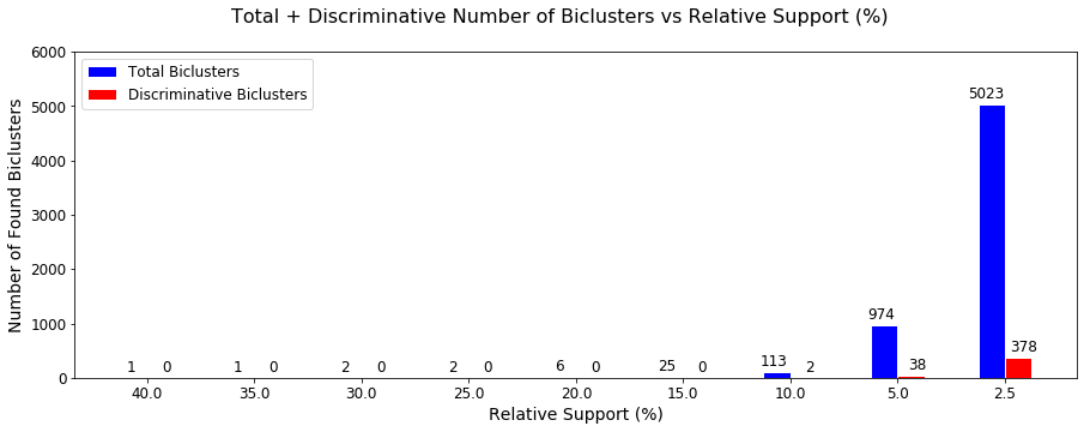


Figure 4.11: Number of Total/Discriminative Biclusters vs Relative Support Percentage [Link].

Thus, as shown in Figure 4.12, the number of discriminative Biclusters per class was determined. Only for the 2.5% Support experiment a significant amount of discriminative Biclusters for each class were present, with this experiment being chosen as the best. At this point we can observe that more discriminative Biclusters were found for the Slow and Fast classes than for the Neutral, possibly by being the classes that are more easily distinguishable. To equalize the number of discriminative Biclusters random sampling was also applied to the most prevalent classes (in this case, 72 for each class).

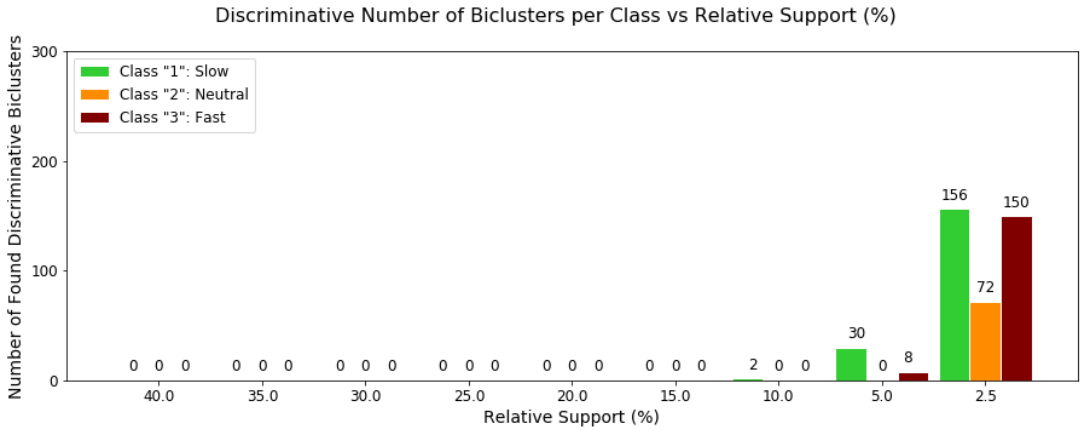


Figure 4.12: Number of Discriminative Biclusters per Class vs Relative Support Percentage [Link].

Table 4.11 summarizes the number of features used in all the Classification experiments for Task 1.

Dataset Version	Number of Features
a) Baseline All Features	198
b) Baseline FS	22
c) Matrix Meta-features	72 * 3 = 216
d) Merged data	216 + 22 = 236

Table 4.11: Descriptive Analysis of Used Data Features for Task 2.

The list of features present in each Baseline dataset can be found in *Appendix C ONWebDUALS dataset features for Task 2*. Additionally, the complete list of discriminative Bicluster patterns for the 2.5% Support experiment can be found in *Appendix D.3.1 Discriminative Bicluster Patterns*.

4.2.2.2 Biclustering-based Classification

The evaluation metrics for the four experiments delineated in *Section 3.1.4 Biclustering-based Classification* are present in Table 4.12.

Metric	Test
Accuracy	0.65 (+/- 0.07)
Precision	0.67 (+/- 0.07)
Recall / Sensitivity	0.65 (+/- 0.07)
F-measure	0.65 (+/- 0.07)
Specificity	0.82 (+/- 0.04)
MCC	0.47 (+/- 0.11)
AUC	0.73 (+/- 0.05)

a) Baseline All Features.

Metric	Test
Accuracy	0.72 (+/- 0.06)
Precision	0.74 (+/- 0.06)
Recall / Sensitivity	0.72 (+/- 0.06)
F-measure	0.73 (+/- 0.06)
Specificity	0.86 (+/- 0.03)
MCC	0.58 (+/- 0.09)
AUC	0.79 (+/- 0.04)

b) Baseline FS.

Metric	Test
Accuracy	0.70 (+/- 0.05)
Precision	0.73 (+/- 0.05)
Recall / Sensitivity	0.68 (+/- 0.06)
F-measure	0.69 (+/- 0.06)
Specificity	0.84 (+/- 0.03)
MCC	0.54 (+/- 0.08)
AUC	0.76 (+/- 0.04)

c) Matrix Subject ID $\times$ Biclusters.

Metric	Test
Accuracy	0.74 (+/- 0.05)
Precision	0.76 (+/- 0.06)
Recall / Sensitivity	0.74 (+/- 0.04)
F-measure	0.74 (+/- 0.05)
Specificity	0.86 (+/- 0.02)
MCC	0.61 (+/- 0.07)
AUC	0.80 (+/- 0.03)

d) Merged Data.

Table 4.12: Evaluation Metrics for RF Classifier - 10-fold CV in Task 2.

For this Task a complementary Python library to Scikit-Learn called *Multiscorer* [64] was used to easily compute 10-fold metrics for the Multiclass Classification case. However, it lacked documentation on how Training metrics could be obtained, explaining why those results are not presented.

From the results presented in Table 4.12 we can see that:

- Some overfitting might be occurring, it was not possible to verify without Training metrics;
- FS (Table 4.12 b)) greatly improved the classification over the other Baseline (Table 4.12 a));
- Using the matrix of subject ids $\times$ discriminative Bicluster ids by itself (Table 4.12 c)) also improved over the Baseline with all features (Table 4.12 a));
- Merging the individual feature space with the Meta-features space (Table 4.12 d)) helped to improve the results on all metrics comparing with the either the Baseline FS (Table 4.12 b)) or the matrix experiment (Table 4.12 c)), clearly being the overall best from the four.

Most Important Features

The features considered most important by the RF models for each Classification experiment can be found below, from Figure 4.13 to Figure 4.16, in descending mean Permutation Imputation order.

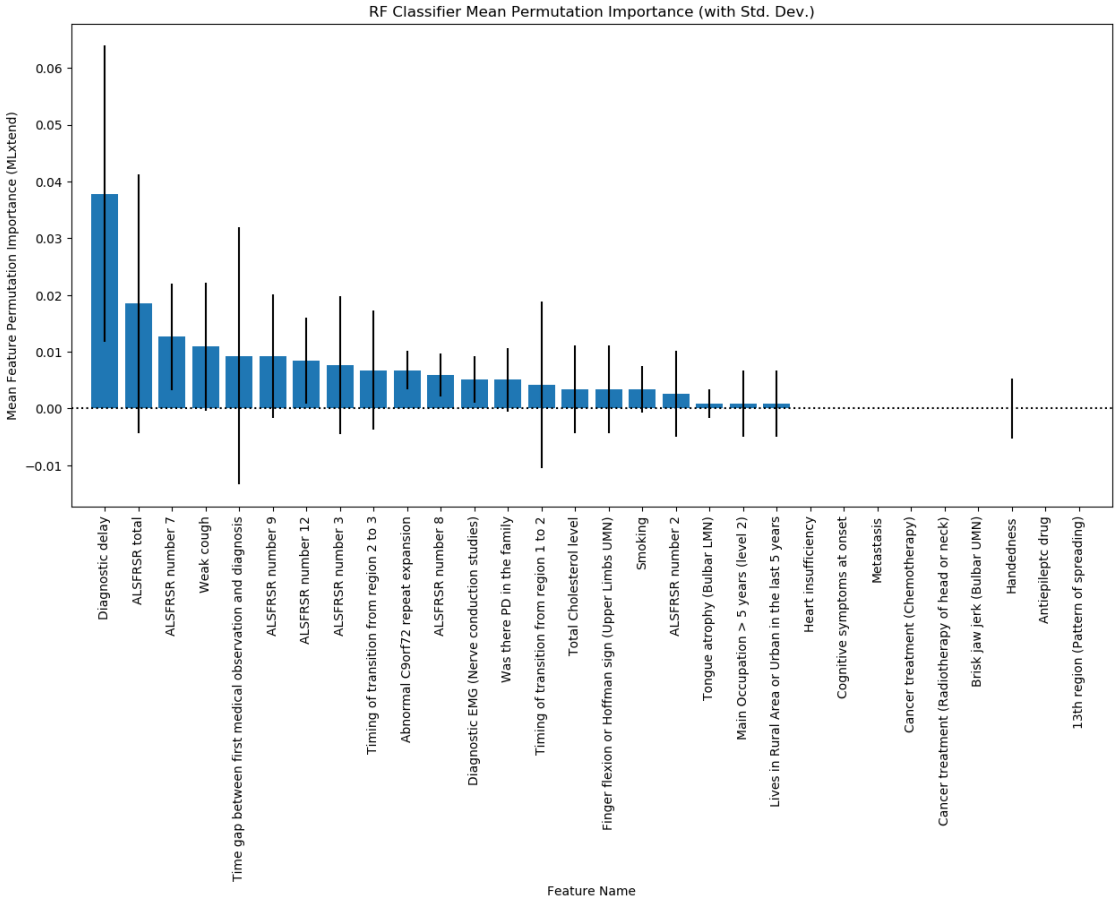


Figure 4.13: Top-30 Most Important Features - a) Baseline All Features for Task 2 [Link].

As before, when more than 30 features are considered for an experiment (as seen in Table 4.11), only the 30 with higher mean Permutation Imputation values are shown due to space limitations. According to Figure 4.13 above for the experiment a) Baseline All Features, the 5 features considered most important were *Diagnostic Delay*, *ALSFRS-R Total*, *ALSFRS-R number 7*, *Weak Cough* and *Time gap between first medical observation and diagnosis*. *Diagnostic Delay*'s importance is around the double for the model compared to the next most important feature, *ALSFRS-R Total*.

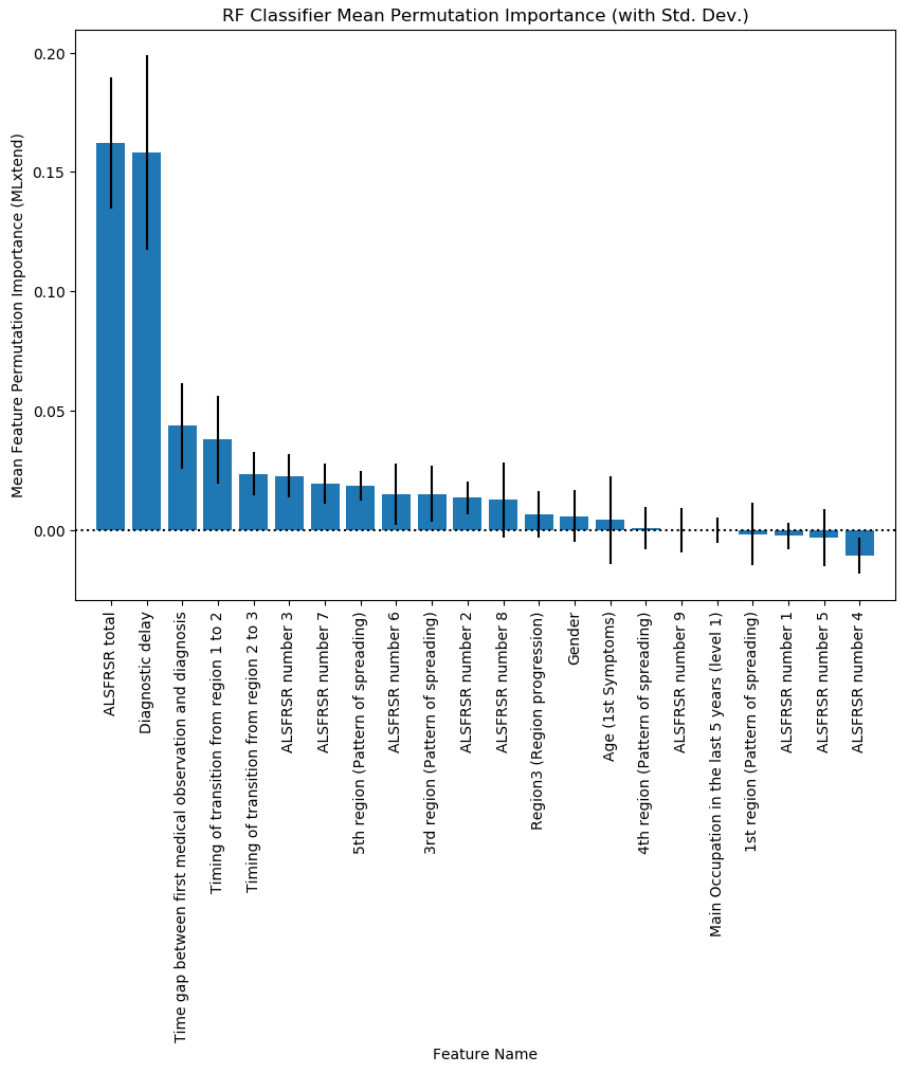


Figure 4.14: Top-30 Most Important Features - b) Baseline FS for Task 2 [Link].

For the experiment b) Baseline FS only 22 features were used in total, and thus all of them are visible in Figure 4.14. Here we can see that most of the 5 most important features stayed the same, even though not in the same order: *ALSFRS-R Total*, *Diagnostic Delay*, *Time gap between first medical observation and diagnosis*, *Timing of transition from region 1 to 2* and *Timing of transition from region 2 to 3*. Additionally, the mean Permutation Importance for the best features are much higher than in Figure 4.13. Moreover, in this figure it is very noticeable why the Classification results (Table 4.12 b)) improved comparing to the Baseline with all features: most of the remaining features were considered useful to the model, mainly the first two (with more than the triple of the relative importance of the subsequent features). This proves that the FS applied was successful for this Task.

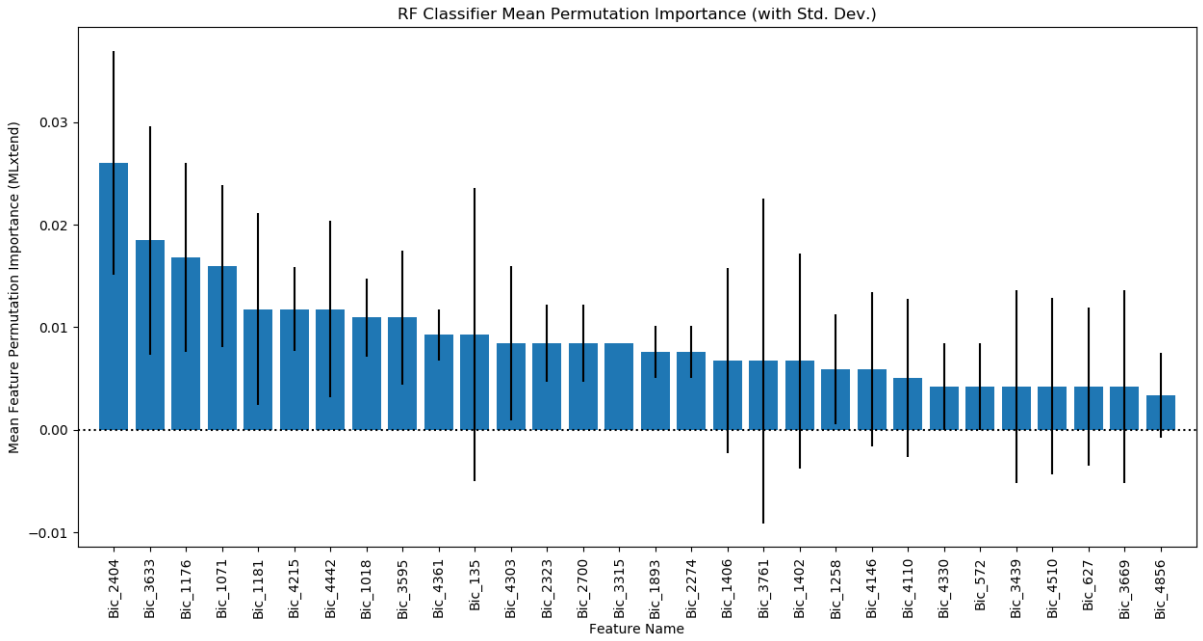


Figure 4.15: Top-30 Most Important Features - c) Matrix Subject ID × Biclusters for Task 2 [Link].

Figure 4.15 contains the mean Permutation Importance values for the top-30 most important features in experiment c) Matrix Subject ID × Biclusters, which considers only Biclusters. As on Task 1, the values are much more uniform and lower than before, probably due to the high number of features/Biclusters (Table 4.11). The patterns of the top-5 Biclusters can be found below in Table 4.13:

Bicluster Id	**Class**	**Rows**	**Pattern**
Bic_2404	Slow	12	{*Diagnostic delay* = [36, 48[months, *ALSFRS-R number 1* = 4, *ALSFRS-R number 2* = 4, *ALSFRS-R number 3* = 4}
Bic_3633	Fast	17	{*Diagnostic delay* = [6, 12[months, *1st region (Pattern of spreading)* = 1>, *ALSFRS-R number 7* = 3}
Bic_1176	Fast	11	{*Diagnostic delay* = [12, 18[months, *ALSFRS-R number 5* = 1, *Time gap between first medical observation and diagnosis* = [12, 18[months}
Bic_1071	Fast	17	{*Diagnostic delay* = [0, 6[months, *Timing of transition from region 1 to 2* = [0, 3[months, *Timing of transition from region 2 to 3* = [0, 3[months}
Bic_1181	Slow	42	{*ALSFRS-R number 1* = 4, *ALSFRS-R number 2* = 4, *ALSFRS-R number 3* = 4, *ALSFRS-R number 6* = 4}

Table 4.13: Top-5 Most Important Bicluster Patterns - c) Matrix Subject ID × Biclusters for Task 2.

In this table we can see that the found patterns generally have a small number of features, much smaller than on Task 1. The appearance of Biclusters of the Slow and Fast classes as the most important for the Classification was expected, given that these are the classes that can be discriminated more easily. Additionally, the *Diagnostic delay* feature appears consistently in almost all of them. The complete list of Bicluster patterns present in Figure 4.15 can be found in *Appendix D.3.2 Top-30 Most Important Bicluster Patterns - c) Matrix Subject ID × Biclusters*.

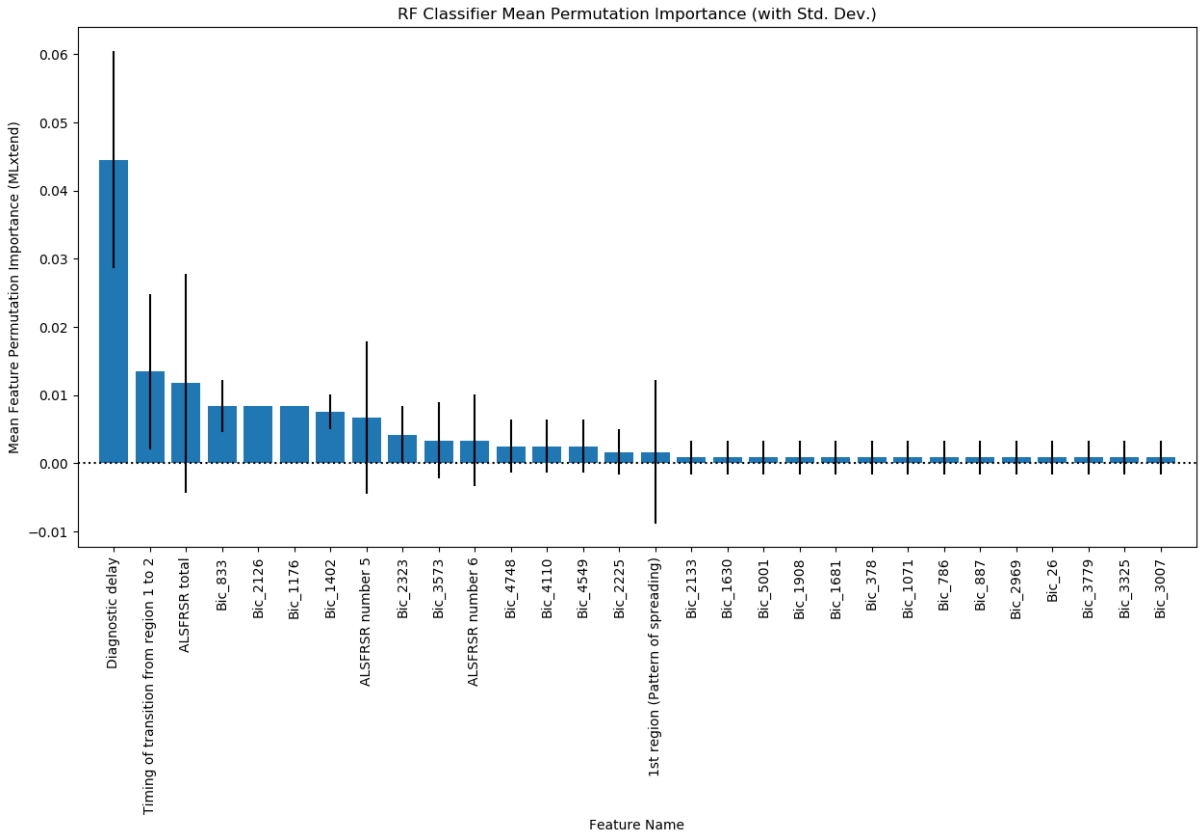


Figure 4.16: Top-30 Most Important Features - d) Merged Data for Task 2 [Link].

Finally, Figure 4.16 reports the mean Permutation Importance values for the top-30 most important features in experiment d) Merged Data. After merging the individual features from b) experiment with the matrix containing the Bicluster patterns from c) experiment, we see that some individual features keep being the most important: *Diagnostic Delay*, *Timing of transition from region 1 to 2* and *ALSFRS-R Total*. Since all of these features have been appointed as relevant on the previous experiments, this proves their importance in solving this group characterization problem. The complete list of Bicluster patterns present in Figure 4.16 can be found in *Appendix D.3.3 Top-30 Most Important Bicluster Patterns - d) Merged Data*.

Nonetheless, some Biclusters are also considered as important for the Classification. Table 4.14 shows the patterns of the top-5 Biclusters for this last experiment.

Bicluster Id	Class	Rows	Pattern
Bic_833	Fast	12	{*Diagnostic delay* = [0, 6[months, *Timing of transition from region 1 to 2* = [0, 3[months, *ALSFRS-R number 1* = 3, *ALSFRS-R number 2* = 3}
Bic_2126	Slow	17	{*1st region (Pattern of spreading)* = 5a>, *ALSFRS-R number 1* = 4, *ALSFRS-R number 2* = 4, *ALSFRS-R number 3* = 4, *Main Occupation in the last 5 years (level 1)* = 10 Pensioner / Out of job}
Bic_1176	Fast	11	{*Diagnostic delay* = [12, 18[months, *ALSFRS-R number 5* = 1, *Time gap between first medical observation and diagnosis* = [12, 18[months}
Bic_1402	Slow	15	{*1st region (Pattern of spreading)* = 5a>, *ALSFRS-R number 1* = 4, *ALSFRS-R number 3* = 4, *ALSFRS-R number 4* = 4, *Main Occupation in the last 5 years (level 1)* = 10 Pensioner / Out of job}
Bic_2323	Fast	21	{*Diagnostic delay* = [0, 6[months, *Timing of transition from region 1 to 2* = [0, 3[months, *ALSFRS-R number 6* = 3}

Table 4.14: Top-5 Most Important Bicluster Patterns - d) Merged Data for Task 2.

The *Diagnostic Delay* and *Timing of transition from region 1 to 2* features regularly appear in the most important Biclusters of the Fast class. Additionally, the ALSFRS-R scale questions present never reach the maximum value of 4, implicating a visible decline in functional performance. These results were confirmed by the clinicians, since shorter diagnostic delay and shorter times of transition between body regions (disease spreading, in short periods of 0-3 months) are is usually associated with faster disease progression [66].

For the Slow class, *1st region (Pattern of spreading)*, *ALSFRS-R number 1*, *ALSFRS-R number 3* and *Main Occupation in the last 5 years (level 1)* are common features. As expected, the ALSFRS-R scale questions present seem to always have the maximum value since Slow progressing patients tend to maintain their functional capabilities longer than the other classes. Slow progressors also tend to have a larger life expectancy, being able to reach retirement age.

Neutral class Biclusters only start to appear after the top-5, confirming that they are considered less discriminative to distinguish between the different progression groups, as postulated before. To try to understand the most important features for this class, some patterns found for Neutral Biclusters are present in Table 4.15.

Bicluster Id	Class	Rows	Pattern
Bic_3573	Neutral	22	{*Gender* = Female, *ALSFRS-R number 3* = 3, *ALSFRS-R number 5* = 3, *Main Occupation in the last 5 years (level 1)* = 10 Pensioner / Out of job}
Bic_4748	Neutral	11	{*Age (1st Symptoms)* = [60, 70[years, *1st region (Pattern of spreading)* = 1>, *ALSFRS-R number 6* = 4}
Bic_5001	Neutral	12	{*Timing of transition from region 1 to 2* = [3, 6[months, *ALSFRS-R number 5* = 3, *ALSFRS-R number 7* = 4}

Table 4.15: Examples of Bicluster patterns for Neutral class - d) Merged Data for Task 2.

The values associated with the ALSFRS-R scale questions present in the patterns for the Neutral class seem to fluctuate between 3 and 4, halfway between the Slow and Fast class, as expected. Particularly, here we can see that the fifth question of the ALSFRS-R scale, *ALSFRS-R number 5*, tends to appear with the value 3, implicating some compromise of the ability of cutting food and handling utensils [7].

4.2.2.3 Class Association Rule Mining

The experiments with Class Association Rules outlined in *Section 3.1.5 Class Association Rule Mining* were able to find rules to characterize all the classes. Table 4.16 contains the minimum thresholds of Support used, the size of the transaction database and the number of rules found (before and after filtering the redundant rules):

Experiment	Class	Support (%)	Transactions	Rules	Non-Redundant Rules
e)	Slow	2.0%	473	75	36
e)	Neutral	2.0%	473	22	14
e)	Fast	2.0%	473	73	49
f)	Slow	2.5%	3064	94	34
f)	Neutral	2.5%	3064	15	15
f)	Fast	2.5%	3064	12	12

Table 4.16: Metrics of Class Association Rule Mining experiments for Task 2.

All non-redundant rules for all classes on both experiments had Lift values above 2 (and even above 3), showing high levels of association between the found patterns and the respective classes. The complete list of non-redundant rules for all experiments and classes can be found in *Appendix D.3.4 Non-Redundant Class Association Rules*.

Given the limited space, the most relevant rules for each Experiment and class were chosen in the same fashion as in Task 1 (from the set of rules with the higher Lift values, the ones with the largest

Support were selected). Thus, for Task 2, the said rules are present in Table 4.17.

Experiment	Class	Rule	Support (%)	Lift
e)	Slow	*ALSFRS-R number 6* = 4 ∧ *ALSFRS-R number 7* = 4 ∧ *ALSFRS-R total* = 46 ⇒ *Class* = Slow	≈ 3.38	≈ 3.15
e)	Neutral	*ALSFRS-R number 7* = 4 ∧ *Diagnostic delay* = [6, 12[months ∧ *Time gap between first medical observation and diagnosis* = [6, 9[months ⇒ *Class* = Neutral	≈ 2.75	≈ 2.49
e)	Fast	*Diagnostic delay* = [0, 6[months ∧ *Timing of transition from region 1 to 2* = [0, 3[months ∧ *Timing of transition from region 2 to 3* = [0, 3[months ⇒ *Class* = Fast	≈ 3.59	≈ 3.56
f)	Slow	*ALSFRS-R number 3* = 4 ∧ *ALSFRS-R number 4* = 4 ⇒ *Class* = Slow	≈ 17.3	≈ 2.46
f)	Neutral	*ALSFRS-R number 2* = 4 ∧ *ALSFRS-R number 5* = 3 ⇒ *Class* = Neutral	≈ 5.35	≈ 3.47
f)	Fast	*Diagnostic delay* = [0, 6[months ∧ *Timing of transition from region 1 to 2* = [0, 3[months ⇒ *Class* = Fast	≈ 7.73	≈ 3.28

Table 4.17: Most Relevant Class Association Rules for Task 2.

The rules obtained in the Baseline experiment (e)) showed higher Lift values for the Slow and Fast classes, while the discriminative Bicluster experiment (f)) showed higher values for Neutral and Fast classes. Regarding the pattern contents, these results confirm what was found in the previous section. For the Fast class, the *Diagnostic Delay* and *Timing of transition from region 1 to 2* features keep appearing regularly in the most relevant rules of the Fast class, implying a great characterizing power of this pattern. For the Slow class, the ALSFRS-R scale questions keep having the maximum value. Finally, for the Neutral class the *ALSFRS-R number 5* feature appears again with the value 3 in the f) experiment. According to these results and those of the previous section, the individual ALSFRS-R questions are important features to characterize, in particular, the Slow and Neutral progression groups. Furthermore, according to expert clinicians, these rules were considered clinically relevant and characterizing of their respective Patient subgroups.

Chapter 5

Conclusions and Future Work

In this thesis a new approach (DMD) based on Biclustering was proposed and used to find disease presentation patterns in the form of Meta-features, in order to characterize patients with ALS as a whole and to distinguish between patient progression groups in ALS.

More specifically, in Task 1 no Meta-features which distinguish the Portuguese ALS Patients from their Controls could be clearly identified. In retrospective, the complexity of this Task was far greater comparing to the second Task, being the subject of investigation of many renowned researchers with very limited success.

However, according to the Classification results using discriminative Biclusters was meaningful in Task 2, implying that considering subsets of features are relevant to find disease presentation patterns between the progression groups. Additionally, the most important individual features to distinguish between the groups were *Diagnostic Delay*, *Timing of transition from region 1 to 2* and *ALSFRS-R Total*.

The Association Rules considered more relevant by expert clinicians for each progression group showed high positive values of Lift, meaning that the Meta-features found on their antecedents were highly associated with each class. Moreover, parts of the rule's antecedents matched the patterns found on the most important Biclusters, indicating greater discriminative power.

These results were confirmed by the clinicians, who highlighted that ALS is an inevitably progressive disease, but the individual rate of clinical deterioration is quite variable. Furthermore, the *Diagnostic Delay*, i.e. the delay from first symptoms until the diagnosis is a recognized prognostic factor in ALS (with a shorter delay associated with faster progression) [66]. Moreover, it has been observed that shorter times of transition between body regions (disease spreading, in short periods of 0-3 months) are also associated with faster progression.

The individual ALSFRS-R questions assessing salivation (*ALSFRS-R number 2*), swallowing (*ALSFRS-R number 3*), handwriting (*ALSFRS-R number 4*), cutting food and handling utensils (*ALSFRS-R number 5*), dressing and personal hygiene (*ALSFRS-R number 6*) and turning on bed and adjusting bed clothes (*ALSFRS-R number 7*) were considered to be important features to characterize the Slow and Neutral progression groups.

In conclusion, the obtained results suggest that our Biclustering-based approach is a promising way to unravel disease presentation patterns and can be applied to similar problems and other diseases. Nonetheless, it is always possible to improve on a given solution, especially if it has identified problems.

In this work the obtained discriminative Biclusters tended to generally have a small number of rows. This happened due to the assumption that all rows in a Bicluster had to be exactly alike, not allowing for any deviation. By relaxing this restriction, taller Biclusters would most likely be found. Nonetheless, in that case the pattern of a Bicluster would have to be found in a different way (e.g. most frequent pattern, or the set of all present patterns), which would increase the complexity of the problem.

Additionally, the used RF models suffered from overfitting due to the high number of Biclusters from the experiments chosen. Thus, some form of filtering/sampling should have been applied to the obtained Biclusters on the to help diminishing that effect, e.g. using the Chi-Square statistic (in a similar way as in *Section 3.1.1 Data Pre-processing*) to choose a number of Biclusters most associated with each class.

Regarding future experiments, we would like to try other forms of Feature Selection with Task 1 data to see if it enhances the results. Additionally, we would like to verify if the alternative patient stratification mentioned in *Section 2.6.4 Progression Groups in ALS* would improve the results for the ALS patients in Task 2. Moreover, other Biclustering algorithms (if available) could be tested in place of BicPAM to check for general improvements.

As for new future approaches we would like to apply the DMD approach to other types of data (e.g. temporal data) by using other Biclustering algorithms with the ability to deal with the said data types. We would also like to suggest possible improvements to BicPAM to better support Biclustering over categorical data, to avoid having to perform the implemented translation steps, allowing to obtain interpretable Biclusters more easily. Finally, another idea was the creation of a complete pipeline, in the form of a full-fledged application like BiGGEsTS (an integrated environment for biclustering analysis of time series gene expression data) [19] that would allow the integrated treatment of the data and then the use of the several techniques employed by the DMD approach, including (if possible) a dynamic way to optimize the diverse thresholds used throughout the workflow.

References

[1] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 04 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq134. URL https://doi.org/10.1093/bioinformatics/btq134. 25, 26

[2] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324. 26

[3] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984. 18

[4] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013. 22, 23, 26, 46

[5] A. Carreiro, A. Ferreira, M. Figueiredo, and S. C . Madeira. Towards a classification approach using meta-biclustering: Impact of discretization in the analysis of expression time series. *Journal of integrative bioinformatics*, 9:207, 07 2012. doi: 10.2390/biecoll-jib-2012-207. 28, 35

[6] A. V. Carreiro, O. Anunciação, J. A. Carriço, and S. C. Madeira. Prognostic prediction through biclustering-based classification of clinical gene expression time series. *Journal of integrative bioinformatics*, 8(3):73–89, 2011. 28, 35

[7] Cedarbaum, Jesse M. and Stambler, Nancy and Fuller, Cynthia and Hilt, Dana and Thurmond, Barbara and Nakanishi, Arline and BDNF ALS Study Group (Phase III). The ALSFRS-R : a revised ALS Functional Rating Scale that incorporates assessments of respiratory function. *Journal of the Neurological Sciences*, 169:13–21, 1999. 6, 72

[8] K. J. Cios and G. William Moore. Uniqueness of medical data mining. *Artif. Intell. Med.*, 26 (1-2):1–24, Sept. 2002. ISSN 0933-3657. doi: 10.1016/S0933-3657(02)00049-0. URL http://dx.doi.org/10.1016/S0933-3657(02)00049-0. 1, 37

[9] de França, Fabrício Olivetti. A Hash-based Co-clustering Algorithm for Categorical Data. *Expert Systems with Applications*, 64:24–35, 2016. doi: 10.1016/j.eswa.2016.07.024. 34

[10] H. Deng, G. C. Runger, and E. Tuv. Bias of importance measures for multi-valued attributes and solutions. *Lecture Notes in Computer Science*, 6792:293–300, 06 2011. doi: 10.1007/978-3-642-21738-8_38. 18

[11] C. B. Do and S. Batzoglou. What is the expectation maximization algorithm? *Nature Biotechnology*, 26:897–899, 08 2008. doi: https://doi.org/10.1038/nbt1406. 34

[12] ENCALS Consortium. ONWebDUALS - JPND Research Project flyer. http://www.neurodegenerationresearch.eu/wp-content/uploads/2015/02/ONWebDUALS.pdf, 2015. [Online; accessed 22-01-2019]. 1, 2

[13] Es, Michael A. Van and Hardiman, Orla and Chiò, Adriano and Al-chalabi, Ammar and Pasterkamp, R. Jeroen and Veldink, Jan H. and Berg, Leonard H. Van Den. Amyotrophic Lateral Sclerosis. *The Lancet*, 390(10107):2084–2098, 2017. doi: 10.1016/S0140-6736(17)31287-4. 5, 6, 56

[14] M. S. Esfahani and E. R. Dougherty. Effect of separate sampling on classification accuracy. *Bioinformatics*, 30(2):242–250, 11 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt662. 22

[15] European Commission. ESCO (European Skills/Competences, qualifications and Occupations) portal, version 1.0.3. https://ec.europa.eu/esco/portal/occupation, 2018. [Online; accessed 22-01-2019]. 40

[16] Fillbrunn, Alexander and Dietz, Christian and Pfeuffer, Julianus and Rahn, René and Landrum, Gregory A. and Berthold, Michael R. KNIME for Reproducible Cross-domain Analysis of Life Science Data. *Journal of Biotechnology*, 261:149–156, 2017. doi: 10.1016/j.jbiotec.2017.07.028. 32, 38

[17] P. Fournier-Viger, C. Lin, A. Gomariz, T. Gueniche, A. Soltani, Z. Deng, and H. T. Lam. The spmf open-source data mining library version 2. In *Proc. 19th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD 2016) Part III*, pages 36–40. Springer LNCS 9853, 2016. 46

[18] Frank, Eibe and Hall, Mark A. and Witten, Ian H. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 4th edition, 2016. XV, 9, 14, 34, 38

[19] Gonçalves, Joana P. and Madeira, Sara C. and Oliveira, Arlindo L. BiGGEsTS: Integrated Environment for Biclustering Analysis of Time Series Gene Expression Data. *BMC Research Notes*, 2: 1–11, 2009. doi: 10.1186/1756-0500-2-124. 76

[20] Gromicho, M. and Pinto, S. and Figueiral, M. and Madeira, Sara C. and Andersen, P. M. and Grosskreutz, J. and Kuzma-Kozakiewicz, M. and Petri, S. and Szacka, K. and Stubendorff, B. and Uysal, H. and Swash, M. and de Carvalho, M. ONWebDUALS Consortium: Spreading in ALS. To Be Published, 2019. 34

[21] M. Hahsler, C. Buchta, B. Gruen, and K. Hornik. *arules: Mining Association Rules and Frequent Itemsets*, 2019. URL https://CRAN.R-project.org/package=arules. R package version 1.6-3. 29

[22] Han, Jiawei and Kamber, Micheline and Pei, Jian. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011. ISBN 0123814790, 9780123814791. XV, XVII, 7, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 29, 30, 32

[23] G. Heinze, C. Wallisch, and D. Dunkler. Variable selection – a review and recommendations for the practicing statistician. *Biometrical Journal*, 60(3):431–449, 2018. doi: 10.1002/bimj.201700067. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201700067. 31

[24] Henriques, Rui. *Learning from High-dimensional Data using Local Descriptive Models*. PhD thesis, Instituto Superior Técnico, Universidade de Lisboa, 2016. 9, 17

[25] Henriques, Rui and Antunes, Cláudia and Madeira, Sara C. A Structured View on Pattern Mining-based Biclustering. *Pattern Recognition*, 48(12):3941–3958, 2015. doi: 10.1016/j.patcog.2015.06.018. 7, 8, 14, 15, 16, 17, 26, 29, 33

[26] Henriques, Rui and Ferreira, Francisco L. and Madeira, Sara C. BicPAMS: Software for Biological Data Analysis with Pattern-based Biclustering. *BMC Bioinformatics*, 18(1):1–16, 2017. doi: 10.1186/s12859-017-1493-3. 34

[27] Henriques, Rui and Madeira, Sara C. BicPAM: Pattern-based Biclustering for Biomedical Data Analysis. *Algorithms for Molecular Biology*, 9(1):27, 2014. doi: 10.1186/s13015-014-0027-z. 9, 12, 32, 33

[28] Henriques, Rui and Madeira, Sara C. Triclustering Algorithms for Three-Dimensional Data Analysis: A Comprehensive Survey. *ACM Computing Surveys*, 51(5):95:1–95:43, 2018. URL http://doi.ieeecomputersociety.org/10.1145/3195833. 7, 8, 9, 13, 14, 16

[29] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014. ISBN 1461471370, 9781461471370. 18

[30] X. Jin, A. Xu, R. Bie, and P. Guo. Machine learning techniques and chi-square feature selection for cancer classification using sage gene expression profiles. In *Proceedings of the 2006*

International Conference on Data Mining for Biomedical Applications, BioDM'06, pages 106–115, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-33104-2, 978-3-540-33104-9. doi: 10.1007/11691730_11. URL http://dx.doi.org/10.1007/11691730_11. 33

[31] R. K Morgan, S. McNally, M. Alexander, R. Conroy, O. Hardiman, and R. Costello. Use of sniff nasal-inspiratory force to predict survival in amyotrophic lateral sclerosis. *American journal of respiratory and critical care medicine*, 171:269–274, 03 2005. doi: 10.1164/rccm.200403-314OC. 41

[32] Kiernan, Matthew C. and Vucic, Steve and Cheah, Benjamin C. and Turner, Martin R. and Eisen, Andrew and Hardiman, Orla and Burrell, James R. and Zoing, Margaret C. Amyotrophic Lateral Sclerosis. *The Lancet*, 377(9769):942–955, 2011. doi: 10.1016/S0140-6736(10)61156-7. 1, 5, 6

[33] R. Kueffner, N. Zach, M. Bronfeld, R. Norel, N. Atassi, V. Balagurusamy, B. Di Camillo, A. Chio, M. Cudkowicz, D. Dillenberger, et al. Stratification of amyotrophic lateral sclerosis patients: a crowdsourcing approach. *Scientific reports*, 9(1):690, 2019. 34

[34] Kuraszkiewicz, Bozenna and Podsiadły-Marczykowska, Teresa and Goszczyńska, Hanna and Piotrkiewicz, Maria. Are There Modifiable Environmental Factors Related to Amyotrophic Lateral Sclerosis? *Frontiers in Neurology*, 9(April):7–10, 2018. doi: 10.3389/fneur.2018.00220. 5, 34

[35] P. G. Laencina, J. L. Sancho-Gómez, and A. R. Figueiras-Vidal. Pattern classification with missing data: A review. *Neural Computing and Applications*, 19:263–282, 03 2010. doi: 10.1007/s00521-009-0295-6. 31, 32

[36] T. C. Landgrebe and R. P. Duin. Approximating the multiclass roc by pairwise analysis. *Pattern recognition letters*, 28(13):1747–1758, 2007. 21

[37] N. Lavrač, B. Cestnik, D. Gamberger, and P. Flach. Decision support through subgroup discovery: Three case studies and the lessons learned. *Machine Learning*, 57(1):115–143, Oct 2004. ISSN 1573-0565. doi: 10.1023/B:MACH.0000035474.48771.cd. 35

[38] G. Lemaître, F. Nogueira, and C. K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017. URL http://jmlr.org/papers/v18/16-365. 46

[39] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu. Feature selection: A data perspective. *ACM Comput. Surv.*, 50(6):94:1–94:45, Dec. 2017. ISSN 0360-0300. doi: 10.1145/3136625. URL http://doi.acm.org/10.1145/3136625. 31

[40] T. Li, C. Zhang, and M. Ogihara. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, 20(15):

2429–2437, 04 2004. ISSN 1367-4803. doi: 10.1093/bioinformatics/bth267. URL https://doi.org/10.1093/bioinformatics/bth267. 18

[41] F. Lofaso, F. Nicot, M. Lejaille, L. Falaize, A. Louis, A. Clement, J.-C. Raphael, D. Orlikowski, and B. Fauroux. Sniff nasal inspiratory pressure: what is the optimal number of sniffs? *European Respiratory Journal*, 27(5):980–982, 2006. ISSN 0903-1936. doi: 10.1183/09031936.06.00121305. URL https://erj.ersjournals.com/content/27/5/980. 41

[42] Logroscino, Giancarlo and Traynor, Bryan J. and Hardiman, Orla and Chiò, Adriano and Mitchell, Douglas and Swingler, Robert J. and Millul, Andrea and Benn, Emma and Beghi, Ettore. Incidence of Amyotrophic Lateral Sclerosis in Europe. *J Neurol Neurosurg Psychiatry*, pages 385–390, 2010. doi: 10.1136/jnnp.2009.183525. 6

[43] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts. Understanding variable importances in forests of randomized trees. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, pages 431–439, USA, 2013. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=2999611.2999660. 25, 26

[44] Madeira, Sara C. and Oliveira, Arlindo L. Biclustering Algorithms for Biological Data Analysis: A Survey. *Transactions on computational biology and bioinformatics*, 1(1):24–45, 2004. XV, 4, 8, 9, 10, 11, 12, 14, 17

[45] Minitab Inc. Minitab 18 Support site - Goodman Kruskal's Statistics. https://support.minitab.com/en-us/minitab/18/help-and-how-to/statistics/tables/supporting-topics/other-statistics-and-tests/what-are-the-goodman-kruskal-statistics/, 2017. [Online; accessed 22-01-2019]. 46

[46] T. M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997. ISBN 0070428077, 9780070428072. XV, 19

[47] Nezhad, Milad Zafar and Zhu, Dongxiao and Sadati, Najibesadat and Yang, Kai and Levi, Phillip. *SUBIC: A Supervised Biclustering Approach for Precision Medicine*, volume 2018-January. IEEE, 2018. doi: 10.1109/ICMLA.2017.00-68. 33

[48] Pensa, Ruggero G. and Robardet, Céline and Boulicaut, Jean-François. *A Bi-clustering Framework for Categorical Data*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005. ISBN 978-3-540-31665-7. 34, 46

[49] Piotrkiewicz, Maria and de Carvalho, Mamede and Grosskreutz, Julian and Kuzma-Kozakiewicz, Magdalena and Petri-Mals, Susanne and Podsiadły-Marczykowska, Teresa and Kuraszkiewicz, Bozenna and Goszczyńska, Hanna and Andersen, Peter M. ONWebDUALS ResearchGate Project page. https://www.researchgate.net/project/ONTology-based-Web-Database-

for-Understanding-Amyotrophic-Lateral-Sclerosis-OnWebDUALS, 2015. [Online; accessed 22-01-2019]. 2

[50] Pires, Sofia. A Supervised Learning Approach for Prognostic Prediction in ALS using Disease Progression Groups and Patient Profiles. Master's thesis, Faculdade de Ciências, Universidade de Lisboa, 2018. 34

[51] Pires, Sofia and Gromicho, Marta and De Carvalho, Mamede and Madeira, Sara C. Predicting Non-Invasive Ventilation in ALS Patients using Stratified Disease Progression Groups. *Proceedings of the sixth Workshop on Data Mining in Biomedical Informatics and Healthcare held in conjunction with IEEE International Conference on Data Mining (ICDM'18), Singapore*, 2018. 34

[52] J. R. Quinlan. Induction of decision trees. *Mach. Learn.*, 1(1):81–106, Mar. 1986. ISSN 0885-6125. doi: 10.1023/A:1022643204877. URL http://dx.doi.org/10.1023/A:1022643204877. 18

[53] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993. ISBN 1-55860-238-0. 18

[54] A. Rajaraman and J. D. Ullman. *Mining of Massive Datasets*. Cambridge University Press, New York, NY, USA, 2011. ISBN 1107015359, 9781107015357. 19

[55] S. Raschka. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *The Journal of Open Source Software*, 3(24), Apr. 2018. doi: 10.21105/joss.00638. URL http://joss.theoj.org/papers/10.21105/joss.00638. 26, 46

[56] V. J. Rayward-Smith. Statistics to measure correlation for data mining applications. *Comput. Stat. Data Anal.*, 51(8):3968–3982, May 2007. ISSN 0167-9473. doi: 10.1016/j.csda.2006.05.025. URL http://dx.doi.org/10.1016/j.csda.2006.05.025. 32

[57] Renton, Alan E. and Chiò, Adriano and Traynor, Bryan J. State of Play in Amyotrophic Lateral Sclerosis Genetics. *Nature Neuroscience*, 17(1):17–23, 2014. doi: 10.1038/nn.3584. 6

[58] L. Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1):1–39, Feb 2010. ISSN 1573-7462. doi: 10.1007/s10462-009-9124-7. URL https://doi.org/10.1007/s10462-009-9124-7. 24

[59] J. A. Sáez, B. Krawczyk, and M. Woźniak. Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. *Pattern Recognition*, 57:164–178, 2016. 25

[60] Santamaria, Rodrigo and Quintales, Luis and Therón, Roberto. Methods to Bicluster Validation and Comparison in Microarray Data. *Intelligent Data Engineering and Automated Learning*, 4881: 780–789, 12 2007. doi: 10.1007/978-3-540-77226-2_78. 13

[61] M. Steinbach, H. Yu, G. Fang, and V. Kumar. Using constraints to generate and explore higher order discriminative patterns. In *Advances in Knowledge Discovery and Data Mining*, pages 338–350, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-20841-6. 26

[62] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25, Jan 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-25. URL https://doi.org/10.1186/1471-2105-8-25. 25, 26

[63] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar. *Introduction to Data Mining (2nd Edition)*. Pearson, 2nd edition, 2018. ISBN 0133128903, 9780133128901. 26, 27

[64] K. S. (username StKyr). Multiscorer library GitHub site. https://github.com/StKyr/multiscorer, 2017. [Online; accessed 22-08-2019]. 66

[65] Veroneze, Rosana and J. Von Zuben, Fernando. Efficient Mining of Maximal Biclusters in Mixed-attribute Datasets. *CoRR*, abs/1710.03289, 2017. 8, 34, 40

[66] H.-J. Westeneng, T. P. Debray, A. E. Visser, R. P. van Eijk, J. P. Rooney, A. Calvo, S. Martin, C. J. McDermott, A. G. Thompson, S. 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. 71, 75

[67] L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pages 856–863. AAAI Press, 2003. ISBN 1-57735-189-4. 32

[68] L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. *J. Mach. Learn. Res.*, 5:1205–1224, Dec. 2004. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1005332.1044700. 30

[69] M. J. Zaki and C.-J. Hsiao. Charm: An efficient algorithm for closed itemset mining. In *Proceedings of the 2002 SIAM international conference on data mining*, pages 457–473. SIAM, 2002. 51, 62

[70] Y. Zhao. *R and Data Mining: Examples and Case Studies*. Academic Press, Elsevier, December 2012. ISBN 978-0-123-96963-7. URL http://www.rdatamining.com/docs/RDataMining-book.pdf. 29

[71] T. Zhu, Y. Lin, and Y. Liu. Synthetic minority oversampling technique for multiclass imbalance problems. *Pattern Recognition*, 72:327–340, 2017. 25

Appendix A

Discretized ONWebDUALS Dataset

Feature Name	Category Value	Category Label
Age (on date of consultation)	1	[0, 10[years
	2	[10, 20[years
	3	[20, 30[years
	4	[30, 40[years
	5	[40, 50[years
	6	[50, 60[years
	7	[60, 70[years
	8	[70, 80[years
	9	[80, 90[years
	10	>= 90 years
Age (1st Symptoms)	1	[0, 10[years
	2	[10, 20[years
	3	[20, 30[years
	4	[30, 40[years
	5	[40, 50[years
	6	[50, 60[years
	7	[60, 70[years
	8	[70, 80[years
	9	[80, 90[years
	10	>= 90 years
Gender	1	Female
	2	Male

Note: some features with the same categories were condensed into one to occupy less space.

Feature Name	Category Value	Category Label
Ethinicity	1	Caucasian
	2	African
	3	Asian
	4	Not feasible
Place of birth	1	Village < 1000 inhabitants
	2	Country towns [1000 - 5000]
	3	Small town [5000 - 20000]
	4	Middle town [20000 – 100000]
	5	Large town > 100000
Place of birth characteristics	1	Rural Area
	2	Urban Area
Mother's Place of birth characteristics	1	Rural Area
	2	Urban Area
Father's Place of birth characteristics	1	Rural Area
	2	Urban Area
Diagnostic delay	1	[0, 6[months
	2	[6, 12[months
	3	[12, 18[months
	4	[18, 24[months
	5	[24, 36[months
	6	[36, 48[months
	7	[48, 60[months
	8	[60, 72[months
	9	[72, 84[months
	10	[84, 96[months
	11	[96, 108[months
	12	[108, 120[months
	13	>= 120 months
Limbs onset	1	Yes
	2	No
Bulbar onset	1	Yes
	2	No
Neck onset	1	Yes
	2	No
Thoracic or Abdominal onset	1	Yes
	2	No

Feature Name	Category Value	Category Label
Respiratory onset	1	Yes
	2	No
Dyscognition onset	1	Yes
	2	No
Generalized onset	1	Yes
	2	No
UMN vs LMN manifestation at onset	1	UMN
	2	LMN
	3	UMN+LMN
Limb onset	1	No
	2	Upper limb
	3	Lower limb
	4	Upper and lower
Predominant side	1	Left
	2	Right
	3	Symmetric
Predominant impairment	1	Distal
	2	Proximal
	3	Distal and proximal
Fasciculations at onset	1	Yes
	2	No
Weight loss	1	Yes
	2	No
Emotional lability at onset	1	Yes
	2	No
Cognitive symptoms at onset	1	Yes
	2	No
Handedness	1	Left
	2	Right
	3	Ambidextrous
Tongue spasticity (Bulbar UMN)	1	Yes
	2	No
Jaw clonus (Bulbar UMN)	1	Yes
	2	No
Brisk jaw jerk (Bulbar UMN)	1	Yes
	2	No

Feature Name	Category Value	Category Label
Tongue atrophy (Bulbar LMN)	1	Yes
	2	No
Tongue fasciculations (Bulbar LMN)	1	Yes
	2	No
Weak orbicularis oris (Bulbar LMN)	1	Yes
	2	No
Facial muscle fasciculations (Bulbar LMN)	1	Yes
	2	No
Masseter atrophy (Bulbar LMN)	1	Yes
	2	No
Hyperreflexia (Upper Limbs UMN)	1	Yes
	2	No
Finger flexion or Hoffman sign (Upper Limbs UMN)	1	Yes
	2	No
Spasticity (Upper Limbs UMN)	1	Yes
	2	No
Atrophy and Weakness (Upper Limbs LMN)	1	Yes
	2	No
Fasciculations at onset (Upper Limbs LMN)	1	Yes
	2	No
Hyporeflexia (Upper Limbs LMN)	1	Yes
	2	No
Hyperreflexia (Lower Limbs UMN)	1	Yes
	2	No
Finger flexion or Babinskis sign (Lower Limbs UMN)	1	Yes
	2	No
Spasticity (Lower Limbs UMN)	1	Yes
	2	No
Atrophy and Weakness (Lower Limbs LMN)	1	Yes
	2	No
Fasciculations at onset (Lower Limbs LMN)	1	Yes
	2	No
Hyporeflexia (Lower Limbs LMN)	1	Yes
	2	No
Neck weakness	1	Yes
	2	No

Feature Name	Category Value	Category Label
Thoracic muscle fasciculations	1	Yes
	2	No
Resting respiratory fatigue	1	Yes
	2	No
Othopnea	1	Yes
	2	No
Paradoxical respiration	1	Yes
	2	No
Weak cough	1	Yes
	2	No
Diagnosis	1	Definite
	2	Probable
	3	Possible
	4	Probable-laboratory supported
	5	PMA
Emotional lability	1	Yes
	2	No
Cognition (Qualitative evaluation)	1	Normal
	2	Abnormal
Depression (Qualitative evaluation)	1	Normal
	2	Abnormal
1st (to 13th) region (Pattern of spreading)	1	1>
	2	1/
	3	2>
	4	2/
	5	2a>
	6	2a/
	7	2b>
	8	2b/
	9	3>
	10	3/
	11	3a>
	12	3a/
	13	3b>
	14	3b/
	15	4>

Feature Name	Category Value	Category Label
	16	4/
	17	4a>
	18	4a/
	19	4b>
	20	4b/
	21	5>
	22	5/
	23	5a>
	24	5a/
	25	5b>
	26	5b/
	27	6>
	28	6/
	29	7>
	30	7/
	31	8>
	32	8/
	33	9>
	34	9/
Region 1 (Region progression)	1	Bulbar
	2	Cervical
	3	Thoracic
	4	Lumbo-sacral
	5	Dyscognition
Region 2 (Region progression)	1	Bulbar
	2	Cervical
	3	Thoracic
	4	Lumbo-sacral
	5	Dyscognition
Timing of transition from Region 1 to 2	1	[0, 3[months
	2	[3, 6[months
	3	[6, 9[months
	4	[9, 12[months
	5	[12, 18[months
	6	[18, 24[months
	7	[24, 36[months
	8	[36, 48[months

Feature Name	Category Value	Category Label
	9	> 48 months
Region 3 (Region progression)	1	Bulbar
	2	Cervical
	3	Thoracic
	4	Lumbo-sacral
	5	Dyscognition
Timing of transition from Region 2 to 3	1	[0, 3[months
	2	[3, 6[months
	3	[6, 9[months
	4	[9, 12[months
	5	[12, 18[months
	6	[18, 24[months
	7	[24, 36[months
	8	[36, 48[months
	9	> 48 months
ALSFRS-R number 1 (to 12)	0	0
	1	1
	2	2
	3	3
	4	4
ALSFRS-R total	0	0
	1	1
	2	2
	3	3
	4	4
	5	5
	6	6
	7	7
	8	8
	9	9
	10	10
	11	11
	12	12
	13	13
	14	14
	15	15

Feature Name	Category Value	Category Label
	16	16
	17	17
	18	18
	19	19
	20	20
	21	21
	22	22
	23	23
	24	24
	25	25
	26	26
	27	27
	28	28
	29	29
	30	30
	31	31
	32	32
	33	33
	34	34
	35	35
	36	36
	37	37
	38	38
	39	39
	40	40
	41	41
	42	42
	43	43
	44	44
	45	45
	46	46
	47	47
	48	48
CK level	1	Above max value (bad)
	2	Below max value (good)
Albumin level	1	Below min value (bad)

Feature Name	Category Value	Category Label
	2	Above min value (good)
Creatinine level	1	Below min value (bad)
	2	Above min value (good)
Total Cholesterol level	1	Above max value (bad)
	2	Below max value (good)
HDL Cholesterol level	1	Below min value (bad)
	2	Above min value (good)
LDL Cholesterol level	1	Above max value (bad)
	2	Below max value (good)
Triglycerides level	1	Above max value (bad)
	2	Below max value (good)
Diagnostic EMG (bulbar region)	1	Normal
	2	Abnormal
Diagnostic EMG (upper limbs)	1	Normal
	2	Abnormal
Diagnostic EMG (lower limbs)	1	Normal
	2	Abnormal
Diagnostic EMG (Nerve conduction studies)	1	Yes
	2	No
Brain MRI	1	Normal
	2	Tumor
	3	Stroke
	4	Multiple sclerosis
	5	Trauma
	6	Hypoxia
Spinal cord MRI Cervical	1	Normal
	2	Tumor
	3	Stenosis
	4	Myelitis
	5	Trauma
	6	Syringomyelia
	7	Other
Spinal cord MRI Thoracic	1	Normal
	2	Tumor
	3	Stenosis
	4	Myelitis

Feature Name	Category Value	Category Label
	5	Trauma
	6	Syringomyelia
	7	Other
Spinal cord MRI Lumbosacral	1	Normal
	2	Tumor
	3	Stenosis
	4	Myelitis
	5	Trauma
	6	Syringomyelia
	7	Other
FVC (% predicted values)	1	< 40 % (very bad)
	2	[40, 60[% (bad)
	3	[60, 80[% (acceptable)
	4	> 80 % (good)
SNIP absolute value (cmH2O)	1	< 40 cmH2O (bad)
	2	[40, 60] cmH2O Female / [40, 70] cmH2O Male (acceptable)
	3	> 60 cmH2O Female / > 70 cmH2O Male (good)
Blood hypertension	1	Yes
	2	No
Diabetes type I or II	1	Yes
	2	No
Hypercholesterolemia	1	Yes
	2	No
Hypertriglyceridemia	1	Yes
	2	No
Hyperthyroidism	1	Yes
	2	No
Hypothyroidism	1	Yes
	2	No
Autoimmune rheumatologic disorder	1	Yes
	2	No
Autoimmune intestinal disorder	1	Yes
	2	No
Stroke	1	Ischemic

Feature Name	Category Value	Category Label
	2	Hemorrhagic
	3	Unknown
	4	No
Heart ischemia	1	Yes
	2	No
Heart arrhytmia	1	Yes
	2	No
Heart insufficiency	1	Yes
	2	No
Primary cancer	1	No
	2	Brain
	3	Neck
	4	Breast
	5	Lung
	6	Other Thoracic
	7	Gastric
	8	Colon
	9	Rectum
	10	Other Abdominal
	11	Prostate
	12	Uterus
	13	Other Pelvic
	14	Spine
	15	UL bone and sarcomas
	16	LL bone and sarcomas
	17	Skin
	18	Leukemia
	19	Lymphoma
	20	Other blood
Metastasis	1	No
	2	Brain
	3	Neck
	4	Breast
	5	Lung
	6	Other Thoracic
	7	Gastric

Feature Name	Category Value	Category Label
	8	Colon
	9	Rectum
	10	Other Abdominal
	11	Prostate
	12	Uterus
	13	Other Pelvic
	14	Spine
	15	UL bone and sarcomas
	16	LL bone and sarcomas
	17	Skin
	18	Leukemia
	19	Lymphoma
	20	Other blood
Cancer treatment (Chemotherapy)	1	Yes
	2	No
Cancer treatment (Radiotherapy)	1	Yes
	2	No
Cancer treatment (Radiotherapy of head or neck)	1	Yes
	2	No
Smoking	1	Smokes/smoked
	2	No
Stopped smoking	1	Yes
	2	No
Tobacco exposure (pack years)	1	[0, 25[% (Q1)
	2	[25, 50[% (Q2)
	3	[50, 75[% (Q3)
	4	> 75 % (Q4, not outlier)
	5	Mild outlier (> Q3 + 1 IQR)
	6	Extreme outlier (> Q3 + 3 IQR)
Psychiatric medication	1	Yes
	2	No
Supplements	1	Yes
	2	No
Riluzole	1	Yes
	2	No
Antiepileptc drug	1	Yes

Feature Name	Category Value	Category Label
	2	No
Statins	1	Yes
	2	No
NSAID	1	Yes
	2	No
Steroids	1	Yes
	2	No
Immunosupressors	1	Yes
	2	No
Abnormal C9orf72 repeat expansion	1	Yes
	2	No
Was there ALS in the family	1	Yes
	2	No
Mother (ALS in the family)	1	Yes
	2	No
Father (ALS in the family)	1	Yes
	2	No
Was there FTD in the family	1	Yes
	2	No
Mother (FTD in the family)	1	Yes
	2	No
Father (FTD in the family)	1	Yes
	2	No
Was there AD in the family	1	Yes
	2	No
Mother (AD in the family)	1	Yes
	2	No
Father (AD in the family)	1	Yes
	2	No
Was there PD in the family	1	Yes
	2	No
Mother (PD in the family)	1	Yes
	2	No
Father (PD in the family)	1	Yes
	2	No
Was there MS in the family	1	Yes

Feature Name	Category Value	Category Label
	2	No
Mother (MS in the family)	1	Yes
	2	No
Father (MS in the family)	1	Yes
	2	No
Other severe disease 1 of mother	1	01 Certain infectious or parasitic diseases
	2	02 Neoplasms
	3	03 Diseases of the blood or blood-forming organs
	4	04 Diseases of the immune system
	5	05 Endocrine, nutritional or metabolic diseases
	6	06 Mental, behavioural or neurodevelopmental disorders
	7	07 Sleep-wake disorders
	8	08 Diseases of the nervous system
	9	09 Diseases of the visual system
	10	10 Diseases of the ear or mastoid process
	11	11 Diseases of the circulatory system
	12	12 Diseases of the respiratory system
	13	13 Diseases of the digestive system
	14	14 Diseases of the skin
	15	15 Diseases of the musculoskeletal system or connective tissue
	16	16 Diseases of the genitourinary system
	17	17 Conditions related to sexual health
	18	18 Pregnancy, childbirth or the puerperium
	19	19 Certain conditions originating in the perinatal period
	20	20 Developmental anomalies
	21	21 Symptoms, signs or clinical findings, not elsewhere classified

Feature Name	Category Value	Category Label
	22	22 Injury, poisoning or certain other consequences of external causes
	23	23 External causes of morbidity or mortality
	24	24 Factors influencing health status or contact with health services
	25	25 Codes for special purposes
	26	26 Traditional Medicine conditions - Module I
Other severe disease 1 of father	1	01 Certain infectious or parasitic diseases
	2	02 Neoplasms
	3	03 Diseases of the blood or blood-forming organs
	4	04 Diseases of the immune system
	5	05 Endocrine, nutritional or metabolic diseases
	6	06 Mental, behavioural or neurodevelopmental disorders
	7	07 Sleep-wake disorders
	8	08 Diseases of the nervous system
	9	09 Diseases of the visual system
	10	10 Diseases of the ear or mastoid process
	11	11 Diseases of the circulatory system
	12	12 Diseases of the respiratory system
	13	13 Diseases of the digestive system
	14	14 Diseases of the skin
	15	15 Diseases of the musculoskeletal system or connective tissue
	16	16 Diseases of the genitourinary system
	17	17 Conditions related to sexual health
	18	18 Pregnancy, childbirth or the puerperium
	19	19 Certain conditions originating in the perinatal period

Feature Name	Category Value	Category Label
	20	20 Developmental anomalies
	21	21 Symptoms, signs or clinical findings, not elsewhere classified
	22	22 Injury, poisoning or certain other consequences of external causes
	23	23 External causes of morbidity or mortality
	24	24 Factors influencing health status or contact with health services
	25	25 Codes for special purposes
	26	26 Traditional Medicine conditions - Module I
Mother alive	1	Yes
	2	No
Mother's cause of death	1	01 Certain infectious or parasitic diseases
	2	02 Neoplasms
	3	03 Diseases of the blood or blood-forming organs
	4	04 Diseases of the immune system
	5	05 Endocrine, nutritional or metabolic diseases
	6	06 Mental, behavioural or neurodevelopmental disorders
	7	07 Sleep-wake disorders
	8	08 Diseases of the nervous system
	9	09 Diseases of the visual system
	10	10 Diseases of the ear or mastoid process
	11	11 Diseases of the circulatory system
	12	12 Diseases of the respiratory system
	13	13 Diseases of the digestive system
	14	14 Diseases of the skin
	15	15 Diseases of the musculoskeletal system or connective tissue
	16	16 Diseases of the genitourinary system

Feature Name	Category Value	Category Label
	17	17 Conditions related to sexual health
	18	18 Pregnancy, childbirth or the puerperium
	19	19 Certain conditions originating in the perinatal period
	20	20 Developmental anomalies
	21	21 Symptoms, signs or clinical findings, not elsewhere classified
	22	22 Injury, poisoning or certain other consequences of external causes
	23	23 External causes of morbidity or mortality
	24	24 Factors influencing health status or contact with health services
	25	25 Codes for special purposes
	26	26 Traditional Medicine conditions - Module I
Father alive	1	Yes
	2	No
Father's cause of death	1	01 Certain infectious or parasitic diseases
	2	02 Neoplasms
	3	03 Diseases of the blood or blood-forming organs
	4	04 Diseases of the immune system
	5	05 Endocrine, nutritional or metabolic diseases
	6	06 Mental, behavioural or neurodevelopmental disorders
	7	07 Sleep-wake disorders
	8	08 Diseases of the nervous system
	9	09 Diseases of the visual system
	10	10 Diseases of the ear or mastoid process
	11	11 Diseases of the circulatory system
	12	12 Diseases of the respiratory system
	13	13 Diseases of the digestive system

Feature Name	Category Value	Category Label
	14	14 Diseases of the skin
	15	15 Diseases of the musculoskeletal system or connective tissue
	16	16 Diseases of the genitourinary system
	17	17 Conditions related to sexual health
	18	18 Pregnancy, childbirth or the puerperium
	19	19 Certain conditions originating in the perinatal period
	20	20 Developmental anomalies
	21	21 Symptoms, signs or clinical findings, not elsewhere classified
	22	22 Injury, poisoning or certain other consequences of external causes
	23	23 External causes of morbidity or mortality
	24	24 Factors influencing health status or contact with health services
	25	25 Codes for special purposes
	26	26 Traditional Medicine conditions - Module I
Regular Physical Exercise	1	Yes
	2	No
Intense (mod. or vig.) Physical Exercise	1	Yes
	2	No
Mild Physical Exercise	1	Yes
	2	No
Head or Neck trauma in the last 5 years	1	Head/Neck
	2	No
Head or Neck trauma more than 5 years ago	1	Head/Neck
	2	No
Other Cervical trauma in the last 5 years	1	Yes
	2	No
Other Thoracic trauma in the last 5 years	1	Yes
	2	No

Feature Name	Category Value	Category Label
Other Lumbosacral trauma in the last 5 years	1	Yes
	2	No
Other Cervical trauma more than 5 years ago	1	Yes
	2	No
Other Thoracic trauma more than 5 years ago	1	Yes
	2	No
Other Lumbosacral trauma more than 5 years ago	1	Yes
	2	No
Cervical Spine surgery in the last 5 years	1	Yes
	2	No
Thoracic Spine surgery in the last 5 years	1	Yes
	2	No
LumboSacral Spine surgery in the last 5 years	1	Yes
	2	No
Cervical Spine surgery more than 5 years ago	1	Yes
	2	No
Thoracic Spine surgery more than 5 years ago	1	Yes
	2	No
LumboSacral Spine surgery more than 5 years ago	1	Yes
	2	No
Upper Limb surgery in the last years	1	Yes
	2	No
Lower Limb surgery in the last 5 years	1	Yes
	2	No
Upper Limb surgery more than 5 years ago	1	Yes
	2	No
Lower Limb surgery more than 5 years ago	1	Yes
	2	No
Abdominal surgery in the last 5 years	1	Yes
	2	No
Abdominal surgery more than 5 years ago	1	Yes
	2	No
Thoracic surgery in the last 5 years	1	Yes
	2	No
Thoracic surgery more than 5 years ago	1	Yes
	2	No

Feature Name	Category Value	Category Label
Pelvic Surgery in the last 5 years	1	Yes
	2	No
Pelvic Surgery more than 5 years ago	1	Yes
	2	No
Head or neck Surgery in the last 5 years	1	Yes
	2	No
Head or neck Surgery more than 5 years ago	1	Yes
	2	No
Main Occupation in the last 5 years (or more than 5 years ago) (level 1)	01	01 Commissioned armed forces officers
	02	02 Non-commissioned armed forces officers
	03	03 Armed forces occupations, other ranks
	10	10 Pensioner / Out of job
	11	11 Chief executives, senior officials and legislators
	12	12 Administrative and commercial managers
	13	13 Production and specialised services managers
	14	14 Hospitality, retail and other services managers
	21	21 Science and engineering professionals
	22	22 Health professionals
	23	23 Teaching professionals
	24	24 Business and administration professionals
	25	25 Information and communications technology professionals
	26	26 Legal, social and cultural professionals
	31	31 Science and engineering associate professionals
	32	32 Health associate professionals

Feature Name	Category Value	Category Label
	33	33 Business and administration associate professionals
	34	34 Legal, social, cultural and related associate professionals
	35	35 Information and communications technicians
	41	41 General and keyboard clerks
	42	42 Customer services clerks
	43	43 Numerical and material recording clerks
	44	44 Other clerical support workers
	51	51 Personal service workers
	52	52 Sales workers
	53	53 Personal care workers
	54	54 Protective services workers
	61	61 Market-oriented skilled agricultural workers
	62	62 Market-oriented skilled forestry, fishery and hunting workers
	63	63 Subsistence farmers, fishers, hunters and gatherers
	71	71 Building and related trades workers, excluding electricians
	72	72 Metal, machinery and related trades workers
	73	73 Handicraft and printing workers
	74	74 Electrical and electronic trades workers
	75	75 Food processing, wood working, garment and other craft and related trades workers
	81	81 Stationary plant and machine operators
	82	82 Assemblers
	83	83 Drivers and mobile plant operator
	91	91 Cleaners and helpers

Feature Name	Category Value	Category Label
	92	92 Agricultural, forestry and fishery labourers
	93	93 Labourers in mining, construction, manufacturing and transport
	94	94 Food preparation assistants
	95	95 Street and related sales and service workers
	96	96 Refuse workers and other elementary workers
Second Occupation in the last 5 years (or more than 5 years ago) (level 1)	01	01 Commissioned armed forces officers
	02	02 Non-commissioned armed forces officers
	03	03 Armed forces occupations, other ranks
	10	10 Pensioner / Out of job
	11	11 Chief executives, senior officials and legislators
	12	12 Administrative and commercial managers
	13	13 Production and specialised services managers
	14	14 Hospitality, retail and other services managers
	21	21 Science and engineering professionals
	22	22 Health professionals
	23	23 Teaching professionals
	24	24 Business and administration professionals
	25	25 Information and communications technology professionals
	26	26 Legal, social and cultural professionals
	31	31 Science and engineering associate professionals
	32	32 Health associate professionals

Feature Name	Category Value	Category Label
	33	33 Business and administration associate professionals
	34	34 Legal, social, cultural and related associate professionals
	35	35 Information and communications technicians
	41	41 General and keyboard clerks
	42	42 Customer services clerks
	43	43 Numerical and material recording clerks
	44	44 Other clerical support workers
	51	51 Personal service workers
	52	52 Sales workers
	53	53 Personal care workers
	54	54 Protective services workers
	61	61 Market-oriented skilled agricultural workers
	62	62 Market-oriented skilled forestry, fishery and hunting workers
	63	63 Subsistence farmers, fishers, hunters and gatherers
	71	71 Building and related trades workers, excluding electricians
	72	72 Metal, machinery and related trades workers
	73	73 Handicraft and printing workers
	74	74 Electrical and electronic trades workers
	75	75 Food processing, wood working, garment and other craft and related trades workers
	81	81 Stationary plant and machine operators
	82	82 Assemblers
	83	83 Drivers and mobile plant operator
	91	91 Cleaners and helpers

Feature Name	Category Value	Category Label
	92	92 Agricultural, forestry and fishery labourers
	93	93 Labourers in mining, construction, manufacturing and transport
	94	94 Food preparation assistants
	95	95 Street and related sales and service workers
	96	96 Refuse workers and other elementary workers
Main Occupation in the last 5 years (or more than 5 years ago) (level 2)	0	0 Armed forces occupations
	1	1 Managers
	2	2 Professionals
	3	3 Technicians and associate professionals
	4	4 Clerical support workers
	5	5 Service and sales workers
	6	6 Skilled agricultural, forestry and fishery workers
	7	7 Craft and related trades workers
	8	8 Plant and machine operators and assemblers
	9	9 Elementary occupations
	10	10 Pensioner / Out of job
Second Occupation in the last 5 years (or more than 5 years ago) (level 2)	0	0 Armed forces occupations
	1	1 Managers
	2	2 Professionals
	3	3 Technicians and associate professionals
	4	4 Clerical support workers
	5	5 Service and sales workers
	6	6 Skilled agricultural, forestry and fishery workers
	7	7 Craft and related trades workers
	8	8 Plant and machine operators and assemblers
	9	9 Elementary occupations

Feature Name	Category Value	Category Label
	10	10 Pensioner / Out of job
Place of living in the last 5 years (or more than 5 years ago)	1	Village < 1000 inhabitants
	2	Country towns [1000 - 5000]
	3	Small town [5000 - 20000]
	4	Middle town [20000 – 100000]
	5	Large town > 100000
Lived in Rural Area or Urban in the last 5 years (or more than 5 years ago)	1	Rural area
	2	Urban area
ALS cases in the neighborhood	1	Yes
	2	No
ALS disease in coworkers	1	Yes
	2	No
ALS disease in friends	1	Yes
	2	No
Time gap between first medical observation and diagnosis	1	[0, 3[months
	2	[3, 6[months
	3	[6, 9[months
	4	[9, 12[months
	5	[12, 18[months
	6	[18, 24[months
	7	[24, 36[months
	8	[36, 48[months
	9	> 48 months
End of Table		

Appendix B

ONWebDUALS dataset features for Task 1

Feature Name	a) Baseline with All Features	b) Baseline FS
Age (on date of consultation)	Yes	Yes
Gender	Yes	Yes
Ethinicity	Yes	
Place of birth	Yes	Yes
Place of birth characteristics	Yes	
Mother's Place of birth characteristics	Yes	Yes
Father's Place of birth characteristics	Yes	
Blood hypertension	Yes	Yes
Diabetes type I or II	Yes	
Hypercholesterolemia	Yes	Yes
Hypertriglyceridemia	Yes	
Hyperthyroidism	Yes	
Hypothyroidism	Yes	
Autoimmune rheumatologic disorder	Yes	
Autoimmune intestinal disorder	Yes	
Stroke	Yes	
Heart ischemia	Yes	
Heart arrhytmia	Yes	
Heart insufficiency	Yes	
Primary cancer	Yes	Yes
Metastasis	Yes	
Cancer treatment (Chemotherapy)	Yes	
Cancer treatment (Radiotherapy)	Yes	

Feature Name	a) Baseline w/All Features	b) Baseline FS
Cancer treatment (Radiotherapy of head or neck)	Yes	
Smoking	Yes	
Stopped smoking	Yes	
Tobacco exposure (pack years)	Yes	
Psychiatric medication	Yes	Yes
Supplements	Yes	
Antiepileptc drug	Yes	
Statins	Yes	
NSAID	Yes	Yes
Steroids	Yes	
Immunosupressors	Yes	
Was there ALS in the family	Yes	Yes
Mother (ALS in the family)	Yes	
Father (ALS in the family)	Yes	
Was there FTD in the family	Yes	
Mother (FTD in the family)	Yes	
Father (FTD in the family)	Yes	
Was there AD in the family	Yes	
Mother (AD in the family)	Yes	
Father (AD in the family)	Yes	
Was there PD in the family	Yes	
Mother (PD in the family)	Yes	
Father (PD in the family)	Yes	
Was there MS in the family	Yes	
Mother (MS in the family)	Yes	
Father (MS in the family)	Yes	
Other severe disease 1 of mother	Yes	Yes
Other severe disease 1 of father	Yes	
Mother alive	Yes	Yes
Mother's cause of death	Yes	Yes
Father alive	Yes	
Father's cause of death	Yes	Yes
Regular Physical Exercise	Yes	Yes
Intense (mod. or vig.) Physical Exercise	Yes	

Feature Name	a) Baseline w/All Features	b) Baseline FS
Mild Physical Exercise	Yes	
Head or Neck trauma in the last 5 years	Yes	
Head or Neck trauma more than 5 years ago	Yes	
Other Cervical trauma in the last 5 years	Yes	
Other Thoracic trauma in the last 5 years	Yes	
Other Lumbosacral trauma in the last 5 years	Yes	
Other Cervical trauma more than 5 years ago	Yes	
Other Thoracic trauma more than 5 years ago	Yes	
Other Lumbosacral trauma more than 5 years ago	Yes	
Cervical Spine surgery in the last 5 years	Yes	
Thoracic Spine surgery in the last 5 years	Yes	Yes
LumboSacral Spine surgery in the last 5 years	Yes	
Cervical Spine surgery more than 5 years ago	Yes	
Thoracic Spine surgery more than 5 years ago	Yes	
LumboSacral Spine surgery more than 5 years ago	Yes	
Upper Limb surgery in the last years	Yes	
Lower Limb surgery in the last 5 years	Yes	
Upper Limb surgery more than 5 years ago	Yes	
Lower Limb surgery more than 5 years ago	Yes	
Abdominal surgery in the last 5 years	Yes	
Abdominal surgery more than 5 years ago	Yes	
Thoracic surgery in the last 5 years	Yes	
Thoracic surgery more than 5 years ago	Yes	
Pelvic Surgery in the last 5 years	Yes	
Pelvic Surgery more than 5 years ago	Yes	
Head or neck Surgery in the last 5 years	Yes	
Head or neck Surgery more than 5 years ago	Yes	
Main Occupation in the last 5 years (level 1)	Yes	Yes
Main Occupation more than 5 years ago (level 1)	Yes	Yes
Second Occupation in the last 5 years (level 1)	Yes	
Second Occupation more than 5 years ago (level 1)	Yes	

Feature Name	a) Baseline w/All Features	b) Baseline FS
Main Occupation in the last 5 years (level 2)	Yes	
Main Occupation more than 5 years ago (level 2)	Yes	
Second Occupation in the last 5 years (level 2)	Yes	
Second Occupation more than 5 years ago (level 2)	Yes	
Place of living in the last 5 years	Yes	Yes
Lived in Rural Area or Urban in the last 5 years	Yes	
Place of living more than 5 years ago	Yes	
Lived in Rural Area or Urban more than 5 years ago	Yes	Yes
ALS cases in the neighborhood	Yes	
ALS disease in coworkers	Yes	
ALS disease in friends	Yes	
End of Table		

Appendix C

ONWebDUALS dataset features for Task 2

Feature Name	a) Baseline with All Features	b) Baseline FS
Age (1st Symptoms)	Yes	Yes
Age (on date of consultation)	Yes	
Gender	Yes	Yes
Ethinicity	Yes	
Place of birth	Yes	
Place of birth characteristics	Yes	
Mother's Place of birth characteristics	Yes	
Father's Place of birth characteristics	Yes	
Diagnostic delay	Yes	Yes
Limbs onset	Yes	
Bulbar onset	Yes	
Neck onset	Yes	
Thoracic or Abdominal onset	Yes	
Respiratory onset	Yes	
Dyscognition onset	Yes	
Generalized onset	Yes	
UMN vs LMN manifestation at onset	Yes	
Limb onset	Yes	
Predominant side	Yes	
Predominant impairment	Yes	
Fasciculations at onset	Yes	
Weight loss	Yes	
Emotional lability at onset	Yes	

Feature Name	a) Baseline w/All Features	b) Baseline FS
Cognitive symptoms at onset	Yes	
Handedness	Yes	
Tongue spasticity (Bulbar UMN)	Yes	
Jaw clonus (Bulbar UMN)	Yes	
Brisk jaw jerk (Bulbar UMN)	Yes	
Tongue atrophy (Bulbar LMN)	Yes	
Tongue fasciculations (Bulbar LMN)	Yes	
Weak orbicularis oris (Bulbar LMN)	Yes	
Facial muscle fasciculations (Bulbar LMN)	Yes	
Masseter atrophy (Bulbar LMN)	Yes	
Hyperreflexia (Upper Limbs UMN)	Yes	
Finger flexion or Hoffman sign (Upper Limbs UMN)	Yes	
Spasticity (Upper Limbs UMN)	Yes	
Atrophy and Weakness (Upper Limbs LMN)	Yes	
Fasciculations at onset (Upper Limbs LMN)	Yes	
Hyporeflexia (Upper Limbs LMN)	Yes	
Hyperreflexia (Lower Limbs UMN)	Yes	
Finger flexion or Babinskis sign (Lower Limbs UMN)	Yes	
Spasticity (Lower Limbs UMN)	Yes	
Atrophy and Weakness (Lower Limbs LMN)	Yes	
Fasciculations at onset (Lower Limbs LMN)	Yes	
Hyporeflexia (Lower Limbs LMN)	Yes	
Neck weakness	Yes	
Thoracic muscle fasciculations	Yes	
Resting respiratory fatigue	Yes	
Othopnea	Yes	
Paradoxical respiration	Yes	
Weak cough	Yes	
Diagnosis	Yes	
Emotional lability	Yes	
Cognition (Qualitative evaluation)	Yes	
Depression (Qualitative evaluation)	Yes	
1st region (Pattern of spreading)	Yes	Yes

Feature Name	a) Baseline w/All Features	b) Baseline FS
2nd region (Pattern of spreading)	Yes	
3rd region (Pattern of spreading)	Yes	Yes
4th region (Pattern of spreading)	Yes	Yes
5th region (Pattern of spreading)	Yes	Yes
6th region (Pattern of Spreading)	Yes	
7th region (Pattern of spreading)	Yes	
8th region (Pattern of spreading)	Yes	
9th region (Pattern of spreading)	Yes	
10th region (Pattern of spreading)	Yes	
11th region (Pattern of spreading)	Yes	
12th region (Pattern of spreading)	Yes	
13th region (Pattern of spreading)	Yes	
Region 1 (Region progression)	Yes	
Region 2 (Region progression)	Yes	
Timing of transition from region 1 to 2	Yes	Yes
Region 3 (Region progression)	Yes	Yes
Timing of transition from region 2 to 3	Yes	Yes
ALSFRSR number 1	Yes	Yes
ALSFRSR number 2	Yes	Yes
ALSFRSR number 3	Yes	Yes
ALSFRSR number 4	Yes	Yes
ALSFRSR number 5	Yes	Yes
ALSFRSR number 6	Yes	Yes
ALSFRSR number 7	Yes	Yes
ALSFRSR number 8	Yes	Yes
ALSFRSR number 9	Yes	Yes
ALSFRSR number 10	Yes	
ALSFRSR number 11	Yes	
ALSFRSR number 12	Yes	
ALSFRSR total	Yes	Yes
CK level	Yes	
Albumin level	Yes	
Creatinine level	Yes	
Total Cholesterol level	Yes	
HDL Cholesterol level	Yes	

Feature Name	a) Baseline w/All Features	b) Baseline FS
LDL Cholesterol level	Yes	
Triglycerides level	Yes	
Diagnostic EMG (bulbar region)	Yes	
Diagnostic EMG (upper limbs)	Yes	
Diagnostic EMG (lower limbs)	Yes	
Diagnostic EMG (Nerve conduction studies)	Yes	
Brain MRI	Yes	
Spinal cord MRI Cervical	Yes	
Spinal cord MRI Thoracic	Yes	
Spinal cord MRI Lumbosacral	Yes	
FVC (% predicted values)	Yes	
SNIP absolute value (cmH2O)	Yes	
Blood hypertension	Yes	
Diabetes type I or II	Yes	
Hypercholesterolemia	Yes	
Hypertriglyceridemia	Yes	
Hyperthyroidism	Yes	
Hypothyroidism	Yes	
Autoimmune rheumatologic disorder	Yes	
Autoimmune intestinal disorder	Yes	
Stroke	Yes	
Heart ischemia	Yes	
Heart arrhytmia	Yes	
Heart insufficiency	Yes	
Primary cancer	Yes	
Metastasis	Yes	
Cancer treatment (Chemotherapy)	Yes	
Cancer treatment (Radiotherapy)	Yes	
Cancer treatment (Radiotherapy of head or neck)	Yes	
Smoking	Yes	
Stopped smoking	Yes	
Tobacco exposure (pack-years)	Yes	
Psychiatric medication	Yes	
Supplements	Yes	

Feature Name	a) Baseline w/All Features	b) Baseline FS
Riluzole	Yes	
Antiepileptc drug	Yes	
Statins	Yes	
NSAID	Yes	
Steroids	Yes	
Immunosupressors	Yes	
Abnormal C9orf72 repeat expansion	Yes	
Was there ALS in the family	Yes	
Mother (ALS in the family)	Yes	
Father (ALS in the family)	Yes	
Was there FTD in the family	Yes	
Mother (FTD in the family)	Yes	
Father (FTD in the family)	Yes	
Was there AD in the family	Yes	
Mother (AD in the family)	Yes	
Father (AD in the family)	Yes	
Was there PD in the family	Yes	
Mother (PD in the family)	Yes	
Father (PD in the family)	Yes	
Was there MS in the family	Yes	
Mother (MS in the family)	Yes	
Father (MS in the family)	Yes	
Other severe disease 1 of mother	Yes	
Other severe disease 1 of father	Yes	
Mother alive	Yes	
Mother’s cause of death	Yes	
Father alive	Yes	
Father’s cause of death	Yes	
Regular Physical Exercise	Yes	
Intense (mod. or vig.) Physical Exercise	Yes	
Mild Physical Exercise	Yes	
Head or Neck trauma in the last 5 years	Yes	
Head or Neck trauma more than 5 years ago	Yes	
Other Cervical trauma in the last 5 years	Yes	
Other Thoracic trauma in the last 5 years	Yes	

Feature Name	a) Baseline w/All Features	b) Baseline FS
Other Lumbosacral trauma in the last 5 years	Yes	
Other Cervical trauma more than 5 years ago	Yes	
Other Thoracic trauma more than 5 years ago	Yes	
Other Lumbosacral trauma more than 5 years ago	Yes	
Cervical Spine surgery in the last 5 years	Yes	
Thoracic Spine surgery in the last 5 years	Yes	
LumboSacral Spine surgery in the last 5 years	Yes	
Cervical Spine surgery more than 5 years ago	Yes	
Thoracic Spine surgery more than 5 years ago	Yes	
LumboSacral Spine surgery more than 5 years ago	Yes	
Upper Limb surgery in the last years	Yes	
Lower Limb surgery in the last 5 years	Yes	
Upper Limb surgery more than 5 years ago	Yes	
Lower Limb surgery more than 5 years ago	Yes	
Abdominal surgery in the last 5 years	Yes	
Abdominal surgery more than 5 years ago	Yes	
Thoracic surgery in the last 5 years	Yes	
Thoracic surgery more than 5 years ago	Yes	
Pelvic Surgery in the last 5 years	Yes	
Pelvic Surgery more than 5 years ago	Yes	
Head or neck Surgery in the last 5 years	Yes	
Head or neck Surgery more than 5 years ago	Yes	
Main Occupation in the last 5 years (level 1)	Yes	Yes
Main Occupation more than 5 years ago (level 1)	Yes	
Second Occupation in the last 5 years (level 1)	Yes	
Second Occupation more than 5 years ago (level 1)	Yes	
Main Occupation in the last 5 years (level 2)	Yes	
Main Occupation more than 5 years ago (level 2)	Yes	
Second Occupation in the last 5 years (level 2)	Yes	

Feature Name	a) Baseline w/All Features	b) Baseline FS
Second Occupation more than 5 years ago (level 2)	Yes	
Place of living in the last 5 years	Yes	
Lived in Rural Area or Urban in the last 5 years	Yes	
Place of living more than 5 years ago	Yes	
Lived in Rural Area or Urban more than 5 years ago	Yes	
ALS cases in the neighborhood	Yes	
ALS disease in coworkers	Yes	
ALS disease in friends	Yes	
Time gap between first medical observation and diagnosis	Yes	Yes
End of Table		

Appendix D

GitHub Repository

Due to space limitations, the data pre-processing workflows, code and additional results produced in the context of this thesis were made available online (GitHub Repository).

D.1 *DMD_approach* GitHub Repository

- **GitHub Repository Link**

D.2 Task 1 Additional Results

Due to space limitations, these results were made available online (GitHub Repository). For Appendices D.2.1, D.2.2 and D.2.3 the Bicluster patterns' format is different from what was seen in the main document: the first line has the features' names and the line below contains the values.

D.2.1 Discriminative Bicluster Patterns

- **GitHub Repository Link**

D.2.2 Top-30 Most Important Bicluster Patterns - c) Matrix Subject ID x Biclusters

- **GitHub Repository Link**

D.2.3 Top-30 Most Important Bicluster Patterns - d) Merged Data

- **GitHub Repository Link**

D.2.4 Non-Redundant Class Association Rules

- **GitHub Repository Link**

D.3 Task 2 Additional Results

Same as above, the Appendices D.3.1, D.3.2 and D.3.3 the Bicluster patterns' format is different from what was seen in the main document: the first line has the features' names and the line below contains the values.

D.3.1 Discriminative Bicluster Patterns

- **GitHub Repository Link**

D.3.2 Top-30 Most Important Bicluster Patterns - c) Matrix Subject ID x Biclusters

- **GitHub Repository Link**

D.3.3 Top-30 Most Important Bicluster Patterns - d) Merged Data

- **GitHub Repository Link**

D.3.4 Non-Redundant Class Association Rules

- **GitHub Repository Link**