UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

JULIANA BARCELLOS MATTOS

A SUPERVISED DESCRIPTIVE LOCAL PATTERN MINING APPROACH TO THE DISCOVERY OF SUBGROUPS WITH EXCEPTIONAL SURVIVAL BEHAVIOUR

Recife

2021

JULIANA BARCELLOS MATTOS

A SUPERVISED DESCRIPTIVE LOCAL PATTERN MINING APPROACH TO THE DISCOVERY OF SUBGROUPS WITH EXCEPTIONAL SURVIVAL BEHAVIOUR

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Inteligência Computacional

Orientador (a): Renato Vimieiro

Coorientador (a): Paulo Salgado Gomes de Mattos Neto

Recife

2021

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

M444s Mattos, Juliana Barcellos

A supervised descriptive local pattern mining approach to the discovery of subgroups with exceptional survival behaviour / Juliana Barcellos Mattos. – 2021.

122 f.: il., fig., tab.

Orientador: Renato Vimieiro.

Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2021.

Inclui referências e apêndice.

1. Inteligência computacional. 2. Mineração de modelos. I. Vimieiro, Renato (orientador). II. Título.

006.31 CDD (23. ed.) UFPE - CCEN 2022 – 52

JULIANA BARCELLOS MATTOS

“A SUPERVISED DESCRIPTIVE LOCAL PATTERN MINING APPROACH TO THE DISCOVERY OF SUBGROUPS WITH EXCEPTIONAL SURVIVAL BEHAVIOUR”

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação. Área de Concentração: Inteligência Computacional

Aprovado em: 10/12/2021.

____

Co-orientador: Prof. Dr. Paulo Salgado Gomes de Mattos Neto

BANCA EXAMINADORA

____

Profa. Dra. Teresa Bernarda Ludermir
Centro de Informática / UFPE

____

Profa. Dra. Gisele Lobo Pappa
Departamento de Ciência da Computação / UFMG

____

Prof. Dr. Renato Vimieiro
Departamento de Ciência da Computação / UFMG
(Orientador)

To the ones that have paid the price of hard medical interventions for the sake of surviving.

To the ones that have struggled with medical uncertainties hoping for the survival of a beloved one.

To the ones that have suffered survival losses waiting for a solution to come.

To the ones that believe that a better survival reality lies in the path of research.

ACKNOWLEDGEMENTS

Throughout the three and a half years that I have been dedicated to this work, there are so many things I must acknowledge the importance of. As I dove into studying methods for extracting factors associated with model responses, I got myself thinking about so many reasons for this work (and all its learning processes) to be so more successful than I would, normally, expect. Here, I would like to acknowledge some people essential in the path that got me here.

I would like to start by thanking my supervisor, Renato. Not rarely, I say that I could not have asked for better supervision. Renato is thoughtful, attentive, and so very patient! It has been a pleasure to learn from him. Alongside Renato, I would like to thank Paulo, my co-supervisor, for his insightful contributions. I also have to thank Eraylson for support materials, useful codes, or even for saving me from debugging headaches sometimes. Apart from those who navigated the paths of technical learning with me, so many others made this process possible.

Firstly, I would like to offer my great acknowledges to my sweet love and great partner in this crazily beautiful life, Renata, for making all the journey here not only possible but also so much better. Thus, I would like to acknowledge my parents, Tuta and Julio. Their efforts through life to provide me with the best resources, opportunities and caring made everything possible. Then, I have to thank Fabricio, my cousin, friend, and one of the extraordinary people I look to. Our endless conversations about life and data science – mostly the latter as a way of considering the first – have helped me in so many steps through life and were so important in the journey from applying for the MSc to here. I also thank my sisters, my (big) family and my dearest friends. They are constant support, help and joy through life.

At last, I would like to acknowledge a great team of health professionals – oncologists, nutritionists, dentists, psychologists, Yoga teachers, and so many more from exams' rooms to laboratories. Their work and dedication have contributed greatly not only to my overall survival but also to the quality of life that allowed me to keep dedicating myself to the things I love – this work included.

ABSTRACT

A variety of works in the literature strive to uncover the factors associated with survival behaviour. However, the computational tools to provide such information are global models designed to predict *if* or *when* a (survival) event will occur. When addressing the problem of explaining differences in survival behaviour, those approaches rely on (assumptions of) predictive features followed by risk stratification. In other words, they lack the ability to discover local exceptionalities in the data and provide new information on factors related to survival. In this work, we aim at providing a computational tool to identify the different (unusual) survival responses that may occur in a population of individuals and provide straightforward information about the circumstances related to such responses. We approach such a problem from the perspective of supervised descriptive pattern mining to discover local patterns associated with different survival behaviours. Hence, we introduce an Exceptional Model Mining (EMM) framework to provide straightforward characterisations of subgroups presenting unusual survival models, given by the Kaplan-Meier estimates. In contrast to the greedy search heuristics prevalent among EMM approaches, we employ stochastic optimisation and introduce the first approach in the literature to explore the Ant-Colony Optimisation (ACO) meta-heuristics for the subgroup search. Thus, we tackle the problem of subgroup redundancy to provide a set of exceptional subgroups that are diverse in their descriptions, coverages and survival models. We conducted experiments on fourteen real-world data sets to assess the performance of our approach. In the results, we show that the framework presented is capable of discovering representative patterns with accurate unusual models and straightforward representations. Moreover, the discovered subgroups potentially capture survival behaviours existent in the data. The approach successfully tackles the problem of subgroup redundancy, providing a set of diverse (unique) exceptional (survival) subgroups. Our framework outperforms the other existent approaches to provide characterisations over unusual survival behaviours regarding the descriptive aspect of its results and diversity of its findings.

Keywords: exceptional model mining; subgroup search; supervised descriptive pattern mining; survival analysis.

RESUMO

Diversos trabalhos na literatura dedicam-se a descobrir fatores associados a comportamentos de sobrevivência. As ferramentas computacionais utilizadas para tal são modelos globais projetados para estimar *se* e *quando* um dado evento de sobrevivência ocorrerá. Em se tratando do problema de explicar diferentes respostas de sobrevivência, as abordagens existentes não são capazes de descobrir excepcionalidades locais nos dados nem prover novos conhecimentos a respeito de fatores associados à sobrevivência, respaldando-se em suposições e a análises estratificadas. Este trabalho tem por objetivo apresentar uma nova ferramenta computacional para identificação e caracterização de diferentes respostas de sobrevivência existentes em uma população de indivíduos. Neste trabalho, o problema enunciado é abordado através da perspectiva da mineração supervisionada de padrões descritivos (em inglês, *supervised descriptive pattern mining*) com o intuito de descobrir padrões locais associados a diferentes comportamentos de sobrevivência. Para tal, é empregada a técnica de mineração de modelos excepcionais (do inglês, *Exceptional Model Mining*) com o objetivo de descrever – de forma simples e concisa – subgrupos que apresentem modelos de sobrevivência (Kaplan-Meier) não usuais. Em contraste às heurísticas 'gulosas' prevalentes na literatura de mineração de modelos excepcionais, a abordagem introduzida neste trabalho explora o uso da meta-heurística de otimização *Ant-Colony Optimisation* na busca por subgrupos. O problema de redundância de padrões também é considerado, objetivando a descoberta de um conjunto de subgrupos que sejam diversos com relação às suas descrições, coberturas e modelos. O desempenho da abordagem apresentada é avaliada em quatorze conjuntos de dados reais. Os resultados mostram que o algoritmo proposto é capaz de descobrir padrões representativos que apresentam modelos precisos e caracterizações de simples compreensão. Adicionalmente, os subgrupos descobertos potencialmente capturam comportamentos de sobrevivência existentes nos dados. A redundância de padrões é abordada de forma bem-sucedida, tal que os resultados retornados apresentam conjuntos de subgrupos que são diversos (únicos) e excepcionais. Quando comparado a outras abordagens existentes na literatura que fornecem caracterizações de comportamentos incomuns de sobrevivência, o algoritmo apresentado se sobressai aos demais tanto em relação ao aspecto descritivo de seus resultados quanto à diversidade de suas descobertas.

Palavras-chaves: análise de sobrevivência; descoberta de subgrupos; mineração de modelos excepcionais; mineração supervisionada de padrões descritivos.

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

LIST OF ABBREVIATIONS AND ACRONYMS

ACO	Ant-Colony Optimisation
EMM	Exceptional Model Mining
KM	Kapplan-Meier (Estimates)
LPM	Local Pattern Mining
SD	Subgroup Discovery

LIST OF SYMBOLS

Ω	Survival data set
A	Set of descriptive attributes
T	Survival Time feature
δ	Survival Censoring feature
I_{ij}	Data set single item representation
$\mathbb{I}$	Set of all data set items (I_{ij})
D	Pattern (subgroup) description
$cov(D)$	Pattern (subgroup) coverage (extent)
G	Subgroup representation for the tuple $(D, cov(D))$
$\mathbb{G}$	Set of subgroups
$\mathcal{I}(G)$	Descriptive itemset – the set of all items represented in a subgroup's description
$\mathcal{C}(G)$	Extensive itemset – the set of all items encompassed by a subgroup's coverage
ϕ	Quality measure for subgroup interestingness assessment
τ	Pheromone trails
η	Heuristic information function

CONTENTS

1 INTRODUCTION

A wide range of real-world problems is built around the analysis of an *event*. In healthcare, one may want to analyse *ICU admission*, *death*, or even *disease remission*. In maintenance researches, the event may be a product *failure*. In churn analysis, clients *dropout*. For a variety of problems under different domains, the solution lies in investigating the occurrence of a specific event of interest. In this work, we focus on the medical domain to address a major challenge of current medicine that is to delineate the circumstances around patient outcome.

In recent years, the field of biomedical sciences has provided robust methods for characterising patients (such as the *Omics* sub-fields, diverse cellular assays, and even mobile health technology) and large-scale biologic databases, revolutionising the possibilities for characterising single individuals. Medicine is, then, following the direction of using this available technology to develop new strategies that better consider individual variability to address the needs of a patient. *Precision Medicine* (KOENIG et al., 2017; COUNCIL et al., 2011), as this emergent medical approach is called, *conveys the principle that although therapeutics were rarely developed for single individuals, subgroups of patients could be defined and targeted in more specific ways* (ASHLEY, 2016). This increasing initiative relies on computational tools capable of generating knowledge from large sets of data and thus bringing new insights to understanding health and diseases.

Although the current possibilities for individual characterisation are countless, the medical community still struggles to identify such subgroups of patients closely related to a prognostic response and the characteristics that describe them. Thus, many recent medical works seek for effective computational methods that can provide a better understanding of factors that interfere in survivability to subdivide populations of patients into more specific and uniform subgroups regarding their survival behaviour.

To this end, most studies resort to the Survival Analysis (SA) (KLEINBAUM, 1998), a branch of study dedicated to analysing and modelling data where the outcome is the *time* until the occurrence of a given *event*. In a case where all individuals under study experience the event, the analysis of the *time-to-event* could be addressed as a regression problem and many methods for data analysis would be applicable. However, the time restrictions of medical studies and the challenges related to patients' follow-up usually result in (some) patients with no information regarding the event. Hence, the Survival Analysis arises to approach specifically such cases when there exists a subset of the data for which the outcome is not observed.

The SA methods most present in the literature focus on two aspects of an event analysis: *if* it will happen (the risk of happening) and *when* it will happen (the time for happening). In other words, the existing survival analysis methods aim to create a predictive model for the data. Therefore, the goal is to turn the data into an accurate prediction machine, i.e. to use all available data to predict outputs that are as close as possible to the data itself and new (unseen) data examples. For that, predictive modelling usually trade-offs the model simplicity (the number of parameters) for the accuracy of a global model – i.e. a model representing the whole data. Hence, a predictive modelling algorithm aims to find an as *accurate* as possible and as *complex* as necessary *global* function.

However, such a predictive perspective presents some drawbacks when addressing the need to provide characterisation (in terms of explanatory variables) over subgroups presenting distinct diagnostic/prognostic (survival) responses. We highlight two aspects of their design that interfere in their ability to discover and describe such subgroups:

1. Predictive methods rely on every (input) feature to maximise the accuracy of the learned model. Thus, when striving to assess the impact of specific factors on (survival) model responses, most approaches restrict their input domains to a set of predefined features that are known (or supposed) to be related to survivability. Hence, predictive approaches usually rely on a priori (expert) knowledge or feature selection techniques to delineate the domain of analysis. However, this leads to the potential loss of interesting information (that is left out of the scope of analysis) and often neglects possible interactions between variables. To address this limitation and encompass broader domains of analysis (i.e. a larger number of explanatory variables), it is usually necessary to raise the complexity of the models. However, the resulting (more) complex models usually lose their ability to provide explicit human-comprehensible insights into the data domain, which is essential when addressing the need to better define more specific groups of patients regarding their survival. Some approaches resort to feature importance or explainability techniques to point out the most prominent factors (regarding the prediction outcome) and improve the model's comprehensibility. Anyhow, those predictive approaches assume that all features considered for the modelling are somehow relevant to the result, failing to identify only specific features related to the model response or even specific features' interactions that result in unexpected outcomes.

2. Even the methods that provide more interpretable results regarding the characterisation of survival responses still focus on global modelling. Which means that, while striving to provide accurate models that fit all individuals, such approaches cannot identify (accurate) models that fit only a subset of the data space. In other words, because predictive methods strive to provide general models, they fail to identify local models, i.e. models that fit only part of the data but vary in their characteristics. Hence, when striving to subdivide populations of patients into different subgroups presenting similar survival behaviour, the predictive methods fail to pinpoint the data subsets where a learned model behave differently from each other or even from the global model observed in the population.

Although predictive models are important in many contexts, with such shortcomings, they fail at identifying features that might affect the outcome for a subgroup of patients. That is, as they attempt to only answer the *if* and *when* questions, they fail to answer questions such as

- *What features are associated with this exceptional prognostic behaviour?*
- *Are there groups of patients with unusual survival responses?*

This situation is not a problem per se. However, the lack of methods that answer these more descriptive questions highlights the gap in the literature.

In the existing literature, similar questions are answered using predictive approaches by stratifying data regarding a variable of interest. For instance, to verify whether a factor (e.g. new treatment or a genetic marker) affects prognosis, patients may be split into groups – test/control or feature strata. Then, each group would be individually modelled, and their differences analysed. Notice, however, that such an approach falls into the two predictive drawbacks listed above: (i) it assumes that the dependent variables (and the scope of interest) are known a priori and (ii) that the different prognostic behaviours (groups) existent in the analysed cohort are encompassed in the (features) stratifications. Those assumptions, however, are not always true.

In some cases, researchers may want to identify these variables that are related to some (un)expected behaviour, as shown by WOLFF et al. (2021). In other cases, the existing definitions of prognostic groups (e.g. a subtype of a specific type of cancer) may not represent precisely all behaviours observed in such a group, as MILIOLI et al.(2017) evince. In fact, many studies in the literature reveal the limitations of the current diseases' markers and reinforce the need for a better characterisation of (more specific) diagnostic and prognostic groups.

To enable the restructuring of current clinical medicine into a more precise approach – with more accurate measures and more effective results, there is a latent need for computational tools capable of identifying unknown behaviours and describing them in a way to enable actions that implement real change. However, to identify and characterise different survival behaviours existing in a population, it is necessary to propose new methods to answer different questions from those *if* and *when* questions that are already being addressed in the literature. Therefore, we pose the following question:

Research question: *Is there an effective approach to identify and characterise multiple and different subsets of a population for which the observed survival behaviour is exceptional (unusual) with respect to a (baseline) expected response?*

Our motivation for tackling this question is as follows.

1.1 MOTIVATION

As previously introduced, the explosion of data associated with individual humans has essentially transformed the possibilities and opportunities we have to characterise patients, understand diseases, and – ultimately – to improve health outcomes. Yet, COUNCIL et al. (2011) had pointed out the necessity of using these data to effectively implement change in the way we conduct real medicine:

> Biomedical research and the practice of medicine, separately and together, are reaching an inflection point: the capacity for description and for collecting data, is expanding dramatically, but the efficiency of compiling, organizing, manipulating these data – and extracting true understanding of fundamental biological processes, and insights into human health and disease, from them – has not kept pace. There are isolated examples of progress: research in certain diseases using genomics, proteomics, metabolomics, systems analyses, and other modern tools has begun to yield tangible medical advances, while some insightful clinical observations have spurred new hypotheses and laboratory efforts. In general, however, there is a growing shortfall: without better integration of information both within and between research and medicine, an increasing wealth of information is left unused (COUNCIL et al., 2011).

There are many challenges associated with the integration of patient data, medical research and clinical medicine. In this work, we are mainly motivated by the latent need for methods capable of 'extracting true understanding of biological processes and insights into human health and disease' from data. As introduced, the existing methods addressing such a problem are population-based approaches relatively inefficient concerning knowledge extraction and may yield conclusions that are not relevant to specific (or even broader) populations.

Hence, what we have as standard medical interventions still rely on traditional protocols that are limited in information content and usability. As the current disease's classifications are primarily based on symptoms and simple forms of laboratory or imaging studies (COUNCIL et al., 2011), such protocols usually consider only limited and over general factors to characterise and distinguish between different prognostic groups. Not rarely, those protocols fail to be an effective solution to some sub-populations of patients or comprise an imprecise general course of treatments.

For instance, patients with Hodgkin Lymphoma are further classified as *early-stage favourable* according to size and sites (number and location) of diseased organs, absence of systemic symptoms and a given laboratory result. Such a group of patients is usually subjected to considerable amounts of irradiation because protocol says that the combined administration of chemotherapy and radiotherapy is the overall preferred treatment (NCCN, 2017). However, such usual clinical decision is conservative and does not consider the fact that the same protocol also includes (and suggests) chemotherapy alone as a viable treatment option (to such group of patients) in order to avoid the long-term risks of radiotherapy – which includes an increased risk for heart disease, pulmonary dysfunction, and the development of secondary cancers, for example.

Current diseases' guidelines fail to consider a larger scope of patient's context – biological, psychological, socio-environmental, and other potentially relevant factors. Many times, such shortfall results in conservative measures (despite evidence that those measures may be extreme or unnecessary), exposing patients to a large number of collateral risks that could be avoided for the sake of improving overall survival. Thus, such guidelines are not designed to exploit and incorporate (rapidly) emerging data or (new) factors relevant to diseases. The recent COVID-19 pandemic is drastically evidence of the implications of such shortfall.

Initially reported as a respiratory infection, the disease caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) turned out to be a systemic condition with very different (and extreme) prognoses across different societies and groups of people. The lack of scientific knowledge about the disease's dynamics aligned to its high spread rate and fast progression culminated in the collapse of many health systems and over 5 million deaths worldwide. The scientific community has put extra research effort to shed light on the uncertainties of this disease and assist in developing effective medicines and governmental decisions. As the pandemic impacts are likely to be felt for a while, broader questions arise in the context of patient care and general health management. Many recent studies seek to discover markers associated with disease's prognosis (WOLFF et al., 2021) aiming to improve treatment of the acute and chronic disease (CROOK et al., 2021), assess the effectiveness of vaccines and medicines (FORNI; MANTOVANI, 2021), or even handle new variants of the virus. In addition to pre-existing health conditions, social, economic, and political aspects are likely to interfere in the disease's dynamics, and the implications of such aspects on patient care are still being investigated as they manifest daily (GOULARTE et al., 2021; PAUL; STEPTOE; FANCOURT, 2021).

Anyhow, when assessing the impact of covariates in the survival outcome, most of those works resort to the predictive approaches of Survival Analysis, falling right into those same drawbacks that we previously introduced. The existing studies are restricted to a limited scope of analysis and struggle to provide novel knowledge. The computational tools we have at our disposal rely on the suppositions of our human insights and experiences. In many cases, however, such suppositions lead to inefficient tests and vague conclusions. For example, when addressing the uncertainty regarding the need for radiotherapy for treating early-stage Hodgkin's lymphoma, RADFORD et al. (2015) randomly assigned patients to receive irradiation or no further treatment. The results they analysed suggested a benefit for combined modality (chemotherapy and radiotherapy) but not necessarily superior to chemotherapy alone. Thus, in the cases the suppositions we make are faulty or inaccurate, such tools fail in shedding light on new insights. For instance, MATTOS et al. (2020a) show that the symptoms most frequently observed in Brazilian COVID-19 patients are associated with better prognoses, contrasting with the results in (ZHENG et al., 2020). The authors also observe differences in comorbidities distribution between Brazilian and US patients (RICHARDSON et al., 2020), which may yield different survival experiences (NEPOMUCENO et al., 2020). Our current methods of investigating survivability comprise trial-and-error processes that do not address the urgent needs of our medical reality and cannot keep pace with the data at our disposal. Ultimately, they rely on data to investigate hypotheses instead of using data to extract the answers we are looking for.

Whether to improve standard medical interventions or to approach new medical challenges more efficiently, as we advance in characterising patients (generating and collecting data), it is evident the need for more powerful tools to restructure the medicine as we have into a more *precise* approach. To that end, we are in need of computational methods capable of describing complex interrelationships associating diseases to each other, and to the vast possibilities of related factors. In other words, to beyond methods capable of predicting disease or a time of survival, we need data-intensive methods capable of discovering and characterising different survival behaviours given all emerging data on human characterisation.

The discovery and understanding of subgroups with distinct responses with relation to their survival can be an essential tool towards new possibilities in medical and bioinformatics researches: for instance, the identification of characteristics that differentiate between groups of patients with different survival rates, the possibility of better prescribing treatment considering unique attributes of a patient, and the recognition of unknown relationships between covariates that affect a patient's prognosis. The following section introduces our proposal to a method that potentially provides new insights on diseases' prognosis and survival outcomes.

1.2 OBJECTIVES AND RESEARCH APPROACH

Our ultimate goal with this work is to provide insightful knowledge over the circumstances related to patient outcome. In other words, we aim at identifying the different prognostic responses that may occur in a study cohort and provide straightforward information about the factors related to such responses. In contrast to the existing predictive approaches that induce survival risk characterisation from predefined features strata, we are concerned with providing a method capable of discovering subsets of the data that present unusual survival behaviour and characterising such subsets in a humanly understandable way.

Apart from the introduced predictive (global) modelling, the exploratory data analysis casts a different perspective over a given data collection, aiming to discover novel insights about the domain in which the data was measured and, thus, boost human expertise. In this sense, local modelling and pattern discovery techniques are designed to value the model simplicity and accuracy over generality (global representation) and thus may selectively focus on parts of the input space where there is a pattern without the obligation of modelling the remainder of the data. In other words, while predictive modelling sacrifices simplicity to provide an accurate model for the whole data collection, a pattern discovery algorithm can report models that are both accurate and simple but only a partial function of the data space (BOLEY, 2017).

Therefore, in this work, we approach the problem introduced through the perspective of *supervised (descriptive) Local Pattern Mining* (LPM) (DUIVESTEIJN, 2013): to find subsets of the data that somehow deviate from the norm, i.e. where something *interesting* is going on. As the name introduces, three aspects of this paradigm make it suitable to approach our research question. The first one is the inherent *descriptive* aspect of LPM tasks: they strive to find *subgroups*, i.e. subsets of the data that can be concisely described in terms of the explanatory data features. Such descriptive aspect gives characterisation (pattern identification) to the interesting data subsets, i.e., the subsets that stand out to the remaining data. Secondly, each subgroup is deemed interesting (relevant) by its own merit, without considering its fit to the remainder of the data and independently from any other finding. This *local* aspect (in contrast to global predictive models) allows not only the identification of data subsets presenting a unique behaviour but also the discovery of overlapping patterns. Lastly, the *supervised* aspect defines the *interestingness* of a subgroup with respect to a *property of interest* – i.e. a property of the (data) population that one is interested in – and, thus, provide local patterns which stand out with relation to a *target concept*.

Subgroup Discovery (SD) (HERRERA et al., 2011; ATZMUELLER, 2015) is one of the earliest tasks of supervised LPM that consists of the discovery of subgroups in populations presenting an unusual distribution of a single target variable (WROBEL, 1997). Understanding that a deviating distribution of one target attribute does not encompass all forms of *interestingness*, the Exceptional Model Mining (EMM) task (LEMAN; FEELDERS; KNOBBE, 2008) is defined as the multi-target generalisation of SD, allowing the representation of more complex target concepts by extending the definition of property of interest from a (single) target variable to a model over several (target) attributes. Hence, given a model (of the data) that best represents our property of interest – and thus constitutes our *target concept*, EMM searches for subgroups of the population where the model fitted to the subgroup substantially differs from a baseline model.

When addressing our research question, the EMM task makes it possible to use the existing (predictive) methods of Survival Analysis as the property of the data we are interested in. Combining such target concept and the local aspect of the task, the EMM potentially values the discovery of data subsets with an unusual survival response. Furthermore, its descriptive aspect provides the characterisation of such subsets that (individually) behave unusually. In contrast to the existing predictive methods, our approach potentially provides more explicit information and new insights on features' interactions and associations with survival response. It also enables a broader scope of analysis by performing multivariate analysis with no need for feature selection. However, to the best of our knowledge, the use of supervised descriptive local pattern mining to discover and characterise subgroups with unusual survival behaviour is still an under-explored area of research.

Finally, the search space can be exponentially large when striving to find subgroups. The combinatorial nature of (local) pattern mining tasks poses a significant challenge concerning the computational cost of the mining algorithms. When focusing on the EMM task, these challenges can be even more significant once the task usually involves the induction of numerical models. Once the complete traversal of the search space becomes infeasible with the increase of data volume and complexity, more sophisticated search strategies are required to make feasible the discovery of local patterns. Hence, *heuristic searches* employ different strategies to favour the search towards regions of the space that are more likely to contain good solutions. However, by directing the search towards the best solutions, most heuristic approaches to subgroup mining usually struggle with the problem of *redundancy*: a variety of patterns that comprise only slight variations of the same (more general) finding (LEEUWEN; KNOBBE, 2012). Hence, to provide meaningful results, the heuristic search strategy needs to achieve a good balance between *exploitation* and *exploration*. In other words, the heuristic search should be able to focus on promising areas while leaving room to extend the search towards several search directions.

The existing approaches to the EMM task, however, resort to the greedy *beam-search* as the heuristic search strategy while striving to tackle the problem of redundancy in their final findings. Another strong line of research to LPM tasks approaches the subgroup search with stochastic optimisation and, thus, employ optimisation meta-heuristics as the search strategy. In this context, evolutionary algorithms have been widely studied in pattern mining applications (VENTURA; LUNA et al., 2016), especially for Subgroup Discovery. The existent methods present great results related to computational aspects and the quality of the discovered patterns and hence pose a competitive approach to supervised descriptive LPM tasks. However, to the best of our knowledge, there is no work on EMM exploring stochastic optimisation algorithms as the search strategy to mine subgroups.

In this work, we propose the discovery of (non-redundant) subgroups presenting unusual survival behaviour. Hence, we present an EMM framework that uses a (predictive) method of Survival Analysis as the target concept that models survival behaviour. We approach the subgroup search as an optimisation problem, aiming to maximise the interestingness of the discovered subgroups – w.r.t. their survival models – while minimising the redundancy of the (final) findings. Thus, we employ the Ant-Colony Optimisation meta-heuristics as the heuristic search of our EMM framework. Therefore, instead of using Survival Analysis models for predictions, we use them as a deviation target in combination with the supervised descriptive local pattern mining task of EMM to provide a set of (diverse) characterisations of subgroups presenting unusual survival behaviours.

We now present this document organisation in the next section.

1.3 DOCUMENT OVERVIEW

This document consists of six chapters, of which this introduction is the first.

Chapter 2 contains a brief theoretical overview of our research area. We introduce the basic concepts of local pattern mining and define an Exceptional Model Mining instance using a non-parametric statistical method of Survival Analysis as the target model. Then, we enunciate the problem of redundancy in sets of subgroups. At last, we present the problem of Optimisation and introduce the basic concepts of ACO meta-heuristic, our choice for subgroup (heuristic) search.

Then, in Chapter 3, we present a brief review of the works in literature that strive to provide knowledge about the factors associated with survival response. We briefly introduce the traditional statistical methods of Survival Analysis and revise the machine learning and data mining approaches to *time-to-event* analysis. As we evince the lack of approaches to provide descriptive knowledge over local aspects of the data, we also provide a brief review on the existing algorithms for supervised descriptive pattern mining, specially the approaches developed for EMM and SD tasks.

We present our first approach to the problem in Chapter 4, the Esmam algorithm: an EMM framework to find itemset induced subgroups. The work was published in MATTOS et al. (2020b), and this chapter consists of the method and the results discussed in the paper.

In Chapter 5, we present the EsmamDS algorithm that builds on the work presented in Chapter 4 to address the problem of redundancy in the final set of discovered subgroups. The approach was previously presented by MATTOS; NETO; VIMIEIRO (2021), and, in this chapter, we provide the method and the results discussed in the paper.

Lastly, we present our final remarks in Chapter 6.

2 THEORETICAL FOUNDATION

As introduced before, *Local Pattern Mining* tasks strive to pinpoint multiple (potentially overlapping) subsets of the data that behave differently than expected. However, the main characteristic of such techniques is that they are not merely looking for data subsets. They search for *subgroups*: *coherent subsets for which we can formulate a concise description in terms of conditions on attributes of the data* (DUIVESTEIJN, 2013). The author argues that such descriptions make the subgroups more actionable: in addition to pinpointing the subjects that behave differently from a norm, we can also pinpoint the factors related to such behaviour.

A different narrative introduces the Exceptional Model Mining through the perspective of Rule Induction (FÜRNKRANZ; GAMBERGER; LAVRAČ, 2012), a traditional data mining technique that aims at learning sets of rules from a given data. *Predictive rule induction* comprises supervised learning techniques to induce prediction/classification models in the form of rule sets. Hence, the individual rules constitute a final global model – ordered or unordered rule set – which is evaluated concerning its accuracy and predictive power. Contrasting to the predictive perspective, *descriptive rule induction*, also known as *association rule mining* (ZHANG; ZHANG, 2002), is a form of exploratory data analysis that typically comprises unsupervised tasks for discovering association relationships or correlations among a set of items. In such a descriptive approach, each rule comprises a pattern representation describing regularities in data and is usually individually evaluated with respect to the uncertainty of the relation it states.

NOVAK; LAVRAČ; WEBB (2009) call *supervised descriptive rule discovery* the family of tasks that lie at the intersection of both predictive and descriptive perspectives, comprising several supervised algorithmic approaches to the discovery of relationships associated with some property of interest – e.g. Exceptional Model Mining, Contrast Set Mining, Emerging Pattern Mining, among others (VENTURA; LUNA, 2018). By inducing descriptive rules from labelled data, such tasks aim to understand the underlying phenomena (according to a target) and not arbitrarily explain the data or predict outcomes for new instances. Therefore, the final set of rules comprises independent patterns that describe the property of interest and are usually evaluated according to a metric of such property.

Rule-based approaches traditionally represent their findings through rules in the form of $antecedent \rightarrow consequent$. The *antecedent* is the pattern description, i.e. the set of conditions over the data attributes that need to be satisfied in order to imply the *consequent*. The form of this latter varies according to the task being executed: it may be a set of items in association rule mining, a target class or target attribute value on predictive induction, or even a property of interest in supervised descriptive rule discovery. As will be further introduced in this chapter, we adopt a rule-based representation with an implicit consequent being a model of the survival behaviour observed in the population delimited by the antecedent. In other words, we assume a representation that is restricted to a *pattern description* (antecedent) that implicitly implies a model (consequent).

At last, rule-based (or pattern-based) approaches, i.e. approaches whose models comprise logical inferences over data attributes, present an advantage over more complex mathematical models in the sense that exists an inherent *interpretability* on their designs and results. There is a vast amount of works on the literature that discuss such a term, its definition, applicability and relevance (CARVALHO; PEREIRA; CARDOSO, 2019). At an abstract level, the aim is the development of data models that can provide meaning in understandable terms to a human, i.e. models that enable a human to verify, interpret, and understand the system's reasoning and how particular decisions are made. In this work, we refer to *interpretability* as the extent to which a model outcome is humanly understandable, in a way that is self-contained and do not need further processing to be fully comprehended (GUIDOTTI et al., 2018). However, although (predictive/descriptive) rule-based models are naturally interpretable (allow easy understanding for human beings) because of the simplicity of their results design, there is still a cognitive limit on how complex a model can be while still also being understandable (LAKKARAJU; BACH; LESKOVEC, 2016).

For now, we proceed with the theoretical foundation of this work. However, the reader should keep in mind this Rule Induction foundation and concepts throughout this document as we revise rule-based approaches and assess their performances regarding the interpretability of their findings. In this chapter, we first define the basic concepts used throughout this document. Then, we introduce the EMM framework to discover subgroups with an unusual survival response using a predictive method of Survival Analysis as the target concept. Moreover, we define the problem of redundancy in subgroup sets. At last, we present the subgroup search strategy we adopt, providing a brief review of the Ant-Colony Optimisation meta-heuristics.

2.1 BASIC CONCEPTS

In order to define the essential elements of a framework to mine exceptional survival behaviours, we first need to define a survival data set. The *survival data*, or *time-to-event* data, is mainly characterised by a phenomenon we call *censoring*: the existence of data subsets that do not present (labeled) information regarding the *time to the event* occurrence. There are different types of censoring, all relating to the moment when the individual was lost to the study. The *right-censoring* is the most common type observed in survival data. It happens when an individual had not yet experienced the event by the time it was last observed in the study; therefore, its time-to-event is unknown. In other words, the right-censoring happens when the (unknown) time for an individual actually suffering the event is future to the time such individual was last observed in the study. For simplicity, from now on, all censoring referred to in this work should be understood as right-censoring.

To define a survival data set, we first provide the toy example in Table 1. We will consider a survival study where *death* is the event, and the survival time T is reported in months. Individuals o are rows and are identified by their numbers. The columns are the set of descriptive attributes $A = \{size, location, type, metastasis\}$ (of size $|A| = 4$) and the model attributes T and δ – over which a survival model can be defined. Table 2 presents the descriptive attributes A_i of the data set in Table 1 and their domains. An individual o is described by a set of $|A|$ values from the domain of each descriptive attribute, has a survival status δ_o and with time T_o. For instance, we have that the individual $\#5 = (\{large, II, malignant, yes\}, 6, 1)$ died 6 months ($T_{\#5} = 6, \delta_{\#5} = 1$) after the time it entered the study. By contrast, the individual

Table 1 – Theoretical foundation – Toy survival data set.

o	$Size$	$Location$	$Type$	$Metastasis$	T	δ
$\#0$	$small$	I	$benign$	no	12	0
$\#1$	$small$	III	$malignant$	no	8	0
$\#2$	$medium$	II	$malignant$	no	4	1
$\#3$	$medium$	III	$malignant$	yes	3	1
$\#4$	$large$	I	$benign$	no	9	0
$\#5$	$large$	II	$malignant$	yes	6	1
$\#6$	$large$	II	$malignant$	no	10	0

Font: The author (2021)

Table 2 – Theoretical foundation – The set of descriptive attributes (and their domains) of the toy survival data set provided in Table 1.

$\boldsymbol{A_i}$	$\boldsymbol{dom(A_i)}$
$A_0 : size$	$\{small, medium, large\}$
$A_1 : location$	$\{I, II, III\}$
$A_2 : type$	$\{benign, malignant\}$
$A_3 : metastasis$	$\{no, yes\}$

Font: The author (2021)

$\#4 = (\{large, I, benign, no\}, 9, 0)$ was observed alive for 9 months ($T_{\#4} = 9, \delta_{\#4} = 0$); after that time, there is no information on whether the individual $\#4$ remained alive or not.

We call *items* the atomic descriptive elements of the data set. Each value v_j in a $dom(A_i)$ constitute a data set *item* $I_{ij} = (A_i, v_j)$. For the set of descriptive attributes defined in Table 2, $I_{01} = (size, medium)$ and $I_{20} = (type, benign)$ are two examples of *items*. By the domains provided, we also have the toy data set in Table 1 comprising a total of 10 items. Hence, we formally define a survival data set as follows.

Definition 1 (Survival dataset $\Omega(A, T, \delta)$). We define a (survival) data set $\Omega(A, T, \delta)$ as a collection of $|\Omega|$ individuals, where A is the set of $|A|$ categorical (descriptive) attributes and T and δ are the target attributes. The numeric attribute $T = \{T_1, T_2, \ldots, T_{|\Omega|}\}$ is the set of survival times of all individuals in the experiment, i.e. the moment in the study timeline when the individual was last observed. The survival status $\delta = \{\delta_1, \delta_2, \ldots, \delta_{|\Omega|}\}$ is a boolean (censoring) attribute to indicate whether the individual experienced the event ($\delta = 1$) or is censored ($\delta = 0$). For censored instances, survival time indicates the time of censoring (last event-free observation in the study). Each individual is therefore a tuple $o = (V, T_o, \delta_o)$, where $V \in dom(A_1) \times dom(A_2) \times \cdots \times dom(A_{|A|})$. The value of an attribute A_i is also denoted by $A_i(o) = v$. The domain of a descriptive attribute in a given data set is denoted by $dom(A_i)$. At last, we define an *item* as a tuple $I_{ij} = (A_i, v_j)$, for $v_j \in dom(A_i)$. Thus, we denote $\mathbb{I} = \bigcup_{A_i} \{I_{ij} = (A_i, v_j) | v_j \in dom(A_i)\}$ the set of all items of the data set. □

We call *description* a conjunction of conditions on the descriptive attributes of the data set. We define that each attribute may be constrained at most once. Thus, a description might have at most $|A|$ conditions. In this work, we define a *condition* as a restriction on the values that the attribute might have. Hence, a description can be seen as an indicator function defining whether an individual in the data set satisfies or not the constraints it imposes. A description D *covers* an observation o if and only if $D(o) = 1$.

Definition 2 (Description). A *description* $D : \Omega \rightarrow \{0, 1\}$ of length $l \leq |A|$ is a pattern given as $D(o) = cond_1(A_i) \wedge cond_2(A_j) \wedge \cdots \wedge cond_l(A_k)$. A condition $cond(A_i)$ (for an arbitrary A_i) is a proposition in the form of $cond(A_i) = A_i(o) \in \mathcal{V}_i$, where $\mathcal{V}_i \subset dom(A_i)$. □

For instance, $D_{ex}(o) = A_1(o) \in \{I, II\} \wedge A_2(o) \in \{benign\}$ is a description of length 2 (it has 2 conditions), where the values for the attribute $location$ (A_1) are restricted to those two values in the first set, and the values for $type$ (A_2) are restricted to that one value in the second set. Considering Table 1, there are two individuals – $\#0$ and $\#4$ – presenting attribute values within the restricted values for each condition in D_{ex}. For those two individuals, D_{ex} returns 1, and we say that D_{ex} covers $\#0$ and $\#4$. Hence, we define $\{\#0, \#4\}$ as the *coverage* of D_{ex}.

Definition 3 (Coverage). The *coverage* (extent) of a description D is the set of individuals it covers, formally given by $cov(D) = \{o \in \Omega \mid D(o) = 1\}$. The *size* of a coverage is $|cov(D)|$. □

As previously stated, the EMM task searches for *subgroups*: *subsets* of the data for which we have concise *descriptions*. A subgroup is, therefore, composed of a description and its coverage.

Definition 4 (Subgroup and Complement). For a given a description D, $G_D = (D, cov(D))$ is a *subgroup*. We call the *complement* of a subgroup G_D the set of individuals not covered by D, formally defined as $\overline{G_D} = \Omega \setminus cov(D)$. □

Hence, $G_{D_{ex}}$ comprises the subset $\{\#0, \#4\}$ of individuals described as $D_{ex}(o) = A_1(o) \in \{I, II\} \wedge A_2(o) \in \{benign\}$. The set of individuals $\{\#1, \#3, \#4, \#5\}$ for which $D_{ex}(o) = 0$ is the complement $\overline{G_{D_{ex}}}$.

We say that a subgroup entails two itemsets. The first one is a function of its description and is called *descriptive itemset*. It contains all data set items represented in a subgroup's description. For instance, the descriptive itemset of $G_{D_{ex}}$ is $\{(location, I), (location, II), (type, benign)\}$. The other itemset a subgroup entails is a function of its coverage and is called *extensive itemset*. It contains the set of items $I_{ij} \in \mathbb{I}$ encompassed by the individuals covered in the subgroup. Given that $cov(D_{ex}) = \{\#0, \#4\}$, the extensive itemset of $G_{D_{ex}}$ is $\{(location, I), (location, II),$

$(type, benign), (size, small), (size, large), (metastasis, no)\}$. Note that this itemset is usually a superset of the descriptive itemset, containing items related to the attributes not restricted in the description in addition to the ones in the descriptive itemset.

Definition 5 (Descriptive Itemset and Extensive Itemset). We call *descriptive itemset* the set of items $\mathcal{I}(G_D) = \bigcup\{(A_i, v)|\forall v \in \mathcal{V}_i\}$, given every arbitrary $cond(A_i) = A_i(o) \in \mathcal{V}_i$ represented in D. Therefore, the descriptive itemset is the set of items $I_{ij} \in \mathbb{I}$ strictly listed in (the conditions of) D. The *extensive itemset* is the set of items related to a subgroup's coverage, given by the function $\mathcal{C} : \mathcal{P}(\Omega) \rightarrow \mathbb{I}$, $\mathcal{C}(G_D) = \bigcup\limits_{A_i}\{(A_i, A_i(o)) \mid o \in cov(D)\}$. Note that $\mathcal{C}(G_D)$ may include even attribute values not listed in the conditions of D. □

We define an empty subgroup $G_\varnothing$ comprising an unrestricted description $D_\varnothing$ for which $cov(D_\varnothing) = \Omega$; hence, $G_\varnothing = (D_\varnothing, \Omega)$. By definition, an arbitrary subgroup G_D always implies a description D and a coverage $cov(D)$. Hence, we may use the expression *subgroup* to refer to one or another interchangeably as the context will make clear its meaning. We will omit the subscript unless necessary. Thus, we assume that a description D is always associated with (and contained by) a subgroup. Hence, we employ G to represent D when convenient.

At last, we introduce the concept of *generality*. Since deviations from the norm are easily achieved in very small subsets of the data, we aim at finding subgroups that encompass as many individuals as possible. Therefore, we define generality in relation to the subgroup's *coverage*. We say a subgroup G_a is more general than a subgroup G_b if and only if $cov(G_a) \supset cov(G_b)$. In this case, we may also say that G_b is more *specific* than G_a. However, note that although generality is defined with respect to the subgroup's coverage, it is dependent on the subgroup's description. Each condition $cond(A_i) = A_i(o) \in \mathcal{V}_i$ imposes a constraint to $dom(A_i)$, restricting the extent of individuals covered by the pattern to the extent delimited by the scope of $\mathcal{V}_i$. Hence, generalisation may be affected by the number of conditions in a description (description length) and the extent of the constraints such conditions impose.

Having defined the basic concepts that will be applied throughout this work, we introduce the Exceptional Model Mining framework to discover subgroups with unusual survival behaviour in the next section.

2.2 EXCEPTIONAL MODEL MINING FRAMEWORK

Exceptional Model Mining (EMM) is a data mining framework that aims to discover concisely described subsets of a data set – subgroups – where a model of interest can be deemed exceptional. In other words, the EMM framework strives to find descriptions D for which a given target model learned from the individuals in $cov(D)$ has parameters that deviate substantially from the parameters of the model learned from a defined baseline group of individuals (DUIVESTEIJN, 2013). There are two possibilities in this case: (i) to compare the characteristics of the model fitted to the subgroup to its *complement*; or (ii) to compare to the *population*, i.e. all individuals in the data set. It is, however, crucial to understand that this choice essentially changes the nature of the task at hand and may lead to different outcomes. Comparing a subgroup to the population implies searching for deviations from the norm. On the other hand, comparing to its complement implies searching for two subsets presenting contrasting behaviour. There is no overall best choice. Sometimes, the real-life problem at hand or mathematical design and computational requirements may lead to a direction. From now on, we refer to a general *baseline* $\mathcal{B}$ as a predefined baseline set of individuals to quantify the exceptionality of a subgroup: the *population* or the *complement*.

Hence, an EMM *instance* is defined by a *model class*, which comprises the (target) property of interest, and a *quality measure*, which quantifies the dissimilarity between two models from the model class. There are several EMM instances defined in the literature (DUIVESTEIJN; FEELDERS; KNOBBE, 2016). However, as we will evince in the review provided on Section 3.2, there is no EMM instance to properly address survival data. In this section, we introduce the EMM framework by defining a model class to represent survival behaviour and a quality measure based on the statistical unusualness of the defined (survival) model class.

Survival behaviour is usually modelled by the survival function $S(t)$ which is a representation of the probability of an individual o surviving up to a time t, i.e., $S(t) = P(T_o > t)$. It presents an initial value $S(0) = 1$ to represent that no individual has suffered (yet) the event at the beginning of the study. Therefore, the probability of surviving past the initial time is one. Throughout the study timeline, the survival function monotonically decreases with t and, theoretically, no individual would survive if the study period increased without limit; therefore, $S(\infty) = 0$.

As will be revised in the next chapter, the survival function may be modelled by different methods. The (statistical) non-parametric Kaplan-Meier (KM) estimator (or product-limit method) (KAPLAN; MEIER, 1958) is one of the simplest methods to estimate survival function, being largely used in survival studies to model survival response. The KM model estimates the survival function by calculating the cumulative survival probability $\hat{S}(t)$ from the observed survival times T and censoring information δ. Considering Ω, we define $\mathcal{T} = \{t_1, t_2, \ldots, t_k | t \in T, k \leq |\Omega|\}$ the set of unique ordered survival times in the data set. The estimated probability $\hat{S}(t_j)$ of surviving past a time $t_j \in \mathcal{T}$ is given by Equation 2.1:

$$\hat{S}(t_j) = \left(\prod_{i=1}^{j-1} \hat{P}(\mathcal{T} > t_i | \mathcal{T} \geq t_i)\right) \cdot \hat{P}(\mathcal{T} > t_j | \mathcal{T} \geq t_j) \equiv \hat{S}(t_{j-1}) \left(\frac{r_j - d_j}{r_j}\right) \tag{2.1}$$

where $\hat{S}(t_{j-1})$ is the (estimated) survival probability at the time t_{j-1}, r_j is the number of individuals at risk (have not suffered the event yet) at time t_j, i.e. $r_j = |\{o \in \Omega \mid T_o \geq t_j\}|$, and d_j is the number of individuals that experienced the event at time t_j, that is $|\{o \in \Omega \mid T_o = t_j \wedge \delta_o = 1\}|$. The plot of the KM survival probabilities $\hat{S}(t)$ against time is called the survival curve and provides a visual assessment of the survival response over time.

Therefore, for each subgroup G under consideration, the KM (survival) model is induced on the target attributes T and δ of the data associated solely with the subgroup's coverage $cov(G)$. Then, it is necessary to evaluate the (target) model fitted on G to determine whether the particular subgroup is exceptional. We say G is exceptional if its model is statistically different from the model fitted to the baseline $\mathcal{B}$.

The *logrank* is a statistical test widely used to verify whether the survival responses of two groups are equivalent or not. It tests the null hypothesis that there is no difference between the groups in the probability of an event occurring at any time point (BLAND; ALTMAN, 2004). For that, the logrank uses the events observed within each group (i.e. the individuals with $\delta_o = 1$) versus the number of events expected to happen.

Let $G \subseteq \Omega$ denote a set of individuals. We define r_j^G the number of individuals in G that are at risk just before $t_j \in \mathcal{T}$. The logrank test assumes that the number of events expected to happen within a group is proportional to the extent of its risk, i.e. to the proportion r_j^G/r_j. Hence, the number E^G of expected events suffered by G over $\mathcal{T}$ is given as follows.

$$E^G = \sum_{\forall t_j \in \mathcal{T}} \frac{r_j^G}{r_j} \times d_j$$

For comparing G to another group $\mathcal{B}$, the logrank test $X^2 \sim \chi_1^2$ is given bellow, where O^G(resp. $O^{\mathcal{B}}$) is the number of observed events in G(resp. $\mathcal{B}$), which can be given by the sum of δ_o for every $o \in G$(resp. $\mathcal{B}$).

$$X^2 = \frac{(O^G - E^G)^2}{E^G} + \frac{(O^{\mathcal{B}} - E^{\mathcal{B}})^2}{E^{\mathcal{B}}}$$

Finally, we define exceptionality in this work by means of the logrank statistic. Hence, being $\mathcal{G}$ the set of all possible subgroups in a data set, we assess the *exceptionality* of a subgroup with a *quality measure* $\phi : \mathcal{G} \to \mathbb{R}$ that quantifies the deviation between the subgroup's KM model and the model fitted to $\mathcal{B}$. For this, we test the hypothesis that the KM curves adjusted for both the subgroup and the baseline are statistically equivalent. Then, we take $1-$ *p-value* of the test as the quality of the subgroup. The stronger the evidence that the null hypothesis should be rejected, the more exceptional the subgroup is. This quality measure is presented in (2.2) where $pval_{G,\mathcal{B}}$ is the *p-value* of the logrank test between the KM curves of subgroup G and the baseline $\mathcal{B}$.

$$\phi(G) = 1 - pval_{G,\mathcal{B}} \tag{2.2}$$

Definition 6 (EMM instance: Target model and Quality measure). We define the Kaplan-Meier (KM) estimates presented in Equation 2.1 as the model class representing survival behaviour and, thus, the EMM target concept. Then, we define the (statistical) quality measure given in Equation 2.2 as the function that maps every subgroup $G \in \mathcal{G}$ to a real number reflecting its exceptionality. □

So far, we have defined an EMM instance with a (target) *model class* representing survival behaviour and a *quality measure* to evaluate subgroups concerning the statistical exceptionality of their survival responses. What is yet to be discussed is how to generate such subgroups and conduct the search to guarantee the discovery of several exceptional subgroups that are unique and representative of the study cohort.

The existing EMM frameworks usually generate subgroups by manipulating descriptions D to maximise the quality measure computed over $cov(D)$ – that is because the set of data subsets for which exists a feasible description is typically smaller than the set of all data subsets[1]. Thus, given a description representation (and its constraints), the computational approaches to mine subgroups usually traverse a search space defined over the descriptive attributes' domain constructing descriptions from atomic elements.

The way of conducting such *subgroup search*, i.e. the strategy employed to traverse the search space while constructing solutions, is a well-known challenge of the broader scope of pattern mining tasks. That is because the increase in data dimensionality and complexity entails two issues (LEEUWEN; KNOBBE, 2011). The first one is that such data leads to huge hypothesis

[1] How much smaller depends on the design of the descriptions. The (descriptive) search space may be exponentially large when representing more complex data types.

spaces, making the solution space's complete traversal infeasible. To tackle such a problem, *heuristic searches* employ different strategies to favour the search towards regions of the space that are more likely to contain good solutions (according to ϕ). This approach, however, bumps into the second issue.

The high dimensionality and cardinality of the data (and dependencies between descriptive attributes) usually produce multiple possible slight variations of a (subgroup's) description. The existence of these multiple variations leads to the problem of redundancy in sets of subgroups: the mining algorithms usually struggle with local minima yielding a large number of slight variations of a particular finding. In other words, the frameworks to discover subgroups strive to assure good quality solutions while delivering a variety of different findings.

In the remainder of this chapter, we first define the problem of redundancy and then present the heuristic search strategy we employ to generate subgroups and tackle such a problem.

2.3 REDUNDANCY IN SUBGROUP SETS

To enunciate the problem of redundancy in sets of subgroups, let us consider the following set of four subgroups based on the toy data set provided in Table 1. In the left column, we provide the description of each subgroup G_i and the subgroup's coverage in the right column.

$\boldsymbol{G_i}$: **(subgroup) Description**	$\boldsymbol{cov(G_i)}$
$G_0 : type \in \{malignant\}$	$\{\#1, \#2, \#3, \#5\}$
$G_1 : type \in \{malignant\} \wedge metastasis \in \{no\}$	$\{\#1, \#2\}$
$G_2 : type \in \{malignant\} \wedge location \in \{II\}$	$\{\#2, \#5\}$
$G_3 : type \in \{malignant\} \wedge metastasis \in \{no\} \wedge location \in \{II\}$	$\{\#2, \#6\}$

Font: The author (2021)

Note that the first subgroup G_0 is the most *general* subgroup. The other ones (namely G_1, G_2, G_3) are actually *specialisations* of G_0, i.e. they comprise subsets of $cov(G_0)$. When observing the subgroups' descriptions, we can also notice that the more specific subgroups are variations of the more general pattern, impelling additional constraints to the domain of descriptive attributes. In other words, they are *refinements* of G_0 subgroup.

Hence, we define *refinement* with respect to the subgroups' descriptions. We say a subgroup G_b is a refinement of a subgroup G_a if D_b imposes all constraints imposed by D_a plus at least one more conjunctive restriction $cond(A_i)$[2]. For instance, the subgroup G_2 is a refinement of G_0 but not of G_1; the subgroup G_3 is a refinement of all other subgroups.

The drawback of refinements when considering the task of subgroup mining is that they usually cover almost the same sets of individuals, which usually leads to similar (target) behaviours. In other words, even though refinements may comprise exceptional subgroups with relation to a baseline $\mathcal{B}$, they usually delineate several subsets of a more general (subgroup's) coverage that do not necessarily present distinct model responses. For instance, if we compare the survival model of G_0 with the models of G_1, G_2 and G_3, we may discover that they present (statistically) equal survival models. When considering high-quality (general) subgroups, their refinements usually also comprise high-quality findings. Thus, once heuristic searches move towards higher quality regions, the majority of subgroup mining approaches end up stuck in local minima, returning a high number of refinements in their final findings. This is the problem of *redundancy in sets of subgroups*: the presence of 'many (slightly) different subgroup descriptions covering many (almost) equal data subsets that present (almost) equal target distribution' (LEEUWEN; KNOBBE, 2012).

Inherent to the problem of redundancy is the lack of *diversity* in the findings. This is because several redundant subgroups actually refer to a single (more general) pattern. In a set of redundant subgroups, the *similar descriptions* (refinements) fail to encompass several regions of the (descriptive) search space, missing out on potential insightful knowledge. Because refinements represent *similar subsets* (coverages) of the data, a large number of individuals remain unrepresented, and no knowledge is provided about their behaviour. Thus, once such similar subgroups usually yield *similar models*, redundant sets of subgroups yield only a few essential discoveries among a potentially larger number of unusual behaviours existent in the (unexplored) data. Redundancy, therefore, arises with the presence of *similarities*. Minimising redundancy, i.e. the similarities between subgroups, potentially diversify – enlarge the range of – the knowledge provided, its extent, and the (hidden) target behaviours we strive to find. Thus, *diversity* is the opposite of redundancy.

Hence, we follow the discussion first presented by LEEUWEN; KNOBBE (2011) and we tackle redundancy considering three (subgroup's) dimensions: *descriptions*, *coverages*, and *models*. Thus, we understand that achieving diversity may require optimising redundancy in more than one dimension at a time, and therefore it should be minimised considering all

[2] Note that *refinements* presuppose no change in the $\mathcal{V}_i$ constraints already imposed by the conditions $cond(A_i) = A_i \in \mathcal{V}_i$ in the more general subgroup's description.

its (three) dimensions simultaneously. For instance, suppose the subgroups $G_a = (\text{'}size \in \{small\}\text{'}, \{\#0, \#1\})$ and $G_b = (\text{'}metastasis \in \{yes\}\text{'}, \{\#3, \#5\})$ are exceptional (w.r.t. a baseline) but present similar models compared to each other. In this case, although diversity was not achieved in the model dimension, one can understand that those two subgroups comprise different characterisations of two distinct subsets that happen to present similar unusual models. It is important to keep in mind that some level of redundancy (in any dimension) in the final set is inevitable – and somehow desired – because the task of searching for subgroups essentially entails intersections between local patterns.

Therefore, to achieve diversity, we define that each subgroup in a set of discovered subgroups should be exceptional concerning the considered baseline and comprise a distinct finding – in all its aspects. Hence, we define diversity in a set of subgroups in terms of those three dimensions of redundancy.

Definition 7 (**Set of Diverse (non-redundant) Subgroups**)**.** In a set $\mathbb{G}$ of *diverse* (non-*redundant*) subgroups, all pairs $G_i, G_j \in \mathbb{G}$ (for $i \neq j$) should substantially differ: in their *descriptions*, in their *coverages*, **and** in their *(survival) models*. □

Finally, to assess redundancy, we define measures to quantify the similarity between pairs of subgroups for each dimension of redundancy.

The *description* similarity $sim_D : \mathcal{P}(\mathbb{I}) \rightarrow [0, 1]$ (Equation 2.3) is computed over the set $\mathcal{I}$ of descriptive items. The *coverage* similarity $sim_C : \mathcal{P}(\Omega) \rightarrow [0, 1]$ (Equation 2.4) is computed over the subgroups' coverage, i.e. set of individuals comprising the subgroups. Hence, given two subgroups, G_a and G_b, we define such measures of similarity as follows.

$$sim_D(G_a, G_b) = \frac{|\mathcal{I}(G_a) \cap \mathcal{I}(G_b)|}{min(|\mathcal{I}(G_a)|, |\mathcal{I}(G_b)|)} \tag{2.3}$$

$$sim_C(G_a, G_b) = \frac{|cov(G_a) \cap cov(G_b)|}{min(|cov(G_a)|, |cov(G_b)|)} \tag{2.4}$$

Both descriptive and coverage similarities are defined as an adaptation of the Jaccard similarity coefficient more sensible to the overlaps between the compared sets. Such metrics are used to gauge the similarity and diversity of sample sets and, in their maximum values, indicate that one set is a subset of the other.

The *model* similarity sim_M is assessed as a boolean function based on the logrank test, where $pval_{G_a,G_b}$ is the p-value of the test between the (KM) models of the two compared subgroups and α is a predefined level of significance. The measure is defined in Equation 2.5.

$$sim_M(G_a, G_b) = pval_{G_a,G_b} > \alpha \tag{2.5}$$

As already discussed, to tackle scalability issues, heuristic strategies to search subgroups suffer from high levels of redundancy – since they lean towards the most promising areas of the search space, which usually comprise many variations of the same finding. To tackle such a problem, several approaches in the literature strive to provide diverse findings while assuring their good quality. In other words, they strive for a good balance between *exploration* and *exploitation*, i.e. the ability to find *multiple local optimum*.

In the next chapter, we will revise the usual heuristic approaches to mine subgroups (especially for the EMM and SD tasks). For now, we proceed to introduce the heuristic search strategy proposed in this work. Thus, next, we will briefly review the optimisation problem and introduce the Ant Colony Optimisation (ACO) meta-heuristic.

2.4 FUNDAMENTALS OF ANT-COLONY OPTIMISATION

From a computational perspective, an *optimisation problem* can be understood as something to be optimised, i.e. something to be made as good or effective as possible. *Optimisation algorithms*, therefore, are methods designed to find the best solution according to a *objective function* among all possible solutions of a given problem (CAVAZZUTI, 2012).

(GENDREAU; POTVIN, 2010) define *meta-heuristics* (TALBI, 2009) as the *solution methods that orchestrate an interaction between local improvement procedures and higher level strategies to create a process capable of escaping from local optima and performing a robust search of a solution space*. In other words, meta-heuristics are high-level methodologies designed to provide an approximate solution to a wide range of optimisation problems without the need to be deeply adjusted to each specific problem. This kind of algorithm has become popular in the past decades mostly because of its capability to provide satisfactory solutions to hard and complex problems (especially combinatorial ones) with reasonable computational resources. The particular strategy employed to achieve balance between exploration and exploitation poses the main difference between the existing meta-heuristics.

Nature-inspired (or bio-inspired) (FLOREANO; MATTIUSSI, 2008) is the name given to the family of optimisation algorithms whose design mimics a natural phenomenon – e.g. biological, physical – in order to solve optimisation problems (FAUSTO et al., 2020). They comprise robust methods for generating solutions regarding an objective function, performing a global search capable of exploring large search spaces without subdividing it or resorting to pruning techniques. There are different taxonomies for bio-inspired algorithms in the literature. Among them, there are two major categories: evolutionary computing (EC) and swarm intelligence (SI). EC is the subarea of optimisation that draws inspiration from the process of natural evolution, being mainly based on the principles of Darwinian evolution and genetics. In contrast to imitating the evolution process of an individual, SI based techniques focus on the interaction of several individuals and their environment, exploiting social and collective behaviour present in groups of animals.

Ant-Colony Optimisation (ACO) is a SI meta-heuristics first introduced by (DORIGO; MANIEZZO; COLORNI, 1991; DORIGO; MANIEZZO; COLORNI, 1996) as an approach to stochastic combinatorial optimisation, and it has been widely used to solve hard optimisation problems throughout the years. This approach is based on the foraging behaviour of some ant species and on the fact that such ants can find the shortest path between their nest and food sources, despite their limited individual capacity for orientation. For a deeper review on this meta-heuristic, we refer some works in the literature (DORIGO; STÜTZLE, 2019; DORIGO; CARO, 1999; DORIGO; STÜTZLE, 2003; DORIGO; BLUM, 2005; DORIGO; BIRATTARI; STUTZLE, 2006).

The main biological inspiration of ACO algorithms comes from the pheromone trail laying-and-following behaviour of real ants. Some species use *pheromone* (a chemical substance) as an indirect form of communication mediated by the environment. While searching for food, ants deposit pheromone on the ground creating a trail that other ants can follow. Moreover, those ants are biologically programmed to follow pheromones, and they tend to follow paths where pheromone concentration is higher. This characteristic of exploiting pheromone trails gives some ant species the ability to discover the shortest path leading to the food.

Analogously to real ants, ACO algorithms implement artificial ants that build solutions and exchange information on the quality of these solutions employing a communication based on pheromone trail. The artificial ant colony constitute an iterative procedure that stochastically *refines* solutions. In other words, each artificial ant constructs a complete solution by iteratively sorting *solution components* (to be added to a partial solution) considering a probabilistic distribution that entails: (1) artificial pheromone trails; and (2) heuristic information about the problem in hand (if available).

The stochasticity in ACO algorithms allows the exploration of a large number of solutions and hence the diversification of the constructed ones. The use of heuristic information helps to guide the search towards more interesting regions of the space, and the pheromone trails allow the algorithm's search experience to bias the solution construction in future iterations, in a way reminiscent of reinforcement learning (SUTTON; BARTO et al., 1998). Moreover, the use of a colony of ants increases the algorithm robustness. In many cases, this collective interaction of a population of agents enables the algorithm to efficiently solve the problem. To define the ACO meta-heuristic, we first define a general model of a combinatorial optimisation problem (COP).

Definition 8 (COP model). A model of a combinatorial optimisation problem consists of:

- a search space $\mathbb{I}$ defined by a finite set of decision (descriptive) attributes $A = \{A_1, \ldots, A_n\}$;
- a set $\mathbb{C}$ of constraints among the variables;
- an objective function $\phi : \wp(\mathbb{I}) \rightarrow \mathbb{R}^+$ to be maximised [3]

We define the search space as the set of all items in the data set $\mathbb{I}$ and the *solution components* as the items $I_{ij} = (A_i, v_j) \in \mathbb{I}$. We say that a set of items that satisfies all constraints in $\mathbb{C}$ is a feasible solution. We constraint solutions to contain at most a single item I_{ij} for each A_i, and to represent at least a minimum number of individuals. Additionally, constraints regarding the redundancy problem were implicitly considered in the design of the search and will be further introduced with the method presented in Chapter 5. As an approximate solution technique, ACO strives to find a good enough solution with reasonable computational resources.

The *pheromone trails* τ are the main element of ACO algorithms, and they provide the amount of pheromone associated with each solution component of the search space in a given iteration of the ant colony. Additionally, a problem-based *heuristic information* function η may be defined to quantify the quality of solution components. Hence, the probabilistic distribution of the solution space is a function of both the pheromone trails τ and heuristic information η (when provided).

Each ant in a colony constructs a feasible solution by relying on such distribution to iteratively assemble solution components. Then, at run-time, such ant updates the pheromone trails (to be used by the next ants) based on the quality of the solution generated. Therefore, τ is time-dependent, varying according to the colony's iteration; when necessary, we will use a subscript τ_{ant} to indicate a specific ant-iteration inside the colony execution. Hence, (candidate)

[3] From the ϕ function first defined as $\phi : \mathcal{G} \rightarrow \mathbb{R}$, we have that a subgroup $G \in \mathcal{G}$ entails (and thus can also be expressed by) the descriptive itemset $\mathcal{I} \subset \mathbb{I}$.

Algorithm 1: General ACO algorithm framework

```
1 Initialisation
2 while (termination condition not met) do
3     ConstructSolutions
4     LocalSearch (optional)
5     PheromoneUpdating
```

solutions are constructed using a parameterised probability distribution over the solution space at the same time that they are used to modify the pheromone values, biasing the search towards regions of the search space that are likely to contain high-quality solutions.

A general framework for ACO meta-heuristic is presented in Algorithm 1. At the beginning of the algorithm (line 1), all pheromone variables are equally initialised with an initial value τ_0. Also, the heuristic variables are defined, and they give the initial probability distribution of the solution space. Then, the main loop of the algorithm (line 2) iterates over three major steps: (i) the ants construct several solutions biased by the pheromone and heuristic information; (ii) the constructed solutions may be improved through an optional local search; and (iii) before the next iteration, the pheromone trails are updated to reflect the ants' search experience.

There is a variety of ACO algorithms proposed in the literature to perform predictive rule induction (FREITAS; PARPINELLI; LOPES, 2009). For instance, the *Ant-Miner* (ant-colony-based data miner) (PARPINELLI; LOPES; FREITAS, 2001) algorithm employs ACO as the search mechanism of the sequential covering strategy to mine a decision list of classification rules. Since its publication, it has been improved, adapted and extended by many contributions, proving to be a versatile approach and easily adaptable to a wide range of applications. In most cases, the proposed contributions employ different pheromone trails and heuristic information functions to seek balance between exploration and exploitation. The Ant-Miner has proven to be competitive regarding both state-of-the-art classification and rule learners algorithms, providing accurate classification models in the form of simple and comprehensible rules (PARPINELLI; LOPES; FREITAS, 2002). When considering the task of Exceptional Model Mining, it could potentially benefit from the Ant-Miner's predictive power – which assures the robustness of the objective function optimisation – and from the descriptive power of its rule-based models.

Before introducing our proposed methods, we revise the literature related to survival risk characterisation and the existing approaches to mine subgroups.

3 LITERATURE REVIEW

In the Chapter 1 of this work, we introduced the problem of identifying groups of patients closely related regarding their survival experience, our motivations to address this problem, and the key idea of our proposal. In Chapter 2, we first provided the theoretical foundations of this work. Thus, we introduced the local pattern mining task of Exceptional Model Mining to uncover unusual survival models associated with subsets (subgroups) of data individuals. Then, we defined the problem of redundancy in sets of subgroups and revised the heuristic search strategy we employ to discover a set of diverse subgroups.

In this chapter, we will provide a review of the related literature. First, we revise the existing methods and approaches used to provide knowledge over survival behaviour – and how they fail to identify and characterise unusual behaviours related only to subsets of a population. Then, we revise the main existing computational approaches to mine subgroups and their strategies to provide diverse findings – as we make clear the lack of approaches that allow the discovery of subgroups with exceptional survival behaviour.

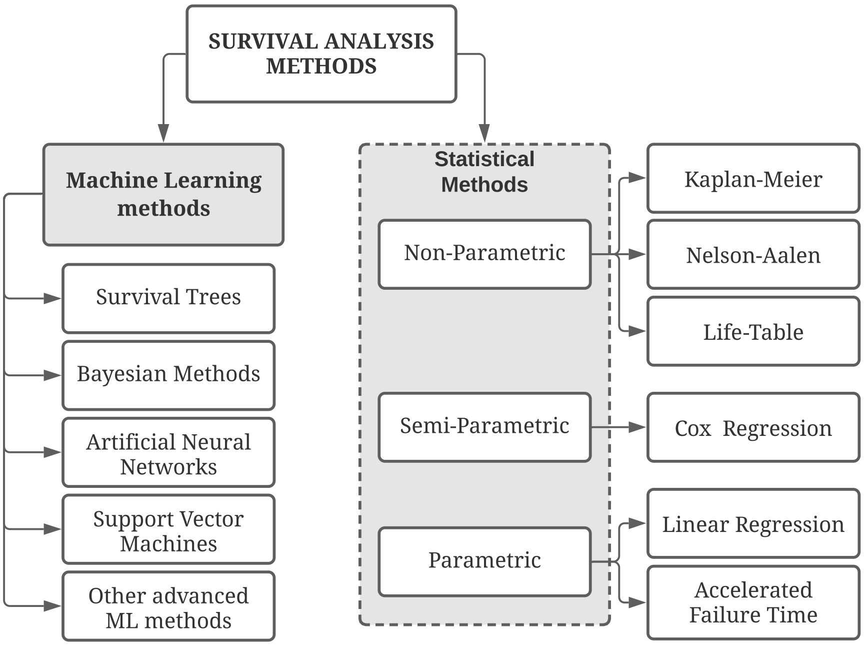
3.1 SURVIVAL RISK CHARACTERISATION

As introduced before, when studying *time-to-event* data, usually, there is a subset of the population under study for which there is no information (label) regarding if and when the event took place. This phenomenon, called *censoring*, is the main characteristic of survival data and the reason why standard predictive methods cannot be directly applied to analyse such data.

Survival Analysis (SA) methods have been developed over the years to assess the probability of an event happening and to model the impact of covariates on the occurrence of such an event. Such methods aim to provide mechanisms to appropriately handle censored data while providing accurate predictive models related to the time-to-event – usually time or risk prediction. WANG; LI; REDDY (2019) classify the SA methods into two broad categories: statistical and machine learning methods. Figure 1 provides a taxonomy of the main general methods developed for SA (some methods can be subdivided into more specific approaches; for a more complete taxonomy, we refer to the provided literature).

The traditional statistical approaches can be subdivided into: (i) non-parametric; (ii) semi-

Figure 1 – Taxonomy of Survival Analysis methods



Font: Adapted from WANG; LI; REDDY (2019)

parametric models; and (iii) parametric models. However, such methods rely on distributional and restrictive (linearity) assumptions that need to be fulfilled to achieve meaningful results. When these assumptions cannot be satisfied (and they are often not satisfied), these methods suffer from inconsistencies and sub-optimal (inaccurate) results. In addition, such methods struggle to model high-dimensional problems (with the need for feature selection), and their resulting models are often of difficult interpretation. For further explanation about the statistical methods, we suggest the aforementioned literature and the reading of a series of four papers: (CLARK et al., 2003a), (BRADBURN et al., 2003b), (BRADBURN et al., 2003a), and (CLARK et al., 2003b).

In order to overcome those limitations of statistical methods, many recent works have adapted machine learning methods to address the challenges of survival data analysis. (WANG; LI; REDDY, 2019) highlight that both statistical and machine learning methods aim at the same goal: *to make predictions of the survival time* (time to the event occurrence) *and estimate the survival probability at the estimated survival time.* Hence, machine learning approaches incorporate those traditional statistical methods into different machine learning techniques, providing more robust predictive survival models. They present the advantage of not imposing distributional assumptions while modelling non-linear relationships and delivering high-quality results. Machine learning methods most commonly used in SA are briefly introduced in the following. For a more thorough review on machine learning methods for SA, we refer to the work of WANG; LI; REDDY (2019).

- *Survival Trees* (GORDON; OLSHEN, 1985; BOU-HAMAD et al., 2011) are an adaptation of classification and regression trees to handle censored data. The ultimate estimator comprises a partition of the explanatory feature space and a Kaplan-Meier estimate for each subset in the partition. Over the years, many tree-based methods have been proposed to SA with the goal of predicting the distribution of the conditional survival function for new data examples, e.g. SEGAL (1988), DAVIS; ANDERSON (1989), LEBLANC; CROWLEY (1992). The idea of survival trees has also been extended to ensemble models, like bagging (HOTHORN et al., 2004) and random forests (ISHWARAN et al., 2008).

- *Bayesian methods* have been applied in the context of survival prediction, providing the probability of the event of interest. Most approaches make use of the Naive Bayesian classifier (KONONENKO, 1993; KONONENKO, 2001; ZUPAN et al., 2000) and Bayesian networks (NEAPOLITAN et al., 2004; LUCAS; GAAG; ABU-HANNA, 2004). They are a useful tool for knowledge representation, capable of inferring predictive models while providing comprehensible explanations and visual representation of features' interactions. Bayesian models are also applied to improve handling censored data (ŠTAJDUHAR; DALBELO-BAŠIĆ; BOGUNOVIĆ, 2009; ŠTAJDUHAR; DALBELO-BAŠIĆ, 2010) and to improve the efficiency of other Survival Analysis methods (RAFTERY; MADIGAN; VOLINSKY, 1996; FARD et al., 2016).

- *Artificial Neural Networks* (ANN) are usually employed to directly predict a subject's survival time or to provide the survival status of a subject (event-occurrence or event-free). Some works associate ANN with partial logistic regression (BIGANZOLI et al., 1998), Bayesian models (LISBOA et al., 2003), and statistical methods (FARAGGI; SIMON, 1995). However, ANN usually lack the transparency of generated knowledge and the ability to explain the decisions (KONONENKO, 2001), which is highly relevant in a wide range of applications.

- *Support Vector Machine* have also been applied to the analysis of survival data to predict the order in which the event happens for a group of samples (BELLE et al., 2007; EVERS; MESSOW, 2008) or to predict survival times (SHIVASWAMY; CHU; JANSCHE, 2007). In literature, there are also work on Support Vector Regression (KHAN; ZUBEK, 2008; BELLE et al., 2011) and on Relevance Vector Machine (WIDODO; YANG, 2011).

- Other machine learning approaches adapted to survival analysis that are found in the literature are *active learning* (VINZAMURI; LI; REDDY, 2014), *transfer learning* (LI et al., 2016b) and *multi-task learning* (LI et al., 2016a).

Most machine learning methods designed to analyse survival data strive to build more accurate models to predict the survival time variate while struggling to handle the challenge of appropriately dealing with censored data. Methods such as ANN and the bayesian ones deliver time predictions in the form of risk/survival scores or probabilities. Tree-based methods and classifiers usually deliver partitions of the data set based on covariates' split criterion and stratifications, aiming to maximise their predictive models' accuracy.

Because they trade explainability for accuracy, most of those approaches cannot provide comprehensible insights over the factors associated with survival outcome. Some approaches resort to explainability techniques to assign a (quantitative) importance value to features depending on their contribution to a prediction (MONCADA-TORRES et al., 2021). Note, however, this means extracting some characterisation from a *global* model. Hence, once machine learning methods aim to optimise a given (predictive) metric, they fail to provide information over more local aspects of the data. When striving to characterise subgroups with respect to their survival behaviour (e.g. "high-risk" and "low-risk" survival groups), such approaches usually set thresholds to the survival time variate or rely on features that are (supposedly) related to the outcome.

When attempting to better explain the factors related to the survival response, some works in the literature employ rule-based approaches due to their simplicity in representing patterns and features' relationships.

BAZAN et al. (2002) propose a rough sets approach to find descriptions of patient groups with different Kaplan-Meier models by inducing a set of decision rules. The rules are induced targeting predefined intervals of a prognostic index based on the (semi-parametric) statistical Cox's PH model. They also compel the observations to artificial classes to search for deviations given a predefined stratification feature. (PATTARAINTAKORN; CERCONE, 2008) propose a rough sets hybrid system to predict the survival time. The approach presents a preprocessing step that uses the survival data and domain knowledge to select the significant risk factors (essential features). For the final rules, the survival time feature is discretised, and the prediction is given in the form of time intervals.

LIU et al. (2004) use a bump hunting (FRIEDMAN; FISHER, 1999) method to characterise high-risk patients. The goal is, then, to subdivide the feature space in regions with a high average value of the target variable. The approach targets the deviance residual (LEBLANC; CROWLEY, 1992) as a substitute for the (censored) survival time feature.

KRONEK; REDDY (2008) propose the Logical Analysis of Survival Data to construct patterns to estimate the survival probability distribution of observations. The approach constructs a set of rules by partitioning the observations regarding their survival status given a specific time. Then, a greedy bottom-up approach is employed to maximise the separability power of the patterns according to a metric. The (predicted) survival function of a new observation is given by averaging the Kaplan-Meier models of all patterns covering such observation (including the estimates over the entire data set).

WRÓBEL (2012) uses a survival tree to generate an ordered set of rules to predict the survival behaviour of new examples. The ruleset is constructed by iteratively learning a survival tree on the uncovered observations and then selecting the rule (the path from a leaf to root) that maximises the difference between the Kaplan-Meier model of the data observations covered by the rule and the remaining observations.

SIKORA et al. (2013) employ the sequential covering strategy (FÜRNKRANZ, 1999) to induce a set of classification rules. The approach generates a partition of the observations into classes regarding their survival status, and a greedy approach is used to induce classification rules from non-censored observations. In (SIKORA et al., 2014), the authors apply the covering strategy together with a weighting scheme for handling censoring.

WRÓBEL; GUDYŚ; SIKORA (2017) present the LR-Rules, a top-down greedy covering algorithm to induce rules for estimating the survival function of new observations. The rule induction process maximises the statistical difference between the Kaplan-Meier model of the observations covered by the rule and the remaining observations. The algorithm iteratively constructs rules by exhaustively searching for the condition whose addition yields the highest separability between the KM models. The conditions to be added are taken from the set of observations currently covered by the rule. The algorithm allows overlapping between rules, and the sequential covering approach relies on a minimum number of previously uncovered observations that need to be encompassed by each new rule. The survival function of a new observation is given as the average survival estimates of all rules it is covered by – or by the population survival model, in case the observation is not covered by any rule in the set.

Although rule-based approaches provide the advantage of delivering interpretive results, the existing approaches still aim at the same goals of machine learning and statistical methods: to predict survival time (distribution) and estimate/classify risks. Thus, they provide global models – the set of rules – that maximises the accuracy of a target prediction.

When striving to characterise differences in survival behaviours, most approaches impel decision classes to the survival data to generate predictive rules. Despite their capability of providing straightforward explanations, such approaches still rely on features' stratification and prior knowledge about feature interactions. From the reviewed literature, only the works presented by (WRÓBEL, 2012) and (WRÓBEL; GUDYŚ; SIKORA, 2017) employ a direct induction from survival data maximising the unusualness of a predictive survival model related to a pattern's coverage. Still, their final rulesets comprise global models that maximise accuracy. Once again, by optimising global predictive metrics, such approaches potentially miss the identification of local patterns. Although we have robust methods to predict survival, the computational tools at our disposal do not look for local (exceptional) aspects in data.

An alternative approach to such predictive methods was proposed by PARK; PARK; YOO (2019) aiming to characterise survival behaviour focusing on local exceptionality detection in contrast to global models. They proposed a Subgroup Discovery (SD) algorithm, RIAS, that comprises a tree-based rule induction approach in which the target is the average survival time deviance. The rule induction tree is built in a general-to-specific method with a depth(best)-first regime, and the subgroup rules are created from the final rule tree. The relevant subgroups are selected by applying a statistical test to assess the deviation between the average survival time of a subgroup and its complement on the data set. The authors propose an SD approach to discover interesting patterns for long-term and short-term survival in breast cancer. To investigate long-term and short-term survival patterns, the authors consider an increase/decrease of a minimum mean difference (delta) in the statistical t-test between subgroup and complement. However, approaching the problem as a standard SD task may lead to the loss of interesting patterns due to outliers in the subgroups. As pointed out by DUIVESTEIJN; FEELDERS; KNOBBE (2016), a single variable is an oversimplified target representation, and more complex models can usually better represent the data. In other words, the deviation of the average survival time may not capture crucial information of the study cohort's survival experience. By generalising the target concept to mathematical models, the EMM task makes it possible to represent survival behaviour through more robust (predictive) models as the (supervised) property of interest.

As discussed in (ATTANASIO, 2019), there is a potential value of supervised LPM (specifically EMM task) as a valid data-driven solution to many medical goals and a gap in the literature on providing such solutions. When specifically addressing the problem of mining local patterns related to a survival response, there are only a few approaches in the literature. Thus, those approaches present significant restrictions in the representation of survival behaviour, which potentially leads to sub-optimal results.

Before introducing our proposal on using EMM to discover and identify unusual survival behaviours, we review the main existing computational approaches to mine subgroups – especially the SD and EMM tasks.

3.2 SEARCHING FOR LOCAL PATTERNS

As previously introduced, the goal of *local pattern mining* is to discover subsets of the data that are interesting somehow. The problem of finding *subgroups* restricts such a problem to discovering data subsets that can be concisely described in terms of a finite universe of attributes, i.e. *descriptive attributes*. With the data explosion, the space of descriptive attributes came to entail large amounts of attributes and a variety of complex data types, leading to exponentially large search spaces. Hence, the strategy used in the search for subgroups is an essential issue for a good performance of computational methods. Over the years, many algorithms have been developed to efficiently traverse search spaces and thus deliver interesting subgroups with satisfactory computational cost.

The earliest approaches to subgroup search resort to the *exhaustive search* strategy to explore all combinatorial space and, thus, deliver the best solutions. From the main exhaustive approaches in literature, we highlight the SD algorithms – most of them comprising adaptations of traditional association rule learning approaches to the search of subgroups: EXPLORA (KLöSGEN, 1996), MIDOS (WROBEL, 1997), APRIORI-SD (KAVŠEK; LAVRAČ; JOVANOSKI, 2003; KAVŠEK; LAVRAČ, 2006), SD-Map (ATZMUELLER; PUPPE, 2006) and SD-Map* (ATZMUELLER; LEMMERICH, 2009), DPSubgroup (GROSSKREUTZ; RÜPING; WROBEL, 2008), and MergeSD (GROSSKREUTZ; RÜPING, 2009). Additionally, the GP-Growth (LEMMERICH; BECKER; ATZMUELLER, 2012) was explicitly developed for the EMM task. There are also SD approaches to big data (e.g. PADILLO; LUNA; VENTURA (2016) and PADILLO; LUNA; VENTURA (2017)) based on MapReduce, allowing the processing of large databases through automatic parallelisation of the computation over a cluster of machines. In order to tackle scalability problems, those algorithms typically rely on pruning techniques or, sometimes, resort to anti-monotonicity restrictions to quality measures to reduce the search space and thus improve efficiency. Still, even with these tricks, exhaustive approaches become infeasible when applied to high dimensional and complex data.

In this context, heuristic strategies arise as an alternative to the exhaustive search once they restrict the search space to fractions that are more likely to contain interesting patterns. The usual heuristic approach among SD and EMM algorithms is the greedy *beam search* strategy (LOWERRE, 1976). It performs a level-wise search similar to the best-first search. However, it selects a predefined number of best candidates (given by a *beam size* parameter) among all partial solutions to keep for the next level. The new candidates are, then, generated from the best candidates kept in the previous level. However, a great disadvantage of this strategy is the lack of diversity in the discovered patterns. By exploring only the (best) parts of the search space, i.e. only some of the best candidates are considered, this strategy often yields sets of redundant patterns. Thus, this strategy may eliminate significant candidates and lack the ability to characterise other potentially interesting subsets of the data. The most popular SD algorithms that employ beam search are the SubgroupMiner (KLÖSGEN; MAY, 2002), SD (GAMBERGER; LAVRAC, 2002), CN2-SD (LAVRAČ et al., 2004) and RSD (LAVRAČ; ŽELEZNỲ; FLACH, 2002; ŽELEZNỲ; LAVRAČ, 2006). Also, the Cortana Subgroup Discovery[1] (MEENG; KNOBBE, 2011) is an open-source Java implementation for both SD and EMM applying a variety of target concepts. The tool supports both nominal and numeric single target (SD) and more complex targets such as regression and correlation (EMM). Additionally, it offers a large variety of quality measures. To tackle the problem of redundancy, some works in the literature apply weighted covering to increase diversity in the final set of subgroups. In the Diverse Subgroup Set Discovery (DSSD) (LEEUWEN; KNOBBE, 2012) approach, the authors also incorporate pattern set selection to the beam search strategy tackling redundancy in subgroups descriptions, coverage or models. For a specific type of redundancy (only one type can be approached), the framework implements a different subgroup selection procedure within the level-wise search – instead of choosing the top-K best subgroups – and in the selection of the final set. Thus, such an approach evaluates sets of subgroups to minimise redundancy, instead of evaluating subgroups individually; note, however, that such a global approach differs from the local aspect of traditional SD/EMM tasks.

[1] Cortana website: https://datamining.liacs.nl/cortana.html

Alternatively, many approaches in the literature employ stochastic optimisation as a subgroup search heuristic strategy, especially *Evolutionary Computing* (EC). The design of this family of algorithms provides a good balance between solution quality and response time, and their flexibility in the solutions' representation is a valuable ally to the descriptive aspect of subgroup mining. Thus, their search operators provide great flexibility in the trade-off between exploration and exploitation. Algorithms based on evolutionary computation have been widely explored to the discovery of interesting subgroups. However, the main approaches in literature are designed to the univariate target of SD task, and they are: the evolutionary algorithms SDIGA (JESUS et al., 2007), GAR-SD (PACHÓN et al., 2011), and EDER-SD (RODRÍGUEZ et al., 2012); the evolutionary programming approach CGBA-SD (LUNA et al., 2013; LUNA et al., 2014); and the multi-objective approaches MESDIF (BERLANGA et al., 2006; JESUS; GONZÁLEZ; HERRERA, 2007) and NMEEF-SD (CARMONA et al., 2009; CARMONA et al., 2010). There are also few approaches that focus on high dimensional data: the MEFASD-BD (PULGAR-RUBIO et al., 2017), a multi-objective evolutionary fuzzy SD algorithm for big data enviroments; and the SSDP (PONTES; VIMIEIRO; LUDERMIR, 2016; LUCAS et al., 2017), a mono-objective evolutionary approach for searching *top-K* subgroups. When approaching the redundancy problem, LUCAS; VIMIEIRO; LUDERMIR (2018) present the SSPD+, an evolutionary approach for SD that aims at providing diversity in top-k subgroups by storing and aggregating redundant subgroups in order to provide more informative results. Despite the advantages that optimisation meta-heuristics provide in searching for subgroups, to the best of our knowledge, no literature explores their use as the heuristic search strategy of the EMM framework. In fact, all presented evolutionary approaches assume a single nominal target. Hence, when considering the analysis of survival data, they would resemble some of the previously revised rule-based approaches, where the survival status is used as a nominal target while striving to incorporate censoring information.

Apart from those three major search strategies already introduced, there are other heuristic contributions to the EMM task in literature. LEEUWEN (2010) introduces the Exception Maximisation and Description Minimisation (EMDM) algorithm. It employs a search strategy that explores structures in the two data subspaces: the descriptive attribute and model spaces. The approach iteratively improves candidate subgroups. Each iteration consists of two steps: Exception Maximisation (EM), which searches for subsets presenting an unusual model, and Description Minimisation (DM), which aims to find a concise description to define a subgroup from the found subset. For the (model) exceptionality measures that the authors provide, all target attributes are assumed to be nominal. MOENS; BOLEY (2014) propose an alternative approach to EMM by extending and adapting a randomised technique to pattern discovery – Controlled Direct Pattern Sampling (CDPS) (BOLEY et al., 2011). The approach defines a sampling process that yields patterns according to a controlled distribution favouring patterns with high frequency and significant model deviation. However, all features in the data set are impelled to be either all numeric or all nominal. KRAK; FEELDERS (2015) present the TGCA (Tree-Constrained Gradient Ascent), a heuristic search strategy that employs numerical optimisation based on gradient ascent. It aims to find subgroups extents exploiting information about the influence of individual records on the quality of a subgroup while assuring that the subgroups can be concisely described. This approach, however, requires a differentiable quality measure.

In addition to the discussion of proper search strategies to handle scalability and computational challenges, another relevant aspect largely discussed in the broad scope of subgroup mining is the problem of redundancy. Apart from those approaches already introduced, other approaches in the literature strive to avoid redundancy in the set of discovered subgroups or, in other words, that strive for diversity in their findings. Early approaches, e.g. (KNOBBE; HO, 2006; BRINGMANN; ZIMMERMANN, 2007), draw inspiration from the domain of feature selection to introduce pattern filtering techniques as a post-processing step following the pattern discovery. The aim is to select smaller (and more comprehensible) sets of informative patterns but with minimal redundancy. This notion of *subgroup set mining* (RAEDT; ZIMMERMANN, 2007; GUNS; NIJSSEN; RAEDT, 2011) – where instead of searching for patterns that individually satisfy local constraints, one should strive to find a small set of patterns that together satisfy global constraints – is largely present in the literature. LEEUWEN; KNOBBE (2011) introduce such notion into the beam-search heuristic to the SD/EMM tasks, which is further explored in the DSSD framework. Other heuristic approaches in SD context relying on the same notion use Monte Carlo Tree Search (BOSC et al., 2018), greedy optimisation (BELFODIL et al., 2019), and Minimum Description Length (PROENÇA; BÄCK; LEEUWEN, 2021).

In this work, we follow the path of stochastic optimisation and propose the use of Ant-Colony Optimisation (ACO) as a heuristic search strategy. In addition to the advantages of the broad family of stochastic meta-heuristics, the ACO design easily incorporates data information to the search process. Moreover, it allows exploring aspects of its own experience to iteratively improve the search. To address the problem of redundancy, we build on the notion of *pattern set mining*. We tackle the problem through two different fronts. We exploit the sequential covering strategy and ACO design mechanisms to iteratively re-weight the space of items and data objects. In this way, we implicitly consider diversity in the pattern search by dynamic weighting the impact of observations and solutions components in future iterations. Moreover, we build on a two-step approach where a (smaller) set of subgroups is selected subsequent to the discovery process to iteratively update the (final) set of subgroups minimising redundancy.

We now proceed to introduce our proposed approaches in the following chapters.

4 ESMAM: EXCEPTIONAL SURVIVAL MODEL ANT-MINER

In this chapter, we address the problem of discovering subgroups of patients with unusual survival behaviour through the perspective of EMM, in contrast to the majority of existent predictive approaches. We aim at providing straightforwards characterisations about the local survival exceptionalities existent in the data.

Therefore, the main goal of this chapter is to present the Exceptional Survival Model Ant Miner (Esmam) algorithm, an EMM framework designed for discovering subgroups with statistically unusual survival models. The Esmam relies on a measure of exceptionality between survival curves (Eq. 2.2) based on the logrank statistical test to guide the subgroup search. Ultimately, it provides a set of exceptional subgroups: a list of descriptions delineating subsets of the data presenting unusual survival responses. In contrast to most EMM frameworks that employ greedy heuristics, we propose the use of Ant-Colony Optimisation (ACO) meta-heuristic (see Section 2.4) as the subgroup search strategy.

The work presented in this chapter was previously published in (MATTOS et al., 2020b). This chapter comprises the method description and empirical evaluation presented in the publication.

4.1 FRAMEWORK

The Esmam algorithm is an adaptation of the well-known classification rule induction algorithm Ant-Miner (PARPINELLI; LOPES; FREITAS, 2002). We adapted the Ant Colony Optimization heuristic to discover subgroups with exceptional KM curves. The Esmam returns a rule-based model comprising a set of (descriptions of) subgroups that are exceptional w.r.t. their *complement*. Its pseudocode is provided in Algorithm 2.

The algorithm is initialised with an empty set of subgroups $\mathbb{G}$ and with a set of uncovered cases $\mathbb{U}$ comprising all observations in the data set. Then, it follows a covering-based approach. In each iteration (lines 4-24), a colony of ants is initialised (line 5) and then constructs several subgroups G_D (lines 8-18). Then the best one (G_{best}) – according to φ (Eq. 2.2) – is selected to be added to the set of discovered subgroups $\mathbb{G}$, and the examples it covers ($cov(G_{best})$) are removed from $\mathbb{U}$ (lines 19-22). A new subgroup G_{best} is only added to the final set if it satisfies a lower quality bound for assuring exceptionality (line 19) and if it comprises a unique subgroup *description* (line 20). This process is repeated while the number of remaining

Algorithm 2: Esmam Framework

```
   Input: $maxU, maxIt, nAnts, nConverg$, $minCov$, $\alpha$
   Output: $\mathbb{G}$ – set of exceptional subgroups
   Data: $\Omega$ – survival data set
1  $\mathbb{G} \leftarrow \varnothing$
2  $\mathbb{U} \leftarrow \Omega$
3  $it \leftarrow 0$
4  while $|\mathbb{U}| > maxU$ or $it < maxIt$ do
5      searchInitialisation($\Omega$)
6      $ant \leftarrow 0$; $converg \leftarrow 0$
7      $G_{-} \leftarrow G_{\varnothing}$; $G_{best} \leftarrow G_{\varnothing}$
8      while $ant \leq nAnts$ or $converg \leq nConverg$ do
9          $D \leftarrow$ buildDescription($\Omega, minCov$)
10         $D \leftarrow$ pruneDescription($D$)
11         pheromoneUpdating($\mathcal{I}(G_D)$)
12         if $\phi(G_D) > \phi(G_{best})$ then
13             $G_{best} \leftarrow G_D$
14         if $\mathcal{I}(G_D) = \mathcal{I}(G_{-})$ then
15             $converg \leftarrow converg + 1$
16         else: $converg \leftarrow 0$
17         $G_{-} \leftarrow G_D$
18         $ant \leftarrow ant + 1$
19     if $\phi(G_{best}) \geq (1 - \alpha)$ then
20         if $\forall G \in \mathbb{G}(\mathcal{I}(G) \neq \mathcal{I}(G_{best}))$ then
21             $\mathbb{G} \leftarrow \mathbb{G} \bigcup \{G_{best}\}$
22             $\mathbb{U} \leftarrow \mathbb{U} \setminus cov(G_{best})$
23     else: break
24     $it \leftarrow it + 1$
25 return: $\mathbb{G}$
```

uncovered observations do not achieve a maximum threshold $maxU$ or until a maximum number of iterations is reached ($maxIt$). In case the ant colony cannot discover significant rules, the algorithm is finalised, and the final set of subgroups is returned (line 23).

Having presented the overall (covering) framework, we now describe the main elements of ACO heuristics, already introduced in Algorithm 1.

The searchInitialisation function is responsible for initialising the pheromone values $\tau_0(I_{ij})$ and heuristic values $\eta(I_{ij})$ associated with each item $I_{ij} \in \mathbb{I}$ at the beginning of a colony ($ant = 0$). As already stated, at the beginning of each colony process, no pheromone has been deposited in the trails yet, and all solution components receive the same amount of pheromone given by $\tau_0(I_{ij}) = |\mathbb{I}|^{-1}$. Following the work in PARPINELLI; LOPES; FREITAS (2002), the heuristic information is given based on Shannon's entropy defined in Equation 4.1.

$$H(W|I_{ij}) = -\sum_{w=1}^{k} P(w|I_{ij}) \cdot \log_2 P(w|I_{ij}) \tag{4.1}$$

We considered an initial partition of the observations in Ω as those with survival time at least as long as the data set average survival time and those with shorter survival time. The quality of an item is then the normalised information gain, obtained by further partitioning observations based on it. The class entropy was computed inducing a partition on the observations according to a condition. Hence, the heuristic value η_{ij} associated to each $I_{ij} \in \mathbb{I}$ is given according to Equation 4.2.

$$\eta(I_{ij}) = \frac{\log_2 k - H(W|I_{ij})}{\sum\limits_{I_{ij} \in \mathbb{I}} \log_2 k - H(W|I_{ij})} \tag{4.2}$$

A new colony is created for each new algorithm iteration (line 4), and the pheromone and heuristic values are re-initialised. Note, however, that once the heuristic values are always computed over the entire data set (the same partition of the observations), they are constant and may be computed just once to spare computational cost. Hence, the colonies perform an independent (stochastic) search in each iteration, always starting from the same (initial) probabilistic distribution given by the configuration of η values.

The subgroup search is then performed by a colony of ants (lines 8-11). Each ant in a colony of $nAnts$ ants delivers a description in a two-step process: the stochastic description construction (line 9); and a local search pruning procedure (line 10). Then, the descriptive items $\mathcal{I}(G_D)$ of the constructed solution are used to update the pheromone trail for the next ant (line 11). This process is repeated for all ants in the colony or until the ants converge to a single solution, i.e. until the colony achieves a minimum threshold ($nConverg$) for identical sequential descriptions (lines 14-16).

The descriptions are induced in a general-to-specific approach and, hence, the refinement function `buildDescription` (line 9) starts from an empty description $D = D_{\varnothing}$ and iteratively generates a more complex one by adding (conjunctive) conditions $cond(A_i) = A_i(o) \in \{v_j\}$ to D. It is important to remember that we constrained the solutions to contain at most a single item I_{ij} for each A_i, and hence the final constructed description is a conjunction of conditions $cond(A_i)$ over singleton sets of $dom(A_i)$.

The addition of conditions is a stochastic procedure that chooses an item given a probability distribution based on both the pheromone and heuristic values. Following the work of PARPINELLI; LOPES; FREITAS (2002), we define the probability P_{ij} of an item I_{ij} to be sorted as given in

Equation 4.3

$$P(I_{ij}) = \frac{x_i \cdot \eta(I_{ij}) \cdot \tau(I_{ij})}{\sum\limits_{I_{ij}} x_i \cdot \eta(I_{ij}) \cdot \tau(I_{ij})}, \text{ for all } I_{ij} \in \mathbb{I} \tag{4.3}$$

where $x_i = 1$ if A_i is not yet represented in the (partial) description D being constructed, and zero otherwise. It is important to notice that this probability is a function of τ_{ant} and, therefore, is time-dependent, assuming different values for each ant in a colony process. This refinement process of iteratively sorting items stops when all A_i are represented in D or when a new condition results in a coverage size $|cov(D)|$ below a minimum threshold $minCov$.

When a full description D is constructed, the pruneDescription function (line 10) is a local search responsible for enhancing both the simplicity and the quality of the final solution. This procedure greedily removes conditions $cond(A_i)$ from D, each time eliminating the condition that leads to the largest improvement in the quality associated with the (pruned) description. The pruning stops when no conditions can be removed without decreasing the quality or when the description already encompasses only a single condition.

At last, the pheromoneUpdating function (line 11) is responsible for computing the ant's search experience in each iteration, generating the pheromone values for the next ant iteration. For the items represented in the final description D, i.e. for the set of items $\mathcal{I}(G_D)$, the pheromone is incremented proportionally to the subgroup's quality, as given in Equation 4.4. For the set of items $\mathbb{I} \setminus \mathcal{I}(G_D)$ not represented in D, an evaporation process is simulated by the normalisation of all τ values in the iteration $(ant + 1)$.

$$\tau_{ant+1}(I_{ij}) = \tau_{ant}(I_{ij}) \cdot (1 + \phi(G_D)) \text{ , for all } I_{ij} \in \mathcal{I}(G_D) \tag{4.4}$$

Finally, we present an analysis of the computational complexity of the Esmam provided in Algorithm 2. This analysis is divided into three parts: (1) the computational complexity of preprocessing the heuristic information; (2) the complexity of a single ant iteration; and (3) the complexity of a single Esmam iteration. Then, we combine the results of these three steps in order to determine the computational complexity of an entire execution of the algorithm. Note that the Esmam builds on the Ant-Miner and, thus, its complexity presents only slight differences from the analysis provided by PARPINELLI; LOPES; FREITAS (2002).

1. *Heuristic information preprocessing*: As previously analysed, the values of all $\eta(I_{ij})$ given by Equation 4.2 are constant throughout the whole algorithm execution and thus are precomputed as a preprocessing step. These values can be computed in a single scan of the data set. So, the time complexity of this step is $\mathcal{O}(|\mathbb{I}| \cdot |\Omega|)$.

2. *Ant iteration* (lines 8-22): Each ant in a colony will perform the following major steps: (i) description construction; (ii) evaluation of candidate descriptions; (iii) description pruning; and (iv) pheromone updating. The computational complexities of these steps are as follows.

   - *Build description* (line 9): The choice of an item to be added to the current description requires the consideration of all possible items with η and τ values already precomputed. Therefore, this step takes $\mathcal{O}(|\mathbb{I}|)$. In order to construct a description, an ant will choose a number of k conditions. Note that the value of k may vary significantly depending on the data set and previously constructed descriptions. Since each attribute can occur at most once in a description, we have that the complexity of building a description is given by $\mathcal{O}(k \cdot |\mathbb{I}|)$, for $k \leq |A|$, being A the set of descriptive attributes.
   - *Solution evaluation*: This process consists of measuring the quality of a description (subgroup), as given by Equation 2.2. This requires matching a description with k conditions with a data set of $|\Omega|$ cases, which takes $\mathcal{O}(k \cdot |\Omega|)$.
   - *Prune description* (line 10): The first pruning iteration requires the evaluation of k new candidate descriptions – each one obtained by removing one of the k conditions from the unpruned description. Each of these evaluations takes on the order of $(|\Omega| \cdot (k-1))$ operations (see the topic of *solution evaluation*). Thus, the first pruning iteration takes on the order of $(|\Omega| \cdot (k-1) \cdot k)$ operations, i.e. $\mathcal{O}(n \cdot k^2)$. The second pruning iteration takes $(|\Omega| \cdot (k-2) \cdot (k-1))$ operations and so on. The entire pruning process is repeated at most k times, so description pruning takes at most $|\Omega| \cdot (k-1) \cdot k + |\Omega| \cdot (k-2) \cdot (k-1) + |\Omega| \cdot (k-3) \cdot (k-2) + \cdots + |\Omega| \cdot (1) \cdot (2)$ operations, which is $\mathcal{O}(k^3 \cdot |\Omega|)$.
   - *Pheromone updating* (line 11): This step consists of increasing the pheromone of the items used in the pruned description, which takes $\mathcal{O}(k)$, and decreasing the pheromone of unused items, which takes $\mathcal{O}(|\mathbb{I}|)$. Since $k < |\mathbb{I}|$, pheromone update takes $\mathcal{O}(|\mathbb{I}|)$.

   Adding up the results derived in the four topics above, a single ant iteration takes $\mathcal{O}(k \cdot |\mathbb{I}|) + \mathcal{O}(k \cdot |\Omega|) + \mathcal{O}(k^3 \cdot |\Omega|) + \mathcal{O}(|\mathbb{I}|)$, which collapses to $\mathcal{O}(k \cdot |\mathbb{I}| + k^3 \cdot |\Omega|)$.

3. *Algorithm single iteration* (lines 4-24): Each iteration of the Esmam can be subdivided into three parts: (i) the pheromone initialisation; (ii) the entire ant-colony loop; and (iii)

the selection method to include a new subgroup into the final set $\mathbb{G}$ of subgroups.

- *Pheromone initialisation* (line 5): Each algorithm iteration starts by defining the values of all $\tau_0(I_{ij})$ (note that the values of η were already initialised in the preprocessing step). This step takes $\mathcal{O}(|\mathbb{I}|)$.
- *Ant-colony loop* (lines 8-18): In the topic above, we have defined the complexity of executing a colony's single (ant) iteration. Thus, to derive the computational complexity for the whole colony execution, the result of the topic above has to be multiplied by the number $nAnts$ of ants (in the worst scenario), taking $\mathcal{O}(nAnts \cdot [k \cdot |\mathbb{I}| + k^3 \cdot |\Omega|])$.
- *Subgroup selection* (lines 19-22): The method to include a newly discovered subgroup into the final set requires the comparison of k conditions with a number of $|\mathbb{G}|$ subgroups that comprise the final set. Hence, we have a complexity of $\mathcal{O}(k \cdot |\mathbb{G}|)$.

Hence, a single iteration of the algorithm takes $\mathcal{O}(|\mathbb{I}|) + \mathcal{O}(nAnts \cdot [k \cdot |\mathbb{I}| + k^3 \cdot |\Omega|]) + \mathcal{O}(k \cdot |\mathbb{G}|)$, which collapses to $\mathcal{O}(nAnts \cdot [k \cdot |\mathbb{I}| + k^3 \cdot |\Omega|] + k \cdot |\mathbb{G}|)$.

Finally, to derive the computational complexity for the whole algorithm execution, we have to multiply $\mathcal{O}(nAnts \cdot [k \cdot |\mathbb{I}| + k^3 \cdot |\Omega|] + k \cdot |\mathbb{G}|)$ by z – the total number of discovered subgroups (note that this number is not necessarily equal to the number $|\mathbb{G}| \leq z$ of subgroups currently in $\mathbb{G}$). Then, we add the computational complexity of the preprocessing step. Therefore, the computational complexity of complete execution of the Esmam algorithm is

$$\mathcal{O}\left(z \cdot \left[nAnts \cdot (k \cdot |\mathbb{I}| + k^3 \cdot |\Omega|) + k \cdot |\mathbb{G}|\right] + |\Omega| \cdot |\mathbb{I}|\right)$$

It should be noted that this complexity depends very much on the values of the number k of conditions per description and the number z of discovered subgroups, which are highly variable for different data sets. Additionally, the size of the descriptive space $\mathbb{I}$ (hence the data set dimensionality and feature's complexity) and the volume of the data set have also a significant impact on the algorithm performance.

When considering the worst-case scenario, the value of k conditions per description equals $|A|$. Thus, since the Esmam copes only with categorical attributes, we can assume that each attribute A takes only a small number of values so that $\mathcal{O}(|\mathbb{I}|)$ can be simplified to $\mathcal{O}(|A|)$. Hence, the formula for worst-case computational complexity is $\mathcal{O}(z \cdot nAnts \cdot |\Omega| \cdot |A|^3)$.

However, we emphasise that this worst-case scenario is unlikely to occur mainly for two reasons. First, in the description pruning step, the factor $\mathcal{O}(k^3 \cdot |\Omega|)$ was derived considering that the pruning process can be repeated k times for all descriptions, which – in practice – is highly unlikely. Second, we considered all descriptions with length $k = |A|$ for the worst-case analysis, which is very unrealistic.

4.2 EXPERIMENTS

We conducted experiments to evaluate our approach based on supervised LPM to discover and characterise subgroups of a population presenting unusual KM (survival) models. We aim at providing simple characterisations capable of representing the majority of the individuals in a population. Thus, we are interested in assessing whether the discovered subgroups represent survival behaviours existent in the data.

We compare the Esmam algorithm with the LR-Rules (WRÓBEL; GUDYŚ; SIKORA, 2017) (revised in Section 3.1), which is a greedy sequential covering algorithm for inducing rules by maximising the difference between the KM models of the rule coverage and its complement. Although it provides a global predictive model, its authors also suggested its application for finding descriptions associated with survival response. To the best of our knowledge, this is the only available computational tool to provide survival behaviour characterisation by inducing patterns directly from the survival data and based on a survival (model) response.

We assess the performance of our ACO-based approach against the LR-Rules greedy search regarding the descriptive aspect of the results and the quality of the discovered survival models. To evaluate the descriptive aspect of our findings, we assess the *interpretability* of the (sets of) subgroups' description and the *representativeness* of their coverages (CARVALHO; PEREIRA; CARDOSO, 2019). Table 3 describes all metrics used in the experiments.

To assess interpretability, i.e. how well humans understand the results provided, we employ two traditional metrics of rule-based approaches: the (average) length of the descriptions ($length_{AV}$); and the number of discovered subgroups ($\#sg$). The rationale is that smaller descriptions (i.e. less conjunctive conditions) are easier to understand. Thus, a smaller number of patterns (subgroups) provides information more comprehensible and actionable. To assess the representativeness of the findings, i.e. the extent of instances covered by the patterns, we assess the average (percentage) subgroup coverage ($sgCov$) and the data set coverage ($dbCov$). As previously introduced, we aim to provide subgroups as large as possible (w.r.t. their coverages) and, thus, a set of subgroups that encompass as many individuals in the data set as possible. At last, to evaluate the quality of the discovered (exceptional) survival models,

Table 3 – Method Esmam – Empirical evaluation metrics

Metrics	**Description**	**Definition**
Interpretability		
#sg	Number of discovered subgroups	$\lvert\mathbb{G}\rvert$
$length_{AV}$	Average subgroup description length	$\sum_{G\in\mathbb{G}} \frac{length(G)}{\lvert\mathbb{G}\rvert}$
Representativeness		
sgCov	Average (percentage) subgroup coverage	$\lvert\Omega\rvert^{-1} \cdot \sum_{G\in\mathbb{G}} \frac{\lvert cov(G)\rvert}{\lvert\mathbb{G}\rvert}$
dbCov	Data set coverage	$\lvert\Omega\rvert^{-1} \cdot \lvert \bigcup_{G\in\mathbb{G}} cov(G)\rvert$
Model Quality		
$IBS_{\mathbb{G}}$	IBS over a set of subgroups	$IBS_{\mathbb{G}} = \sum_{G\in\mathbb{G}} IBS_G$

Font: The author (2021)

we assess the integrated Brier score (IBS), which is a measure of the error between the KM estimated survival model $\hat{S}(t)$ and the cohort's real survival experience.

The Brier score (BS) (GRAF et al., 1999) measures the square difference between an observation's survival status δ and its estimated survival probability $\hat{S}(t)$, in a given time $T^* \in T$. The BS value for an observation o (incorporating censoring) is given by Equation 4.5, where $\hat{C}(t)$ is the KM estimate of the censoring distribution, obtained from estimating the survival function for $\delta = (1 - \delta)$. The IBS, given by Equation 4.6, is the score integrated over all survival times T and for n observations. The IBS_G associated to a subgroup G is calculated considering all $n = |cov(G)|$ individuals comprising the subgroup. Hence, the $IBS_{\mathbb{G}}$ calculated over a set of subgroups $\mathbb{G}$ is the sum of IBS_G for all $G \in \mathbb{G}$.

$$BS_o(T^*) = \begin{cases} \frac{1}{\hat{C}(T_o)} \left(0 - S(T^*)\right)^2 & \text{if } T_o \leq T^*, \delta_o = 1 \\ \frac{1}{\hat{C}(T^*)} \left(1 - S(T^*)\right)^2 & \text{if } T_o > T^* \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

$$IBS = \frac{1}{max(T)} \int_0^{max(T)} \left(\frac{1}{n} \sum_{i=1}^{n} BS_o(T^*) \right) dT^* \tag{4.6}$$

Next, we describe the process of empirical evaluation and analyse the results. Some contents – like the algorithm implementation, the data sets used in the experiments, configurations and results are available on Esmam repository.

4.2.1 Data sets and Experimental setup

We conduct experiments with 14 real-world survival data sets from the medical domain. These data sets were used as benchmark data in many survival analysis studies. Besides being frequently used as benchmark data, they are also of particular interest for us since we believe that one of the most suitable applications for our method is in the medical domain. In this case, we can also observe the behaviour of our approach in this domain. Table 4 contains the list of these data sets together with a brief characterisation.

The patterns we analyse are searched in the space of the domains of attributes A_i (i.e. the set of items $\mathbb{I}$). Then, the survival behaviour is analysed through the deviation of the target concept over the target features T, δ. As the Esmam algorithm is not adapted to process high-dimensional data, we selected data sets of low dimensionality. All numerical descriptive attributes were discretised with K-Means into five interval categories. Pre-processing of the data was employed to remove observations containing missing values (and features with a high level of missing data).

In Table 5, we provide the configuration of the algorithms compared. For the choice of Esmam parameters, we assumed the ACO framework setup defined by the authors in

Table 4 – Data sets – Characteristics of 14 survival data sets used in the experimental study: the number of observations ($|\Omega|$), the number of descriptive attributes ($|A_i|$), the number of descriptive attributes A_i that were discretised ($|A_i^d|$), the number of items in the data set ($|\mathbb{I}|$), the proportion of censored observations ($\%cens$), the *Subject of Research*, and the survival event description ($Event$)

data set (Ω)	$\lvert\Omega\rvert$	$\lvert A_i\rvert$	$\lvert A_i^d\rvert$	$\lvert\mathbb{I}\rvert$	%cens	Subject of research	Event
actg320	1151	11	3	39	91.66	HIV-infected patients	AIDS death/diagnosis
breast-cancer	196	80	78	269	73.98	Node-Negative breast cancer	distant metastasis
cancer	168	7	5	29	27.98	Advanced lung cancer	death
carcinoma	193	8	1	28	27.46	Carcinoma of the oropharynx	death
gbsg2	686	8	5	31	56.41	Breast cancer	recurrence
lung	901	8	0	23	37.40	Early lung cancer	death
melanoma	205	5	3	28	72.20	Malignant melanoma	death
mgus	176	8	6	30	6.25	Monoclonal gammopathy	death
mgus2	1338	7	5	23	29.90	Monoclonal gammopathy	death
pbc	276	17	10	61	59.78	Primary biliary cirrhosis	death
ptc	309	18	1	71	93.53	Papillary thyroid carcinoma	recurrence/progression
uis	575	9	4	33	19.30	Drug addiction treatment	return to drug use
veteran	137	6	3	23	6.57	Lung cancer	death
whas500	500	14	6	46	57.00	Worcester Heart Attack	death

Font: The author (2021)

Table 5 – Method Esmam – Information on the algorithms compared in the empirical evaluation

Algorithm	Search Strategy	Target Concept	Baseline ($\mathcal{B}$)	Source	Parameter(Value)
Esmam	ACO	KM model	Complement	Esmam repository	$maxU(0)$, $maxIt(3000)$, $nAnts(3000)$, $minCov(10)$, $nConverg(10)$, $\alpha(0.05)$
LR-Rules	Sequential covering	KM model	Complement	LR-Rules repository	–

Font: The author (2021)

(PARPINELLI; LOPES; FREITAS, 2002) as a robust configuration to provide satisfactory results regarding the optimisation task and descriptive aspects of the findings. We, however, force the complete representation of the data set – i.e. we set the number of maximum uncovered observations allowed to zero – so the results could be comparable with the LR-Rules. For the LR-Rules algorithm, we adopted the default parameters defined in the available implementation.

The experimental procedure presented in this section is conducted by running both the Esmam and the LR-Rules once on each of the 14 data sets already introduced. Then, statistical analysis of the results was performed by the Wilcoxon signed ranks test, using a significance level of 5%. We employ the test to assess whether or not the two compared approaches present statistically similar performances (null hypothesis) regarding the proposed metrics. We consider both the interpretability and the model quality minimisation metrics. For the representativeness measures, we consider maximisation. Additionally, we also analyse some individual discovered subgroups to evaluate whether our EMM approach can discover interesting survival patterns and retrieve essential characteristics from the data. Next, we present and analyse the results we achieved.

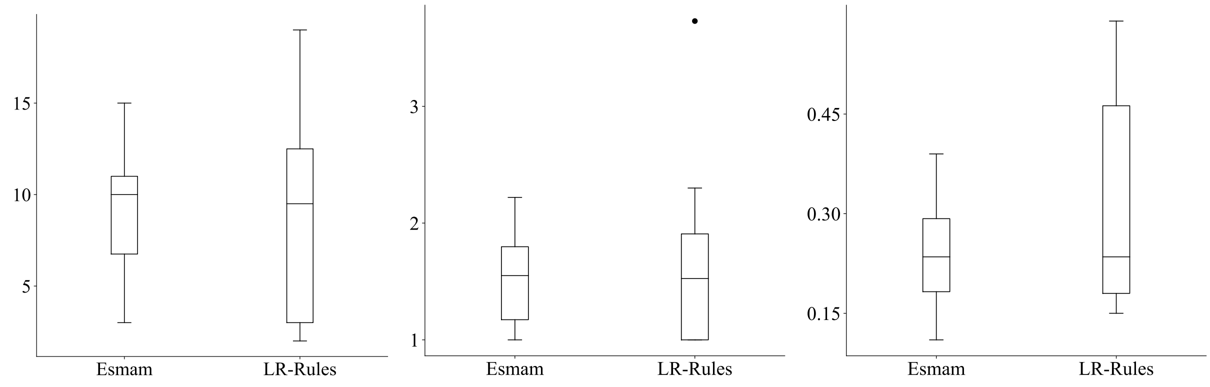
4.2.2 Results analysis

The results for both Esmam and LR-Rules algorithms on each data set are presented in Table 6. The Esmam algorithm returned sets of subgroups with, on average, 9.43 descriptions of (average) length 1.52 (condition), compared to the LR-Rules' average of 8.93 discovered subgroups of length 1.63. We notice then that Esmam was able of generating compact results concerning both the size of the subgroup set (number of returned patterns) and the length of the subgroups' descriptions. The coverage of Esmam subgroups was, on average, 25% of the total cases in the data sets, comprising patterns that neither cover the majority of the cases

Table 6 – Method Esmam – Evaluation metrics computed over the results provided by the Esmam and LR-Rules algorithms: the number of discovered subgroups ($\#sg$), the average subgroup description length ($length_{AV}$), the average (percentage) subgroup coverage (cov±std, $sgCov$), the data set coverage($dbCov$), and integrated Brier score on the rule set ($IBS_{\mathbb{G}}$). Bold values represent the best results.

Metrics	$\#sg$		$length_{AV}$		$sgCov$		$dbCov$		$IBS_{\mathbb{G}}$	
Algorithms	Esmam	LR-Rules	Esmam	LR-Rules	Esmam	LR-Rules	Esmam	LR-Rules	Esmam	LR-Rules
actg320	**9**	15	**2.22**	3.73	0.26 ± 0.16	0.15 ± 0.16	1.00	1.00	**0.43**	0.45
breast-cancer	**11**	19	**1.18**	1.95	0.24 ± 0.15	0.17 ± 0.10	0.94	**0.98**	0.01	0.01
cancer	11	**9**	2.00	**1.78**	0.17 ± 0.16	0.25 ± 0.16	0.74	**1.00**	0.08	**0.07**
carcinoma	10	**3**	1.80	**1.00**	0.27 ± 0.20	0.33 ± 0.30	0.99	0.99	0.06	**0.02**
gbsg2	14	**10**	**1.79**	2.30	0.19 ± 0.24	0.22 ± 0.23	1.00	1.00	0.13	**0.11**
lung	9	**7**	**1.00**	1.14	0.35 ± 0.13	0.35 ± 0.14	1.00	1.00	0.12	**0.09**
melanoma	6	**2**	1.00	1.00	0.39 ± 0.17	0.50 ± 0.06	1.00	1.00	0.02	**0.01**
mgus	13	**11**	**1.62**	1.73	0.11 ± 0.08	0.18 ± 0.11	0.71	**1.00**	0.01	0.01
mgus2	**6**	18	**1.17**	1.50	0.18 ± 0.09	0.17 ± 0.09	0.70	**1.00**	**0.38**	1.25
pbc	11	**3**	1.36	**1.00**	0.20 ± 0.28	0.59 ± 0.37	1.00	1.00	0.03	**0.02**
ptc	4	**2**	1.50	**1.00**	0.30 ± 0.36	0.50 ± 0.42	1.00	1.00	0.33	**0.08**
uis	15	**13**	**2.00**	2.08	0.23 ± 0.18	0.22 ± 0.17	0.99	**1.00**	0.36	**0.27**
veteran	**10**	11	1.60	**1.55**	0.18 ± 0.08	0.18 ± 0.08	0.84	**1.00**	0.16	**0.09**
whas500	3	**2**	1.00	1.00	0.38 ± 0.23	0.50 ± 0.19	1.00	1.00	0.04	**0.03**

Figure 2 – Method Esmam – Boxplots of the Esmam and LR-Rules algorithms results for: (a) the number of discovered subgroups; (b) the average description length; and (c) the average (percentage) subgroup coverage.



(a) $\#sg$

(b) $length_{AV}$

(c) $sgCov$

nor very small groups.

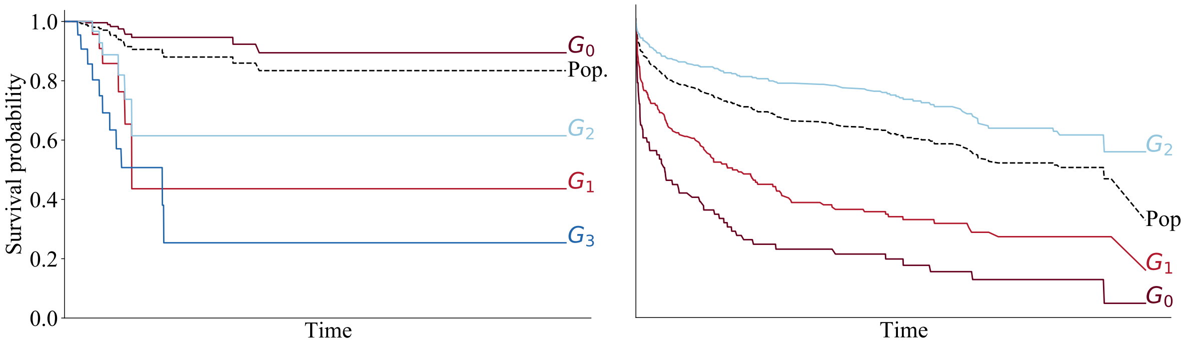
Figure 2 shows the boxplots of the performance of both algorithms with relation to the $\#sg$, $length_{AV}$ and $sgCov$ metrics. We notice that comparing to LR-Rules, Esmam results presented smaller variability. When evaluating the coverage of the data set (*dbCov*), Esmam showed greater variability, presenting in some cases, a higher percentage of observations that remained not covered by any subgroup. For the $IBS_{\mathbb{G}}$ results, Esmam algorithm presented an average of 0.15 comparing to 0.18 presented by LR-Rules. One could understand the $IBS_{\mathbb{G}}$ as a measure of the quadratic error between the survival estimates of the observations covered by a subgroup and their true survival status. Therefore, we notice that the Esmam algorithm was able to discover more homogeneous subgroups concerning survival response. Finally, for a level of significance of 5%, the Wilcoxon test showed statistically significant difference between Esmam and LR-Rules performances only in terms of the *dbCov* criterion (p-value $= 0.036$).

Besides, to evaluate Esmam final models in terms of the subgroups discovered by our EMM framework, we assess whether the induced subgroups present statistically significant survival models. Figure 3 presents the KM curves for the subgroup set discovered for *ptc* and *whas500* data sets. The plots additionally include the cohort's KM model, given by the *Pop.* curve. It is possible to observe the significant difference between the survival curve of the study cohort in comparison to the curves induced over the subgroups discovered by Esmam, indicating that the algorithm is able to identify local patterns with significant distinct survival response. In a more detailed analysis of the individual discovered patterns, we found that the algorithm was able to retrieve information on attributes that stratify the data into different survival experiences.

In the *actg320* data set, the *strat2* variable represents the counting of cells with expression of the CD4 protein, dividing the observations into low/high ($strat2 = 0$ /$strat2 = 1$) counting – where a low counting imply a higher risk for the patient. Among the nine resultant subgroups induced on this data set, the algorithm recovered such information presenting the following two subgroup descriptions: $\mathbf{G_7}$: $\{strat2 = 0\}$ and $\mathbf{G_8}$: $\{strat2 = 1\}$. Figure 4a presents the KM plot of both subgroups reflecting the expected survival behaviour.

In the *lung* data set, the $stage1 = \{1, 2, 3\}$ variable reflects the overall stage of lung cancer, for $stage1 = 1$ earlier than $stage1 = 3$. For the set discovered on this data set, the Esmam algorithm returned also nine subgroups, two of them: $\mathbf{G_2} : \{stage1 = 3\}$ and $\mathbf{G_6} : \{stage1 = 1\}$. Figure 4b presents the KM curves for both rules, showing that the

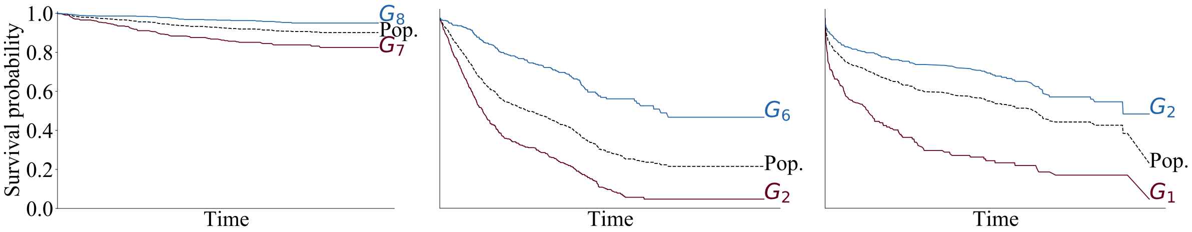
Figure 3 – Method Esmam – Analysis of the discovered subgroups for (a) *ptc*, and (b) *whas500* data sets. The *Pop.* curves represent the KM estimates on the entire cohort.



(a) ptc

(b) whas500

Figure 4 – Method Esmam – Analysis of individual subgroups induced for the following data sets: (a) *actg320*, (b) *lung*, and (c) *whas500*; the *Pop.* curves represent the KM estimates of the study cohort



(a) actg320

(b) lung

(c) whas500

survivability is better for early lung cancer stage.

In the *whas500* data set, the *chf* variable stands for *congestive heart complications*, dividing the observations into a group of patients that present complications and the ones that do not. The Esmam algorithm returned a set comprising three subgroups, two of them: $\mathbf{G_1}$: $\{chf = True\}$ and $\mathbf{G_2}$: $\{chf = False\}$. Figure 4c present the plot of both rules, showing that the presence of heart complications decreases the chances of survival.

4.3 DISCUSSIONS AND LIMITATIONS

In this chapter, we introduced our approach to discovering subgroups with unusual survival behaviour based on supervised descriptive pattern mining, in contrast to the predictive approaches existent in literature. We presented the Esmam (Exceptional Survival Model Ant Miner) algorithm, an EMM framework that uses ACO meta-heuristic for the subgroup search process. The algorithm returns a set of descriptions comprising subgroups presenting statistically unusual KM (survival) models compared to their complement on the data set. This proposed algorithm is the first approach for the EMM task to explore a bio-inspired meta-heuristic as the search strategy.

We evaluated our proposal by assessing its capability to return simple and representative descriptive findings and discover interesting patterns. Therefore, we tested our ACO-based approach to the discovery of local survival exceptionalities on 14 data sets. The performance of Esmam was evaluated in comparison to the LR-Rules algorithm – a greedy covering rule induction algorithm for survival data analysis. Our approach achieved competitive results concerning the simplicity of the final set of characterisations and the generality of the discovered patterns, performing similarly to LR-Rules with relation to the number of discovered subgroups, length of the induced descriptions, and subgroups' coverage. The only statistical difference observed between the performance of both algorithms was in the data set coverage. While the LR-Rules induces the full coverage of the data set, the Esmam algorithm presented higher percentages of data observations unrepresented by any discovered subgroup. When analysing the quality of the discovered survival patterns, the Esmam algorithm delivered survival models as accurate as the ones delivered by the LR-Rules predictive approach. The low $IBS_{\mathbb{G}}$ also indicates that the rules discovered by the Esmam comprise homogeneous subgroups with respect to survival response. When assessing the algorithm's capability of discovering unusual local survival behaviour, we notice that the Esmam was able to discover (statistically) significant subgroups and identify data characteristics that interfere with survival experience.

This approach, however, presents many limitations. First is the heuristic information that is constant throughout the algorithm execution. The constant heuristic adopted allocates the colonies always at the same initial place in the solution space, lacking potential *exploration*. Second, the subgroups in the final set are only restricted to the uniqueness of their descriptions (Algorithm 2, line 20). In other words, the only condition to accept a discovered subgroup as a final solution is that it has not already been accepted as so. Those two limitations together contribute to the problem of redundancy. By starting the search (always) with the same probability distribution, the Esmam algorithm potentially explores the proximity of the same regions of the search space. Thus, as revised in Section 2.3, pattern refinements comprise close locations of the search space and present close quality. Hence, the Esmam search is potentially confined among refinements. In most of the results achieved by the Esmam algorithm, we observed that it returns a variety of subgroups that are exceptional w.r.t. the defined baseline but which are, actually, many refinements of a more general subgroup. Consequently, the final set of discovered subgroups lacks diversity and comprises the characterisation of only a few interesting behaviours. Lastly (for now), the approach restricts the subgroup's exceptionality to the comparison with its complement. As already discussed, this implies searching for a dichotomy rather than a norm deviation. Although both tasks yield great applicability, they pose two different analyses with potentially different results. Essentially, the EMM task is concerned with deviations from a norm, and so are we addressing the discovery of survival behaviours that are unusual – unexpected with respect to an (expected) known behaviour.

In the next chapter, we address those issues by proposing a new EMM method for finding diverse subgroups with unusual survival behaviour.

5 ESMAMDS: A MORE DIVERSE SEARCH

In this chapter, we describe the Exceptional Survival Model Ant-Miner Diverse Search (EsmamDS). This EMM framework extends the work presented in Chapter 4 to provide a set of more diverse subgroups. This method tackles redundancy in three different dimensions: description, coverage, and survival model. Such problem is addressed on two fronts: (i) by enhancing the exploration power of the search through a new design of the ACO heuristic information function; and (ii) by minimising the redundancy in the final set of discovered subgroups with a new subgroup selection method. The EsmamDS also allows the user to choose between complement or population to compare subgroups with. Thus, we explore the description language structure to increase the generality and expressivity of the final set of patterns. The work presented in this chapter was previously introduced in (MATTOS; NETO; VIMIEIRO, 2021).

5.1 FRAMEWORK

The EsmamDS framework employs the Kaplan-Meier Estimates as the target model and the quality measure based on the logrank defined in Equation 2.2 to mine subgroups with exceptional survival model. It provides the choice of the baseline model to compare subgroups with as a user-defined parameter. The framework's pseudocode is provided in Algorithm 3.

Analogously to the Esmam framework presented in Chapter 4, the EsmamDS algorithm is initialised with an empty set of discovered subgroups $\mathbb{G}$, an empty subgroup $G_{\varnothing} = (D_{\varnothing}, \Omega)$, and an initial set of uncovered observations $\mathbb{U}$ containing all individuals in the data set. Then, following a covering-based approach (lines 4-12), it iteratively searches for subgroups (given a baseline $\mathcal{B}$) until all observations in Ω are covered by $\mathbb{G}$ at least once, or until the algorithm achieves a maximum stagnation threshold $maxStag$ (a number of consecutive iterations with no change in $\mathbb{U}$). In each iteration, a new colony of ants is responsible for discovering a single subgroup.

Thus, the EsmamDS framework consists of three major steps: (i) the initialisation of the probabilistic elements of the ACO search (line 5); (ii) the subgroup search, which returns the best subgroup G discovered by a complete colony of ants (line 6); and (iii) the update of the final subgroup set $\mathbb{G}$ considering the discovered subgroup G (line 7). We now discuss each of

Algorithm 3: EsmamDS Framework

```
Input: $\mathcal{B}$ – baseline for subgroup comparison,
$\alpha$ – level of significance, $maxStag$ – maximum stagnation of the algorithm,
$\wp^S = \{nAnts, nConverg, minCov\}$ – hyperparameters of the subgroup search,
$\wp^{DS} = \{L, W\}$ – hyperparameters of the diverse search
Output: $\mathbb{G}$ – set of exceptional subgroups
Data: $\Omega$ – survival data set
$\mathbb{G} \leftarrow \varnothing$; $G \leftarrow \varnothing$
$\mathbb{U} \leftarrow \Omega, \Delta U \leftarrow 0$
$stag \leftarrow 0$
while $\mathbb{U} \neq \varnothing$ and $stag \leq maxStag$ do
    searchInitialisation($G, \mathbb{G}, \mathbb{U}, \wp^{DS}$)
    $G \leftarrow$ subgroupSearch($\wp^S$)
    $\mathbb{G} \leftarrow$ subgroupSetUpdating($G, \mathbb{G}$)
    $\Delta U \leftarrow |\mathbb{U}| - |\bigcup_{G_k \in \mathbb{G}} cov(G_k)|$
    $\mathbb{U} \leftarrow \bigcup cov(G_k)$
    if $\Delta U = 0$ then
        $stag \leftarrow stag + 1$
    else: $stag \leftarrow 0$
return: $\mathbb{G}$
```

these steps in detail.

5.1.1 Search Initialisation

The `searchInitialisation` function is responsible for the initialisation of the pheromone values τ and heuristic values η associated with each item $I_{ij} \in \mathbb{I}$.

By the ACO design, the pheromone trails are initialised with the same amount of pheromone, i.e. with an equal probability of being chosen. We follow our definition for the Esmam initialisation in Chapter 4 and define the initial configuration of the pheromone values as $\tau_0(I_{ij}) = |\mathbb{I}|^{-1}$. Once the probabilistic choice of solution components is defined by both pheromone and heuristic values, and given that all pheromone values are equal at the beginning of each colony, we have that the heuristic information associated with the items define the initial probability distribution of the search space.

In the method presented in Chapter 4, we proposed an entropy-based heuristic information computed always over the same partition of the observations (comprising the entire data set), what results in constant heuristic values. In contrast to such static heuristic information, here, we initialise each colony with a different (initial) probabilistic distribution by proposing a dynamic function that depends on the state of the algorithm.

In other words, η is a function of the subgroup G discovered in a given algorithm iteration (it), the set $\mathbb{G}$ of subgroups currently selected in such iteration, and the current set $\mathbb{U}$ of individuals not covered by such $\mathbb{G}$. We use $\eta_{it}(I_{ij})$ to refer to the heuristic value associated with the item I_{ij} in a given algorithm iteration it (one algorithm iteration is the while loop in Algorithm 3, lines 4-12). Note that the heuristic function η and the pheromone trails τ follow different dynamics: the former is constant throughout an iteration it while the latter varies within each colony (Algorithm 3, line 6).

We define the (dynamic) heuristic information function $\eta_{it} : \mathbb{I} \to [0,1]$, $\eta_{it}(I_{ij}) = \eta_H(I_{ij}) \cdot \eta_L(I_{ij}) \cdot \eta_W(I_{ij})$ in terms of three components. Analogously to the heuristic presented in Chapter 4, we use information theory to provide a problem-dependent quantification of the relevance associated with the items in the search space (η_H). In addition, we propose to use information from both the descriptions (η_L) and coverages (η_W) of the discovered subgroups to improve search exploration.

The entropy-based component η_H provides a quantification of the discriminative power of the items regarding survivability, and it is defined in Equation 5.1 as

$$\eta_H(I_{ij}) = \frac{\log_2 k - H(W|I_{ij})}{\sum\limits_{I_{ij} \in \mathbb{I}} \log_2 k - H(W|I_{ij})} \tag{5.1}$$

where $H(W|I_{ij}) = -\sum_{w=1}^{k} P(w|I_{ij}) \cdot \log_2 P(w|I_{ij})$ is the Shannon's entropy. Note that this is the same definition of the Esmam heuristic information provided in Section 4.1 (Eq. 4.2). However, instead of computing such measure always considering the entire data set, here, in each new algorithm iteration (for each new colony), we consider only the individuals not covered by the current $\mathbb{G}$. Hence, each new colony is initialised with a different probabilistic distribution dependent on the current set of discovered subgroups. This way, we prioritise the survival experience of individuals that are not yet represented in our findings. Hence, we consider an initial partition of the individuals in $\mathbb{U}$ as those with survival time at least as long as the average survival time in $\mathbb{U}$, and those with shorter survival time. The more uniformly distributed an item is across those two survival groups (i.e. if it does not discriminate between the considered survival partition), the smaller is its heuristic quantification. By contrast, $\eta_H(I_{ij})$ assumes maximum value when the item is entirely associated with a single survival group.

The second component, the descriptive attenuation η_L, uses the descriptive itemset $\mathcal{I}(G)$ of all discovered subgroups G to guide the search towards unvisited (or more rarely visited) items in the search space. The rationale is that more discriminative items may bias the search, and, thus, we should penalise them somehow to promote diversity. Such a penalisation is based on the logistic function as presented in Equation 5.2,

$$\eta_L(I_{ij}) = 1 - \frac{1}{1 + e^{-(c(I_{ij})-L)}} \tag{5.2}$$

where $c(I_{ij})$ is the number of times I_{ij} was encompassed by the description of previously discovered subgroups. The parameter L adjusts the penalisation of an item regarding its usage, defining the value of $c(I_{ij})$ for which the heuristic value η_{ij} decreases by half. Therefore, we have that the more an item appears in the descriptions discovered by the subgroup search, the smaller becomes its (*a priori*) probability of being explored by future ant colonies.

Lastly, the weighted covering component η_W uses the subgroups in $\mathbb{G}$ to guide the search towards items describing observations less represented in the final set. For that, we make use of a score based on multiplicative weighted covering proposed by LEEUWEN; KNOBBE (2012), and presented in Equation 5.3,

$$\eta_W(I_{ij}) = \frac{1}{|cov(G_{I_{ij}})|} \sum_{o \in cov(G_{I_{ij}})} W^{g(o,\mathbb{G})} \tag{5.3}$$

where $G_{I_{ij}}$ is the subgroup for which $\mathcal{I}(G_{I_{ij}}) = \{I_{ij}\}$, $g(o, \mathbb{G}) = |\{G \in \mathbb{G} | o \in G\}|$ is the number of subgroups in $\mathbb{G}$ that contain an observation o, and $W \in (0, 1]$ is the weight parameter. Hence, the less often the observations described by I_{ij} are covered by subgroups in $\mathbb{G}$, the more likely it is for the item to be visited in future iterations of the algorithm.

It is important to notice that η_L takes into consideration the descriptions of all subgroups already discovered by the algorithm – whether or not they are included in $\mathbb{G}$. On the other hand, η_W only considers the coverage of the subgroups currently in $\mathbb{G}$.

Finally, each new colony (in each new algorithm iteration) is initially allocated in a different region of the search space (a different a priori probabilistic distribution) given the heuristic information that considers the search experience (the optimum solution) of the past ant colonies.

5.1.2 Subgroup Search

In the EsmamDS framework (Algorithm 3), once the pheromone and heuristic values are initialised (line 5), the subgroupSearch function implements the ACO search and returns a single subgroup (line 6). It is similar to the search implemented by the Esmam (Algorithm 2, lines 6-18). The pseudocode of the function is presented in the Algorithm 4.

Algorithm 4: EsmamDS: Subgroup Search

```
   Input: $nAnts$ – size of the ant colony,
   $nConverg$ – number of similar patters for convergence,
   $minCov$ – minimum subgroup coverage
   Output: $G_{best}$ – discovered subgroup
1  Function subgroupSearch($nAnts, nConverg, minCov$):
2      $ant \leftarrow 0$; $converg \leftarrow 0$
3      $G_{-} \leftarrow G_{\varnothing}$; $G_{best} \leftarrow G_{\varnothing}$
4      while $t \leq nAnts$ or $converg \leq nConverg$ do
5          $D \leftarrow$ buildDescription($\Omega, minCov$)
6          $D \leftarrow$ pruneDescription($D$)
7          pheromoneUpdating($\mathcal{I}(G_D)$)
8          if $\phi(G_D) > \phi(G_{best})$ then
9              $G_{best} \leftarrow G_D$
10         if $\mathcal{I}(G_D) = \mathcal{I}(G_{-})$ then
11             $converg \leftarrow converg + 1$
12         else: $converg \leftarrow 0$
13         $G_{-} \leftarrow G_D$
14         $t \leftarrow t + 1$
15     return: $G_{best}$
```

In the subgroup search, each ant in a colony of $nAnts$ ants delivers (builds and prunes) a complete description D (lines 5-6), which is used to update the pheromone trail for the next ant iteration (line 7). This process is repeated for all ants in the colony or until the ants converge to a solution (line 4), given a minimum threshold $nConverg$ for identical[1] sequential descriptions (lines 10-12). The best subgroup G_{best} (according to the quality measure ϕ) discovered within the colony is, then, returned. Having described the overall pipeline of the subgroup search function, we proceed to detail its main procedures: the (stochastic) description construction, the description pruning, and the pheromone updating.

The `buildDescription` (line 5) is a refinement function that starts from an empty description $D_{\varnothing}$ and iteratively assembles conjunctive conditions $cond(A_i) = A_i(o) \in \{v_j\}$ by sorting items I_{ij} belonging to the observations covered by the current partial D, given the probability distribution defined in Equation 5.4,

$$P(I_{ij}) = \frac{x_i \cdot \eta(I_{ij}) \cdot \tau(I_{ij})}{\sum\limits_{I_{ij}} x_i \cdot \eta(I_{ij}) \cdot \tau(I_{ij})}, \text{ for all } I_{ij} \in \mathcal{C}(G_D) \tag{5.4}$$

where $x_i = 1$ if A_i is not yet represented in the (partial) description D being constructed, and zero otherwise. This refinement process of iteratively sorting items stops when all A_i

[1] Note that two subgroups G_a and G_b are considered identical if their descriptions impose exactly the same constraints, i.e. if $\mathcal{I}(G_a) = \mathcal{I}(G_b)$.

are represented in D or when a new condition results in a coverage size $|cov(D)|$ below a minimum threshold $minCov$. Note that, similarly to the Esmam algorithm, the final constructed description D is a conjunction of conditions over singleton sets of $dom(A_i)$.

After a complete description D is constructed, the `pruneDescription` is a generalisation function (line 6) that greedily removes conditions $cond(A_i)$ from D, each time eliminating the condition that leads to the largest improvement in the quality of the resultant subgroup. The pruning stops when no conditions can be removed without decreasing the quality or when the description already encompasses only a single condition.

Finally, the `pheromoneUpdating` function (line 7) computes the pheromone values for the next ant iteration similarly to presented for the Esmam (Eq. 4.4). For the items represented in the final description D, the pheromone is incremented proportionally to the subgroup's quality and, for the items not represented in D, the evaporation process is simulated by the normalisation of τ_{ant+1} values for all $I_{ij} \in \mathbb{I}$. The rule for pheromone updating is given in Equation 5.5

$$\tau_{ant+1}(I_{ij}) = \tau_{ant}(I_{ij})\left(1 + \phi(G_D)\right), \text{ for all } I_{ij} \in \mathcal{I}(G_D) \tag{5.5}$$

5.1.3 Subgroup Set Updating

In each iteration of the EsmamDS (Algorithm 3), after a (new) subgroup G is discovered (line 6), the final set of subgroups $\mathbb{G}$ is updated considering the inclusion of such subgroup (line 7). Hence, the `subgroupSetUpdating` method is a recursive function that adjusts the subgroup set to (i) minimise both descriptive and model redundancies and (ii) maximise the coverage of subgroups, i.e. improve their generalisation.

By allowing descriptions to constrain attributes on a set of values, generalisation may be achieved by extending descriptions to subsume a set of subgroups. In other words, the structure of our description language allows generalisation operations to take place between subgroups. Hence, we first introduce two operations to provide a more general subgroup by combining two subgroups' descriptions: (i) the *root* operator, that provides a common generalisation between two descriptions; and (ii) the *merge* operator, that unifies two different descriptions into a more general one. Lets consider the three descriptions as follows:

$$\begin{array}{llll} D_1: & A_i \in \{v_{i1}\} & \wedge & A_j \in \{v_{j1}, v_{j2}\} \\ D_2: & A_i \in \{v_{i1}, v_{i2}\} & \wedge & A_j \in \{v_{j2}, v_{j3}\} \\ D_3: & A_i \in \{v_{i3}\} & \wedge & A_k \in \{v_{k1}\} \end{array}$$

Considering the pair (D_1, D_2), the *root* operation yield a new description $D_r : A_i \in \{v_{i1}\} \wedge A_j \in \{v_{j2}\}$. When employing the *merge* operation, we have the generalisation $D_m : A_i \in \{v_{i1}, v_{i2}\} \wedge A_j \in \{v_{j1}, v_{j2}, v_{j3}\}$. Note, however, that the description D_3 does not have a *root* with neither D_1 nor D_2 because they do not present a common attribute constrain. Additionally, it cannot be merged with either D_1 or D_2 because they constrain different attributes.

Formally, we have that, given two subgroups G_a and G_b:

- $root(G_a, G_b) = \mathcal{I}(G_a) \cap \mathcal{I}(G_b)$ provided that the intersection exists; and
- $merge(G_a, G_b) = \mathcal{I}(G_a) \cup \mathcal{I}(G_b)$ if and only if the attributes A_i represented in both G_a and G_b descriptions are exactly the same.

Note that the root and the merge are *generalisation* operators that perform over two aspects of a description: (i) the conjunctive conditions $cond(A_i) = A_i \in \mathcal{V}_i$; and (ii) the extent of $\mathcal{V}_i$ constraints. Their difference is, hence, in the aspects they manipulate. The *root* operator generalises a description manipulating both those aspects by eliminating conditions and altering the sets of restricted domains. In contrast, the *merge* operator provides generalisation by manipulating only the last aspect by enlarging the extent of the (already) restricted domains.

Additionally, we define that G_b *is-in* G_a if $\mathcal{I}(G_b) \subseteq \mathcal{I}(G_a)$ given that the attributes A_i represented in both their descriptions are exactly the same. For example, we have that a subgroup described as $D_{in} = A_i \in \{v_{i1}\} \wedge A_j \in \{v_{j3}\}$ *is-in* the subgroup represented by D_2. In case G_b *is-in* G_a, we also have that G_b is a specialisation of G_a. Note that while the *refinements* are specialisations concerning the number of constraints $cond(A_i) = A_i \in \mathcal{V}_i$ (assuming no change in the domains already constrained in a more general description), the *is-in* relation allows the identification of specialisations concerning the extent of the constrained A_i domains – i.e. the extent of $\mathcal{V}_i$. Hence, we assume that the two descriptions being compared must constrain the same attributes but to different extents.

The pseudocode of the `subgroupSetUpdating` function is provided in Algorithm 5. It gives the rules that determine which subgroups will constitute the current final set (remember that

Algorithm 5: EsmamDS: Subgroup Set Updating

```
   Input: $G_{new}$ – subgroup, $\mathbb{G}$ – set of subgroups
   Output: $\mathbb{G}$ – set of exceptional subgroups
 1 Function subgroupSetUpdating($G_{new}, \mathbb{G}, \alpha$):
 2     if $\phi(G_{new}, \mathcal{B}) < 1 - \alpha$: then
 3         return: $\mathbb{G}$
 4     for $G \in \mathbb{G}$ do
 5         if $G_{new}, G$ have different models or strictly different attributes then
 6             next
 7         else
 8             if  $G_{new}$ is-in $G$: then
 9                 return: $\mathbb{G}$
10             if $G$ is-in $G_{new}$ then
11                 $\mathbb{G}' \leftarrow$ subgroupSetUpdating($G_{new}, \mathbb{G} \setminus G, \alpha$)
12                 if $G_{new} \in \mathbb{G}'$: return: $\mathbb{G}'$
13                 else: return: $\mathbb{G}$
14             if not $root(G_{new}, G)$ nor $merge(G_{new}, G)$ then
15                 next
16             else
17                 $G_r \leftarrow root(G_{new}, G)$
18                 $G_m \leftarrow merge(G_{new}, G)$
19                 $\mathbb{G}' \leftarrow$ subgroupSetUpdating($G_r, \mathbb{G}, \alpha$)
20                 $\mathbb{G}' \leftarrow$ subgroupSetUpdating($G_m, \mathbb{G}', \alpha$)
21                 if neither $G_r, G_m \in \mathbb{G}'$: next
22                 if only $G_r \in \mathbb{G}'$ then
23                     if $(G_r, G_{new})$ models are different: next
24                     else: return: $\mathbb{G}'$
25                 if only $G_m \in \mathbb{G}'$ then
26                     if $(G_m, G_{new})$ models are different: next
27                     else: return: $\mathbb{G}'$
28                 if both $G_r, G_m \in \mathbb{G}'$ then
29                     if $G_{new}$ model differs from both $G_r, G_m$: next
30                     else: return: $\mathbb{G}'$
31     return: $\mathbb{G} \leftarrow \mathbb{G} \cup \{G_{new}\}$
```

this function will update the final set at each iteration of the algorithm, after the discovery of a new subgroup (see Algorithm 3, line 7).

After a new candidate subgroup G_{new} is returned by the subgroupSearch procedure, it is added to the set $\mathbb{G}$ if it satisfies a lower quality bound for assuring exceptionality (line 2) **and** if (for all subgroups $G \in \mathbb{G}$) (lines 4-6):

- G_{new} and G have statistically different survival models, i.e. if $sim_M(G_{new}, G) = 0$ (see

Eq. 2.5); **or**

- G_{new} represents only attributes A_i not represented in G, i.e. their conjunctive conditions $cond(A_i)$ restrict completely different attributes.

In other words, we immediately select the exceptional subgroups that present a previously unobserved behaviour or a completely new characterisation (description).

The contrary case to directly incorporate G_{new} to $\mathbb{G}$ (line 7) happens when a pair (G_{new}, G), for any $G \in \mathbb{G}$, present similar models and some resemblance in description. In other words, such pair of subgroups potentially represent a common subset of the data (because their descriptions are similar in some level) and, thus, they manifest similar behaviours. Note that this configures redundancy, which we want to avoid. Hence, we further process those cases presenting similarities in description and model to minimise redundancy and improve the generalisation of the final patterns.

Finally, the `subgroupSetUpdating` function provided in Algorithm 5 implements the following pipeline. For each new subgroup G_{new} under consideration, if it satisfies a lower quality bound (lines 2-3), it will be compared to each subgroup $G \in \mathbb{G}$ (line 4) to verify whether it satisfies the two conditions given above (lines 5-6). For the (contrary) case, when both description and model similarities exist between the pair (G_{new}, G) (line 7), the following procedure is implemented:

1. We first assess whether G_{new} or G are a specialisation of one another. For that, we use the *is-in* relation to keep in $\mathbb{G}$ the more general subgroup among those two (lines 8-13). Note that the function is recursively applied for replacing a subgroup in the final set for a new one to guarantee its diversity inside the set.
2. Next, if neither one of the subgroups is a specialisation, we assess whether it is possible to generalise both G_{new} and G into a more general pattern using the *root* and *merge* generalisation operators (lines 14-15).
3. When a generalisation is possible (lines 16-30), we recursively update $\mathbb{G}$ considering the G_r root generalisation and/or the G_m merge generalisation. Here, we keep testing G_{new} against the next $G \in \mathbb{G}$ in two cases: (i) if neither generalisations can be added to $\mathbb{G}$; or (ii) if the model of G is different from the model of the added generalisation(s). In the first case, although (G_{new}, G) present description and model similarities, they present some differences that cannot be unified in a single pattern; hence, we choose to keep such differences. In the second case, although generalisation considering both (G_{new}, G) is provided, G_{new} now consists of a specification with a distinct behaviour; hence, we

choose to keep the diversity of the model responses.

Having presented the EsmamDS algorithm, we describe the experiments we conducted to evaluate our approach and the results we achieved.

5.1.4 Computational complexity

Having presented the EsmamDS algorithm provided in Algorithm 3, we analyse its computational complexity. A single iteration of the algorithm (lines 4-12) can be divided into three major processes that we will analyse individually: (1) the computational complexity of the search initialisation; (2) the complexity of the subgroup search, i.e. the complexity of the ant-colony loop; and (3) the complexity of the method for updating the final set $\mathbb{G}$ of subgroups. Then, we combine the results of these three processes to determine the computational complexity of the algorithm execution. It is important to remember that the EsmamDS builds on the Esmam algorithm introduced on Chapter 4. Hence, the complexity of similar methods already derived for the Esmam analysis will only be referred to its complete explanation.

1. *Search initialisation* (line 5): Each algorithm iteration starts by computing the τ and η values. The step of defining the values of all $\tau_0(I_{ij})$ takes $\mathcal{O}(|\mathbb{I}|)$. Differently from the Esmam approach, here, $\eta(I_{ij})$ values are dynamic and need to be recomputed on each new algorithm's iteration. The EsmamDS heuristic information is a function of three components that we analyse in the following:

   - The component η_H (Equation 5.1) can be computed in a single scan of the data set. So, the time complexity of this step is $\mathcal{O}(|\mathbb{I}| \cdot |\Omega|)$.
   - The component η_L (Equation 5.2) requires the scanning of k conditions of the latest discovered subgroup (description). Hence, the time complexity of this step is $\mathcal{O}(|\mathbb{I}| \cdot k)$.
   - The component η_W (Equation 5.3) requires scanning the data set and all subgroups comprising the current set $\mathbb{G}$. Thus, this step takes $\mathcal{O}(|\mathbb{I}| \cdot |\Omega| \cdot |\mathbb{G}|)$.

   Adding up the results derived for the three components above and the complexity for initialising the pheromone values, we have that the computational complexity for the search initialisation is $\mathcal{O}(|\mathbb{I}| \cdot [k + |\Omega| \cdot |\mathbb{G}| + |\Omega|])$.

2. *Subgroup search* (Algorithm 3, line 6; Algorithm 4): The subgroup search of EsmamDS is equivalent to the ant-colony loop introduced in Section 4.1. Each ant in a colony of (at most) $nAnts$ performs the major processes: (i) description construction; (ii) evaluation of candidate descriptions; (iii) description pruning; and (iv) pheromone updating. Combining the complexity of those four steps, we have that the execution of the whole colony loop takes $\mathcal{O}(nAnts[k \cdot |\mathbb{I}| + k^3 \cdot |\Omega|])$.

3. *Subgroup set updating* (Algorithm 3, line 7; Algorithm 5): The method to include a newly discovered subgroup into the final set requires the comparison of k conditions with a number of $|\mathbb{G}|$ subgroups. The method, however, is recursive. In the worst case, a new subgroup G_{new} is compared to all subgroups in $\mathbb{G}$, yielding two different generalisations – *root* and *merge*. Then, the method updates $\mathbb{G}$ testing the inclusion of both generalisations provided, which may happen at most $2 \cdot |\mathbb{G}|$ times. Hence, the method for updating the final set of subgroups takes $\mathcal{O}(2 \cdot k \cdot |\mathbb{G}|^2)$.

Finally, to derive the computational complexity of complete algorithm execution, we have to add the complexities of those three topics above and multiply it by the total number z of discovered subgroups. Therefore, the computational complexity of the EsmamDS algorithm is

$$\mathcal{O}\left(z \cdot \left[|\mathbb{I}| \cdot (k + |\Omega| \cdot |\mathbb{G}| + |\Omega|) + nAnts(k \cdot |\mathbb{I}| + k^3 \cdot |\Omega|) + 2 \cdot k \cdot |\mathbb{G}|^2\right]\right)$$

To simplify, we may consider that, in the worst-case scenario, the value of k conditions per description is equal to $|A|$. Thus, we can replace $|\mathbb{I}|$ by $|A|$ if we consider that the categorical attributes take only a small number of values. Hence, the formula for worst-case computational complexity is

$$\mathcal{O}\left(z \cdot |A| \left[nAnts \cdot |\Omega| \cdot |A|^2 + |\mathbb{G}|^2\right]\right)$$

When comparing to the Esmam approach presented in Chapter4, the diverse search of EsmamDS adds complexity in the order of $|A| \cdot |\mathbb{G}|^2$. Hence, the size of the set of subgroups has a greater impact on the time complexity of this version. Thus, the EsmamDS performance depends significantly on the data set volume, dimensionality, and feature complexity.

5.2 EXPERIMENTS

We conducted experiments to evaluate our proposed approach for mining local patterns associated with exceptional survival behaviours. In addition to provide comprehensible characterisation of subgroups presenting unusual KM models, we aim at providing a less redundant and more expressive set of patterns than we achieved with the method proposed in Chapter 4.

We compare the EsmamDS algorithm with state of the art methods in the literature that provide characterisation over unusual survival behaviour: (i) the Esmam algorithm, the ACO heuristic approach to mine unusual survival models presented in Chapter 4 (which served as the base for this approach); (ii) the beam-search heuristic for mining subgroups, considering either a single target (SD) and a model target (EMM); (iii) the *DSSD-CBSS* algorithm (LEEUWEN; KNOBBE, 2012), a beam-search approach to mine a diverse set of subgroups adapted to use a similar target as ours; and (iv) the *LR-Rules* (WRÓBEL; GUDYŚ; SIKORA, 2017), a rule-based covering algorithm for predicting survival response (although the ultimate goal of this method is to build a predictive model, we decided to include it in our study as its authors also suggested its application for finding descriptions).

Table 7 presents all metrics used in the results evaluation. We, once again, assess the results with respect to their descriptive aspects using the metrics of *interpretability* and *representativeness* introduced in Section 4.2. Additionally, we assess the findings concerning two new aspects: exceptionality and redundancy.

The metric $\mathcal{E}$ of exceptionality evaluates the unusualness of the survival models discovered in a set of subgroups. For that, we assess the similarity between each subgroup to the baseline model and provide the proportion of exceptional models in the discovered set. This metric ranges from zero (no exceptional models in a set) to one (all discovered models are exceptional). Redundancy is assessed – for descriptions (ρ_D), coverage (ρ_C), and survival models (ρ_M) – as the normalised sum of the similarity measures (sim_D, sim_C and sim_M, respectively) for all (unordered) pairs of subgroups in the set. Note that the similarity metrics are comparisons between pairs of subgroups (the baseline $\mathcal{B}$ can be considered a subgroup to compare with). By contrast, the metrics of exceptionality, interpretability, representativeness and redundancy are global metrics for a *set of subgroups*.

Additionally, we also evaluate coverage redundancy using the CR measure (LEEUWEN; KNOBBE, 2012). Such measure quantifies the extent of the deviation between the coverage of the subgroups in a set $\mathbb{G}$ from a uniform (cover) distribution. Being $g(o, \mathbb{G})$ the number of subgroups in $\mathbb{G}$ that cover an observation o, we have that the expected number of times for a random observation to be covered is $\hat{g} = |\mathbb{G}|^{-1} \sum_{o \in \Omega} g(o, \mathbb{G})$. Then, the CR is defined as

Table 7 – Method EsmamDS – Empirical evaluation metrics

Metrics	Description	Definition
Interpretability		
$\#sg$	Number of discovered subgroups	$\vert\mathbb{G}\vert$
$length_{AV}$	Average subgroup description length	$\sum_{G\in\mathbb{G}} \frac{length(G)}{\vert\mathbb{G}\vert}$
Representativeness		
sgCov	Average (percentage) subgroup coverage	$\vert\Omega\vert^{-1} \cdot \sum_{G\in\mathbb{G}} \frac{\vert cov(G)\vert}{\vert\mathbb{G}\vert}$
dbCov	Data set coverage	$\vert\Omega\vert^{-1} \cdot \vert \bigcup_{G\in\mathbb{G}} cov(G)\vert$
Exceptionality		
$\mathcal{E}$	Proportion of exceptional subgroups	$\sum_{G\in\mathbb{G}} \frac{sim_M(G,\mathcal{B})}{\vert\mathbb{G}\vert}$
Similarity		
sim_D	Description similarity	$sim_D(G_a,G_b) = \frac{\vert\mathcal{I}(G_a)\cap\mathcal{I}(G_b)\vert}{min(\vert\mathcal{I}(G_a)\vert,\vert\mathcal{I}(G_b)\vert)}$ (Eq. 2.3)
sim_C	Coverage similarity	$sim_C(G_a,G_b) = \frac{\vert cov(G_a)\cap cov(G_b)\vert}{min(\vert cov(G_a)\vert,\vert cov(G_b)\vert)}$ (Eq. 2.4)
sim_M	Model similarity	$sim_M(G_a,G_b) = pval_{G_a,G_b} > \alpha$ (Eq. 2.5)
Redundancy		
ρ_D	Description redundancy	$\rho_D = \binom{\mathbb{G}}{2}^{-1} \sum_{G_a,G_b\in\mathbb{G},G_a\neq G_b} sim_D(G_a,G_b)$
ρ_C	Coverage redundancy	$\rho_C = \binom{\mathbb{G}}{2}^{-1} \sum_{G_a,G_b\in\mathbb{G},G_a\neq G_b} sim_C(G_a,G_b)$
ρ_M	Model redundancy	$\rho_M = \binom{\mathbb{G}}{2}^{-1} \sum_{G_a,G_b\in\mathbb{G},G_a\neq G_b} sim_M(G_a,G_b)$
CR	Cover Redundancy	Equation 5.6

Font: The author (2021)

presented in Equation 5.6.

$$CR = \frac{1}{|\Omega|} \sum_{o\in\Omega} \frac{|g(o,\mathbb{G}) - \hat{g}|}{\hat{g}} \tag{5.6}$$

High values of this measure indicate that the observations contained in the subgroups of $\mathbb{G}$ are covered more than expected. In other words, a large number of subgroups in the set cover the same observations. Hence, low values of CR indicate more diversity/less redundancy between subgroups.

Next, we describe the process of empirical evaluation and analyse the results. Some contents like EsmamDS implementation, the data sets used in the tests, configurations and results are available on EsmamDS repository.

5.2.1 Experimental Setup

We conduct experiments with 14 real-world survival data sets from the medical domain. These data sets were previously presented in Subsection 4.2.1 and are described in Table 4. The preprocessing of the data was employed to remove observations containing missing values (and features with a high level of missing data). All numerical descriptive attributes were discretised using equal-frequency discretisation into five interval categories.

We conducted experiments to assess the performance of the EsmamDS considering both baselines $\mathcal{B}$ for subgroup comparison – *population* and *complement* – and evaluated them separately. In the empirical evaluation, we performed 30 executions for each data set due to its stochastic nature. Throughout the analysis of the results, we identify an arbitrary experiment execution n with the identification '*exp. n*' (that stands for experiment number n). The proper configuration of the algorithm was defined with a randomised search considering the following parameters' values:

- $nAnts = \{100, 200, 500, 1000, 3000\}$;
- $minCov = \{0.01, 0.02, 0.05, 0.1\}$;
- $nConverg = \{5, 10, 30\}$;
- $maxStag = \{20, 30, 40, 50\}$;
- $L = \{1, 3, 5, 10\}$;
- $W = 0.9$ (according to employed by LEEUWEN; KNOBBE (2012) in their cover-based subgroup selection method);
- $\alpha = 0.05$

We sampled 10% of the total number of combinations and then executed the EsmamDS for three data sets (namely: actg320, breast-cancer and ptc). The best configuration (for each baseline) was chosen by ordering all configuration samples from the random search according to their average performance for the following metrics' order: ρ_D, ρ_C, CR, ρ_M, $sgCov$, $dbCov$, $length_{AV}$ and $\#sg$.

For the configuration of the other approaches, for each data set and according to a baseline, we used the results achieved by the EsmamDS in the empirical evaluation to adjust the following three parameters:

- $(minCov)$ Minimum coverage: defined by the same parameter value chosen for the EsmamDS;
- (bs) Beam-size (or maximum number of discovered subgroups): given by the average number of subgroups discovered by the EsmamDS in the 30 experiments;
- $(maxDepth)$ Rule-depth (or refinement/search depth): given by the average of the the maximum description length achieved during the EsmamDS execution in the 30 experiments.

Table 8 displays the configuration for each baseline of all the algorithms compared. The beam-search approaches were executed using the PySubgroup package (LEMMERICH; BECKER, 2018) given two types of targets: the survival time T single numeric target and the KM model target. For the beam-search approaches that consider single target, we employed a quality measure given as $1 - pval_{\hat{T}_{G,\mathcal{B}}}$, being $pval_{\hat{T}_{G,\mathcal{B}}}$ the *p-value* of the bilateral t-Test for the survival time for comparing a subgroup G to the baseline $\mathcal{B}$. The DSSD-CBSS algorithm was executed using Cortana with the quality measure defined as the t-Test. The remaining approaches employ

Table 8 – Method EsmamDS – Information on the algorithms compared in the empirical evaluation. The $bs, maxDepth$ parameters were configured for each data set as defined in the text. The remaining user-defined configuration of the frameworks were kept as default. The specifics of all configurations are provided here.

	Algorithm	Search Strategy	Target Concept	Source	Parameter(Value)
Population	EsmamDS-pop	ACO	KM model	EsmamDS repository	$\alpha(0.05)$, $nAnts(100)$, $minCov(0.1)$, $nConverg(5)$, $maxStag(40)$, $W(0.9)$, $L(5)$
	Esmam-pop	ACO	KM model	Esmam repository	$\alpha(0.05)$, $nAnts(100)$, $minCov(0.1)$, $nConverg(5)$, $maxStag(40)$
	BS-EMM-pop	Beam Search	KM model	PySubgroup package	$minCov(0.1)$, bs, $maxDepth$
	BS-SD-pop	Beam Search	Survival time	PySubgroup package	$minCov(0.1)$, bs, $maxDepth$
	DSSD-CBSS	Beam Search	Survival time	Cortana package	*search strategy(Cover-based beam selection)*, $minCov(0.1)$, bs, $maxDepth$, $time(\infty)$
Complement	EsmamDS-cpm	ACO	KM model	EsmamDS repository	$\alpha(0.05)$, $nAnts$ (100), $minCov(0.05)$, $nConverg(5)$, $maxStag(40)$, $W(0.9)$, $L(10)$
	Esmam-cpm	ACO	KM model	Esmam repository	$\alpha(0.05)$, $nAnts(100)$, $minCov(0.05)$, $nConverg(5)$, $maxStag(40)$
	BS-EMM-cpm	Beam Search	KM model	PySubgroup package	$minCov(0.05)$, bs, $maxDepth$
	BS-SD-cpm	Beam Search	Survival time	PySubgroup package	$minCov(0.05)$, bs, $maxDepth$
	LR-Rules	Sequential covering	KM model	LR-Rules repository	–

Font: The author (2021)

the measure defined in Equation 2.2.

Finally, statistical analysis of the results was performed by a (paired) Friedman test followed by a Nemenyi posthoc test. We performed the Friedman test to compare whether or not the compared approaches present statistically similar performances (null hypothesis) regarding the proposed metrics. When Friedman's null hypothesis is rejected, we proceed with the Nemenyi test to validate which approaches stand out in their performances. We executed the tests using EsmamDS (and Esmam) complete sample of 420 results for each metric (30 experiments on 14 data sets). For the remaining deterministic algorithms, we paired the results for each data set by repeating them 30 times. We consider both exceptionality and representativeness maximisation metrics; for the others, we consider minimisation. Thus, we assessed the tests using a level of significance of 5%.

5.2.2 Results Analysis

In this section, we present and analyse the results achieved by the algorithms. Table 9 contains the average performance for all evaluation metrics (except for the metrics of similarity between pairs of subgroups, which will be approached later in this section). We performed the Friedman test for interpretability, representativeness, and redundancy metrics. We rejected the null hypothesis that the algorithms present similar performances for all tested metrics. Hence, we employed the Nemenyi posthoc test for each metric to assess the differences between the performances. The average rank used in the Nemenyi test for each metric is also provided in the table. Tables presenting the performance of the compared approaches over each data set

Table 9 – Method EsmamDS – Evaluation metrics computed over the results provided by the compared approaches: the metrics' average over all data sets (Avg.); and the mean rank of the Nemenyi post-hoc test computed for each metric (Rank). Bold values represent the best results.

	Algorithms	$\mathcal{E}$	$\#sg$		$length_{AV}$		$sgCov$		$dbCov$		ρ_D		ρ_C		CR		ρ_M	
		Avg.	Avg.	Rank	Avg.	Rank	Avg.	Rank	Avg.	Rank	Avg.	Rank	Avg.	Rank	Avg.	Rank	Avg.	Rank
Population	EsmamDS-pop	**1.00**	**5.05**	**2.28**	1.52	1.93	**0.28**	**1.37**	**0.87**	**1.14**	**0.03**	**1.23**	**0.23**	**1.25**	**0.31**	**1.09**	**0.26**	**1.22**
	Esmam-pop	**1.00**	7.31	3.51	**1.46**	**1.74**	0.21	2.33	0.64	2.25	0.17	2.11	0.39	2.39	0.51	2.39	0.56	2.40
	BS-EMM-pop	**1.00**	5.50	3.07	2.22	3.57	0.16	4.12	0.33	4.14	0.54	3.86	0.67	4.07	0.70	4.27	0.84	3.91
	BS-SD-pop	0.81	5.50	3.07	2.46	3.33	0.17	3.39	0.39	3.40	0.46	3.34	0.54	3.33	0.62	3.21	0.69	3.10
	DSSD-CBSS	0.71	5.50	3.07	2.63	4.43	0.17	3.79	0.27	4.07	0.69	4.46	0.73	3.95	0.74	4.05	1.00	4.36
Complement	EsmamDS-cpm	**1.00**	**5.36**	**2.31**	**1.22**	**1.78**	**0.34**	**1.57**	0.98	2.13	**0.02**	**1.54**	0.33	2.33	**0.29**	**1.78**	**0.24**	**1.84**
	Esmam-cpm	**1.00**	5.70	2.58	1.27	1.84	0.25	2.92	0.71	3.18	0.10	2.23	0.40	2.95	0.46	3.28	0.45	3.20
	BS-EMM-cpm	**1.00**	6.00	3.10	2.16	3.98	0.19	4.08	0.50	4.00	0.51	4.37	0.67	4.30	0.63	4.05	0.66	4.19
	BS-SD-cpm	0.82	6.00	3.10	2.56	4.29	0.18	3.93	0.50	4.10	0.41	4.07	0.49	3.50	0.59	3.90	0.52	3.35
	LR-Rules	0.94	9.43	3.91	1.86	3.11	0.28	2.50	**1.00**	**1.59**	0.15	2.66	**0.32**	**1.84**	0.32	1.99	0.30	2.34

EsmamDS and Esmam results were averaged over 30 experiments for each data set.

are provided in the Appendix A. We now discuss each of these results in detail.

The $\mathcal{E}$ metric of exceptionality reveals the unusualness of the survival behaviour associated with the discovered subgroups. We inspected each subgroup to verify if its survival model is exceptional (statistically different) compared to the survival model of the considered baseline. We observe that the methods that employ the subgroup discovery approach rather than exceptional model mining, namely BS-SD-pop, DSSD-CBSS and BS-SD-cpm, do not assure the discovery of exceptional models. In addition, the predictive sequential covering LR-Rules also delivers some subgroups that are not unusual.

To analyse such a result, confront the two tasks of Subgroup Discovery (SD) and Exceptional Model Mining (EMM) by comparing two sets of discovered subgroups: one provided by the BS-SD-pop algorithm performing the SD task, and a set of subgroups delivered by our approach, the EsmamDS-pop, performing the EMM task. It is important to remember that, in the SD task, a subgroup is deemed exceptional if its mean survival time is statistically different ($p\text{-}value \leq 0.05$) from the baseline's average survival (time). Such unusualness is assessed with the (bilateral) t-Test comparing both means – subgroup and baseline. For the EMM task, exceptionality is assessed with the logrank test, and a subgroup is considered exceptional if its survival model is statistically different from the baseline model.

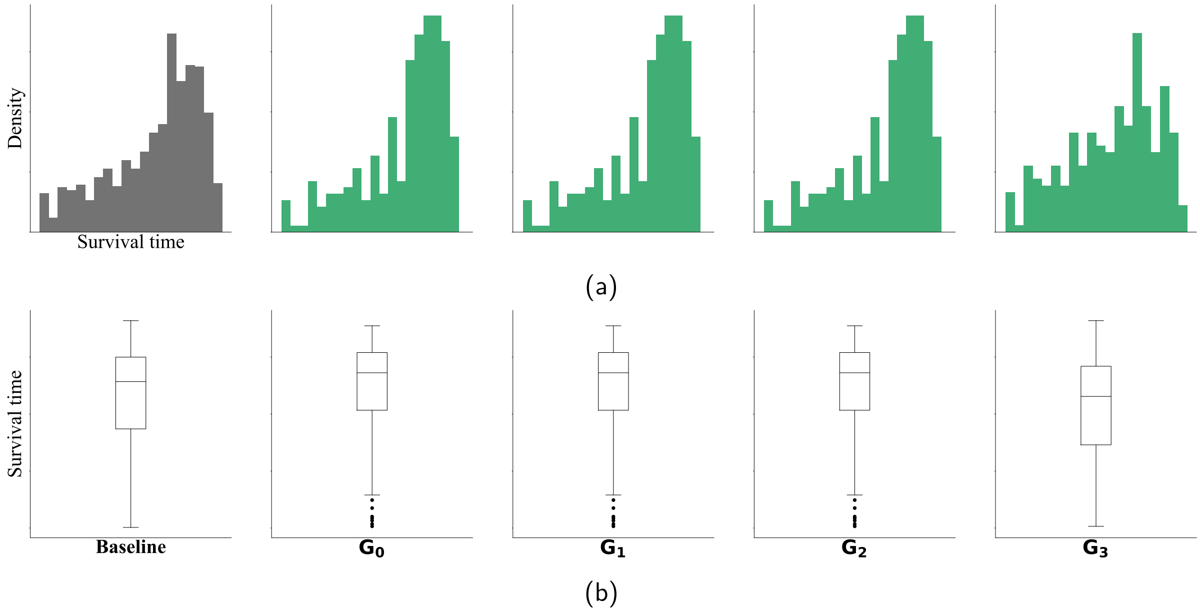
Hence, we first present the subgroups discovered by the BS-SD-pop in Table 10. There, we provide the p-*value* of both statistical tests mentioned above. In that way, for each subgroup provided, we can identify whether it is exceptional according to each task – EMM and SD. Analysing the table, we can observe that all subgroups deemed exceptional in the SD task are not exceptional when considering their survival models. In other words, although SD subgroups present an unusual distribution of the survival time, their models may not be unusual. In Figure 5, we show the distribution of the survival time over the baseline and the subgroups

Table 10 – Method EsmamDS – Set of subgroups discovered by the BS-SD-pop algorithm in the *actg320* data set (*exp. 0*): the subgroups' description (Subgroups G_i), the p-value of the logrank test between the subgroup and baseline ($logrank(G_i, \mathcal{B})$), and the p-value of the t-Test for comparing the the survival times of the subgroup and baseline ($tTest(G_i, \mathcal{B})$)

Subgroups $\mathbf{G_i}$	$\mathbf{logrank(G_i, \mathcal{B})}$ (*p-value*)	$\mathbf{tTest(G_i, \mathcal{B})}$ (*p-value*)
$G_0 : sex \in \{1\} \wedge raceth \in \{1\} \wedge hemophil \in \{0\} \wedge txgrp \in \{2\} \wedge tx \in \{1\}$	0.100	0.005
$G_1 : sex \in \{1\} \wedge raceth \in \{1\} \wedge hemophil \in \{0\} \wedge txgrp \in \{2\}$	0.100	0.005
$G_2 : sex \in \{1\} \wedge raceth \in \{1\} \wedge hemophil \in \{0\} \wedge tx \in \{1\}$	0.100	0.005
$G_3 : raceth \in \{2\} \wedge hemophil \in \{0\} \wedge ivdrug \in \{1\}$	0.859	0.006

Font: The author (2021)

Figure 5 – Method EsmamDS – Analysis of the survival time distribution over the subgroups discovered by the BS-SD-pop algorithm (provided in Table 10) on the *actg320* data set (*exp. 0*). The left column of plots provides the distribution over the baseline $\mathcal{B}$-population. The distributions are displayed in (a) histograms, and (b) boxplots. The distributions' identification is provided below the plots. Axis labels are on the left. The green color indicates a distribution with mean survival time statistically different from the baseline average survival.



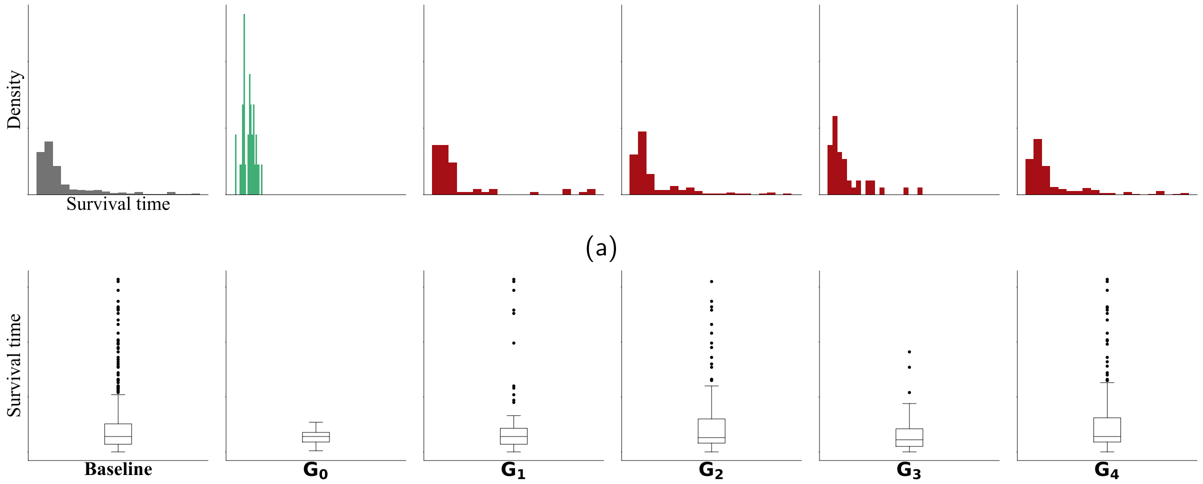
displayed in the table. The green colour indicates the distributions for which the t-Test null hypothesis was rejected, i.e. the subgroups deemed exceptional by the SD task. Analysing the figures, we observe that even slight variations in the distribution of the survival time (target) feature may comprise a statistically unusual distribution. In this sense, the occurrence of outliers in a subgroup may significantly bias its average survival time while not necessarily yielding a different survival model, as Table 10 evinces.

Table 11 – Method EsmamDS – Set of subgroups discovered by the EsmamDS-pop algorithm in the *ptc* data set (*exp. 0*): the subgroups' description (Subgroups G_i), the p-value of the logrank test between the subgroup and baseline ($logrank(G_i, \mathcal{B})$), and the p-value of the t-Test for comparing the the survival times of the subgroup and baseline ($tTest(G_i, \mathcal{B})$)

Subgroups $\mathbf{G_i}$	$\mathbf{logrank(G_i, \mathcal{B})}$ (*p-value*)	$\mathbf{tTest(G_i, \mathcal{B})}$ (*p-value*)
$G_0 : risk\_group \in \{Intermediate\} \wedge path\_n\_stage \in \{N1b\}$ $\wedge\, wgs\_status \in \{No\}$	0.002	0.041
$G_1 : mrna\_cluster \in \{5\} \wedge mirna\_cluster \in \{6\}$	0.011	0.430
$G_2 : path\_n\_stage \in \{N0\} \wedge tumor\_status \in \{tumor\_free\}$	0.012	0.730
$G_3 : risk\_group \in \{Intermediate\} \wedge sex \in \{Male\}$ $\wedge\, histological\_type \in \{Classical\} \wedge wgs\_status \in \{No\}$	0.004	0.222
$G_4 : sex \in \{Female\} \wedge lowpass \in \{No\} \wedge wgs\_status \in \{No\}$ $\wedge\, tumor\_status \in \{tumor\_free\}$	0.048	0.250

Font: The author (2021)

Figure 6 – Method EsmamDS – Analysis of the survival time distribution over the subgroups discovered by the EsmamDS-pop algorithm (provided in Table 11) on the *ptc* data set (*exp. 0*). The left column of plots provides the distribution over the baseline $\mathcal{B}$-population. The distributions are displayed in (a) histograms, and (b) boxplots. The distributions' identification is provided below the plots. Axis labels are on the left. The green color indicates a distribution with mean survival time statistically different from the baseline average survival, while the red color indicates a distribution with average survival similar to the baseline.



(a)

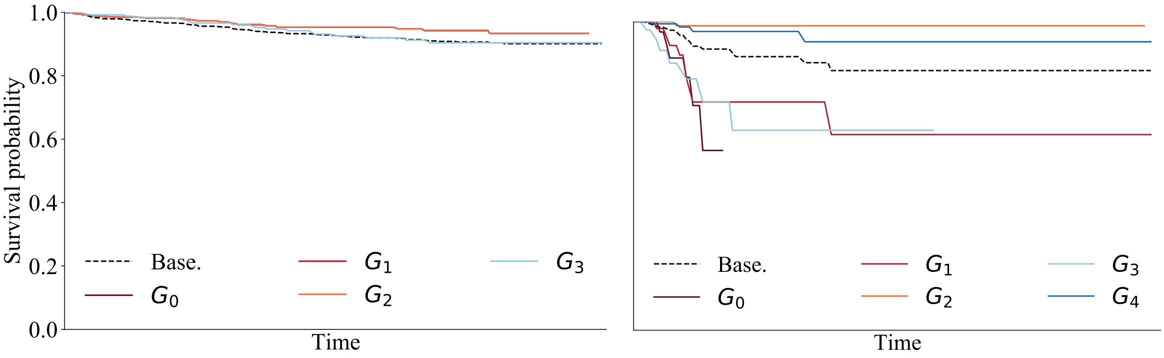
(b)

Analogously, we provide a set of subgroups discovered by the EsmamDS-pop algorithm in Table 11, and their survival time distribution in Figure 6. The red colour indicates the distributions for which the mean survival time can be considered statistically equal to the baseline's average survival. We observe that the subgroup G_0 is the only one with unusual distribution compared to the baseline and, therefore, the only subgroup deemed exceptional by the SD task. However, when assessing the subgroups' survival models (with the logrank test results shown in the table), we have that all subgroups are indeed exceptional. Additionally to the fact that a deviating mean survival does not necessarily imply an unusual survival behaviour, we have that the contrary is also true: exceptional survival responses do not necessarily imply an unusual survival time distribution.

Hence, to assess our ultimate goal of discovering subgroups with unusual survival behaviour, we present in Figure 7 the survival models of the subgroups comprising the two sets analysed above. In Figure 7a, we observe the subgroups discovered by the SD approach. We see that such subgroups which do not present exceptional models present survival responses very similar to the baseline response (*Base.*). On the other hand, the subgroups discovered by our EMM approach (Figure 7b) present a variety of behaviours that are distinct from the baseline – some subgroups presenting better survivability, some worse.

Ultimately, survival models, or *time-to-event* models, provide the distribution of the *event*

Figure 7 – Method EsmamDS – Survival curves of the set subgroups discovered by: (a) the BS-SD-pop algorithm on the *actg320* data set (provided in Table 10), and (b) the EsmamDS-pop algorithm on the *ptc* data set (provided in Table 11).



(a) SD approach

(b) EMM approach

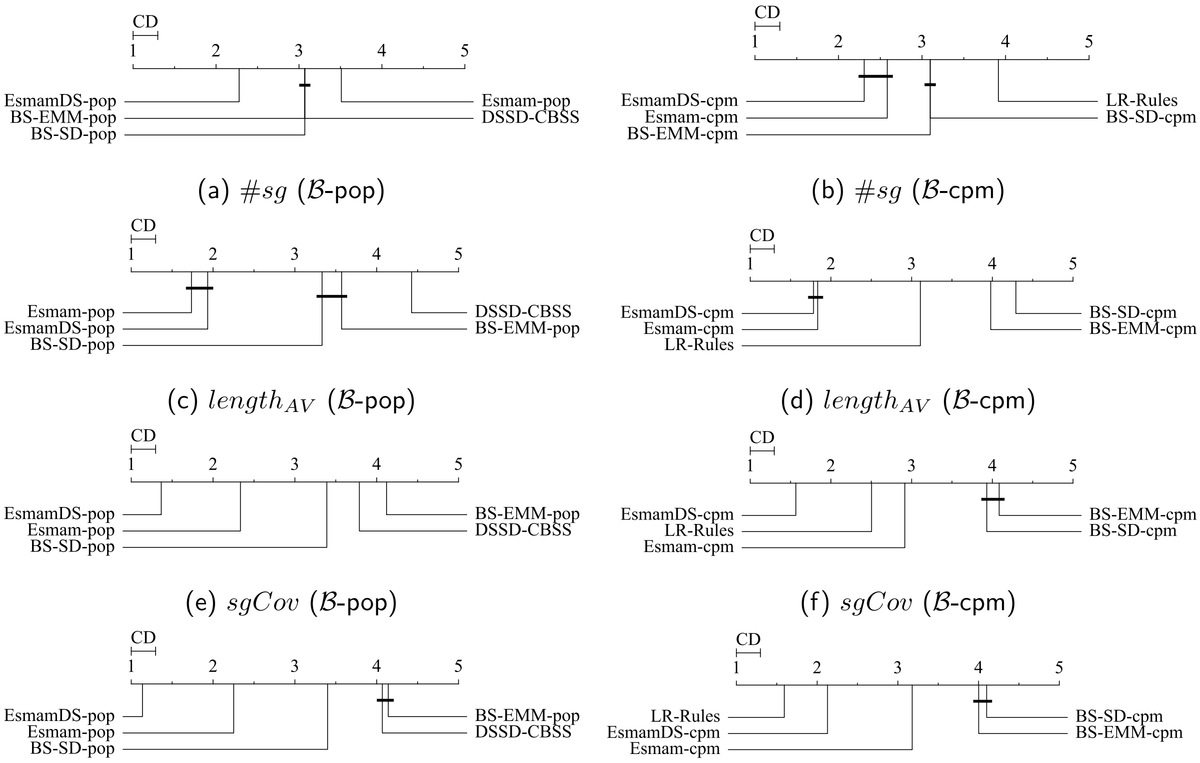
over the whole period of study – i.e. over the survival time feature. Hence, to model survival data and thus analyse survival behaviour, it is necessary to consider not only the survival times but also the censoring information – which is essential information regarding the event occurrence. By disregarding the censoring variable, what is being analysed is merely the time that patients were observed in the study (its distribution) and not the event itself – whether or not it occurred and its distribution over time. In this sense, a deviating distribution of the survival times does not imply a deviating distribution of the event (over time).

Since SD algorithms rely solely on the (single target) survival time to assess survival exceptionality, they ignore crucial information to analyse survival data. Hence, although SD approaches (and the predictive LR-Rules) provide insights on divergences in survivability, when aiming to characterise survival behaviours (models), it is crucial to properly represent the target concept to be optimised. In this sense, the EMM approach plays an important role in behavioural analysis.

Next, we analyse the results for the remaining metrics presented in Table 9, which were assessed using the Friedman and Nemenyi statistical tests. First, we assess the descriptive aspects of the results with the metrics of interpretability and representativeness. Figure 8 provides the Critical Distance (CD) diagrams for the Nemenyi post-hoc test showing the results of the statistical comparison of all approaches against each other by the mean ranks of populations. The approaches that are not significantly different are connected by a horizontal bar.

We notice that the EsmamDS (for both baselines) outperforms all other approaches in the *interpretability* of their results, presenting a (statistically) significant reduction of 31% in $\#sg$

Figure 8 – Method EsmamDS – Critical Distance (CD) diagrams for the Nemenyi post-hoc test of the interpretability and representativeness metrics.



(a) $\#sg$ ($\mathcal{B}$-pop)

(b) $\#sg$ ($\mathcal{B}$-cpm)

(c) $length_{AV}$ ($\mathcal{B}$-pop)

(d) $length_{AV}$ ($\mathcal{B}$-cpm)

(e) $sgCov$ ($\mathcal{B}$-pop)

(f) $sgCov$ ($\mathcal{B}$-cpm)

(g) $dbCov$ ($\mathcal{B}$-pop)

(h) $dbCov$ ($\mathcal{B}$-cpm)

when comparing EsmamDS-pop to Esmam-pop, 43% comparing EsmamDS-cpm to the LR-Rules, and 6% comparing to Esmam-cpm (with no statistical difference)[2]. Furthermore, no statistical difference was observed for the $length_{AV}$ metric between EsmamDS and its predecessor Esmam. The EsmamDS also outperforms the compared approaches in the *representativeness* of its patterns ($sgCov$), presenting an average coverage from 7 to 16% higher than the others. Regarding the representativeness of the data set ($dbCov$), the EsmamDS(-*cpm*) is outperformed only by the covering LR-Rules, with a difference of 2%. It is essential to notice that the LR-Rules stopping criteria includes covering all the individuals in the data set, which justifies the difference in coverage. Hence, we have that our approach delivers smaller sets of patterns ($\#sg$) that comprise more compact characterisations ($length_{AV}$) while providing more generalisation ($sgCov$) and characterising the large majority of the data records ($dbCov$).

Comparing our Diverse Search approach to its predecessor Esmam, we observe that the description language constraining attributes on a set of values (instead of a single value) and the subgroup selection method that prioritises generalisation yield more informative subgroups.

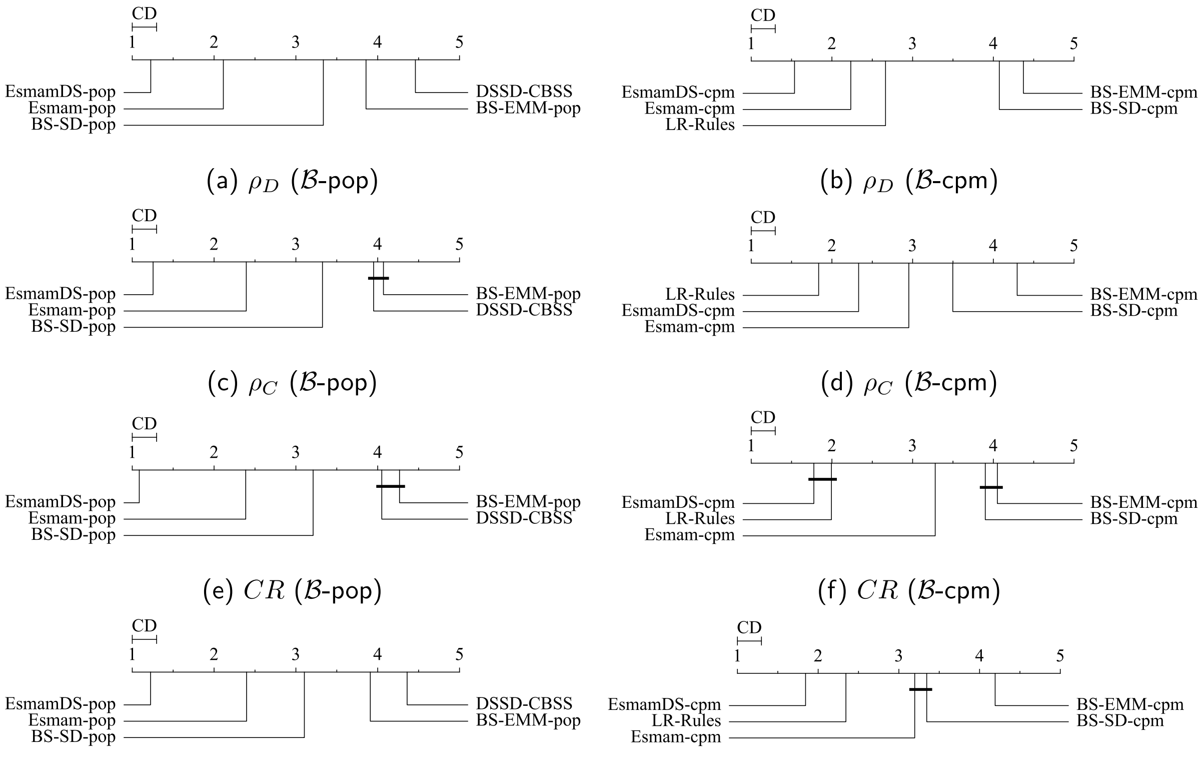
[2] The beam-search approaches – BS-EMM, BS-SD and DSSD-CBSS – assume a fixed number of discovered subgroups, defined as described in Section 5.2

That is because the EsmamDS yields subgroups with statistically broader coverage but with similar description lengths. Also, the heuristic function improving search exploration and the subgroup selection method for minimising redundancy yields fewer patterns that better represent the data (an increase of over 20% of the data set coverage).

Next, we provide the CD diagram of the redundancy metrics in Figure 9. We observe that our approach is outperformed only by the covering LR-Rules in the coverage redundancy ρ_C (a difference of 0.01 in the metric average) but similar in the CR performance. It is important to notice that the LR-Rules naturally induces rules for disjoint subsets of individuals. At each algorithm iteration, a new pattern has to cover a minimum number of previously uncovered individuals from the complete data set. This characteristic justifies the low redundancy in coverage. By contrast, the EsmamDS outperforms all other approaches regarding the redundancy in the final findings, providing a reduction of 80-95% on the levels of descriptive redundancy (ρ_D) and 20-74% the levels of model redundancy (ρ_M). To better analyse the performance of the compared approaches with respect to the problem of redundancy in sets of subgroups, we will use the results achieved on the *pbc* data set (*exp. 0*).

In Figure 10, we present the plot of the similarity metrics (sim_D, sim_C and sim_M)

Figure 9 – Method EsmamDS – Critical Distance (CD) diagrams for the Nemenyi post-hoc test of the redundancy metrics.



(a) ρ_D ($\mathcal{B}$-pop)

(b) ρ_D ($\mathcal{B}$-cpm)

(c) ρ_C ($\mathcal{B}$-pop)

(d) ρ_C ($\mathcal{B}$-cpm)

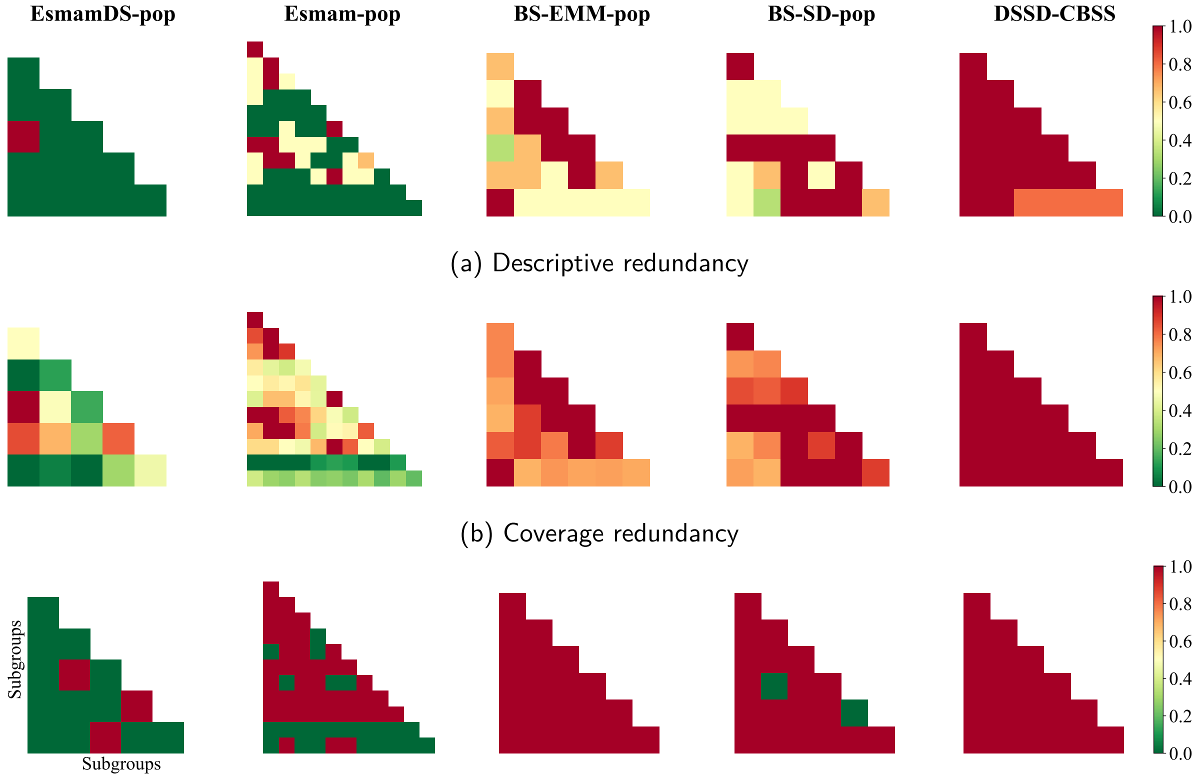
(e) CR ($\mathcal{B}$-pop)

(f) CR ($\mathcal{B}$-cpm)

(g) ρ_M ($\mathcal{B}$-pop)

(h) ρ_M ($\mathcal{B}$-cpm)

Figure 10 – Method EsmamDS – Plot of the similarity metrics between all pairs of subgroups comprising the final set of discovered subgroups delivered by each compared approach on the *pbc* data set (*exp. 0*). Each column of plots provides the results of the approach indicated in the column's title.



(a) Descriptive redundancy

(b) Coverage redundancy

(c) Model redundancy

between all pairs of subgroups comprising the final discoveries of each approach. In that way, we can provide a visual assessment of the levels of similarity – and, thus, redundancy – present in the discovered sets of subgroups. It is interesting to notice that the descriptive (Equation 2.3) and the coverage (Equation 2.4) similarity measures are adaptations of the Jaccard similarity more sensible to subsets – refinements in the descriptive domain. Hence, maximum descriptive/coverage similarity may represent set equality or a subset relation.

Let us consider $\mathbb{G}$ the set of subgroups G discovered by a given approach. We computed the similarity measures for all $(G_i, G_j) \in \mathbb{G}$, for $i \neq j$, and then we plotted such results in a triangular heatmap matrix $|\mathbb{G}| \times |\mathbb{G}|$. Therefore, each index on the matrix represents the level of similarity between two different subgroups within the same discovered set. From the plots, we have that the subgroups discovered by the EsmamDS are more diverse, i.e. present low similarity compared to the others within its set. As a result, we observe that the EsmamDS consistently provides sets of subgroups that present lower redundancy in description and coverage while delivering a larger variety of exceptional survival models.

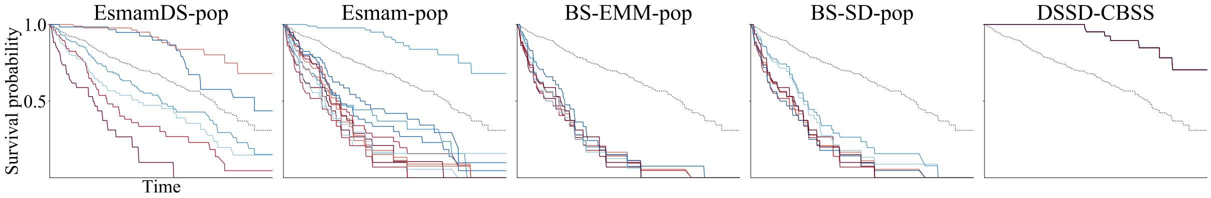
Thus, we observe high levels of redundancy in beam-search approaches, which is an inherent problem of this search heuristic. By prioritising a number of best solutions in each search level, this heuristic leans towards regions of the search space that usually comprise many variations of the same finding. Hence, its subgroup sets usually comprise many subgroup refinements, presenting higher levels of descriptive similarity, highly overlapping coverages and a large number of statistically similar model responses (see Figure 10). Such a high number of refinements among the final findings is also observed through the presence of larger descriptions with smaller coverages (see Figure 8), indicating the specialisations of a more general pattern. The redundancy resultant from refinements also reflects low data representation, which is less than 50% of the data set cases for this family of algorithms (see Table 9). Additionally, redundant patterns yield poor diversity (uniqueness) of survival behaviours, with over 50% of the discovered survival models being similar to each other (ρ_M metric in Table 9).

Figure 10 also allows an important observation: subgroups with no similarity in their descriptions, i.e. subgroups that represent entirely different areas of the descriptive space, sometimes present some inherent redundancy in the population they represent or in the behaviours observed on such populations. By providing more diversify of subgroups, i.e. less similarity between the descriptive patterns provided and the data subsets they describe, our approach eliminates redundant information but preserves the potential information from the inherent overlaps (or similarities) between different exceptional local patterns. Thus, such diversification reflects on the variety (diversity) of the survival behaviours we are able to find and characterise. When assessing the survival models discovered by each approach, the model redundancy metric ρ_M (see Table 9 and Figure 9) shows that the EsmamDS attains higher diversity by achieving lower proportions of similar (pairs of) models.

The impacts of improving subgroup diversity may be observed directly from the survival models uncovered by each approach. In Figure 11, we present the KM survival curves associated to the sets of subgroups discovered on the *pbc* data set (the same sets analysed in Figure 10).

Throughout the plots, it is possible to observe that the survival models provided by our approach (left plot) are more distinct from each other and capture a wider range of

Figure 11 – Method EsmamDS – Survival curves of the subgroups discovered by the $\mathcal{B}$*-population* algorithms on the *pbc* data set (*exp. 0*). The dotted line is the survival curve fitted on the data set.



survival behaviours – from lower to higher response curves. Compared to the other results, the EsmamDS is capable of summarising a range of several similar behaviours while providing more representative patterns (see $\#sg$ metric in Figure 8a and 8b). When assessing the discovered behaviours, we observe that our approach better generalises the patterns into more distinct behaviours, providing clearer and more actionable information.

Finally, we are interested in analysing whether the other approaches also identify the patterns discovered by the EsmamDS and to which extent our approach can identify the others' patterns. In other words, we are interested in assessing the similarity between the set of subgroups delivered by the EsmamDS and those discovered by each compared approach. For that matter, we compute the similarity measures – sim_D, sim_C and sim_M – for all combinations of subgroups in both sets, and we present such analysis in a heatmap matrix, where the rows represent the subgroups discovered by the EsmamDS algorithm and the columns are the unique subgroups discovered by the compared approach.

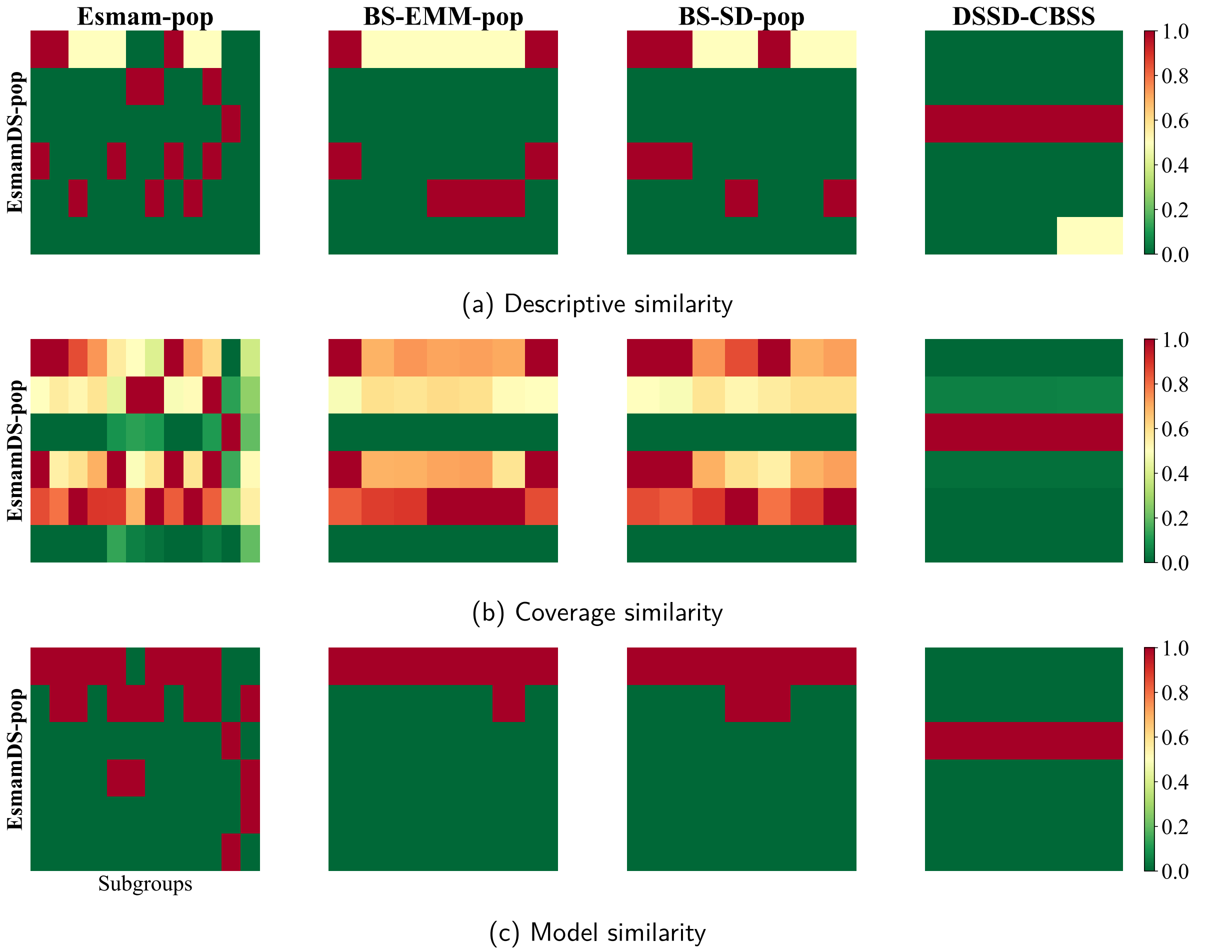
Figure 12 presents such comparison for the results achieved on the *pbc* data set and analysed above (see Figure 10 and 11). It is important to notice the horizontal patterns from the plots, which show the similarity between a single EsmamDS subgroup and all subgroups found by the other approach. In other words, horizontal patterns of high similarity indicate that a single subgroup discovered by our approach may actually typify several subgroups on the compared set.

To exemplify such analysis, we use the comparison between the EsmamDS-pop and the DSSD-CBSS (the last column of heatmaps in Figure 12). From the plot, we can observe that all subgroups in the DSSD-CBSS set (columns) are somehow similar to a single EsmamDS subgroup (third row). Such subgroups' descriptions are as follows, where G_2 is the EsmamDS subgroup and G^* are the DSSD-CBSS subgroups.

$$\begin{array}{lllllllll}
G_2: & sb \in \{0\} & & & & & & & \\
G_0^*: & sb \in \{0\} & \wedge & hp \in \{0\} & albumin \in \{3\} & spiders \in \{0\} & & & \\
G_1^*: & sb \in \{0\} & \wedge & hp \in \{0\} & albumin \in \{3\} & spiders \in \{0\} & \wedge & ascites \in \{0\} \\
G_2^*: & sb \in \{0\} & \wedge & hp \in \{0\} & albumin \in \{3\} & spiders \in \{0\} & \wedge & edema \in \{0\}
\end{array}$$

From the above descriptions, we can observe that all subgroups discovered by the DSSD-CBSS algorithm are refinements of a single EsmamDS-pop subgroup G_2. From Figure 12b and Figure 12c, we observe that the refinements also comprise subsets of G_2 coverage presenting similar survival behaviour. Most of the time, such horizontal patterns in the heatmaps' results

Figure 12 – Method EsmamDS – Similarity measures between the EsmamDS subgroup set and the $\mathcal{B}$-population approaches on the *pbc* data set (*exp. 0*). Each column of plots provides is a comparison with a different approach identified by the column's title. The y-axis of the plots are the subgroups discovered by the EsmamDS-pop, and the x-axis is the subgroups discovered by the compared approach.



(a) Descriptive similarity

(b) Coverage similarity

(c) Model similarity

indicate the presence of subgroup refinements that can be generalised by a single (or few) subgroup(s) in the EsmamDS set without loss of information. Thus, we observe that the sets of subgroups delivered by EsmamDS somehow encompass the subgroups returned by the other algorithms, usually representing the majority of the findings delivered by the other algorithms in a more compact and general way while providing higher diversity of the discovered patterns and models.

5.3 DISCUSSIONS

This chapter presents the EsmamDS, an EMM framework based on Ant-Colony Optimisation (ACO) meta-heuristics for mining subgroups presenting unusual survival behaviour. The EsmamDS builds on the Esmam algorithm (see Chapter 4) to tackle the problem of pattern redundancy. The presented approach explores the ACO design to improve search exploration and introduces a new subgroup selection method to provide a set of diverse subgroups minimising redundancy in description, coverage and model. Additionally, EsmamDS explores the description language to improve the generalisation and comprehensibility of the discovered patterns.

The EsmamDS was confronted with other approaches to mine subgroups with disparities in survivability on 14 survival data sets. We first assessed the results regarding the exceptionality of the discovered patterns. We showed that the SD approaches to find subgroups with deviating mean survival time do not guarantee the exceptionality (unusualness) of their survival models. Thus, we reinforce the fundamental argumentation of Exceptional Model Mining that states the need for more complex target designs to represent more complex properties of interest. When addressing the problem of investigating unexpected survival responses, we should target a proper survival model instead of simply targeting the time to the event occurrence. Then, we assessed the EsmamDS performance regarding its final findings' interpretability, representativeness, and redundancy. We compared its results with its predecessor Esmam, the usual beam-search strategy and an approach that tackles redundancy, and a sequential covering search that maximises the unusualness of KM models. The experiments showed that the EsmamDS yields smaller sets of subgroups, with more straightforward and informative characterisations, capable of representing most of the data observations.

When considering the Esmam limitations discussed in Section 4.3 regarding the lack of diversity among the final findings, the EsmamDS improvements on the description language, search exploration and redundancy minimisation yielded satisfactory results. By increasing the generality and expressivity of the patterns, the EsmamDS provides characterisations with roughly the same length of Esmam ones but representing statistically larger data subsets. As a result, the final sets of subgroups provided by the EsmamDS have roughly the same size as the Esmam sets but with significantly larger data representativeness. This significant increase in the data set coverage also relates to the new ACO design, which improves exploration. The pattern's expressivity that yielded more extensive data coverage is achieved by manipulating discovered subgroups (descriptions). Thus, the diverse search that the EsmamDS introduces is capable of discovering patterns regarding data subsets not represented in the Esmam findings. Our proposed method for subgroup selection yielded a considerable decrease in the levels of descriptive, coverage and model redundancy. The EsmamDS hardly contains refinements on its final findings, which is observed in very low levels of descriptive redundancy. The reduction in the number of refinements among the final findings also reflects lower levels of coverage and model redundancy compared with the other approaches.

Thus, the set of subgroups discovered by this approach usually encompasses the subgroups delivered by the others, but with more general and compact representations, while discovering subgroups that the others do not uncover. The EsmamDS minimises the presence of subgroup refinements in the final set by allowing their occurrence only if they present distinct (unique) survival distribution, providing subgroups that are more diverse concerning their description and coverage while delivering a variety of interesting survival models.

Finally, more thorough experimentation needs to be conducted to assess the impact of the parameters on the search performance and algorithm convergence and assess the impact of factors such as dimensionality and data volume on the algorithm's performance and results. In the following (and final) chapter, we discuss the limitations of this approach and the ways of extending it in future works.

6 CONCLUSIONS AND FINAL REMARKS

We investigated in this work the problem of identifying the factors related to a given event of interest. For any problem surrounding the analysis of an event, i.e. any Survival Analysis (SA) problem, we represent *survival behaviour* through the Kaplan-Meier (KM) non-parametric statistical method of (predictive) SA. Then, we provide a set of (diverse) characterisations of unusual behaviours. In this sense, we contributed to the area of Survival Analysis by approaching it in a descriptive perspective in contrast to the predictive perspective prevalent in the SA literature. The methods we presented here complement the existing SA approaches and aim to fulfil the need for computational tools capable of extracting human-comprehensible and insightful knowledge over subsets of individuals that behave unusually (w.r.t. their survival responses). Although our proposal is applicable to many domains where the problem relies on the investigation of an event occurrence, we focus our efforts on the need of investigating patient outcome.

We resorted to supervised local pattern mining, specifically the Exceptional Model Mining (EMM) task, to discover subsets of the data concisely described that present a deviating (target) survival model. The EMM has evolved from the traditional task of Subgroup Discovery (SD) to represent more complex forms of targets (rather than a single target attribute). Although there are several instances of EMM in the literature defining different target models and evaluation metrics, to the best of our knowledge, the methods we presented here are the first (and only) approaches to explore the use of EMM along with Survival Analysis to model the data. There are, though, several applications of SD and other supervised (and unsupervised) local pattern mining approaches to the domain of medical and biomedical research (HERRERA et al., 2011). Our approach, however, specifically targets survival behaviour in a more complex and informative format without resorting to class labels or stratification. In this sense, we believe this work is a valuable contribution to medical research.

Hence, we introduce an EMM framework that uses the KM estimated survival function as a target model and discovers subgroups presenting statistically unusual survival models. In contrast to the prevalent greedy heuristic searches employed in the EMM literature, we approach the problem of pattern search with stochastic optimisation. Such an approach is already vastly explored for SD tasks but not yet to EMM applications. We, thus, introduce two approaches. In Chapter 4, we introduced the Esmam algorithm that uses Ant-Colony

Optimisation (ACO) meta-heuristic in a sequential covering approach to mine subgroups with exceptional KM models. This approach, however, presents the major drawback of providing sets of subgroups that are highly redundant. Such limitation motivated the second approach. In Chapter 5, we built on the Esmam algorithm to tackle the problem of redundancy in subgroup's set introducing the EsmamDS algorithm. It (implicitly) considers redundancy in the search optimisation process and includes a subgroup selection method that minimises coverage, description and model redundancy. Thus, it explores the description language structure to enhance the generality and expressivity of the final set of patterns.

With the work first presented in Chapter 4 and further enhanced in Chapter 5, we conclude that we were successful in answering our research question. The framework we introduce here was able to provide several characterisations of diverse subgroups that present unusual survival behaviour. The Esmam results show that our approach can discover representative patterns with accurate unusual models and straightforwardly represent them. We also observe that the discovered subgroups potentially capture survival behaviours (known to be) existent in the data. In the EsmamDS results, we show that our approach successfully tackles the problem of subgroup redundancy, providing a set of diverse (unique) exceptional survival subgroups. Thus, the enhanced representation of its patterns yields simpler and more expressive characterisations. Our ACO-based subgroup search outperformed the traditional beam-search in all evaluated aspects, including (and mainly) the redundancy of the final findings. We also show that our EMM approach outperforms the SD task in the exceptionality of the discovered patterns, comprising a more suitable approach to investigate survival behaviour. When considering the compared approaches to provide characterisations over unusual survival behaviours, we show that the EsmamDS results are more straightforward and complete. It is capable of discovering subgroups that are equivalent to the compared approaches (in a simpler and more general way) while uncovering patterns not discovered by the compared approaches.

We believe we made a valuable contribution to investigating factors related to survival response. Rather than using predictive global models to test hypotheses about survival risk, we provide a solution capable of retrieving unusual survival behaviours existent in the data. However, despite the promising results provided in this document, the presented work is just the beginning of investigating a new computational perspective to Survival Analysis problems. There is still a way to go before this method is mature enough to address large scale real problems, such as the COVID-19 data or Omics data sets. To fully achieve our goal of providing insightful knowledge over the circumstances related to patient outcome, some limitations of our approach need to be addressed in order to expand this application towards some open problems. These are the topics of the next sections.

6.1 LIMITATIONS

The first limitation we consider is regarding the descriptive power of our patterns. Since we limited the scope of our search to a combinatorial (discrete) problem, we impel the descriptive attributes to be nominal features. As a result, it is necessary to employ pre-processing discretisation of the data, resulting in considerable loss of information. A variety of works in the literature address the problem of pattern mining in continuous domains – both in the scope of the ACO search (SOCHA, 2004; SWAMINATHAN, 2006; SOCHA; DORIGO, 2008; OTERO; FREITAS; JOHNSON, 2008; OTERO; FREITAS; JOHNSON, 2009) and the subgroup search (MEENG; KNOBBE, 2021). Additionally, the descriptive patterns introduced in the EsmamDS are generated by manipulating simple conjunctive representations. Adapting the pattern induction process to mine disjunctive patterns directly from data may yield interesting results.

Another significant limitation that needs to be tackled is the statistical relevance of the findings. As subgroups are deemed exceptional based on repeated statistical tests, some findings will eventually be false statistical discoveries. Some recent works in literature investigate such a problem and propose solutions to guarantee the statistical robustness of discovered subgroups (DUIVESTEIJN; KNOBBE, 2011; PROENÇA; BÄCK; LEEUWEN, 2021).

Finally, we point out that the approaches presented here are not adapted to efficiently search large volumes of data and high-dimensional domains. Given such characteristics, our approaches struggle to find relevant patterns and, thus, do not converge. (DORIGO; STÜTZLE, 2019) highlight the importance of an appropriate design of the pheromone model and heuristic information in achieving a good balance between exploration and exploitation and ultimately assuring a satisfactory performance over time. Thus, the heuristic information may be used to dynamically bias the probabilistic construction of solutions allowing to consider attribute interaction in the pattern search. Other additional improvement may be implemented in the ant-colony dynamic to improve the final results (MARTENS et al., 2007).

6.2 OPEN PROBLEMS AND FUTURE WORKS

Now that we highlighted the strengths and weaknesses of our methods, we finally present some open problems and some possible new research lines that may extend this work. Hence, we return to the context that motivates this work to briefly consider the possibilities ahead.

As we introduced, the technological development and data explosion is revolutionising modern medicine. The capacity for patient characterisation and for collecting data is expanding dramatically, which opens new possibilities for computational development. In recent years, electronic health records have been increasingly implemented worldwide, in addition to advanced exams and medical machines that collect an enormous amount of patient data. More recently, the challenges encountered in handling the COVID-19 pandemic revealed the latent need for proper data collection and better data structuration. In response, governments and research institutes made a great effort to collect and organise a large amount of data (in the number of cases, features, and data types). This data still being collected enables the structuring of longitudinal data sets, essential to survival studies. Clinical, biological and demographic aspects of a patient that are easily represented in a table (data set) are only a part of all possible information to characterise patients. We have a large amount of imaging and sound exams, temporal data, and many other more complex data types, comprising important information on patients and study cohorts that should be considered while striving for factors associated with prognosis.

Hence, what we see as the mainline of research to extend this work is considering types of data other than tabular ones. One way of approaching this is to enrich the ACO subgroup search procedure (precisely the heuristic information) to consider correlations between features and incorporate more complex data that cannot be considered in the domain of descriptive attributes without loss of comprehensibility.

Another direction to expand this work is further developing the algorithm design to improve performance. It could be addressed by new heuristic information, a new design of the pheromone updating rule and the rule of probabilistic transition (the rule that defines the probability distribution of the search space). Thus, the performance of the algorithm could be improved with auto-adaptative ACO meta-parameters and parallelisation. Additionally, the ACO meta-heuristic is only one suitable alternative to subgroup heuristic search. Other heuristic approaches are being investigated in the literature, especially when considering large and complex data. Moreover, we point out the possibility to improve the representativeness of the target concept. In this work, we use the Kaplan-Meier Estimates to represent survival behaviour. However, as we had revised, it comprises one of the simplest models to represent the probability of surviving over time, suffering from several limitations arising from its statistical design. There are other options in the literature to model survival data that could better capture the (unusual) behaviours we are interested in. Finally, other quality measures can be investigated using other methods for comparing survival curves apart from the logrank statistical test.

REFERENCES

ASHLEY, E. A. Towards precision medicine. *Nature Reviews Genetics*, Nature Publishing Group, v. 17, n. 9, p. 507–522, 2016.

ATTANASIO, C. *Exceptional Incidence Distribution Mining on a Nationwide Cancer Registry: a Descriptive Approach*. Dissertação (Mestrado) — Eindhoven University of Technology, 2019.

ATZMUELLER, M. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 5, n. 1, p. 35–49, 2015.

ATZMUELLER, M.; LEMMERICH, F. Fast subgroup discovery for continuous target concepts. In: SPRINGER. *International Symposium on Methodologies for Intelligent Systems*. [S.l.], 2009. p. 35–44.

ATZMUELLER, M.; PUPPE, F. Sd-map–a fast algorithm for exhaustive subgroup discovery. In: SPRINGER. *European Conference on Principles of Data Mining and Knowledge Discovery*. [S.l.], 2006. p. 6–17.

BAZAN, J.; OSMÓLSKI, A.; SKOWRON, A.; ŚLÇEZAK, D.; SZCZUKA, M.; WROBLEWSKI, J. Rough set approach to the survival analysis. In: SPRINGER. *International Conference on Rough Sets and Current Trends in Computing*. [S.l.], 2002. p. 522–529.

BELFODIL, A.; BELFODIL, A.; BENDIMERAD, A.; LAMARRE, P.; ROBARDET, C.; KAYTOUE, M.; PLANTEVIT, M. Fssd-a fast and efficient algorithm for subgroup set discovery. In: IEEE. *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. [S.l.], 2019. p. 91–99.

BELLE, V. V.; PELCKMANS, K.; HUFFEL, S. V.; SUYKENS, J. A. Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artificial intelligence in medicine*, Elsevier, v. 53, n. 2, p. 107–118, 2011.

BELLE, V. V.; PELCKMANS, K.; SUYKENS, J.; HUFFEL, S. V. Support vector machines for survival analysis. In: *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007)*. [S.l.: s.n.], 2007. p. 1–8.

BERLANGA, F.; JESUS, M. J. D.; GONZÁLEZ, P.; HERRERA, F.; MESONERO, M. Multiobjective evolutionary induction of subgroup discovery fuzzy rules: a case study in marketing. In: SPRINGER. *Industrial Conference on Data Mining*. [S.l.], 2006. p. 337–349.

BIGANZOLI, E.; BORACCHI, P.; MARIANI, L.; MARUBINI, E. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach. *Statistics in medicine*, Wiley Online Library, v. 17, n. 10, p. 1169–1186, 1998.

BLAND, J. M.; ALTMAN, D. G. The logrank test. *BMJ*, BMJ Publishing Group Ltd, v. 328, n. 7447, p. 1073, 2004. ISSN 0959-8138. Disponível em: <https://www.bmj.com/content/328/7447/1073>.

BOLEY, M. *The Power of Saying "I don't know"— An Introduction to Subgroup Discovery and Local Modeling*. 2017. <http://www.realkd.org/subgroup-discovery/the-power-of-saying-i-dont-know-an-introduction-to-subgroup-discovery-and-local-modeling/#more-377>. Accessed: 2022-01-18.

BOLEY, M.; LUCCHESE, C.; PAURAT, D.; GÄRTNER, T. Direct local pattern sampling by efficient two-step random procedures. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2011. p. 582–590.

BOSC, G.; BOULICAUT, J.-F.; RAÏSSI, C.; KAYTOUE, M. Anytime discovery of a diverse set of patterns with monte carlo tree search. *Data mining and knowledge discovery*, Springer, v. 32, n. 3, p. 604–650, 2018.

BOU-HAMAD, I.; LAROCQUE, D.; BEN-AMEUR, H. et al. A review of survival trees. *Statistics Surveys*, The author, under a Creative Commons Attribution License, v. 5, p. 44–71, 2011.

BRADBURN, M.; CLARK, T.; LOVE, S.; ALTMAN, D. Survival analysis part iii: multivariate data analysis–choosing a model and assessing its adequacy and fit. *British journal of cancer*, Nature Publishing Group, v. 89, n. 4, p. 605, 2003.

BRADBURN, M. J.; CLARK, T. G.; LOVE, S.; ALTMAN, D. Survival analysis part ii: multivariate data analysis–an introduction to concepts and methods. *British journal of cancer*, Nature Publishing Group, v. 89, n. 3, p. 431, 2003.

BRINGMANN, B.; ZIMMERMANN, A. The chosen few: On identifying valuable patterns. In: IEEE. *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. [S.l.], 2007. p. 63–72.

CARMONA, C. J.; GONZÁLEZ, P.; JESÚS, M. J. del; HERRERA, F. Non-dominated multi-objective evolutionary algorithm based on fuzzy rules extraction for subgroup discovery. In: SPRINGER. *International Conference on Hybrid Artificial Intelligence Systems*. [S.l.], 2009. p. 573–580.

CARMONA, C. J.; GONZÁLEZ, P.; JESUS, M. J. del; HERRERA, F. Nmeef-sd: non-dominated multiobjective evolutionary algorithm for extracting fuzzy rules in subgroup discovery. *IEEE Transactions on Fuzzy Systems*, IEEE, v. 18, n. 5, p. 958–970, 2010.

CARVALHO, D. V.; PEREIRA, E. M.; CARDOSO, J. S. Machine learning interpretability: A survey on methods and metrics. *Electronics*, v. 8, n. 8, 2019. ISSN 2079-9292. Disponível em: <https://www.mdpi.com/2079-9292/8/8/832>.

CAVAZZUTI, M. *Optimization methods: from theory to design scientific and technological aspects in mechanics*. [S.l.]: Springer Science & Business Media, 2012.

CLARK, T.; BRADBURN, M.; LOVE, S.; ALTMAN, D. Survival analysis part i: basic concepts and first analyses. *British journal of cancer*, Nature Publishing Group, v. 89, n. 2, p. 232, 2003.

CLARK, T.; BRADBURN, M.; LOVE, S.; ALTMAN, D. Survival analysis part iv: further concepts and methods in survival analysis. *British journal of cancer*, Nature Publishing Group, v. 89, n. 5, p. 781, 2003.

COUNCIL, N. R. et al. *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease*. [S.l.]: National Academies Press, 2011.

CROOK, H.; RAZA, S.; NOWELL, J.; YOUNG, M.; EDISON, P. Long covid—mechanisms, risk factors, and management. *Bmj*, British Medical Journal Publishing Group, v. 374, 2021.

DAVIS, R. B.; ANDERSON, J. R. Exponential survival trees. *Statistics in Medicine*, Wiley Online Library, v. 8, n. 8, p. 947–961, 1989.

DORIGO, M.; BIRATTARI, M.; STUTZLE, T. Ant colony optimization. *IEEE computational intelligence magazine*, IEEE, v. 1, n. 4, p. 28–39, 2006.

DORIGO, M.; BLUM, C. Ant colony optimization theory: A survey. *Theoretical computer science*, Elsevier, v. 344, n. 2-3, p. 243–278, 2005.

DORIGO, M.; CARO, G. D. Ant colony optimization: a new meta-heuristic. In: IEEE. *Proceedings of the 1999 congress on evolutionary computation-CEC99 (Cat. No. 99TH8406)*. [S.l.], 1999. v. 2, p. 1470–1477.

DORIGO, M.; MANIEZZO, V.; COLORNI, A. Positive feedback as a search strategy. ., Citeseer, 1991.

DORIGO, M.; MANIEZZO, V.; COLORNI, A. Ant system: optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, IEEE, v. 26, n. 1, p. 29–41, 1996.

DORIGO, M.; STÜTZLE, T. The ant colony optimization metaheuristic: Algorithms, applications, and advances. In: *Handbook of metaheuristics*. [S.l.]: Springer, 2003. p. 250–285.

DORIGO, M.; STÜTZLE, T. Ant colony optimization: overview and recent advances. In: *Handbook of metaheuristics*. [S.l.]: Springer, 2019. p. 311–351.

DUIVESTEIJN, W. *Exceptional Model Mining*. Tese (Doutorado) — Faculteit der Wiskunde en Natuurwetenschappen, Leiden Institute of Advanced Computer Science (LIACS), Faculty of Science, Leiden University, 9 2013.

DUIVESTEIJN, W.; FEELDERS, A. J.; KNOBBE, A. Exceptional model mining. *Data Mining and Knowledge Discovery*, Springer, v. 30, n. 1, p. 47–98, 2016.

DUIVESTEIJN, W.; KNOBBE, A. Exploiting false discoveries – statistical validation of patterns and quality measures in subgroup discovery. In: *2011 IEEE 11th International Conference on Data Mining*. [S.l.: s.n.], 2011. p. 151–160.

EVERS, L.; MESSOW, C.-M. Sparse kernel methods for high-dimensional survival data. *Bioinformatics*, Oxford University Press, v. 24, n. 14, p. 1632–1638, 2008.

FARAGGI, D.; SIMON, R. A neural network model for survival data. *Statistics in medicine*, Wiley Online Library, v. 14, n. 1, p. 73–82, 1995.

FARD, M. J.; WANG, P.; CHAWLA, S.; REDDY, C. K. A bayesian perspective on early stage event prediction in longitudinal data. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 28, n. 12, p. 3126–3139, 2016.

FAUSTO, F.; REYNA-ORTA, A.; CUEVAS, E.; ANDRADE, Á. G.; PEREZ-CISNEROS, M. From ants to whales: metaheuristics for all tastes. *Artificial Intelligence Review*, Springer, v. 53, n. 1, p. 753–810, 2020.

FLOREANO, D.; MATTIUSSI, C. Bio-inspired artificial intelligence. *Ch*, v. 5, p. 335–396, 2008.

FORNI, G.; MANTOVANI, A. Covid-19 vaccines: where we stand and challenges ahead. *Cell Death & Differentiation*, Nature Publishing Group, v. 28, n. 2, p. 626–639, 2021.

FREITAS, A. A.; PARPINELLI, R. S.; LOPES, H. S. Ant colony algorithms for data classification. In: *Encyclopedia of Information Science and Technology, Second Edition*. [S.l.]: IGI Global, 2009. p. 154–159.

FRIEDMAN, J. H.; FISHER, N. I. Bump hunting in high-dimensional data. *Statistics and Computing*, Springer, v. 9, n. 2, p. 123–143, 1999.

FÜRNKRANZ, J. Separate-and-conquer rule learning. *Artificial Intelligence Review*, Springer, v. 13, n. 1, p. 3–54, 1999.

FÜRNKRANZ, J.; GAMBERGER, D.; LAVRAČ, N. *Foundations of rule learning*. [S.l.]: Springer Science & Business Media, 2012.

GAMBERGER, D.; LAVRAC, N. Expert-guided subgroup discovery: Methodology and application. *Journal of Artificial Intelligence Research*, v. 17, p. 501–527, 2002.

GENDREAU, M.; POTVIN, J.-Y. *Handbook of Metaheuristics*. 2nd. ed. [S.l.]: Springer Publishing Company, Incorporated, 2010. ISBN 1441916636.

GORDON, L.; OLSHEN, R. A. Tree-structured survival analysis. *Cancer treatment reports*, v. 69, n. 10, p. 1065–1069, 1985.

GOULARTE, J. F.; SERAFIM, S. D.; COLOMBO, R.; HOGG, B.; CALDIERARO, M. A.; ROSA, A. R. Covid-19 and mental health in brazil: Psychiatric symptoms in the general population. *Journal of Psychiatric Research*, Elsevier, v. 132, p. 32–37, 2021.

GRAF, E.; SCHMOOR, C.; SAUERBREI, W.; SCHUMACHER, M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, Wiley Online Library, v. 18, n. 17-18, p. 2529–2545, 1999.

GROSSKREUTZ, H.; RÜPING, S. On subgroup discovery in numerical domains. *Data mining and knowledge discovery*, Springer, v. 19, n. 2, p. 210–226, 2009.

GROSSKREUTZ, H.; RÜPING, S.; WROBEL, S. Tight optimistic estimates for fast subgroup discovery. In: SPRINGER. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. [S.l.], 2008. p. 440–456.

GUIDOTTI, R.; MONREALE, A.; RUGGIERI, S.; TURINI, F.; GIANNOTTI, F.; PEDRESCHI, D. A survey of methods for explaining black box models. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 51, n. 5, aug 2018. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3236009>.

GUNS, T.; NIJSSEN, S.; RAEDT, L. D. Evaluating pattern set mining strategies in a constraint programming framework. In: SPRINGER. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. [S.l.], 2011. p. 382–394.

HERRERA, F.; CARMONA, C. J.; GONZÁLEZ, P.; JESUS, M. J. D. An overview on subgroup discovery: foundations and applications. *Knowledge and information systems*, Springer, v. 29, n. 3, p. 495–525, 2011.

HOTHORN, T.; LAUSEN, B.; BENNER, A.; RADESPIEL-TRÖGER, M. Bagging survival trees. *Statistics in medicine*, Wiley Online Library, v. 23, n. 1, p. 77–91, 2004.

ISHWARAN, H.; KOGALUR, U. B.; BLACKSTONE, E. H.; LAUER, M. S. et al. Random survival forests. *The annals of applied statistics*, Institute of Mathematical Statistics, v. 2, n. 3, p. 841–860, 2008.

JESUS, M. J. D.; GONZÁLEZ, P.; HERRERA, F.; MESONERO, M. Evolutionary fuzzy rule induction process for subgroup discovery: a case study in marketing. *IEEE Transactions on Fuzzy Systems*, IEEE, v. 15, n. 4, p. 578–592, 2007.

JESUS, M. J. del; GONZÁLEZ, P.; HERRERA, F. Multiobjective genetic algorithm for extracting subgroup discovery fuzzy rules. In: IEEE. *2007 IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making*. [S.l.], 2007. p. 50–57.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, Taylor & Francis, v. 53, n. 282, p. 457–481, 1958.

KAVŠEK, B.; LAVRAČ, N. Apriori-sd: Adapting association rule learning to subgroup discovery. *Applied Artificial Intelligence*, Taylor & Francis, v. 20, n. 7, p. 543–583, 2006.

KAVŠEK, B.; LAVRAČ, N.; JOVANOSKI, V. Apriori-sd: Adapting association rule learning to subgroup discovery. In: SPRINGER. *International Symposium on Intelligent Data Analysis*. [S.l.], 2003. p. 230–241.

KHAN, F. M.; ZUBEK, V. B. Support vector regression for censored data (svrc): a novel tool for survival analysis. In: IEEE. *2008 Eighth IEEE International Conference on Data Mining*. [S.l.], 2008. p. 863–868.

KLEINBAUM, D. G. Survival analysis, a self-learning text. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, Wiley Online Library, v. 40, n. 1, p. 107–108, 1998.

KLöSGEN, W. Explora: A multipattern and multistrategy discovery assistant. In: _____. *Advances in Knowledge Discovery and Data Mining*. USA: American Association for Artificial Intelligence, 1996. p. 249–271. ISBN 0262560976.

KLÖSGEN, W.; MAY, M. Spatio-temporal subgroup discovery. In: *Mining Spatio-Temporal Information Systems*. [S.l.]: Springer, 2002. p. 149–168.

KNOBBE, A. J.; HO, E. K. Pattern teams. In: SPRINGER. *European Conference on Principles of Data Mining and Knowledge Discovery*. [S.l.], 2006. p. 577–584.

KOENIG, I. R.; FUCHS, O.; HANSEN, G.; MUTIUS, E. von; KOPP, M. V. What is precision medicine? *European respiratory journal*, Eur Respiratory Soc, v. 50, n. 4, 2017.

KONONENKO, I. Inductive and bayesian learning in medical diagnosis. *Applied Artificial Intelligence an International Journal*, Taylor & Francis, v. 7, n. 4, p. 317–337, 1993.

KONONENKO, I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, Elsevier, v. 23, n. 1, p. 89–109, 2001.

KRAK, T. E.; FEELDERS, A. Exceptional model mining with tree-constrained gradient ascent. In: SIAM. *Proceedings of the 2015 SIAM International Conference on Data Mining*. [S.l.], 2015. p. 487–495.

KRONEK, L.-P.; REDDY, A. Logical analysis of survival data: prognostic survival models by detecting high-degree interactions in right-censored data. *Bioinformatics*, Oxford University Press, v. 24, n. 16, p. i248–i253, 2008.

LAKKARAJU, H.; BACH, S. H.; LESKOVEC, J. Interpretable decision sets: A joint framework for description and prediction. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 1675–1684. ISBN 9781450342322. Disponível em: <https://doi.org/10.1145/2939672.2939874>.

LAVRAČ, N.; KAVŠEK, B.; FLACH, P.; TODOROVSKI, L. Subgroup discovery with cn2-sd. *Journal of Machine Learning Research*, v. 5, n. Feb, p. 153–188, 2004.

LAVRAČ, N.; ŽELEZNỲ, F.; FLACH, P. A. Rsd: Relational subgroup discovery through first-order feature construction. In: SPRINGER. *International Conference on Inductive Logic Programming*. [S.l.], 2002. p. 149–165.

LEBLANC, M.; CROWLEY, J. Relative risk trees for censored survival data. *Biometrics*, JSTOR, p. 411–425, 1992.

LEEUWEN, M. V. Maximal exceptions with minimal descriptions. *Data Mining and Knowledge Discovery*, Springer, v. 21, n. 2, p. 259–276, 2010.

LEEUWEN, M. V.; KNOBBE, A. Non-redundant subgroup discovery in large and complex data. In: SPRINGER. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. [S.l.], 2011. p. 459–474.

LEEUWEN, M. V.; KNOBBE, A. Diverse subgroup set discovery. *Data Mining and Knowledge Discovery*, Springer, v. 25, n. 2, p. 208–242, 2012.

LEMAN, D.; FEELDERS, A.; KNOBBE, A. Exceptional model mining. In: SPRINGER. *Joint European conference on machine learning and knowledge discovery in databases*. [S.l.], 2008. p. 1–16.

LEMMERICH, F.; BECKER, M. pysubgroup: Easy-to-use subgroup discovery in python. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. [S.l.: s.n.], 2018. p. 658–662.

LEMMERICH, F.; BECKER, M.; ATZMUELLER, M. Generic pattern trees for exhaustive exceptional model mining. In: SPRINGER. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. [S.l.], 2012. p. 277–292.

LI, Y.; WANG, J.; YE, J.; REDDY, C. K. A multi-task learning formulation for survival analysis. In: ACM. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S.l.], 2016. p. 1715–1724.

LI, Y.; WANG, L.; WANG, J.; YE, J.; REDDY, C. K. Transfer learning for survival analysis via efficient l2, 1-norm regularized cox regression. In: IEEE. *2016 IEEE 16th International Conference on Data Mining (ICDM)*. [S.l.], 2016. p. 231–240.

LISBOA, P. J.; WONG, H.; HARRIS, P.; SWINDELL, R. A bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer. *Artificial intelligence in medicine*, Elsevier, v. 28, n. 1, p. 1–25, 2003.

LIU, X.; MININ, V.; HUANG, Y.; SELIGSON, D. B.; HORVATH, S. Statistical methods for analyzing tissue microarray data. *Journal of biopharmaceutical statistics*, Taylor & Francis, v. 14, n. 3, p. 671–685, 2004.

LOWERRE, B. T. *The HARPY speech recognition system*. [S.l.], 1976.

LUCAS, P. J.; GAAG, L. C. Van der; ABU-HANNA, A. Bayesian networks in biomedicine and health-care. *Artificial intelligence in medicine*, Elsevier Science Publishers Ltd., v. 30, n. 3, p. 201–214, 2004.

LUCAS, T.; SILVA, T. C.; VIMIEIRO, R.; LUDERMIR, T. B. A new evolutionary algorithm for mining top-k discriminative patterns in high dimensional data. *Applied Soft Computing*, Elsevier, v. 59, p. 487–499, 2017.

LUCAS, T.; VIMIEIRO, R.; LUDERMIR, T. Ssdp+: A diverse and more informative subgroup discovery approach for high dimensional data. In: IEEE. *2018 IEEE Congress on Evolutionary Computation (CEC)*. [S.l.], 2018. p. 1–8.

LUNA, J. M.; ROMERO, J. R.; ROMERO, C.; VENTURA, S. Discovering subgroups by means of genetic programming. In: SPRINGER. *European Conference on Genetic Programming*. [S.l.], 2013. p. 121–132.

LUNA, J. M.; ROMERO, J. R.; ROMERO, C.; VENTURA, S. On the use of genetic programming for mining comprehensible rules in subgroup discovery. *IEEE transactions on cybernetics*, IEEE, v. 44, n. 12, p. 2329–2341, 2014.

MARTENS, D.; BACKER, M. D.; HAESEN, R.; VANTHIENEN, J.; SNOECK, M.; BAESENS, B. Classification with ant colony optimization. *IEEE Transactions on Evolutionary Computation*, IEEE, v. 11, n. 5, p. 651–665, 2007.

MATTOS, J. B.; NETO, P. S.; VIMIEIRO, R. Esmamds: A more diverse exceptional survival model mining approach. *arXiv preprint arXiv:2109.02610*, 2021.

MATTOS, J. B.; SILVA, E.; NETO, P. M.; VIMIEIRO, R. Clinical risk factors of icu & fatal covid-19 cases in brazil. In: SBC. *Anais do VIII Symposium on Knowledge Discovery, Mining and Learning*. [S.l.], 2020. p. 33–40.

MATTOS, J. B.; SILVA, E. G.; NETO, P. S. de M.; VIMIEIRO, R. Exceptional survival model mining. In: SPRINGER. *Brazilian Conference on Intelligent Systems*. [S.l.], 2020. p. 307–321.

MEENG, M.; KNOBBE, A. Flexible enrichment with cortana–software demo. In: *Proceedings of BeneLearn*. [S.l.: s.n.], 2011. p. 117–119.

MEENG, M.; KNOBBE, A. For real: a thorough look at numeric attributes in subgroup discovery. *Data Mining and Knowledge Discovery*, Springer, v. 35, n. 1, p. 158–212, 2021.

MILIOLI, H. H.; TISHCHENKO, I.; RIVEROS, C.; BERRETTA, R.; MOSCATO, P. Basal-like breast cancer: molecular profiles, clinical features and survival outcomes. *BMC medical genomics*, BioMed Central, v. 10, n. 1, p. 19, 2017.

MOENS, S.; BOLEY, M. Instant exceptional model mining using weighted controlled pattern sampling. In: SPRINGER. *International Symposium on Intelligent Data Analysis*. [S.l.], 2014. p. 203–214.

MONCADA-TORRES, A.; MAAREN, M. C. van; HENDRIKS, M. P.; SIESLING, S.; GELEIJNSE, G. Explainable machine learning can outperform cox regression predictions and provide insights in breast cancer survival. *Scientific Reports*, Nature Publishing Group, v. 11, n. 1, p. 1–13, 2021.

NCCN, N. C. C. N. *NCCN Clinical Practice Guidelines in Oncology Hodgkin Lymphoma Version 1.2017*. [S.l.]: Journal of the National Comprehensive Cancer Network, 2017. v. 15.

NEAPOLITAN, R. E. et al. *Learning bayesian networks*. [S.l.]: Pearson Prentice Hall Upper Saddle River, NJ, 2004. v. 38.

NEPOMUCENO, M. R.; ACOSTA, E.; ALBUREZ-GUTIERREZ, D.; ABURTO, J. M.; GAGNON, A.; TURRA, C. M. Besides population age structure, health and other demographic factors can contribute to understanding the covid-19 burden. *Proceedings of the National Academy of Sciences*, National Academy of Sciences, v. 117, n. 25, p. 13881–13883, 2020. ISSN 0027-8424. Disponível em: <https://www.pnas.org/content/117/25/13881>.

NOVAK, P. K.; LAVRAČ, N.; WEBB, G. I. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, v. 10, n. Feb, p. 377–403, 2009.

OTERO, F. E.; FREITAS, A. A.; JOHNSON, C. G. cant-miner: an ant colony classification algorithm to cope with continuous attributes. In: SPRINGER. *International Conference on Ant Colony Optimization and Swarm Intelligence*. [S.l.], 2008. p. 48–59.

OTERO, F. E.; FREITAS, A. A.; JOHNSON, C. G. Handling continuous attributes in ant colony classification algorithms. In: IEEE. *2009 IEEE Symposium on Computational Intelligence and Data Mining*. [S.l.], 2009. p. 225–231.

PACHÓN, V.; MATA, J.; DOMÍNGUEZ, J. L.; MAÑA, M. J. Multi-objective evolutionary approach for subgroup discovery. In: SPRINGER. *International Conference on Hybrid Artificial Intelligence Systems*. [S.l.], 2011. p. 271–278.

PADILLO, F.; LUNA, J. M.; VENTURA, S. Subgroup discovery on big data: exhaustive methodologies using map-reduce. In: IEEE. *2016 IEEE Trustcom/BigDataSE/ISPA*. [S.l.], 2016. p. 1684–1691.

PADILLO, F.; LUNA, J. M.; VENTURA, S. Exhaustive search algorithms to mine subgroups on big data using apache spark. *Progress in Artificial Intelligence*, Springer, v. 6, n. 2, p. 145–158, 2017.

PARK, J. V.; PARK, S. J.; YOO, J. S. Finding characteristics of exceptional breast cancer subpopulations using subgroup mining and statistical test. *Expert Systems with Applications*, Elsevier, v. 118, p. 553–562, 2019.

PARPINELLI, R. S.; LOPES, H. S.; FREITAS, A. A. An ant colony based system for data mining: applications to medical data. In: CITESEER. *Proceedings of the genetic and evolutionary computation conference (GECCO-2001)*. [S.l.], 2001. p. 791–797.

PARPINELLI, R. S.; LOPES, H. S.; FREITAS, A. A. Data mining with an ant colony optimization algorithm. *IEEE transactions on evolutionary computation*, IEEE, v. 6, n. 4, p. 321–332, 2002.

PATTARAINTAKORN, P.; CERCONE, N. A foundation of rough sets theoretical and computational hybrid intelligent system for survival analysis. *Computers & Mathematics with Applications*, Elsevier, v. 56, n. 7, p. 1699–1708, 2008.

PAUL, E.; STEPTOE, A.; FANCOURT, D. Attitudes towards vaccines and intention to vaccinate against covid-19: Implications for public health communications. *The Lancet Regional Health-Europe*, Elsevier, v. 1, p. 100012, 2021.

PONTES, T.; VIMIEIRO, R.; LUDERMIR, T. B. Ssdp: A simple evolutionary approach for top-k discriminative patterns in high dimensional databases. In: IEEE. *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*. [S.l.], 2016. p. 361–366.

PROENÇA, H. M.; BÄCK, T.; LEEUWEN, M. van. Robust subgroup discovery. *CoRR*, abs/2103.13686, 2021. Disponível em: <https://arxiv.org/abs/2103.13686>.

PULGAR-RUBIO, F.; RIVERA-RIVAS, A.; PÉREZ-GODOY, M. D.; GONZÁLEZ, P.; CARMONA, C. J.; JESUS, M. D. Mefasd-bd: multi-objective evolutionary fuzzy algorithm for subgroup discovery in big data environments-a mapreduce solution. *Knowledge-Based Systems*, Elsevier, v. 117, p. 70–78, 2017.

RADFORD, J.; ILLIDGE, T.; COUNSELL, N.; HANCOCK, B.; PETTENGELL, R.; JOHNSON, P.; WIMPERIS, J.; CULLIGAN, D.; POPOVA, B.; SMITH, P. et al. Results of a trial of pet-directed therapy for early-stage hodgkin's lymphoma. *New England Journal of Medicine*, Mass Medical Soc, v. 372, n. 17, p. 1598–1607, 2015.

RAEDT, L. D.; ZIMMERMANN, A. Constraint-based pattern set mining. In: SIAM. *proceedings of the 2007 SIAM International conference on Data Mining*. [S.l.], 2007. p. 237–248.

RAFTERY, A. E.; MADIGAN, D.; VOLINSKY, C. T. Accounting for model uncertainty in survival analysis improves predictive performance. *Bayesian statistics*, v. 5, p. 323–349, 1996.

RICHARDSON, S.; HIRSCH, J. S.; NARASIMHAN, M.; CRAWFORD, J. M.; MCGINN, T.; DAVIDSON, K. W.; BARNABY, D. P.; BECKER, L. B.; CHELICO, J. D.; COHEN, S. L. et al. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with covid-19 in the new york city area. *Jama*, 2020.

RODRÍGUEZ, D.; RUIZ, R.; RIQUELME, J. C.; AGUILAR-RUIZ, J. S. Searching for rules to detect defective modules: A subgroup discovery approach. *Information Sciences*, Elsevier, v. 191, p. 14–30, 2012.

SEGAL, M. R. Regression trees for censored data. *Biometrics*, JSTOR, p. 35–47, 1988.

SHIVASWAMY, P. K.; CHU, W.; JANSCHE, M. A support vector approach to censored targets. In: IEEE. *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. [S.l.], 2007. p. 655–660.

SIKORA, M.; MIELCAREK, M.; KAŁWAK, K. et al. Application of rule induction to discover survival factors of patients after bone marrow transplantation. *Journal of Medical Informatics & Technologies*, v. 22, p. 35–53, 2013.

SIKORA, M. et al. Censoring weighted separate-and-conquer rule induction from survival data. *Methods of information in medicine*, Schattauer GmbH, v. 53, n. 02, p. 137–148, 2014.

SOCHA, K. Aco for continuous and mixed-variable optimization. In: SPRINGER. *International Workshop on Ant Colony Optimization and Swarm Intelligence*. [S.l.], 2004. p. 25–36.

SOCHA, K.; DORIGO, M. Ant colony optimization for continuous domains. *European journal of operational research*, Elsevier, v. 185, n. 3, p. 1155–1173, 2008.

ŠTAJDUHAR, I.; DALBELO-BAŠIĆ, B. Learning bayesian networks from survival data using weighting censored instances. *Journal of biomedical informatics*, Elsevier, v. 43, n. 4, p. 613–622, 2010.

ŠTAJDUHAR, I.; DALBELO-BAŠIĆ, B.; BOGUNOVIĆ, N. Impact of censoring on learning bayesian networks in survival modelling. *Artificial intelligence in medicine*, Elsevier, v. 47, n. 3, p. 199–217, 2009.

SUTTON, R. S.; BARTO, A. G. et al. *Reinforcement learning: An Introduction*. [S.l.]: MIT press Cambridge, 1998. v. 135.

SWAMINATHAN, S. *Rule induction using ant colony optimization for mixed variable attributes*. Tese (Doutorado) — Texas Tech University, 2006.

TALBI, E.-G. *Metaheuristics: from design to implementation*. [S.l.]: John Wiley & Sons, 2009. v. 74.

VENTURA, S.; LUNA, J. M. *Supervised descriptive pattern mining*. [S.l.]: Springer, 2018.

VENTURA, S.; LUNA, J. M. et al. *Pattern mining with evolutionary algorithms*. [S.l.]: Springer, 2016.

VINZAMURI, B.; LI, Y.; REDDY, C. K. Active learning based survival regression for censored data. In: ACM. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. [S.l.], 2014. p. 241–250.

WANG, P.; LI, Y.; REDDY, C. K. Machine learning for survival analysis: A survey. *ACM Computing Surveys (CSUR)*, ACM, v. 51, n. 6, p. 110, 2019.

WIDODO, A.; YANG, B.-S. Application of relevance vector machine and survival probability to machine degradation assessment. *Expert Systems with Applications*, Elsevier, v. 38, n. 3, p. 2592–2599, 2011.

WOLFF, D.; NEE, S.; HICKEY, N. S.; MARSCHOLLEK, M. Risk factors for covid-19 severity and fatality: a structured literature review. *Infection*, Springer, v. 49, n. 1, p. 15–28, 2021.

WRÓBEL, Ł. Tree-based induction of decision list from survival data. *Journal of Medical Informatics & Technologies*, v. 20, 2012.

WRÓBEL, Ł.; GUDYŚ, A.; SIKORA, M. Learning rule sets from survival data. *BMC bioinformatics*, BioMed Central, v. 18, n. 1, p. 285, 2017.

WROBEL, S. An algorithm for multi-relational discovery of subgroups. In: SPRINGER. *European Symposium on Principles of Data Mining and Knowledge Discovery*. [S.l.], 1997. p. 78–87.

ŽELEZNỲ, F.; LAVRAČ, N. Propositionalization-based relational subgroup discovery with rsd. *Machine Learning*, Springer, v. 62, n. 1-2, p. 33–63, 2006.

ZHANG, C.; ZHANG, S. *Association rule mining: models and algorithms*. [S.I.]: Springer-Verlag, 2002.

ZHENG, Z.; PENG, F.; XU, B.; ZHAO, J.; LIU, H.; PENG, J.; LI, Q.; JIANG, C.; ZHOU, Y.; LIU, S. et al. Risk factors of critical & mortal covid-19 cases: A systematic literature review and meta-analysis. *Journal of Infection*, Elsevier, 2020.

ZUPAN, B.; DEMŠAR, J.; KATTAN, M. W.; BECK, J. R.; BRATKO, I. Machine learning for survival analysis: a case study on recurrence of prostate cancer. *Artificial intelligence in medicine*, Elsevier, v. 20, n. 1, p. 59–75, 2000.

APPENDIX A – ESMAMDS COMPLEMENTARY RESULTS

In this appendix, we provide complementary results for the EsmamDS empirical evaluation presented in Section 5.2. We present each evaluated metric individually, with the results specified by data set. It is important to notice that the EsmamDS and Esmam results were averaged over 30 experiments.

A.1 EXCEPTIONALITY ASSESSMENT ($\mathcal{E}$)

Table 12 – Appendix: EsmamDS complementary results – Metric $\mathcal{E}$ ($\mathcal{B}$-*population* approaches).

	Metric	$\mathcal{E}$				
	Algorithms	EsmamDS-pop	Esmam-pop	BS-EMM-pop	BS-SD-pop	DSSD-CBSS
Data sets	actg320	1.00	1.00	1.00	0.00	1.00
	breast-cancer	1.00	1.00	1.00	1.00	0.80
	cancer	1.00	1.00	1.00	1.00	0.00
	carcinoma	1.00	1.00	1.00	1.00	1.00
	gbsg2	1.00	1.00	1.00	1.00	0.67
	lung	1.00	1.00	1.00	1.00	1.00
	melanoma	1.00	1.00	1.00	0.33	0.00
	mgus2	1.00	1.00	1.00	1.00	1.00
	mgus	1.00	1.00	1.00	1.00	1.00
	pbc	1.00	1.00	1.00	1.00	1.00
	ptc	1.00	1.00	1.00	0.00	0.00
	uis	1.00	1.00	1.00	1.00	1.00
	veteran	1.00	1.00	1.00	1.00	0.50
	whas500	1.00	1.00	1.00	1.00	1.00

Table 13 – Appendix: EsmamDS complementary results – Metric $\mathcal{E}$ ($\mathcal{B}$-*complement* approaches).

	Metric	$\mathcal{E}$				
	Algorithms	EsmamDS-cpm	Esmam-cpm	BS-EMM-cpm	BS-SD-cpm	LR-Rules
Data sets	actg320	1.00	1.00	1.00	0.00	1.00
	breast-cancer	1.00	1.00	1.00	1.00	0.81
	cancer	1.00	1.00	1.00	0.83	1.00
	carcinoma	1.00	1.00	1.00	1.00	1.00
	gbsg2	1.00	1.00	1.00	1.00	1.00
	lung	1.00	1.00	1.00	1.00	1.00
	melanoma	1.00	1.00	1.00	0.67	1.00
	mgus2	1.00	1.00	1.00	1.00	0.79
	mgus	1.00	1.00	1.00	1.00	0.73
	pbc	1.00	1.00	1.00	1.00	1.00
	ptc	1.00	1.00	1.00	0.00	1.00
	uis	1.00	1.00	1.00	1.00	1.00
	veteran	1.00	1.00	1.00	1.00	0.82
	whas500	1.00	1.00	1.00	1.00	1.00

A.2 NUMBER OF DISCOVERED SUBGROUPS ($\#sg$)

Table 14 – Appendix: EsmamDS complementary results – Metric $\#sg$ ($\mathcal{B}$-*population* approaches).

	Metric	$\#sg$				
	Algorithms	EsmamDS-pop	Esmam-pop	BS-EMM-pop	BS-SD-pop	DSSD-CBSS
Data sets	actg320	8.80	3.27	4.00	4.00	4.00
	breast-cancer	24.63	10.00	10.00	10.00	10.00
	cancer	6.03	3.53	4.00	4.00	4.00
	carcinoma	5.03	5.90	6.00	6.00	6.00
	gbsg2	2.60	5.07	6.00	6.00	6.00
	lung	4.53	3.90	4.00	4.00	4.00
	melanoma	5.07	2.40	3.00	3.00	3.00
	mgus2	6.03	5.17	6.00	6.00	6.00
	mgus	3.73	3.90	4.00	4.00	4.00
	pbc	11.50	6.37	7.00	7.00	7.00
	ptc	10.17	5.70	6.00	6.00	6.00
	uis	4.23	6.07	7.00	7.00	7.00
	veteran	3.90	3.47	4.00	4.00	4.00
	whas500	6.03	5.97	6.00	6.00	6.00

Table 15 – Appendix: EsmamDS complementary results – Metric $\#sg$ ($\mathcal{B}$-*complement* approaches).

	Metric	$\#sg$				
	Algorithms	EsmamDS-cpm	Esmam-cpm	BS-EMM-cpm	BS-SD-cpm	LR-Rules
Data sets	actg320	2.60	4.00	3.00	3.00	8.00
	breast-cancer	8.30	15.10	9.00	9.00	16.00
	cancer	5.07	5.27	6.00	6.00	11.00
	carcinoma	4.27	2.93	5.00	5.00	3.00
	gbsg2	6.47	1.73	7.00	7.00	18.00
	lung	5.23	7.10	6.00	6.00	7.00
	melanoma	2.87	4.87	3.00	3.00	3.00
	mgus2	6.20	6.87	7.00	7.00	19.00
	mgus	6.90	6.50	7.00	7.00	11.00
	pbc	7.30	4.73	8.00	8.00	4.00
	ptc	3.07	10.43	4.00	4.00	4.00
	uis	7.07	3.10	8.00	8.00	14.00
	veteran	5.30	4.30	6.00	6.00	11.00
	whas500	4.37	2.90	5.00	5.00	3.00

A.3 AVERAGE DESCRIPTION SIZE ($length_{AV}$)

Table 16 – Appendix: EsmamDS complementary results – Metric $length_{AV}$ ($\mathcal{B}$-*population* approaches).

	Metric	$length_{AV}$				
	Algorithms	EsmamDS-pop	Esmam-pop	BS-EMM-pop	BS-SD-pop	DSSD-CBSS
Data sets	actg320	1.47	1.15	4.00	4.00	4.00
	breast-cancer	1.94	2.30	3.80	4.60	4.70
	cancer	1.70	1.72	1.75	1.50	2.00
	carcinoma	1.89	1.84	1.83	1.83	2.00
	gbsg2	1.01	1.15	1.67	1.50	1.67
	lung	1.00	1.10	1.00	3.00	2.50
	melanoma	1.34	1.02	2.00	1.33	2.00
	mgus2	1.01	1.13	1.17	1.33	2.00
	mgus	1.29	1.15	2.00	1.50	1.75
	pbc	1.71	1.64	2.86	2.29	4.71
	ptc	2.05	2.62	2.33	5.50	2.67
	uis	1.28	1.51	1.86	1.86	2.14
	veteran	1.32	1.14	2.00	1.50	2.00
	whas500	1.41	1.77	2.83	2.67	2.67

Table 17 – Appendix: EsmamDS complementary results – Metric $length_{AV}$ (*$\mathcal{B}$-complement* approaches).

	Metric	$length_{AV}$				
	Algorithms	EsmamDS-cpm	Esmam-cpm	BS-EMM-cpm	BS-SD-cpm	LR-Rules
Data sets	actg320	1.09	1.32	3.33	2.67	2.75
	breast-cancer	1.66	1.37	5.67	5.78	4.44
	cancer	1.25	1.73	1.83	1.83	1.82
	carcinoma	1.14	1.34	1.80	1.60	1.00
	gbsg2	1.16	1.00	1.86	1.71	1.50
	lung	1.12	1.00	1.00	1.33	1.14
	melanoma	1.02	1.17	1.33	1.67	1.33
	mgus2	1.10	1.03	1.14	1.29	2.00
	mgus	1.32	1.46	3.00	2.00	2.82
	pbc	1.20	1.18	1.50	2.12	1.50
	ptc	1.47	1.74	1.75	6.75	1.25
	uis	1.38	1.02	2.62	2.00	2.07
	veteran	1.13	1.37	2.00	2.50	1.36
	whas500	1.04	1.00	1.40	2.60	1.00

A.4 SUBGROUP COVERAGE REPRESENTATIVENESS ($sgCov$)

Table 18 – Appendix: EsmamDS complementary results – Metric $sgCov$ (*$\mathcal{B}$-population* approaches).

	Metric	$sgCov$				
	Algorithms	EsmamDS-pop	Esmam-pop	BS-EMM-pop	BS-SD-pop	DSSD-CBSS
Data sets	actg320	0.30	0.41	0.15	0.23	0.19
	breast-cancer	0.18	0.20	0.11	0.12	0.10
	cancer	0.13	0.15	0.13	0.16	0.21
	carcinoma	0.19	0.22	0.15	0.15	0.20
	gbsg2	0.19	0.27	0.15	0.17	0.20
	lung	0.35	0.42	0.37	0.33	0.33
	melanoma	0.28	0.47	0.18	0.19	0.11
	mgus2	0.20	0.20	0.18	0.17	0.17
	mgus	0.21	0.20	0.17	0.18	0.13
	pbc	0.18	0.29	0.12	0.14	0.20
	ptc	0.18	0.26	0.11	0.11	0.12
	uis	0.19	0.24	0.15	0.15	0.15
	veteran	0.19	0.30	0.13	0.17	0.13
	whas500	0.20	0.29	0.13	0.15	0.10

Table 19 – Appendix: EsmamDS complementary results – Metric $sgCov$ ($\mathcal{B}$-*complement* approaches).

	Metric	$sgCov$				
	Algorithms	EsmamDS-cpm	Esmam-cpm	BS-EMM-cpm	BS-SD-cpm	LR-Rules
Data sets	actg320	0.46	0.40	0.10	0.19	0.21
	breast-cancer	0.31	0.33	0.05	0.08	0.13
	cancer	0.31	0.15	0.14	0.13	0.20
	carcinoma	0.40	0.32	0.27	0.32	0.33
	gbsg2	0.29	0.20	0.10	0.13	0.17
	lung	0.39	0.33	0.35	0.42	0.35
	melanoma	0.46	0.35	0.40	0.20	0.40
	mgus2	0.26	0.19	0.17	0.17	0.16
	mgus	0.25	0.16	0.11	0.16	0.20
	pbc	0.32	0.15	0.20	0.29	0.45
	ptc	0.42	0.22	0.06	0.05	0.44
	uis	0.30	0.20	0.10	0.14	0.25
	veteran	0.26	0.18	0.10	0.08	0.19
	whas500	0.37	0.37	0.50	0.17	0.41

A.5 DATA SET REPRESENTATIVENESS ($dbCov$)

Table 20 – Appendix: EsmamDS complementary results – Metric $dbCov$ ($\mathcal{B}$-*population* approaches).

	Metric	$dbCov$				
	Algorithms	EsmamDS-pop	Esmam-pop	BS-EMM-pop	BS-SD-pop	DSSD-CBSS
Data sets	actg320	1.00	0.99	0.19	0.45	0.19
	breast-cancer	0.81	0.92	0.22	0.24	0.33
	cancer	0.30	0.39	0.27	0.29	0.40
	carcinoma	0.52	0.89	0.22	0.43	0.47
	gbsg2	0.40	0.93	0.34	0.58	0.56
	lung	0.85	0.97	0.90	0.33	0.33
	melanoma	0.77	1.00	0.22	0.54	0.22
	mgus2	0.87	0.99	0.74	0.71	0.20
	mgus	0.47	0.66	0.22	0.39	0.17
	pbc	0.69	0.95	0.20	0.22	0.20
	ptc	0.65	0.84	0.23	0.17	0.12
	uis	0.61	0.89	0.40	0.40	0.20
	veteran	0.54	0.82	0.25	0.42	0.26
	whas500	0.55	0.96	0.24	0.24	0.11

Table 21 – Appendix: EsmamDS complementary results – Metric $dbCov$ ($\mathcal{B}$-*complement* approaches).

	Metric	$dbCov$				
	Algorithms	EsmamDS-cpm	Esmam-cpm	BS-EMM-cpm	BS-SD-cpm	LR-Rules
Data sets	actg320	1.00	1.00	0.10	0.50	1.00
	breast-cancer	0.98	0.99	0.06	0.13	1.00
	cancer	0.99	0.37	0.30	0.36	1.00
	carcinoma	0.98	0.74	0.96	0.96	0.99
	gbsg2	0.98	0.31	0.21	0.48	1.00
	lung	1.00	0.84	0.94	1.00	1.00
	melanoma	1.00	0.91	1.00	0.57	1.00
	mgus2	0.97	0.86	0.76	0.74	1.00
	mgus	0.97	0.55	0.22	0.39	0.99
	pbc	0.99	0.43	1.00	0.94	1.00
	ptc	0.96	0.98	0.08	0.07	1.00
	uis	0.97	0.61	0.20	0.40	0.99
	veteran	0.90	0.55	0.23	0.21	1.00
	whas500	1.00	0.85	1.00	0.24	1.00

A.6 SUBGROUP DESCRIPTION REDUNDANCY (ρ_D)

Table 22 – Appendix: EsmamDS complementary results – Metric ρ_D ($\mathcal{B}$-*population* approaches).

	Metric	ρ_D				
	Algorithms	EsmamDS-pop	Esmam-pop	BS-EMM-pop	BS-SD-pop	DSSD-CBSS
Data sets	actg320	0.17	0.00	0.74	0.63	0.83
	breast-cancer	0.10	0.02	0.65	0.51	0.33
	cancer	0.31	0.08	0.25	0.33	0.33
	carcinoma	0.28	0.07	0.67	0.33	0.60
	gbsg2	0.00	0.00	0.40	0.27	0.40
	lung	0.00	0.00	0.00	0.94	1.00
	melanoma	0.23	0.00	1.00	0.00	0.33
	mgus2	0.00	0.00	0.07	0.07	0.87
	mgus	0.23	0.00	0.92	0.42	0.83
	pbc	0.24	0.02	0.71	0.75	0.96
	ptc	0.28	0.11	0.52	0.74	0.91
	uis	0.16	0.03	0.43	0.48	0.90
	veteran	0.12	0.06	0.42	0.25	0.33
	whas500	0.31	0.08	0.78	0.71	1.00

Table 23 – Appendix: EsmamDS complementary results – Metric ρ_D ($\mathcal{B}$-*complement* approaches).

	Metric	ρ_D				
	Algorithms	EsmamDS-cpm	Esmam-cpm	BS-EMM-cpm	BS-SD-cpm	LR-Rules
Data sets	actg320	0.01	0.18	0.89	0.44	0.40
	breast-cancer	0.00	0.05	0.86	0.70	0.10
	cancer	0.11	0.35	0.43	0.30	0.27
	carcinoma	0.00	0.08	0.50	0.35	0.00
	gbsg2	0.01	0.00	0.64	0.38	0.04
	lung	0.00	0.00	0.00	0.10	0.00
	melanoma	0.00	0.17	0.33	0.00	0.33
	mgus2	0.00	0.01	0.05	0.10	0.07
	mgus	0.00	0.17	0.92	0.49	0.39
	pbc	0.03	0.04	0.21	0.38	0.17
	ptc	0.03	0.20	0.75	0.73	0.00
	uis	0.06	0.00	0.83	0.46	0.25
	veteran	0.03	0.16	0.53	0.70	0.09
	whas500	0.01	0.00	0.25	0.67	0.00

A.7 SUBGROUP COVERAGE REDUNDANCY (ρ_C)

Table 24 – Appendix: EsmamDS complementary results – Metric ρ_C ($\mathcal{B}$-*population* approaches).

	Metric	ρ_C				
	Algorithms	EsmamDS-pop	Esmam-pop	BS-EMM-pop	BS-SD-pop	DSSD-CBSS
Data sets	actg320	0.39	0.22	0.91	0.50	1.00
	breast-cancer	0.38	0.23	0.69	0.62	0.35
	cancer	0.58	0.28	0.50	0.67	0.45
	carcinoma	0.44	0.25	0.75	0.36	0.46
	gbsg2	0.26	0.23	0.52	0.37	0.49
	lung	0.55	0.38	0.49	1.00	1.00
	melanoma	0.44	0.11	1.00	0.12	0.33
	mgus2	0.11	0.01	0.18	0.18	0.98
	mgus	0.33	0.11	1.00	0.33	0.90
	pbc	0.51	0.31	0.84	0.89	1.00
	ptc	0.41	0.30	0.55	0.74	1.00
	uis	0.17	0.22	0.47	0.46	0.95
	veteran	0.26	0.29	0.53	0.42	0.33
	whas500	0.55	0.32	0.95	0.94	1.00

Table 25 – Appendix: EsmamDS complementary results – Metric ρ_C (*$\mathcal{B}$-complement* approaches).

	Metric	ρ_C				
	Algorithms	EsmamDS-cpm	Esmam-cpm	BS-EMM-cpm	BS-SD-cpm	LR-Rules
Data sets	actg320	0.16	0.36	1.00	0.33	0.66
	breast-cancer	0.37	0.48	0.98	0.79	0.19
	cancer	0.41	0.67	0.69	0.48	0.37
	carcinoma	0.28	0.36	0.54	0.34	0.00
	gbsg2	0.30	0.26	0.66	0.40	0.19
	lung	0.43	0.57	0.39	0.45	0.35
	melanoma	0.22	0.46	0.33	0.08	0.33
	mgus2	0.23	0.15	0.17	0.23	0.15
	mgus	0.25	0.39	1.00	0.41	0.33
	pbc	0.52	0.61	0.66	0.55	0.48
	ptc	0.42	0.51	0.95	0.82	0.41
	uis	0.33	0.01	0.94	0.41	0.41
	veteran	0.27	0.27	0.56	0.69	0.22
	whas500	0.39	0.39	0.48	0.87	0.33

A.8 SUBGROUP COVERAGE REDUNDANCY (CR)

Table 26 – Appendix: EsmamDS complementary results – Metric CR (*$\mathcal{B}$-population* approaches).

	Metric	CR				
	Algorithms	EsmamDS-pop	Esmam-pop	BS-EMM-pop	BS-SD-pop	DSSD-CBSS
Data sets	actg320	0.41	0.21	0.81	0.55	0.81
	breast-cancer	0.62	0.39	0.79	0.76	0.68
	cancer	0.70	0.61	0.73	0.71	0.60
	carcinoma	0.52	0.37	0.78	0.57	0.53
	gbsg2	0.60	0.35	0.66	0.45	0.52
	lung	0.45	0.30	0.43	0.67	0.67
	melanoma	0.43	0.10	0.78	0.46	0.78
	mgus2	0.29	0.05	0.34	0.31	0.80
	mgus	0.54	0.34	0.78	0.61	0.83
	pbc	0.63	0.34	0.80	0.78	0.80
	ptc	0.52	0.38	0.77	0.83	0.88
	uis	0.40	0.36	0.61	0.60	0.80
	veteran	0.46	0.23	0.75	0.58	0.74
	whas500	0.54	0.35	0.76	0.76	0.89

Table 27 – Appendix: EsmamDS complementary results – Metric CR ($\mathcal{B}$-*complement* approaches).

	Metric	CR				
	Algorithms	EsmamDS-cpm	Esmam-cpm	BS-EMM-cpm	BS-SD-cpm	LR-Rules
Data sets	actg320	0.13	0.30	0.90	0.50	0.58
	breast-cancer	0.33	0.38	0.94	0.87	0.41
	cancer	0.39	0.67	0.70	0.64	0.49
	carcinoma	0.22	0.41	0.41	0.34	0.01
	gbsg2	0.32	0.69	0.79	0.52	0.29
	lung	0.32	0.44	0.30	0.33	0.34
	melanoma	0.19	0.39	0.25	0.43	0.25
	mgus2	0.26	0.34	0.35	0.40	0.32
	mgus	0.26	0.48	0.78	0.61	0.39
	pbc	0.40	0.64	0.56	0.53	0.25
	ptc	0.23	0.50	0.92	0.93	0.21
	uis	0.30	0.39	0.80	0.61	0.37
	veteran	0.35	0.47	0.77	0.79	0.32
	whas500	0.34	0.37	0.31	0.76	0.29

A.9 SUBGROUP MODEL REDUNDANCY (ρ_M)

Table 28 – Appendix: EsmamDS complementary results – Metric ρ_M ($\mathcal{B}$-*population* approaches).

	Metric	ρ_M				
	Algorithms	EsmamDS-pop	Esmam-pop	BS-EMM-pop	BS-SD-pop	DSSD-CBSS
Data sets	actg320	0.44	0.16	1.00	1.00	1.00
	breast-cancer	0.82	0.46	1.00	1.00	1.00
	cancer	1.00	0.60	1.00	1.00	1.00
	carcinoma	0.97	0.51	1.00	0.47	1.00
	gbsg2	0.73	0.34	1.00	0.67	1.00
	lung	0.43	0.11	0.33	1.00	1.00
	melanoma	0.46	0.11	1.00	0.67	1.00
	mgus2	0.01	0.00	0.07	0.07	1.00
	mgus	0.27	0.20	1.00	0.33	1.00
	pbc	0.67	0.24	1.00	0.90	1.00
	ptc	0.71	0.40	1.00	1.00	1.00
	uis	0.17	0.17	0.52	0.52	1.00
	veteran	0.64	0.16	1.00	0.50	1.00
	whas500	0.57	0.24	0.80	0.60	1.00

Table 29 – Appendix: EsmamDS complementary results – Metric ρ_M ($\mathcal{B}$-*complement* approaches).

	Metric	ρ_M				
	Algorithms	EsmamDS-cpm	Esmam-cpm	BS-EMM-cpm	BS-SD-cpm	LR-Rules
Data sets	actg320	0.12	0.31	1.00	1.00	0.39
	breast-cancer	0.40	0.56	1.00	1.00	0.39
	cancer	0.29	0.61	0.67	0.40	0.33
	carcinoma	0.24	0.57	0.60	0.40	0.33
	gbsg2	0.29	0.73	0.71	0.33	0.35
	lung	0.36	0.62	0.27	0.27	0.29
	melanoma	0.22	0.42	0.33	0.33	0.33
	mgus2	0.02	0.05	0.10	0.05	0.29
	mgus	0.26	0.49	0.71	0.43	0.42
	pbc	0.23	0.38	0.54	0.36	0.00
	ptc	0.22	0.66	1.00	1.00	0.17
	uis	0.15	0.02	0.86	0.36	0.30
	veteran	0.24	0.55	1.00	1.00	0.35
	whas500	0.36	0.52	0.40	0.40	0.33