Machine Learning manuscript No.
(will be inserted by the editor)

EsmamDS

A more diverse exceptional survival model mining approach

Juliana Barcellos Mattos · Paulo S. G. de Mattos Neto · Renato Vimieiro

Received: date / Accepted: date

Abstract Insert your abstract here. Include keywords, PACS and mathematical subject classification numbers as needed. Insert your abstract here. Include keywords, PACS and mathematical subject classification numbers as needed. Insert your abstract here. Include keywords, PACS and mathematical subject classification numbers as needed. Insert your abstract here. Include keywords, PACS and mathematical subject classification numbers as needed. Insert your abstract here. Include keywords, PACS and mathematical subject classification numbers as needed. Insert your abstract here. Include keywords, PACS and mathematical subject classification numbers as needed. Insert your abstract here. Include keywords, PACS and mathematical subject classification numbers as needed. Insert your abstract here. Include keywords, PACS and mathematical subject classification numbers as needed. Insert your abstract here. Include keywords, PACS and mathematical subject classification numbers as needed. Insert your abstract here. Include keywords, PACS and mathematical subject classification numbers as needed. Insert your abstract here. Include keywords, PACS and mathematical subject classification numbers as needed. Insert your abstract here. Include keywords, PACS and mathematical subject classification numbers as needed.

Keywords First keyword · Second keyword · More

J.B. Mattos
E-mail: julianabarcellosmattos@gmail.com

P.S.G.M. Neto
Centro de Informática (CIn), Universidade Federal de Pernambuco (UFPE)
E-mail: psgmn@cin.ufpe.br

R. Vimieiro
Departamento de Ciência da Computação (DCC), Universidade Federal de Minas Gerais (UFMG)
E-mail: rvimieiro@dcc.ufmg.br

1 Introduction

For many years, the medical interventions are based on standard treatments protocols suited for a large group of patients listed under the same general medical condition, e.g. some cancer, diabetes, Alzheimer's disease. Often, such approaches are aggressive measures, with intense physiological reactions, a large amount of emotional and psychological stress, and - not rare - low guarantees of effectiveness. In a more recent context of the COVID-19, the medical community worldwide has been struggling to uncover the factors associated with the disease's prognosis to assure effective measures in a medical crisis.

In recent years, the development of biologic databases and methods for characterising patients (e.g. genomics and diverse cellular assays) - in addition to other available information, such as clinical, behavioural, physiological, and even environmental aspects - has essentially changed the possibilities for characterising individuals in their individuality – not as a group. Therefore, an emergent medical approach aims at considering individual variability to the development of personalised medical treatments suited for small and more specific groups of patients. Such personalised interventions may better address the needs of a patient as an individual, improving treatment responses and survival outcomes. Although the current possibilities for individual characterisation are countless, the medical community still struggles to identify subgroups of patients that present unusual behaviour and the characteristics that describe such subgroups – and are thus associated with different survival responses.

A variety of medical researches strive to identify factors associated with the survival response of patients. When looking into the literature, many studies reveal the limitations of the current diseases' markers, reinforcing the need for a better characterisation of diagnostic and prognostic groups. Hence, a large number of studies strive to better understand the hidden patterns that affect survival.

To this end, most studies resort to the Survival Analysis, which aims at estimating the time for the occurrence of a given event of interest, e.g. death, disease recurrence. In other words, the methods of Survival Analysis are essentially predictive approaches, comprising global models to either predict the time for an event to occur or classify patients into risk groups. The approaches that strive to characterise patients' groups regarding their survival behaviour - and define risk groups - usually resort to feature stratification or even impel patients to fit specific classes. By resorting to predefined predictive variables with respect to survival behaviour, such methods rely on assumptions about survival response interactions or already known features' dependencies.

Therefore, when approaching the problem of characterising individuals regarding their survival behaviour, the existing methods cannot identify the factors associated with survival response and, thus, lack the ability to shed light on new and possibly unknown survival factors. In contrast to those predictive approaches, a new look at this problem may comprehend the data as a potential composition of different subsets that present distinct survival behaviour and, thus, strive to uncover patterns associated with survival response. When comparing to global predictive models, our premise is that the discovery of local patterns is a more suitable approach to the characterisation of distinct subgroups. Instead of constructing survival risk characterisation from defined predictive features, our goal

is to discover and describe the features' interactions - patterns - that yield unusual survival behaviours.

Therefore, we approach the problem of discovering and characterising subgroups of patients with distinct survival response through the perspective of Supervised Descriptive Pattern Mining. It comprises the set of pattern mining tasks that use local patterns to characterise/distinguish two or more data subsets and, hence, provide descriptive knowledge from labelled data. Subgroup Discovery (SD) [51,2] is one of its traditional tasks and aims at discovering and describing subsets of the population that are as large as possible and for which the distribution of a target variable over the subset can be considered exceptional. In Park *et al.* [91], the authors employ Subgroup Discovery to uncover interesting patterns related to long-term and short-term survival in breast cancer based on the distribution of the survival time of the patients.

However, one can understand that a single variable is an oversimplified target representation, and more complex models can usually better represent the data. In other words, the deviation of one target attribute may not capture important information of the study cohort's survival experience. Therefore, our second premise is that survival behaviour can be better represented by a model target rather than a numerical one. Hence, we propose the use of Exceptional Model Mining (EMM) [70], a multi-target generalisation of Subgroup Discovery that searches for subsets of the population that present a deviating mathematical model.

In [83], we introduced the Esmam algorithm, an EMM framework based on Ant-Colony Optimisation designed for discovering subgroups with unusual survival behaviour. The algorithm returns a set of descriptions (characterisations) of subsets of the data that present a distinctive survival model. To the best of our knowledge, there is no previous work on literature that addresses the problem of characterising different survival groups by searching for local exceptionalities related to a survival model. And although EMM has been widely explored in the study of unusual behaviours, the study of survival behaviour is still an unexplored field. The Esmam algorithm is also the first EMM approach to explore bio-inspired metaheuristics as the pattern mining search strategy, in contrast to the state-of-the-art greedy beam-search algorithm. Its results indicate that the approach can discover significant subgroups related to survival response and identify factors that interfere with survival experience - while delivering competitive results concerning the complexity and generality of the subgroups' set. In this manuscript, we build on the work presented in [83] with the hypothesis that a more diverse search yield fewer redundant patterns and that decreasing the redundancy of discovered subgroups improves the effectiveness of the search for unusual survival behaviours.

This paper presents the Exceptional Survival Model AntMiner - Diverse Search (EsmamDS) algorithm, an extension of the work presented in [83] with a more powerful search exploration. Instead of a static information-based heuristic function, the EsmamDS relies on a dynamic heuristic that improves search exploration in two dimensions of redundancy: the pattern's description and the patterns' coverage. Additionally, we employ a new subgroup selection procedure to minimise redundancy in the final set, considering both the subgroups' characterisation and the subgroups' models. Furthermore, with a new description language, the EsmamDS subgroups can better represent generality and yield more straightforward and easy-to-comprehend descriptive models. The proposed algorithm was tested against the state-of-the-art SD and EMM algorithms on 14 survival datasets. We

analyse the results in terms of the complexity, generality and redundancy of the subgroups' sets. The remainder of this work is organised as follows: In Section 2, we first introduce the main concepts of Survival Analysis along with a literature review on the main approaches to analyse survival and provide risk characterisation (subsection 2.1). Then, we formalise Exceptional Model Mining, introduce the problem of patterns' redundancy, and review the existing heuristic searches for EMM and SD tasks (subsection 2.2). Section 4 presents the EsmamDS algorithm, the main contribution of this work, followed by Section 5, which presents the empirical evaluation of our proposal. Finally, in Section 6, we draw some conclusions and present directions to extend this proposal.

2 Literature Review

2.1 Fundamentals of Survival Analysis

Survival Analysis is originally a sub field of Statistics dedicated to analyse and model data where the outcome is the *time* until the occurrence of a given *event* and for which the event outcome is not always known. What we call *event* is any designated experience of interest that may happen to a subject under study, e.g. a device failure, the time a user remains in a website, a patients' death, response to treatment, recovery. 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, specifically in medical studies, the time available to data collecting and the challenges related to patients' follow-up usually result in subjects of the study for which there is no information regarding the event. The Survival Analysis arises to approach specifically such cases when exists a subset of the data for which the occurrence of the *event* - and therefore the time of its occurrence - is not observed.

The *survival data*, or *time-to-event* data, is mainly characterised by a phenomenon we call *censoring*: that is the existence of data subsets that do not present [labeled] information regarding the event occurrence. There are different types of censoring and all relate to the moment when the subject was lost to the study. The *right-censoring* is the most common type observed in survival data. It happens when a subject is lost to the study observation before it has actually experienced the event. In other words, the right-censoring happens when the time when a subject would have suffered the event is past the last time it was observed in study. For simplicity, from now on, all censoring referred in this work should be understood as right-censoring.

For most real-life applications, the goal of Survival Analysis is to estimate the time of occurrence of the designated event. The methods designed to analyse survival data usually strive to build more accurate models to predict the time of event while struggling to handle the challenge of appropriately deal with censored data. Such methods can be divided into: (i) the traditional statistical methods - non-parametric, semi-parametric and parametric - that suffer the drawback of distributional assumptions; (ii) the machine learning methods, that deal with non-linear relations and usually comprise complex mathematical models; and (iii) data mining approaches, that strive for extracting interesting and understandable knowledge from the data while delivering accurate predictive models. In Section 3

we revise the existing literature on the different approaches of survival data analysis and how they relate to the characterisation of different survival responses. In the remainder of this section, we define the Survival Analysis concepts encompassed in this work.

Survival Data Let $\Omega(\mathbb{A}, T, \delta)$ be a survival dataset with size of $|\Omega|$ observations (examples, instances, records). We call o^i the ith observation in Ω, and each observation can be represented as a vector $o^i = (A_1^i, A_2^i, \ldots, A_{|A|}^i, T^i, \delta^i)$. We define $\mathbb{A} = \{A_1, A_2, \ldots, A_{|A|}\}$ as the set of descriptive attributes. $T = \{T^1, T^2, \ldots, T^{|\Omega|}\}$ is the numeric *time*-to-event feature, also referred as *survival time*, and it represents the time since the beginning of the study until the last time T^i that a subject o^i was observed in the study. Finally, the boolean *censoring* feature $\delta = \{\delta^1, \delta^2, \ldots, \delta^{|\Omega|}\}$, also called *survival status*, indicates whether the subject o^i experienced is censored ($\delta^i = 0$) or has experienced the event ($\delta^i = 1$). Therefore, for censored instances ($\delta^i = 0$), i.e. subjects for whom we do not know the occurrence of the event, the survival time T^i indicates the last time the subject was observed [event-free] in the study. For the subjects who suffered the event, $\delta^i = 1$, T^i indicates the time of event occurrence. We say an subject has *survived* if it has not experienced the event. Therefore, from now on, all expressions related to an individual's *survival* refer to the occurrence [or not] of the event.

Survival Function and the Kaplan-Meier Model The survival function $S(t)$ is the usual representation for the probability of an individual o_i to survive up to a specific future time t, i.e., $S(t) = P(T^i > t)$. It presents initial value, $S(t = 0) = 1$ to represent the fact that no subject has suffered (yet) the event at the beginning of the study, and, therefore, the probability of surviving past the initial time is one. Through the time-line of the study, the function monotonically decreases with t and, theoretically, no subject would survive if the study period increased without limit; therefore, $S(t = \infty) = 0$.

The survival function is usually estimated by the Kaplan-Meier (KM) survival estimates [55] – or product-limit method. It comprises an statistical non-parametric model that estimates the survival function by calculating the cumulative survival probability $\hat{S}(t)$ from the observed survival times T for censored and uncensored observations. 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 of Ω. The estimated probability $\hat{S}(t_j)$ of surviving past a time $t_j \in \mathcal{T}$ is given by Equation 1:

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

where $\hat{S}(t_{j-1})$ is the probability of being alive (i.e. not suffer the event) at the time interval $[t_{j-1}, t_j)$, r_j is the number of patients alive just before t_j, and d_j the number of events that happened at the time interval $[t_j, t_{j+1})$. The KM survival curve – or simply survival model or KM model – is the plot of the KM survival probabilities $\hat{S}(t)$ against time, and it provides the visual assessment of the survival probability response over time.

2.2 Discovery of Exceptional Local Behaviours

Data Mining was first defined by [41] as the step in the Knowledge Discovery in Databases process concerned with the development of computational methods capable of extracting and enumerating the data patterns – i.e. substructures that represent some type of homogeneity and regularity in data [116]. Historically, the goal of pattern discovery process can be divided into two different perspectives: (1) *predictive*, whose objective is to find patterns and models capable of predicting the behavior of new data examples; and (2) *descriptive*, that is concerned with finding patterns that better represent the data in a human-understandable form. We call *Supervised Descriptive Pattern Mining* (SDPM) the intersection of both pattern discovery perspectives, emerging from techniques that aim at providing straightforward representation from labeled data [87,115]. As a result, the main goal is to understand the underlying phenomena (according to a target) and not to arbitrarily explain the data nor predict any behaviour.

Subgroup Discovery [51,2] is one of the early tasks of SDPM defined by [122] as *the discovery of interesting subgroups in populations, where interestingness is defined as distributional unusualness with respect to a certain property of interest*, and for which the *property of interest* comprises a single target variable. Understanding that a deviating distribution of one target attribute does not encompass all forms of *interestingness*, the Exceptional Model Mining (EMM) task [70] is defined as the multi-target generalization of Subgroup Discovery for which the *property of interest* assumes the form of a mathematical model. Therefore, adapting from the SD task definition stated by [123], one could say that *given a population of individuals and a property (model) of those individuals that we are interested in, the task of EMM is then to discover the subgroups of the population that are statistically "most interesting", i.e. are as large as possible and have the most unusual statistical (distributional) characteristics with respect to the model of interest.* In other words, given a target model, EMM task searches for subgroups of the data for which the model fitted to the subgroup differs substantially from the same model fitted to the entire data set (or to the subgroup complement).

In [84], the authors highlight that the central concept of SD (and its generalisation EMM), the *subgroups*, essentially comprise a *description* and a *coverage*, i.e. a subset of data records defined through a description in terms of the available data attributes. The absence of a description would configure the discovery of subsets rather than the discovery of subgroups. Such descriptive aspect is inherent to the discovery of subgroups and it is what gives characterisation/identification to the subsets that behave exceptionally. In contrast to the global models prevalent in predictive paradigms, the subgroups are evaluated independently from each other. Such local aspect allows not only the identification of small data subsets presenting exceptional behaviour but also the discovery of overlapping local patterns. Although inherent to the SD/EMM tasks, high levels of description/coverage overlapping between the subgroups should be avoided for the sake of diversity. Finally, both tasks optimise their subgroup's search considering a target concept. This supervised aspect allows the discovery of local patterns with unusual target distributions, where the unusualness is quantified by a predefined quality measure.

Such innate dynamic is essentially related to those three aspects that differentiate both SD and EMM tasks from other Data Mining paradigms – specially the predictive ones: SD/EMM are *descriptive*, *local*, *supervised*, pattern mining

paradigms. As result, they provide straightforward descriptions of local patterns to characterise/distinguish multiple (potentially overlapping) interesting subsets accordingly to a target concept. In the remainder of this section, we define the fundamental concepts and quality measures of both SD and EMM, formulate the problem of redundancy, review the existing literature for computational approaches to the search of subgroups and, finally, provide a review on the recent applications of SD and EMM in real-world problems.

2.2.1 Basic Concepts and Definitions

We assume that the data records to be analysed are characterised by a set of *descriptive* attributes $\mathbb{A} = \{A_1, A_2, \ldots, A_{|A|}\}$ and a set of *target* – or *model* – attributes $\mathbb{M} = \{M_1, \ldots, M_m | m \geq 1\}$ (note that the SD task is the case when $m = 1$). Therefore, in this work we assume the data as a survival dataset $\Omega(\mathbb{A}, \mathbb{M})$ of the size of $|\Omega|$ records (instances, examples), where $\mathbb{M} = \{T, \delta\}$ for the survival time T and the survival status δ defined in the previous section. In this work, $\mathbb{A}$ is taken from a nominal domain. Furthermore, we define the set of *items* $\mathcal{I}_{ij}$ from Ω as $\mathbb{I}^{\Omega} = \{(A_i, V_i^j)\}$, $\forall A_i \in \mathbb{A}$ and $\forall V_i^j \in Domain(A_i)$.

The *description* $\mathcal{D}^{\mathcal{S}}$ of a subgroup $\mathcal{S}$ is taken from a *description language* $\mathcal{L}$ usually defined in terms of $\mathbb{A}$ and of free choice within EMM framework.

Definition 1 (Description) A *description* is a function $\mathcal{D} : \mathbb{A} \rightarrow \{0, 1\}$, where $\mathcal{D}$ *covers* an observation o^i if and only if $\mathcal{D}(A_1^i, \ldots, A_{|A|}^i) = 1$.

In most related works, the description language $\mathcal{L}$ comprises a conjunction of conditions over the set of attributes $\mathbb{A}$. In this work, we propose a slightly different language that employs both conjunctions and disjunctions over the set of items $\mathcal{I}$. Such language will be addressed inside the section dedicated to present our proposed approach. However, for now, we assume $\mathcal{L}$ as a conjunction of *terms* $\mathcal{T}_{ij}$ in the form of $(A_i = V_i^j)$ for $A_i \in \mathbb{A}$ and $V_i^j \in Domain(A_i)$. Note that although the *terms* t_{ij} of a description $\mathcal{D}$ are constructed over the items $\mathcal{I}_{ij} \in \mathbb{I}^{\Omega}$, they impose an operation between A_i and $Domain(A_i)$ – in this work, the operation is restricted to equality since $\mathbb{A}$ is nominal. We call the *length* $|\mathcal{D}|$ of a description the number of *terms* $\mathcal{T}_{ij}$ it encompasses. Such *term* structure will be revisited in Section 4 with the proposal of a different language $\mathcal{L}$.

We define the *refinement* of a description $\mathcal{D}$ as the induction of more complex descriptions from $\mathcal{D}$ by the addition of *items*. We also define *generalisation* of a description as the contrary of refinement, i.e. the induction of simpler patterns by removing items from a description. Note that for a description language based on conjunctions of items, the refinement of a subgroup always yield smaller coverages; i.e a smaller description (imposes less restrictions) is a more general description. That is not the case for the language we define in Section 4.

From the definition of *description*, we have that the restrictions imposed by its terms delineate a subset of the data. And, therefore, we can define a *coverage* $\mathcal{C}^{\mathcal{S}}$ of a subgroup $\mathcal{S}$ in terms of its description $\mathcal{D}^{\mathcal{S}}$.

Definition 2 (Coverage) A *coverage* $\mathcal{C}$ corresponding to a description $\mathcal{D}$ is the set of observations that are covered by $\mathcal{D}$, i.e. $\mathcal{C} = \{o^i \in \Omega | \mathcal{D}(A_1^i, \ldots, A_{|A|}^i) = 1\}$.

We define the *size* $|\mathcal{C}|$ of a coverage as the number of observations it encompasses. Thus, we define the coverage *generality* with respect to the size: a coverage encompassing a larger number of observations is a more general coverage. Note that the descriptions are related to the domain of the attributes $\mathbb{A}$, while the coverage is related to the domain of $|\Omega|$ observations. However, one can understand that the conditions in a description delineate a data subset, and, therefore, more general (less restricted) descriptions lead to more general (larger) coverages.

Finally, we define *subgroups* as an entity comprising both a *description* and a *coverage*. From now on, we assume all descriptions and coverages are associated to a subgroup, and thus the former's imply the latter, and the latter entails the former's.

Definition 3 (Subgroup) A subgroup $\mathcal{S}$ comprises a unique description $\mathcal{D}$ delineating a coverage $\mathcal{C}$ over which a *target concept* is deemed *unusual*.

In order to evaluate a subgroup it is necessary a defined *target concept* $\mathcal{M}$ based on the target attributes $\mathbb{M}$ and a *quality measure* $\phi : \mathcal{D} \rightarrow \mathbb{R}$ that quantifies the deviation between the target $\mathcal{M}^{\mathcal{S}}$ induced over the subgroup's coverage and a defined *baseline* target. In other words, the quality measure needs to quantify the difference between two observed target distributions: it is necessary to compare the target concept distribution observed inside a subgroup to another distribution. Literature offers two possibilities: to compare the subgroup's distribution to either the distribution of its *complement* $\overline{\mathcal{S}} = \{o^i \in \Omega | \mathcal{D}^{\mathcal{S}}(A_1^i, \ldots, A_{|A|}^i) = 0\}$ (i.e. the set of observations not covered by $\mathcal{S}$) or to the distribution on the whole dataset Ω, i.e. the distribution of the *population* $\mathcal{P} = \{o^i \in \Omega\}$.

Literature indicates that there is no general correct choice of whether to compare a subgroup to the *population* or to its *complement*. However, it is important to understand that this choice is crucial to the result, and thus different choices may lead to different outcomes. Notice the conceptual difference of this choice. Evaluate a subgroup with respect to the behaviour of the entire population implies that one is searching for subgroups that deviate from the norm. By contrast, to compare a subgroup to the behaviour of its complement implies that one is searching for a partitioning of the population into two subgroups displaying clearly contrasting behaviour (i.e., schisms in the dataset). One can understand that the choice of a baseline target concept to which compare the subgroups essentially changes the nature of the task in hand: when searching for deviations from a possibly heterogeneous norm, it makes more sense to compare a subgroup to the population, whereas comparing to its complement is more pertinent in finding dichotomies [33]. We implemented our approach to perform both baseline comparisons - *population* and *complement* - and in the experiments we evaluate separately both baselines. From now on, every time we refer to a general *baseline* $\mathcal{B}$, we refer to a predefined baseline target concept to quantify the exceptionality of a subgroup $\mathcal{S}$: the *population* $\mathcal{P}$ or its *complement* $\overline{\mathcal{S}}$.

Finally, what differentiate the SD and EMM frameworks from each other is the definition of the *target concept* and *quality measure*.

SD Framework As previously stated, the Subgroup Discovery task defines the *target concept* as single model attribute (i.e. $\mathbb{M} = \{M_1\}$), and searches for subgroups with unusual distribution of such attribute. For the SD task performed in this work,

the assume its *target concept* $\mathcal{M}_{SD}$ as the single numeric survival time attribute T.

Once a target concept is defined, an appropriate quality measure should quantify the unusualness of the target distribution. A variety of works present different quality measures for different types of SD targets (e.g. numeric, binary, nominal) [51,2]. In [35], the authors highlight that target deviations are more easily found in very small coverages, advising to the importance of considering the coverage size while assessing the quality of a subgroup. The authors also suggest the use of a quality measure based on a statistical test avoid extremely small subgroups. It is important to highlight that to properly assess statistical significance of subgroups the problem of multiple comparisons should be addressed.

Therefore, for SD task, we define a statistical-based quality measure that tests the null hypothesis that the average survival time $\overline{\mathcal{M}}^{\mathcal{S}}_{SD}$ observed in a subgroup $\mathcal{S}$ is equal to the baseline average $\overline{\mathcal{M}}^{\mathcal{B}}_{SD}$ against the alternative of both averages being different.Hence, the quality measure $\phi_{SD}(\mathcal{S})$ of a subgroup $\mathcal{S}$, given by Equation 2, is based on the bilateral t-Test statistical test, where p_{tTest} is the p-value of the test between the survival times of the subgroup $\mathcal{M}^{\mathcal{S}}_{SD}$ and the baseline target $\mathcal{M}^{\mathcal{B}}_{SD}$.

$$\phi_{SD} = 1 - p_{tTest} \tag{2}$$

EMM Framework The Exceptional Model Mining task generalises the SD approach by admitting multiple model attributes (i.e. $\mathbb{M} = \{M_1, \ldots, M_m | m > 1\}$) and, therefore, defining the *target concept* as a model induced over $\mathbb{M}$. Given a model that best represents the data and, therefore, configures the property of interest (target), the quality measure should be able to quantify unusualness in the model distribution.

A variety of model targets have been defined in literature [35], e.g. correlation [30] and Bayesian networks [36]. However, to the best of our knowledge, there is no EMM framework implementation on literature that provides a Survival Analysis model as the target concept. In Section 3 we revise the existing Survival Analysis models – statistical methods, machine learning approaches and rule-base models – as it becomes clearer the lack of local descriptive approaches to the problem of identifying unusual survival behaviours. Therefore, we define an EMM instance for which the *target concept* $\mathcal{M}_{EMM}$ is the Kaplan-Meier Estimates (i.e. KM model) (Eq. 1). We also employ a statistical-based quality measure using the *logrank* test.

The *Logrank* test [96] is the most widely used method for assessing the statistical difference between two survival curves. It tests the null hypothesis that there is no overall difference between two KM models, against the alternative that the compared models are different survival distributions. Essentially, the logrank is a chi-squared test that, for each group under consideration, takes into account the *observed* number of events (indicated by the survival status δ) and the number of events that are *expected* to happen. In the logrank, the *expected* number of events in a group in relation to the total number of observed events in all compared groups is considered to be the proportional to the extent of its risk, i.e. to the proportion of the subjects at risk (not yet suffered the event) that belongs to the group. In other words, for N subjects at risk under evaluation by the logrank, the test defines that the expected number of event for a group of $n \in N$ subjects is equal to n/N of the number of events observed in N.

Finally, we define the quality measure ϕ_{EMM} for the EMM framework as given in Equation 3, where $p_{logrank}$ is the p-value of the logrank test between the KM model of the subgroup $\mathcal{M}^{\mathcal{S}}_{EMM}$ and the considered baseline model $\mathcal{M}^{\mathcal{B}}_{EMM}$.

$$\phi_{EMM} = 1 - p_{logrank} \tag{3}$$

2.2.2 Redundancy in Subgroup Sets

Ultimately, both SD and EMM tasks are essentially one same framework: for each subgroup $\mathcal{S}$ under evaluation, a *target concept* $\mathcal{M}^{\mathcal{S}}$ distribution is learned on $\mathcal{C}^{\mathcal{S}}$ and is then evaluated with a designated *quality measure* ϕ in order to determine whether or not this particular subgroup is exceptional. In order to induce (construct) such subgroups, the existing frameworks usually traverse the description space – i.e. the universe of items $\mathcal{I}^{\Omega}$ – building subgroups from an initially empty description $\mathcal{D}^{\varnothing}$ (note that $\mathcal{D}^{\varnothing} \Rightarrow \mathcal{C}^{\mathcal{P}}$) and incrementally refining it.

The way of conducting the *subgroup search*, i.e. the strategy to induce subgroups through the traversal of a search (description) space, is a well known challenge of the broader scope of pattern mining tasks. That is because the recent high increase in data dimensionality and complexity entails two issues [113]. The first one is that such data leads to huge hypothesis spaces making infeasible the complete traversal of the solution space. To tackle such problem, *heuristic searches* employ different strategies to favor the search towards regions of the space that are more likely to contain good solutions (next section will revise the existing *subgroup search* strategies in literature). This solution, however, bumps into the second issue. The high dimensionality and cardinality of the data, and dependencies between descriptive attributes usually produce multiple possible refinements of a subgroup [114]. This is the problem of *redundancy*: the multiple refinements of a more general subgroup, i.e. the large number of slightly variations of a particular interesting finding, that usually comprise roughly the same description representing almost the same coverage with only slightly changes in the target distribution.

To better enunciate the problem of redundancy in sets of subgroups, we start by considering a hypothetical cancer dataset described by a number of features, including the following:

size:	small, medium, large
location:	I, II
type:	malignant, benign
metastasis:	yes, no

And lets consider the following set of discovered subgroups $\mathcal{S}^i$:

$$\begin{array}{lllllll}
\mathcal{S}^1: & size = large & & & & & \\
\mathcal{S}^2: & size = large & \text{AND} & type = malignant & & & \\
\mathcal{S}^3: & size = large & \text{AND} & location = II & & & \\
\mathcal{S}^4: & size = large & \text{AND} & metastasis = yes & & & \\
\mathcal{S}^5: & size = large & \text{AND} & type = malignant & \text{AND} & location = II & \\
\mathcal{S}^6: & size = large & \text{AND} & type = malignant & \text{AND} & metastasis = yes &
\end{array}$$

Note that the entire set of subgroups is refinements of the more general subgroup $\mathcal{S}^1$, and that there are domains of the data that not even were encompassed in the final set. Once heuristic approaches lean towards the most promising areas of the search space, most works on literature suffer from the high levels of redundancy in the final set, since the most promising areas are usually a large number of variations of the same finding. To deal with such problem, most heuristic approaches strive to achieve a good balance between *exploration* and *exploitation*, i.e. seek for the ability to find multiple and *diverse* local optima. Some works in literature strive to avoid *redundancy* in the set of discovered subgroups by aiming to provide *diversity* in their final findings [61,15,114,113,12,98].

Considering *redundancy* as the presence of *many (slightly) different subgroup descriptions implying many (almost) equal subgroup coverages that have (almost) equal similarity* [114], we follow the works in [113,114] and consider redundancy in three dimensions: in between *descriptions*, in between *coverages* and in between *models* (only in case of EMM). In [113] and [114], the authors present such dimensions as *degrees* of redundancy, being each subsequent degree more strict than its predecessor (description, coverage and target distribution, respectively). Therefore, dealing with each stricter degree of redundancy implies the guarantee of diversity in the previous degree. However, if we consider those three aspects of subgroups – description, coverage and target distribution – as independent dimensions of redundancy (instead of ordered degrees), then we also can understand that the (univariate) minimisation of redundancy in one dimension may lead to avoid diversity in the others. As an example, two subgroups with distinct descriptions, distinct coverages and unusual model distribution (with respect to a baseline model) may present similar models in comparison to each other. And although diversity was not achieved in the latest *degree*, one can understand that those two subgroups comprise distinct (*diverse*) characterisations of two exceptional subgroups that happen (for some reason or not) to have similar *unusual* behaviours.

Therefore, in contrast to the definition in [113,114], we understand that achieving diversity may require optimising redundancy in more than one dimension at a time, and therefore redundancy should be considered in all its dimensions simultaneously. It is important to understand that, although we aim at achieving results that minimise redundancy in its three dimensions, our primary goal is to discover (as many as possible) descriptions of subgroups that are deemed exceptional in comparison to a baseline. If we understand that the task of searching for subgroups entails overlaps between local patterns, and if we understand that such local patterns – the subgroups – entail those three dimensions of redundancy, then one can understand that some redundancy (in any dimension) in the final set is inevitable and, somehow, desired.

Hence, differently from the enunciation in [113,114] that defines diversity as a disjunction of three *degrees* of redundancy, we define diversity in a set of subgroups as a conjuntion of three different *dimensions*. It is also important to pinpoint another aspect that set our approach apart: the authors in [113,114] address redundancy *globally*, i.e. a strict degree of redundancy is optimised regarding the entire set of subgroups. In contrast, we approach redundancy *locally*, analysing all three redundancy dimensions from each local exceptionality individually with relation to each other (discovered) local exceptionalities. Therefore, instead of defining a *Diverse Subgroup Set*, we define a *Set of Diverse Subgroups*.

Definition 4 (Set of Diverse Subgroups) In a set $\mathbb{S}$ of *diverse* (non-*redundant*) subgroups $\mathcal{S}$, all pairs $\mathcal{S}^i, \mathcal{S}^j \in \mathbb{S}$ (for $i \neq j$) should substantially differs:

- in the subgroup *descriptions* $\mathcal{D}$, and
- in the subgroups *coverages* $\mathcal{C}$, and
- in the subgroups *survival models* $\mathcal{M}_{EMM}$

Note that, as our goal is to provide subgroups presenting *unusual survival behaviour*, we define the third dimension of redundancy with relation to the KM survival model induced over the subgroup coverage, i.e. we define such dimension with relation to the EMM target concept. And although SD approaches are also encompassed in this work (and for them the analysis of this third dimension would not apply), this analysis makes sense when considering our premise that the survival behaviour can be better represented by a model target (EMM) rather than by a single numeric one (SD). As we are interested specifically on the EMM target concept (i.e. the KM survival models $\mathcal{M}_{EMM}$), from now on, we will refer to it simply as $\mathcal{M}$, unless explicit provided otherwise. The main contribution of this work builds on the work of presented in [83], providing a set of more diverse subgroups by: (1) improving the exploration mechanisms of our heuristic search approach – an bio-inspired optimisation meta-heuristic – to better explore the spaces of *description* and *coverage* in the induction of the subgroups; and (2) implementing a subgroup selection method that avoids redundancy in all its three dimensions.

For quantifying redundancy in a set of subgroups, we first define measures of *similarity* between pairs of subgroups, for each redundancy dimension.

Description Similarity. Let $\mathbb{I}^{\Omega}$ be the complete set of items $\mathcal{I}$ associated to a dataset Ω, i.e. the *descriptive space*. We define $\mathbb{I}^{\mathcal{S}}$ the set of items that belong to a subgroup $\mathcal{S}$, i.e. the set of items encompassed by the *terms* $\mathcal{T}_{ij}$ in the subgroup's description $\mathcal{D}^{\mathcal{S}}$. Therefore, we define the similarity $\varsigma_{\mathcal{D}}$ between two subgroups descriptions $\mathcal{D}^{\mathcal{S}^A}, \mathcal{D}^{\mathcal{S}^B}$ on the basis of the Jaccard index, as given in Equation 4.

$$\varsigma_{\mathcal{D}}(\mathcal{S}^A, \mathcal{S}^B) = \frac{\mathbb{I}^{\mathcal{S}^A} \cap \mathbb{I}^{\mathcal{S}^B}}{\mathbb{I}^{\mathcal{S}^A} \cup \mathbb{I}^{\mathcal{S}^B}} \tag{4}$$

Coverage Similarity. We define similarity in the coverage dimension, $\varsigma_{\mathcal{C}}$ (given in Equation 5), as the Jaccard index between two subgroups coverages $\mathcal{C}^{\mathcal{S}^A}$ and $\mathcal{C}^{\mathcal{S}^B}$, i.e. we assess coverage similarity on the basis of the observations $o^i \in \Omega$ covered by the subgroups.

$$\varsigma_{\mathcal{C}}(\mathcal{S}^A, \mathcal{S}^B) = \frac{\mathcal{C}^{\mathcal{S}^A} \cap \mathcal{C}^{\mathcal{S}^B}}{\mathcal{C}^{\mathcal{S}^A} \cup \mathcal{C}^{\mathcal{S}^B}} \tag{5}$$

Model Similarity. We define the similarity between two subgroups survival models, $\varsigma_{\mathcal{M}}$ (given in Equation 6), as a boolean function based on the logrank test between the subgroups' models $\mathcal{M}^{\mathcal{S}^A}$ and $\mathcal{M}^{\mathcal{S}^B}$, where $p_{logrank}$ is the p-value of the test and α is a predefined level of significance.

$$\varsigma_{\mathcal{M}}(\mathcal{S}^A, \mathcal{S}^B) = f : p_{logrank} > \alpha \mapsto \{0, 1\} \tag{6}$$

Finally, given $\mathbb{S} = \{\mathcal{S}^1, \ldots, \mathcal{S}^n\}$ a set of n subgroups $\mathcal{S}^i$, we define $\rho^{\mathbb{S}} = \{\rho_{\mathcal{D}}, \rho_{\mathcal{C}}, \rho_{\mathcal{M}}\}$ as the metrics for quantifying *redundancy* (in each one of its dimensions) in a *set of subgroups* by computing the normalised sum of the similarity measures ($\varsigma_{\mathcal{D}}, \varsigma_{\mathcal{C}}$ and $\varsigma_{\mathcal{M}}$) for all combinations of two subgroups in the set. For the sake of simplicity, we omit the $\mathbb{S}$ from the redundancy metrics representation; but one should keep in mind that such are global metrics over a set of subgroups.

Therefore, we define the *description redundancy* $\rho_{\mathcal{D}}$, the *coverage redundancy* $\rho_{\mathcal{C}}$ and the *model redundancy* $\rho_{\mathcal{M}}$ according to Equation 7, where $\varsigma = \{\varsigma_{\mathcal{D}}, \varsigma_{\mathcal{C}}, \varsigma_{\mathcal{M}}\}$ (respectively), $\mathrm{C}_{\mathbb{S},2}$ is the number of all pair-combinations from the subgroups in $\mathbb{S}$ and $\mathbb{C}_{\mathbb{S},2}$ the set of all pair-combinations. For $\rho_{\mathcal{D}}(\rho_{\mathcal{C}})$, one can understand such metrics as the average description (coverage) similarity between the subgroups in a given set $\mathbb{S}$: the smaller they are, the more distinct the subgroups in $\mathbb{S}$ are from each other with relation to their description and to the data examples they characterise. As for model redundancy $\rho_{\mathcal{M}}$, one can understand that such a metric comprises the [normalised] number of $(\mathcal{S}^i, \mathcal{S}^j)$ pair-combinations of subgroups for which the survival models are deemed similar. Therefore, it indicates the percentage of redundant models present in $\mathbb{S}$: the lower this metrics is, the more distinct are the survival models encompassed by the subgroups in the set.

$$\rho_{\mathcal{D}}, \rho_{\mathcal{C}}, \rho_{\mathcal{M}} = \frac{1}{\mathrm{C}_{\mathbb{S},2}} \sum_{\mathcal{S}^i, \mathcal{S}^j \in \mathbb{C}_{\mathbb{S},2}} \varsigma(\mathcal{S}^i, \mathcal{S}^j) \tag{7}$$

Finally, we also evaluate the coverage redundancy between the subgroups in $\mathbb{S}$ by the means of the *Cover Redundancy* (CR) measure introduced by [114]. The authors assume that a *maximally diverse set of subgroups would uniformly cover all observations in a dataset*. Therefore, the CR measures the extent of the deviation between the cover distribution of the observations covered by $\mathbb{S}$ from the uniform cover distribution. Given a dataset Ω and a set $\mathbb{S}$ with $|\mathbb{S}|$ subgroups, the *cover count* of an observation $o^i \in \Omega$ is defined as $c(o^i, \mathbb{S}) = \sum_{\mathcal{S} \in \mathbb{S}} \mathcal{D}^{\mathcal{S}}(o^i)$, i.e. the number of times o^i is covered by any subgroup in $\mathbb{S}$. The *expected cover count* $\hat{c}$ of any random observation $o^i \in \Omega$ is given by the total amount of *cover counts* on Ω avaraged by the number of subgroups in $\mathbb{S}$, i.e. $\hat{c} = \frac{1}{|\mathbb{S}|} \sum_{o^i \in \Omega} c(o^i, \mathbb{S})$. Hence, the Cover Redundancy (CR) measure over a given set of subgroups $\mathbb{S}$ in a given dataset Ω is defined in Equation 8, and the larger measure is, the larger is the deviation from the uniform cover distribution. For the sake of simplicity, we suppress $\mathbb{S}$ and Ω from the metrics representation.

$$CR^{\Omega}(\mathbb{S}) = \frac{1}{|\Omega|} \sum_{o^i \in \Omega} \frac{|c(o^i, \mathbb{S}) - \hat{c}|}{\hat{c}} \tag{8}$$

The authors in [114] argue that because we aim at finding exceptionalities, we cannot expect all tuples to be uniformly covered and, therefore, such measure is not informative. However, when comparing sets of subgroups *(roughly) the same size and for the same dataset*, lower values of CR indicates lower numbers of observations whose *cover counts* largely deviate from the *expected count*. In other words, a lower CR indicates that the set of subgroups presents lesser observations covered a number of times far greater than the expected, i.e. indicates that the subgroup set is more diverse/less redundant.

2.2.3 Subgroup Search

Search for subgroups of a population that are exceptional with relation to a property of interest is an important tool in exploratory data analysis. The strategy used in the search for subgroups is an essential issue for a good performance of computational methods. Over the years, a variety of algorithms have been developed with the goal of efficiently traversing search spaces and thus delivering interesting subgroups with satisfactory computational cost. The existing approaches can be broadly divided into three major categories according to the strategy employed for searching subgroups: (1) exhaustive search based approaches, (2) beam search based approaches, and (3) evolutionary computing approaches. This section is dedicated to introduce these main search strategies. As EMM can be understood as a generalization of SD task with a more complex target concept, this work will present the main SD algorithms as well as the EMM ones. This review was guided mainly by the works of [50], [18], [51] and [35]. For a more detailed overview on the specifics of the algorithms to be presented, we refer the aforementioned literature.

Exhaustive Search The exhaustive search strategy explores all combinatorial space, ensuring the delivery of 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 [59]; MIDOS [122]; APRIORI-SD [57,56]; SD-Map [4] and SD-Map* [3]; DPSubgroup [49]; MergeSD [48]; and the GP-Growth [71] developed specifically to EMM task. There are also SD approaches to big data (e.g. AprioriK-SD and PFP-SD [89,90]) based on MapReduce [21], that allows the processing of large databases through automatic parallelization 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. In this work, we focus on comparing our EMM heuristic approach with the state-of-the-art existing heuristic approaches.

Beam Search The beam search strategy [77] is an usual heuristic approach among SD and EMM algorithms. It performs a level-wise search similar to best-first search; however, on each search level, the algorithm selects a predefined number of best candidates (determined by *beam size* parameter) among all partial solutions to keep as candidates for the next level. The new candidates are, then, generated from the best candidates kept in the previous level. The most popular SD algorithms that employ beam search are the SubgroupMiner [60], SD [45], CN2-SD [66] and RSD [67,125]. Also, the Cortana Subgroup Discovery[1] is an open source Java implementation for both SD and EMM applying a variety of target concepts, e.g single nominal and numeric target for SD and EMM model classes such as regression and correlation [35]. By exploring only parts of the search space, this strategy

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

may result in the elimination of significant candidates and one of its greatest disadvantages is the lack of diversity in the discovered patterns. Since only some of the best candidates are considered, beam search algorithms usually yield sets of patterns comprising a large number of refinements of a more general description, providing characterisation over the (roughly) same subset of the subjects under analysis and lacking the ability of characterising other potentially interesting subsets of the data. As a manner of dealing with such redundancy, some algorithms apply weighted coverage to increase diversity in the final set of subgroups. In [114], the authors propose to handle such redundancy problem by evaluating sets of subgroups instead of evaluating subgroups by their individual merit – note that such approach differs from the local aspect of traditional SD/EMM tasks. Therefore, the authors present the Diverse Subgroup Set Discovery (DSSD), a beam search approach that handles both single and multiple target attributes. The DSSD handles three degrees of redundancy in sets of subgroups (descriptions, covers and models): for a specific type of redundancy (only on type can be approached), the framework implements a different subgroup selection procedure within the beam search level-wise search (instead of choosing the *top K* best subgroups), and in the selection of the final set. In this work, we evaluate one DSSD variant, the fixed-size Cover-Based Subgroup Selection (CBSS^{k}) which aims at minimizing coverage overlap in the final set of subgroups. Such variant employs a score based on multiplicative weighted covering to dynamically weight the quality of the candidates subgroups given the coverage of the subgroups already selected, and delivers a set of $k \in \mathbb{Z}^+$ subgroups.

Evolutionary Computing The extraction of subgroups associated with a target concept can be treated as an optimization problem – i.e. the problem of finding the best solution among all possible solutions for a given *something* to be improved [19]. Evolutionary Computing (EC) is an area of computer science research that draws inspiration from the process of natural evolution – mainly from the principles of Darwinian evolution and genetics – to design methodologies able to provide an approximately solutions to a wide range of optimization problems, i.e. *evolutionary meta-heuristics*. Algorithms based on evolutionary computation have been widely explored to the discovery of interesting subgroups presenting good results This because they constitute a robust optimization method, performing a global search with the capability to explore large search spaces without subdividing it or resorting to pruning techniques. Their search operators and flexibility in design provide a good balance between solution quality and response time – even for large search spaces – and the flexibility in the solutions' representation is a valuable ally to the descriptive aspect of SD/EMM. Additionally, the multi-objective evolutionary approaches allow the simultaneously optimisation of different measures which may be desirable in the search for subgroups. The state-of-the-art evolutionary approaches to SD task are: the evolutionary algorithms SDIGA [22], GAR-SD [88], EDER-SD [103]; the evolutionary programming approach CGBA-SD [81,82]; and the multi-objective approaches MESDIF [9,54] and NMEEF-SD [16,17]. Literature also present few approaches that focus on the problem of searching for subgroups in high dimensional data: the MEFASD-BD [99] is a multi-objective evolutionary fuzzy SD algorithm for big data enviroments; the SSDP [97,79] is a mono-objective evolutionary approach for searching *top-K* subgroups; and the SSDP+ [80] that builds on the SSDP to provide a more diverse set of discovered

subgroups. To the best of our knowledge, there is no evolutionary approach – or any other optimisation meta-heuristic – on literature to provide the discovery of subgroups based on multi-variable target concepts, i.e. there is no work on literature that explores the use of optimisation meta-heuristics as the search engine of EMM framework.

Apart from those three major search strategies presented, there are in literature other heuristic contributions to EMM task. The Exception Maximisation and Description Minimisation (EMDM) [69] approach employs a search strategy that explores structures in the two data subspaces: the descriptive attributes space and the model space. The approach iteratively improves candidate subgroups, and each iteration consists of two steps: *exception maximization*, that searches for subsets presenting unusual model; and *description minimization*, that aims at finding a concise description to define a subgroup from the found subset. In [85], the authors propose an alternative approach by extending and adapting a randomized technique to pattern discovery – Controlled Direct Pattern Sampling (CDPS) [11]. The approach defines a sampling process that yields patterns according to a controlled distribution that favors patterns with high frequency and large model deviation. Finally, the Tree-Constrained Gradient Ascent (TGCA) [64] is a heuristic search strategy designed to exploit information about the influence of individual records on the quality of a subgroup while assuring that the subgroups can be concisely described.

2.2.4 Applications of Subgroup Mining

There is a wide range of contributions in the specialised literature that strive for descriptive knowledge associated with a property of interest. [51] present several SD contributions in fields such as medical domain, bioinformatics, marketing, e-learning, among others. The complex target concept introduced by EMM task expands even more the possibilities in data analysis applications. [72] employ EMM for detecting interpretable subgroups with exceptional transition behavior in sequential data. In [34] the authors apply EMM task in the context of product testing. In [31] the EMM framework is employed to discover exceptional student behavior. [124] propose an algorithm based on the EMM framework to identify and name regions for which there is higher controversy among different classifiers. [1] introduces an EMM approach to discover interesting breast cancer incidence patterns from the Netherlands Cancer Registry (NCR). [8] apply EMM to the problem of discovering exceptional (dis)agreement within groups in behavioral data while [7] approach the problem of discovering (dis)agreement between groups, both works included on [6] that approaches EMM for behavioral data analysis. [102] apply EMM in the context of object-relational data. In [32], the EMM framework is employed to find subgroups with exceptional spatio-temporal behavioral patterns.

Although EMM task has been applied to a wide range of contributions in different domains being valuable in the analysis of different types of data, as will be further discussed in the next section, there is still little research investigating the use of EMM to contributions in survival data analysis.

3 Related Work: Survival risk characterisation

Over the years, many Survival Analysis methods have been developed for assessing the probability of an event occurrence and for modelling the impact of covariates on the occurrence of such event. The authors in [118] broadly classify the survival analysis methods into two categories: statistical methods and machine learning based methods. An overall taxonomy of Survival Analysis methods is provided in the aforementioned work.

Statistical Methods. The traditional statistical approaches can be subdivided into: (i) non-parametric models, e.g. Kaplan-Meier, Nelson-Aalen and Life-Table methods; (ii) semi-parametric models, that are methods based on the Cox Proportional-Hazards (PH) model [14]; and (iii) parametric models, mainly the Accelerated Failure Time (AFT) models. However, such methods are often of difficult interpretation and rely on distributional and restrictive assumptions that need to be fulfilled in order to achieve meaningful results. When these assumptions cannot be satisfied (and they often are not satisfied), the statistical methods suffer from inconsistencies and sub-optimal (inaccurate) results. In addition, such methods struggle to model high-dimensional problems, since the feature selection to include in the model is not a simple choice. 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. The authors in [118] also 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.* The machine learning approaches, however, incorporate those traditional statistical methods for survival analysis in different machine learning techniques, providing more robust predictive survival models.

Machine Learning Methods. In contrast to the limitations of statistical methods, machine learning techniques do not impose distributional assumptions while presenting the advantages of modelling non-linear relationships and delivering high-quality results. Here, we present a brief review of the machine learning methods commonly used in survival analysis, according to the survey presented by [118].

Survival Trees [46][13] are an adaptation of classification and regression trees to handle censored data, where 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 survival analysis (e.g. [104], [20], [68]) with the goal to predict the distribution of the conditional survival function for new data examples [100]. The idea of survival trees has also been extended to ensemble models, like bagging [52] and random forests [53].

Bayesian methods have also been applied in the context of survival prediction, providing the probability of the event of interest. Most approaches make use of Naive bayesian classifier [63][62][126] and Bayesian networks [86][78]. 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 [109][108] and to improve the efficiency of other Survival Analysis methods [101][39].

Artificial Neural Networks are usually employed to directly predict a subject's survival time or to provide the survival status of a subject (that can be represented by the survival or hazard probabilities). There are also works that associate ANN with partial logistic regression [10], Bayesian models [75], and statistical methods [38] However, ANN usually lack the transparency of generated knowledge and the ability to explain the decisions [63], which is highly relevant in a wide range of applications.

Support Vector Machine have also been applied to the analysis of survival data, with the goal of predicting the order in which the event happens for a group of samples [111], [37] or to predict survival times [105] In literature, there are also work on Support Vector Regression [58][112] and on Relevance Vector Machine [119].

There are, still, other machine learning approaches adapted to survival analysis that are found in the literature, e.g the use of **active learning** [117], **transfer learning** [74] and **multi-task learning** [73]. Finally, the machine learning methods designed to analyse survival data strive to build more accurate models to predict survival time – and survival risk – while struggling to handle the challenge of appropriately dealing with censored data.

A wide range of real-world problems under different research domains are built around the analysis of an *event* – Survival Analysis. In healthcare, one can want to analyse *re-hospitalisation* posterior to discharge. In reliability and maintenance researches, one may be interested in product *failure*. In churn analysis, one may want to analyse the clients *dropout*. In literature, there is a variety of works that apply Survival Analysis in different domain problems, e.g. bioinformatics, crowdfunding, student retention, among others [118]. In other words, for a variety of real-world problems, the solution lies in the investigation and analysis of an specific event of interest.

As already presented in this section, the Survival Analysis methods most present in literature focus in two aspects of an *event* analysis: *if* it will happen (the *risk* of happening) and *when* it will happen (the *time* for happening). Therefore, such approaches strive for better knowing the probability of suffering an event and the most probable time for it to happen, i.e. strive for accurate predictive models for *time* and *risk* prediction. For some (or many) problems, however, there are other two important questions to be posed in the analysis of an event: *why* it happens and *what* makes it happens. Why basal-like breast cancer patients present lower survival *times*? Which factors are associated to a higher *risk* of hospitalisation among COVID-19 patients?

Although some of the aforementioned machine learning approaches are able to deliver some comprehension on data partitions characterisation or final estimated scores, they are not designed to the description of the data. Therefore, they usually set thresholds to survival time or rely on features that are already know for being related to the event in order to shed light over different (survival) risks, i.e. to better understand and characterise the reasons and factors related to whether and when an event happens.

When striving to analyse *why* and *what (makes)* an event happens, a variety of works in literature resort to pattern mining approaches aiming to provide un-

derstandable knowledge from the data. The majority of existing approaches make use of rule-based algorithms due to their simplicity in the representation of patterns and features' relationships hidden in data. As we focus our motivation on the struggles of medical researches, we also restrict the literature review on rule-based methods of Survival Analysis to such research domain.

Rule-based Methods (on medical domain). In [5], the authors propose the application of rough sets theory [94][95] to induce a set of decision rules with the goal of finding descriptions of patient groups with different survival estimates. The rules are induced targeting predefined intervals of a prognostic index (PI) that is based on the statistical Cox's PH model. They also compel the observations to artificial classes to search for deviations given a predefined stratification feature. [93] propose an rough sets hybrid system to predict the survival time; the time feature is discretised and the prediction is given in the form of time intervals.

In [76], a bump hunting [43] method is used to characterise high-risk patients. The approach generates rules by searching regions in the feature space with a high average value of the response variable. As the target, the study uses the deviance residual [68] as a substitute for the censored survival time.

In [65], the authors propose the Logical Analysis of Survival Data (LASD) with the goal of constructing patterns to estimate the survival probability distribution of observations. The approach construct a set of rules by partitioning the observations regarding their survival status given a certain time, and then employing a greedy bottom-up approach that maximises the separability power of the resulting pattern according to a defined metric. Hence, they predict the survival function of an observation by averaging the KM estimates of all patterns covering such observation (including the estimates over the entire dataset).

In [120], a survival tree is used to generate an ordered set of rules with the goal of predicting the survival behaviour of new examples. The ruleset is constructed through a separate-and-conquer approach by iteratively learning a survival tree on the still uncovered observations and then selecting the rule that maximises the difference between the KM model fitted on the rule's coverage and the KM model fitted to the remaining cases using the logrank statistical test.

In [106], the authors use sequential covering strategy [44] 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 [107], the authors apply the covering strategy together with a weighting scheme for handling censored observations.

Wróbel *et al.* [121] present the LR-Rules, a top-down greedy covering algorithm to induce accurate models for estimating the survival function of new observations. The rule induction process is guided by a quality measure based on the logrank test between the KM model of the rule coverage and its complement. The algorithm iteratively constructs rules by exhaustively searching for the condition whose addition to the rule yields the highest quality, i.e. the highest separability between KM models (rule and complement). The conditions to be added are taken from the set of observations currently covered by the rule, and the algorithm seeks generality by setting a threshold for minimum rule coverage based on a minimum number of previously uncovered observations that need to be encompassed by the rule. Hence, 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. In this work, we evaluate the LR-Rules as an exhaustive covering search strategy against the SD/EMM state-of-the-art beam-search and our EMM proposed bio-inspired metaheuristics search in the task of providing descriptions of subsets of the data that present distinctive KM models with relation to the subset's complement.

When compared with machine learning approaches, rule-based approaches deliver results that are easier to explain and to understand. However, the works still aim to predict one's survival probability (distribution) in time or to classify risks. Although they have the advantage (over most machine learning techniques) of delivering comprehensible explanation over the data, they are posing the same questions: *if* and *when* an *event* will take place. Therefore, such approaches strive to provide the global model (set of rules) that maximises the accuracy of a target prediction. And despite their capability of providing straightforward explanations, when striving to characterise differences in survival behaviours, such approaches still rely on features' stratification and already known features' interactions.

We, however, pose different questions: *why*, or *what (makes)*, an event (distribution) happens (to be different from the expected). We wanna know the factors related to unusual survival distributions. In contrast to a global predictive model, we aim at a set of local models (patterns, rules, subgroups) whose target distribution substantially deviates from a baseline. And here is where lies the first premise of this work: that to better answer the reasons of the occurrence of unusual survival distributions, the (supervised) discovery of local patterns is a more suitable approach then the existing predictive approaches.

In [91], the authors propose a SD approach to discover interesting patterns for long-term and short-term survival in breast cancer. They presented a tree-based rule induction approach that uses the survival time average as target variable and offers the choice of minimisation/maximisation. 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 subgroups survival time average and the average of its complement. To investigate patterns of long-term and short-term survival, the authors consider a (increase or decrease) of a minimum mean difference in the statistical test between subgroup and complement. However, once again, we argue that – when compared to the average survival time – the survival function better describes the survival behaviour (survival response over time) of the study population. Finally, here we bump into the second premise of this work: that the survival behaviour can be better represented by a model target (survival model) rather than by a single numeric one (average survival time).

To the best of our knowledge, there is no work in literature that combines both our premises and, hence, strives to characterise unexpected *survival behaviours* through the discovery of *local patterns*. Therefore, our approach rely on (predictive) Survival Analysis methods to model the data, but instead of using such models to predictions, we use them as deviation target in combination with the supervised descriptive local pattern mining task of EMM to search for exceptional models. Therefore, for any problem surrounding the analysis of an event – i.e. any Survival Analysis problem – in any research domain, we represent *survival behaviour* through the Kaplan-Meier non-parametric statistical method of (predictive) Sur-

vival Analysis, and then provide a set of (diverse) characterisations of unusual behaviours.

4 EsmamDS: Exceptional Survival Model Ant-Miner - Diverse Search

The Exceptional Survival Model Ant-Miner Diverse Search (EsmamDS) is an EMM framework that extends our previous work presented in [83] to provide a set of more diverse subgroups, where diversity is sought regarding the subgroup's description, coverage and model exceptionality. As stated by [33], an instance of Exceptional Model Mining is defined by a *model class* over the targets and a *quality measure* over the model. Subgroups are generated following a *search strategy*, and then, for each subgroup under consideration, the *model class* is induced over only the data related to the subgroup. The *quality measure* is finally employed to individually evaluate subgroups with regard to their model characteristics (against a baseline model), and the most interesting ones are provided in a final set of (exceptional) subgroups.

We define the EsmamDS instance with a *model class* given by the Kaplan-Meier Estimates (KM model) and the *quality measure* based on the logrank statistical test, as defined in Equation 3. Instead of defining a baseline model to compare subgroups with, our framework provide this choice as a parameter, and therefore provides a set of subgroups exceptional with relation to the population or to their complement – according to the user's choice.

Throughout this section we describe the EsmamDS algorithm. First, we define a new description language in Subsection 4.1 that possibilitates a more flexible and general representation (description $\mathcal{D}$) of subgroups. Then, we present the search strategy of our EMM approach in Subsection 4.2. We briefly introduce the bio-inspired Ant Colony Optimisation (ACO) meta-heuristic, our choice for an heuristic approach to the search of subgroups. Then we present our previous work, the Esmam algorithm, that is the base framework for building our *Diverse Search* – we highlight here that the Esmam code in this work presents slightly differences from the original code in [83] that are adjustments in the algorithm convergence. And, finally, we present the core collaboration of this work: (i) a new heuristic function of ACO framework, which strives for exploration in the *description* and *coverage* dimensions; and (ii) a new subgroup selection method that strives to minimise redundancy in all its three dimensions.

4.1 New Description Language

In Section 2.2, we defined an usual description language $\mathcal{L}$ comprising a conjunction of *terms* $\mathcal{T}_{ij}^{\mathcal{L}} : (A_i = V_i^j)$ (for simplicity: $\mathcal{T}^{\mathcal{L}}$) built over the descriptive space – i.e. the dataset Ω complete set of items $\mathcal{I}_{ij} \in \mathbb{I}^{\Omega} = \{(A_i, V_i^j)\}\ \forall A_i \in \mathbb{A}$ and $\forall V_i^j \in Domain(A_i)$, where the set of descriptive attributes $\mathbb{A}$ is taken from a nominal domain.

Therefore, a description $\mathcal{D} : \mathbb{A} \to \{0, 1\}$ (of size $n \leq |\mathbb{A}|$) taken from the description language $\mathcal{L}$ is given as bellow.

$$\mathcal{D}_{\mathcal{L}} : \mathcal{T}_1^{\mathcal{L}} \wedge \cdots \wedge \mathcal{T}_n^{\mathcal{L}} \cong (A_i = V_i^j)_1 \wedge \cdots \wedge (A_i = V_i^j)_n$$

From Definition 2, we have that the set of observations o^i covered by $\mathcal{D}_\mathcal{L}$ can be given by $\mathcal{C} = \{o^i \in \Omega | \mathcal{D}(\mathcal{T}_1^\mathcal{L} \wedge \cdots \wedge \mathcal{T}_n^\mathcal{L}) = 1\}$, where each term $\mathcal{T}^\mathcal{L}$ is a restriction over a different attribute domain that takes any $A_i \rightarrow \{0, 1\}$. Therefore, each attribute may only be represented once (i.e. by a single term) in a description. Because $\mathcal{T}^\mathcal{L}$ imposes an equality operation over (single) *values*, descriptions $\mathcal{D}_\mathcal{L}$ can only encompass one item $\mathcal{I}_{ij}$ for each term representing an attribute. In other words, a description in $\mathcal{L}$ can only represent single fractions (a single nominal value) of the attribute domains. As consequence, if one understands that the generalisation of a subgroup is the enlargement of its coverage, then we have that $\mathcal{L}$ allows generalisation only through the removal of terms $\mathcal{T}^\mathcal{L}$ from $\mathcal{D}_\mathcal{L}$. However, decreasing the amount of descriptive restrictions is not the only way of striving for generalisation.

Therefore, we propose to aggregate generality to subgroups by allowing descriptions to capture *disjuntive* conditions on single attributes. Instead of representing single items, we propose a more flexible description language that allows the structure of a term to represent *sets* of items. In other words, instead of imposing an equality operator over values in $Domain(A_i)$, we propose a $\mathcal{T}_{ij}$ structure that imposes a *disjunction* over a *set* of values in $Domain(A_i)$.

EsmamDS Description Language. We define the EsmamDS description language $\mathcal{L}^{DS}$ comprising a conjunction of *terms* $\mathcal{T}_{ij}^{\mathcal{L}^{DS}}$ – for simplicity, $\mathcal{T}^{\mathcal{L}^{DS}}$ – of the form $(A_i = \{V_i^j | V_i^j \in Domain(A_i)\})$ built over the descriptive space $\mathbb{I}^\Omega$. Therefore, a description $\mathcal{D} \rightarrow \{0, 1\}$ (of size $n \leq |\mathbb{A}|$) taken from the description language $\mathcal{L}^{DS}$ is given as bellow.

$$\mathcal{D}_{\mathcal{L}^{DS}} : \mathcal{T}_1^{\mathcal{L}^{DS}} \wedge \cdots \wedge \mathcal{T}_n^{\mathcal{L}^{DS}} \cong (A_i = \{V_i^j\})_1 \wedge \cdots \wedge (A_i = \{V_i^j\})_n$$

Note that the general conjunctive form of $\mathcal{D}$ is preserved and coverage still can be given by $\mathcal{C} = \{o^i \in \Omega | \mathcal{D}(\mathcal{T}_1^{\mathcal{L}^{DS}} \wedge \cdots \wedge \mathcal{T}_n^{\mathcal{L}^{DS}}) = 1\}$, where each term $\mathcal{T}^{\mathcal{L}^{DS}}$ is a restriction over a different attribute domain that takes any $A_i \rightarrow \{0, 1\}$. However, instead of imposing restriction to a single attribute's value as $\mathcal{T}^\mathcal{L} : (A_i = V_i^j) \cong \mathcal{I}_{ij}$, we adopt a term structure that imposes one same restriction to one or more values in an attribute's domain, i.e. $\mathcal{T}^{\mathcal{L}^{DS}} : (A_i = V_i^j) \vee \cdots \vee (A_i = V_i^z) \cong \mathcal{I}_{ij} \vee \cdots \vee \mathcal{I}_{iz}$. Note, also, that each attribute is still only represented once in a description, i.e. each term in a description represents a different feature.

Finally, we have that $\mathcal{L}^{DS}$ allows generalisation not only through the removal of terms $\mathcal{T}$ from $\mathcal{D}$, but also by enlarging the extent of an attribute's representation in $\mathcal{T}$. This last added functionality, provides not only more flexibility in describing more general subgroups, but also allows generalisation operations to take place between subgroups – e.g *merge* and *root* operations that will be properly approached in the description of the subgroup selection method in the next subsection. In this sense, from the example dataset provided in Section 2.2.2, given the three descriptions bellow,

$$\mathcal{D}_{\mathcal{L}^{DS}} : location = I \; AND \; size = \{medium, large\}$$

$$\mathcal{D}_\mathcal{L}^A : location = I \; AND \; size = medium$$

$$\mathcal{D}_\mathcal{L}^B : location = I \; AND \; size = large$$

the description $\mathcal{D}_{\mathcal{L}^{DS}}$ not only is more general than both descriptions in $\mathcal{L}$ as also encompasses both descriptions in a single one. In other words, our proposed description language $\mathcal{L}^{DS}$ provides generalisation power to the subgroups' descriptions while aggregating simplicity to the final set of discovered subgroups.

4.2 Subgroup Search Strategy

As revised in Section 2.2.3, the state-of-the-art heuristic approaches to EMM task rely on the greedy beam-search algorithm, and despite the good results achieved by evolutionary approaches to SD, no bio-inspired optimisation meta-heuristic has been explored in EMM applications.

Bio-inspired (or nature-inspired) [42] is the name given to the family of optimization algorithms which design mimics a natural phenomenon – e.g. biological, physical, etc. – in order to solve optimization problems [40]. Along the large area of Evolutionary Computing, the other major category of bio-inspired algorithms is the Swarm Intelligence (SI). In contrast to imitate the evolution process of an individual, SI based techniques focus on the interaction of several individuals and their environment, exploiting social and collective behavior present in groups of animals. The present work explores the use of the swarm-based Ant-Colony Optimization (ACO) algorithm.

In the remainder of this section, we briefly present a theoretical review on ACO meta-heuristic, and then present the EsmamDS algorithm – an EMM framework that employs ACO as search heuristic.

Ant-Colony Optimisation (ACO). The ACO meta-heuristics was first introduced [26,27] as an approach to stochastic combinatorial optimization and has been widely used to solve hard optimization problems throughout the years. This optimization approach is based on the foraging behavior of some ant species and on the fact that such ants are able to 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 literature, e.g. [29,25,28,24, 23].

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 which can be followed by other ants, that tend to follow paths where pheromone concentration is higher. This characteristic of exploiting pheromone trails give some ant species the capability to discover the shortest path leading to the food. In an experiment to assess such behaviour, the authors in [47] perform an experiment where ants can freely move throughout two different paths guiding from the nest to a food source, one shorter than the other. What was observed is that, at an initial moment (when there is no pheromone on the paths), the ants choose randomly between the two options, i.e. the two paths can be chosen with equal probability (except for stochastic oscillations). But then, the ants following the shorter path arrive first at the food source and, as consequence, in the way back to nest, the higher level of pheromone in that path biases the ants' decision in its favor. As a result, pheromone accumulates faster on the shorter path and,

eventually, it becomes the preferable route to all ants. The studied showed that, by interacting with each other via pheromone trail, the colony of ants is capable of selecting the shortest route between nest and food with great reliability.

Analogously to real ants, ACO algorithms implement artificial ants that build solutions and exchange information on the quality of these solutions employing communication based on pheromone trail. If we understand such solutions as subgroup descriptions $\mathcal{D} \rightarrow \mathcal{I}_{ij}$, than we have the descriptive space $\mathbb{I}^{\Omega}$ as our (complete) solution search space. Moreover, we have that each item $\mathcal{I}_{ij} \in \mathbb{I}^{\Omega}$ is a possible *solution component* to be incorporated in a solution. Therefore, the artificial ant colony constitute a iterative procedure that stochastically constructs solutions given a probability distribution associated to the *solution components*. Each artificial ant constructs a complete solution by iteratively sorting solution components considering a probabilistic distribution that entails: (1) artificial pheromone trails, and (2) a 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 the most promising solutions 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 [110]). Moreover, the use of a colony of ants increases the algorithm robustness and, in many cases, this collective interaction of a population of agents is what enables the algorithm to efficiently solve the problem.

The EsmamDS algorithm builds on the framework of its predecessor, Esmam [83], which is an adaptation of the Ant-Miner [92] – a well-known ACO classification rule induction algorithm – to provide a set of discovered subgroups (descriptions) whose coverage present exceptional KM models. The Exceptional Survival Model Ant-Miner framework is provided in Algorithm 1.

The algorithm is initialised with an empty set of subgroups and with a set of uncovered observations comprising all cases in the dataset. Following a covering-based approach, it iteratively (lines 9-27) builds a entire new colony of ants where each ant constructs a complete subgroup's description $\mathcal{D}^{t}$. For each best description $\mathcal{D}^{best}$ (according to ϕ_{EMM}, given a baseline $\mathcal{B}$) constructed by a colony, its inclusion in the final subgroup set $\mathbb{S}$ is evaluated according to a subgroup selection method (line 21). For each new subgroup $\mathcal{S}$ inserted in $\mathbb{S}$, the observations it covers (i.e. $\mathcal{C}^{\mathcal{S}}$) are removed from the set of uncovered observations $\mathbb{U}$ (line 23). This process stops when the set of discovered subgroups $\mathbb{S}$ covers all observations in Ω at least once or when the algorithm achieves a stagnation maximum threshold ($ITS2STAGNATION$) – i.e. achieves a number of consecutive iterations with no change in $\mathbb{U}$.

In each iteration, we start from an empty description (and coverage $\mathcal{C}^{\mathcal{P}}$) (line 6), an empty best subgroup with minimum quality (line 7), and an initial configuration of the probabilistic distribution of the solution components. As stated previously, the ants will construct descriptions by traversing the descriptive space sorting items $\mathcal{I}_{ij}$ given a probability distribution $P(\tau(\mathcal{I}), \eta(\mathcal{I})) : \mathcal{I}_{ij} \rightarrow (0, 1)$, where τ and η are the pheromone and heuristic values associated to a item, respectively, and we will approach their definition latter. However, previously to the execution of each ant colony, the pheromone associated to all items $\mathcal{I}$ is equally initiated (with the value of $\frac{1}{|\mathbb{I}^{\Omega}|}$) and an initial heuristic value is set to each $\mathcal{I} \in \mathbb{I}^{\Omega}$ (line 8). Then,

Algorithm 1: Exceptional Survival Model Ant-Miner Framework

Input: The subgroup's baseline comparison $\mathcal{B}$, the level of significance α; the ACO hyper-parameters $ITS2STAGNATION, N\_ANTS, N\_CONVERG$ and MIN_COV; and the *Diverse Search* hyper-parameters $\mathcal{P}^{DS}$
Output: $\mathbb{S}$: set of exceptional subgroups $\mathcal{S}$
Data: A survival dataset Ω

1 $\mathbb{S} \leftarrow \varnothing$ `// the initially empty set of subgroups`
2 $\mathbb{U}, \mathbb{U}^{-} \leftarrow \Omega$ `// the sets of (previously) uncovered observations`
3 $stag = 0$
4 **while** $\mathbb{U} \neq \varnothing$ ***and*** $stag \leq ITS2STAGNATION$ **do**
5 $ant\_index, converg\_index = 1$
6 $\mathcal{D}^{t-1} \leftarrow \varnothing$
7 $\mathcal{S}^{best} \leftarrow \varnothing$
8 `InitialiseTrails`$(\mathbb{I}^{\Omega}, \Omega, \mathbb{U}, \mathbb{S}, \mathcal{P}^{DS})$
9 **while** $ant\_index < N\_ANTS$ ***or*** $converg\_index < N\_CONVERG$ **do**
10 $\mathcal{D}^{t} \leftarrow$ `ConstructDescription`$(\mathcal{D}^{\varnothing}, \mathbb{I}^{\Omega}, MIN\_COV)$
11 $\mathcal{D}^{t} \leftarrow$ `PruneDescription`$(\phi_{EMM}(\mathcal{D}|\mathcal{B}), \mathcal{D}^{t})$
12 `PheromoneUpdating`$(\mathcal{D}^{t}, \mathbb{I}^{\Omega})$
13 **if** $\mathbb{I}^{\mathcal{D}^{t}} == \mathbb{I}^{\mathcal{D}^{t-1}}$ **then**
14 $converg\_index += 1$
15 **else**
16 $converg\_index = 1$
17 **if** $\phi_{EMM}(\mathcal{D}^{t}) > \phi_{EMM}(\mathcal{S}^{best})$ **then**
18 $\mathcal{S}^{best} \leftarrow \mathcal{D}^{t}, \phi_{EMM}(\mathcal{D}^{t})$
19 $\mathcal{D}^{t-1} \leftarrow \mathcal{D}^{t}$
20 $ant\_index += 1$
21 **if** *CanAddSubgroup*$(\mathcal{S}^{best}, \mathbb{S}, \alpha)$ **then**
22 $\mathbb{S} \leftarrow \mathbb{S} \cup \{\mathcal{S}^{best}\}$
23 $\mathbb{U} \leftarrow \mathbb{U} \setminus \mathcal{C}^{\mathcal{S}^{best}}$
24 **if** $\mathbb{U}^{-} - \mathbb{U} == 0$ **then**
25 $stag += 1$
26 **else**
27 $stag = 0$
28 **return:** $\mathbb{S}$

in the colony loop (lines 9-20), each ant constructs (line 10) and prunes (line 11) a candidate description $\mathcal{D}^{t}$, updates the pheromone associated to each $\mathcal{I} \in \mathbb{I}^{\Omega}$ (line 12) and checks the convergence of the colony (lines 13-20).

The descriptions $\mathcal{D}^{t}$ are constructed (line 10) in a general-to-specific approach by iteratively adding items to a initially empty description $\mathcal{D}^{\varnothing}$ until there is no more items to be added or until an item addition results in a rule coverage bellow a threshold (MIN_COV). Its is important to highlight that the description construction follows the format of descriptive language $\mathcal{L}$, and therefore, only one item per attribute is allowed to be added. The disjuctive form of the description terms $\mathcal{T}$ proposed in $\mathcal{L}^{DS}$ is only incorporated to the solutions in the subgroup selection method (line 21) – through operations between descriptions that will be introduced latter in this section. In EsmamDS algorithm, the items to be added to a description are *sorted* from the set of items that belong to the observations it covers, i.e. from $\mathbb{I}^{\mathcal{C}^{\mathcal{D}}}$ (note that the set of possible items for the initially empty

description is $\mathbb{I}^{\Omega}$). This configuration differs from the implementation of Esmam, that always sorts the items from the complete set of items $\mathbb{I}^{\Omega}$. The quality associated to a description (considering a baseline $\mathcal{B}$ for subgroup comparison) is given by $\phi_{EMM}(\mathcal{D}|\mathcal{B})$ (or $\phi_{EMM}(\mathcal{S}|\mathcal{B})$ since $\mathcal{S}$ entails $\mathcal{D}$).

Then, a pruning procedure (line 11) iteratively removes items from the description, each time eliminating the item that leads to the largest improvement in the quality associated to the (pruned) description. The pruning stops when no items can be removed without decreasing the quality, or when the description already encompasses only one item.

This process is repeated for all N_ANTS ants in the colony or until the ants converge to a (best) solution. For the convergence of the colony (lines 13-20), each generated (pruned) description is compared to the solution of the (immediately) previously ant, and $N\_CONVERG$ gives a maximum threshold for identical sequential descriptions (so convergence is achieved).

The pheromone updating process in line 12 is responsible for incrementing the amount of pheromone associated to the items encompassed on the (pruned) $\mathcal{D}^t$ (i.e. $\mathbb{I}^{\mathcal{D}^t}$) according to Equation 9,

$$\tau_{ij}(t+1) = \frac{1}{\sum\limits_{\forall \mathcal{I}_{ij} \in \mathbb{I}^{\Omega}} \tau_{ij}(t+1)} \cdot \begin{cases} \tau_{ij}(t) + \phi_{EMM}(\mathcal{D}^t|\mathcal{B}) \cdot \tau_{ij}(t), \ iif \ \mathcal{I}_{ij} \in \mathbb{I}^{\mathcal{D}^t} \\ \tau_{ij}(t), \ iif \ \mathcal{I}_{ij} \notin \mathbb{I}^{\mathcal{D}^t} \end{cases} \tag{9}$$

where $\tau_{ij}(t+1)$ is the (updated) amount of pheromone associated to $\mathcal{I}_{ij}$ in the next ant iteration. For the items not encompassed on $\mathcal{D}^t$, the evaporation process is simulated by the normalisation of τ values in $(t+1)$.

The stochasticity of ACO algorithms takes place in the "biased-random" choice each ant makes for constructing their path, in other words, in *sorting* items to be added to solutions. This sorting procedure – i.e. the probabilistic choice of an item to be added to the current partial description – depends on both the pheromone (τ_{ij}) and an heuristic (η_{ij}) values associated with each item $\mathcal{I}_{ij}$. Equation 10 gives the probability rule associated to the items $\mathcal{I}_{ij}$. Note that the probability only exists for the set of items obtained from the coverage of the current description in construction (i.e. $\mathbb{I}^{\mathcal{C}^{\mathcal{D}^t}}$) and, therefore, it is computed previously to each *sorting*.

$$P_{ij} = \frac{\eta_{ij} \cdot \tau_{ij}(t)}{\sum\limits_{\forall \mathcal{I}_{ij}} \eta_{ij} \cdot \tau_{ij}(t)}, \ \forall \ \mathcal{I}_{ij} \in \mathbb{I}^{\mathcal{C}^{\mathcal{D}^t}} \tag{10}$$

In the Esmam algorithm, we employed a static heuristic function η_{ij}^{H} based on Shannon's entropy (Equation 11). We considered an initial partition of the observations as those with survival time at least as long as the cohort's 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. The proposed heuristic function is given in Equation 12.

$$H(W|\mathcal{I}_{ij}) = -\sum_{w=1}^{k} P(w|\mathcal{I}_{ij}) \cdot \log_2 P(w|\mathcal{I}_{ij}) \tag{11}$$

$$\eta_{ij}^{H} = \frac{\log_2 k - H(W|\mathcal{I}_{ij})}{\sum\limits_{\forall \mathcal{I}_{ij} \in \mathbb{I}^{\Omega}} \log_2 k - H(W|\mathcal{I}_{ij})} \tag{12}$$

Note that because Esmam employs a static heuristic, it may be computed only once in the begining of the algorithm. Therefore, the initialisation of the probabilistic search space in line 8 is reduced to the (equally) initialisation of the pheromone amounts associated to each item in $\mathbb{I}^{\Omega}$. As we will introduce next in this section, on EsmamDS, we propose an dynamic heuristic function that considers the best colony's solution ($\mathcal{S}^{best}$) to bias the initial probabilistic search space aiming to increase the exploration power of the ACO search. Therefore, for EsmamDS, the trails initialisation in line 8 perform both the pheromone and the heuristic initialisations and, for the latter, additional parameters $\mathcal{P}^{DS}$ (to be introduced) are required.

Finally, besides heuristic function, the other fundamental change incorporated by EsmamDS lies in the subgroup selection method (line 21). The Esmam algorithm only constraints to the inclusion of a subgroup in the final set $\mathbb{S}$ are: a lower bound to the quality of the new best subgroup ($\phi_{EMM}(\mathcal{S}^{best}) \geq 1 - \alpha$) and the subgroup's description $\mathcal{D}^{\mathcal{S}^{best}}$ to be different from every subgroup $\mathcal{S}^{i} \in \mathbb{S}$. In the EsmamDS algorithm, we propose a selection method that performs operations between descriptions aiming at improving the generalisation of the final subgroups and provide diversity in the final set considering all three dimensions of redundancy.

4.3 The Diverse Search Approach

In this section we present the new heuristic function employed in the ACO search and the new method for selecting subgroups for the final set.

EsmamDS Heuristic Function. As the problem of redundancy leads to many (almost) similar descriptions representing (almost) similar coverages, we presume that improving the search (for exceptional models) exploration towards less visited spaces of both the descriptive ($\mathbb{I}^{\Omega}$) and the coverage ($\mathcal{P} = \{o^i \in \Omega\}$) spaces potentially increases diversity in the final set of subgroups. We, therefore, resort to the search mechanisms of ACO meta-heuristics. As at the beginning of each ant colony all items $\mathcal{I}_{ij}$ have equal pheromone values, one can understand that the heuristic value η_{ij} associated to $\mathcal{I}_{ij}$ comprises the *a priori* probability. In other words, it is the heuristic function that define the initial probability distribution and allocates the colony of ants in (strategic) initial positions in the search space.

In constrast to the static heuristic proposed in [83], we propose a dynamic heuristic function that uses the best solution of each ant colony to penalise the solution components $\mathcal{I}$ that were already visited and to penalise the observations already covered by $\mathbb{S}$. Therefore, although each new colony stochastically constructs their on paths through the pheromone trails they build, each new colony is also (initially) biased towards less explored regions of the solution space. We, therefore, propose a new heuristic function that incorporates not only the information-based heuristic, but also a descriptive logistic attenuation and a weighted covering approach.

The new proposed heuristic value η_{ij}^{DS} associated to an item $\mathcal{I}_{ij}$ is presented in Equation 13, and it comprises three components: $\eta_{ij}^{\Delta H}$ is the dynamic (Δ) entropy-based heuristic component; η_{ij}^{L}, the description attenuation component; and η_{ij}^{W} the weighted covering component.

$$\eta_{ij}^{DS} = \frac{\eta_{ij}^{\Delta H} \cdot \eta_{ij}^{L} \cdot \eta_{ij}^{W}}{\sum\limits_{\forall \mathcal{I}_{ij} \in \mathbb{I}^{\Omega}} \eta_{ij}^{\Delta H} \cdot \eta_{ij}^{L} \cdot \eta_{ij}^{W}} \tag{13}$$

The dynamic entropy-based heuristic component $\eta_{ij}^{\Delta H}$ is defined as presented in Equation 12, with the difference that it is computed over the set of remaning uncovered observations $\mathbb{U}$ (instead of considering the entire dataset Ω – as is the case of its static version η_{ij}^{H}). This information-based component is (still) the heuristic quantification of the items quality. The other two newly incorporated components are meant to attenuate the quality of the items already visited in early solutions.

The description attenuation component η_{ij}^{L} is based on logistic function and strives for diversity purely on the basis of the (subgroup) descriptions already discovered by early ant colonies. Equation 14 presents the proposed component, where $c(\mathcal{I}_{ij})$ is the counting of the item $\mathcal{I}_{ij}$ presence in the best subgroups $\mathcal{S}^{best}$ found by the colonies, and L_{TH} is the x-value of the logistic sigmoid midpoint. In other words, L_{TH} define the value threshold for $c(\mathcal{I}_{ij})$ so $\eta_{ij}^{L} = 0.5$, i.e. defines the usage of $\mathcal{I}_{ij}$ so its probability decreases in a half. L_{TH} is defined as a hyper-parameter of EsmamDS diverse search ($L_{TH} \in \mathcal{P}^{DS}$). Therefore, we have that the more an item appears in the best solutions of the ACO search, the smaller becomes its (*a priori*) probability of being explored by the next colonies.

$$\eta_{ij}^{L} = 1 - \frac{1}{1 + e^{-(c(\mathcal{I}_{ij}) - L_{TH})}} \tag{14}$$

Finally, with the η_{ij}^{W} component we strive to improve coverage diversity by guiding the search towards the the observations less encompassed in the final subgroups' set $\mathbb{S}$. We make use of a score based on multiplicative weighted covering [66] proposed by [114] in their Cover Based Subgroup Selection (DSSD-CBSS) approach – which is also evaluated in this work. This score is employed to weigh the heuristic value associated to the items $\mathcal{I}_{ij}$ considering the frequency each observation covered by such item is represented in the set $\mathbb{S}$ of discovered subgroups. The score is presented in Equation 15, where $|\mathcal{C}^{\mathcal{I}_{ij}}|$ is the size of term's $\mathcal{I}_{ij}$ coverage (i.e. the number of observations $o^i \in \Omega$ the item covers), $c(o^i)$ is the counting of how many times the observation o^i is covered by any subgroup $\mathcal{S} \in \mathbb{S}$, and $W_{TH} \in (0, 1]$ is the weight parameter – and along with L_{TH} comprise the set of EsmamDS diverse search hyper-parameters ($\mathcal{P}^{DS} = \{L_{TH}, W_{TH}\}$). Therefore, the more the observations covered by an item $\mathcal{I}_{ij}$ is encompassed by the final set of subgroups, lesser is its probability of being explored in future iterations of the algorithm.

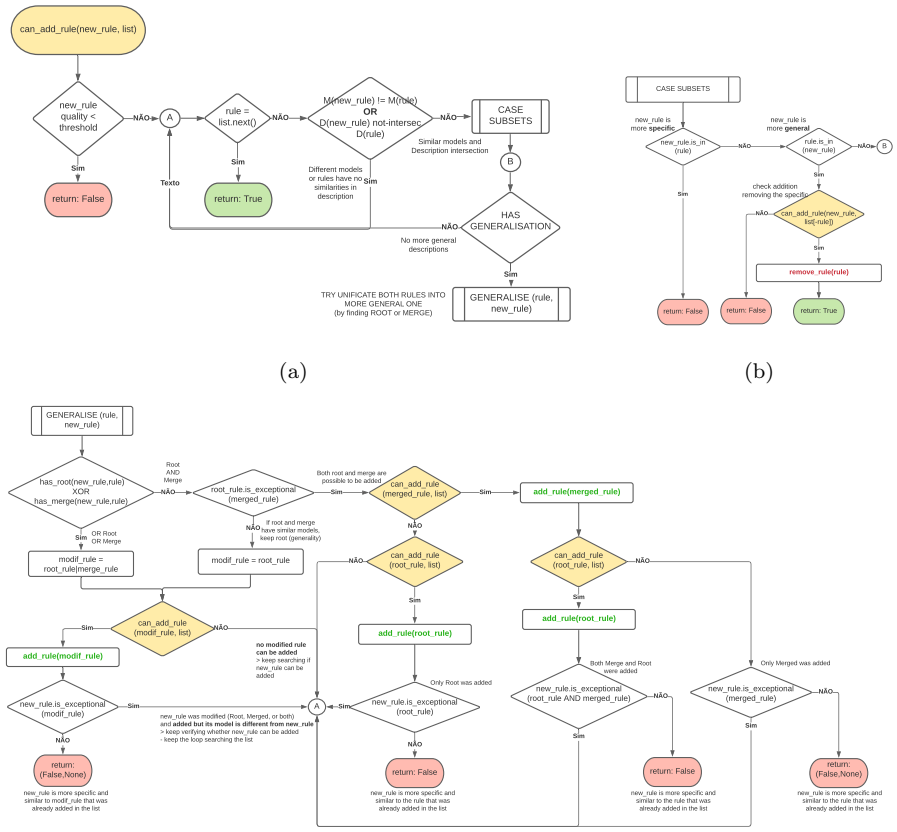
$$\eta_{ij}^{W} = \frac{1}{|\mathcal{C}^{\mathcal{I}_{ij}}|} \sum_{\forall o^i \in \mathcal{C}^{\mathcal{I}_{ij}}} {W_{TH}}^{c(o^i)} \tag{15}$$

It is important to notice that η_{ij}^{L} takes into consideration all best subgroups $\mathcal{S}^{best}$ discovered by each ant colony, while η_{ij}^{W} considers only the ones that are

actually selected to be added to the final set (the selection method is introduced next).

EsmamDS Subgroup Selection. For selecting the subgroups to compose the final set, the EsmamDS algorithm implements an extended (and more complex) subgroup selection method (Algorithm 1, line 21) than its predecessor Esmam. In addition to impose a lower quality bound ($\phi_{EMM}(\mathcal{S}) \geq 1 - \alpha$) constraint and assure that will be no duplicates in the descriptions $\mathcal{S}^{\mathcal{D}} \in \mathbb{S}$, EsmamDS selection method also handles redundant patterns while assuring generality. The complete fluxogram of the implemented function in line 21 of Algorithm 1 is presented in Figure 1.

Figure 1a presents the overall process of the selection method. The new candidate subgroups are immediately added to $\mathbb{S}$ when they satisfy the quality constraint **and** if: (i) its model differs from all models $\mathcal{S}^{\mathcal{M}} \in \mathbb{S}$; **or** (ii) its description *completely* differs from all descriptions $\mathcal{S}^{\mathcal{D}} \in \mathbb{S}$. In other words, a candidate subgroup is compared to each subgroup $\mathcal{S}^{i} \in \mathbb{S}$ and it is immediately accepted if it comprises



(a) (b)

(c)

Fig. 1: CanAddSubgroup fluxogram

a completely new description and (exceptional) model. The exception for this default acceptance is the cases in which the candidate subgroup somehow resembles another already in the final set with relation to their model and description. In other words, we handle the cases where the pair $(\mathcal{S}^{new}, \mathcal{S}^i)$ (for $\mathcal{S}^i \in \mathbb{S}$) present (statistically) similar models and $\mathbb{I}^{\mathcal{S}^{new}} \cap \mathbb{I}^{\mathcal{S}^i}$ exists.

We first assess refinement cases, i.e. the cases where the subgroups in a pair $(\mathcal{S}^{new}, \mathcal{S}^i)$ are one subset of the other's description. Figure 1b presents the fluxogram of the procedure to handle refinements keeping generality in $\mathbb{S}$. In the cases where refinements do not apply (the subgroups are not subsets from one another's), we strive for generalisation through two different procedures: *root* and *merge*.

As stated previously, we understand that subgroup generalisation my be achieve through two different procedures: (i) by removing items from a description, or (ii) by incorporating items to a description through disjunctive conditions.

The *root* generalisation relies on the first procedure, and aims at finding a common more general description between two descriptions. Therefore, given $\mathbb{I}^{\mathcal{S}^{best}}$ and $\mathbb{I}^{\mathcal{S}^i}$ the sets of items encompassed by $(\mathcal{S}^{best}, \mathcal{S}^i)$ descriptions, we say that there is a root subgroup $\mathcal{S}^{root}(\mathcal{S}^{best}, \mathcal{S}^i)$ such that $\mathbb{I}^{\mathcal{S}^{root}} = \mathbb{I}^{\mathcal{S}^{best}} \cap \mathbb{I}^{\mathcal{S}^i}$ iif $\exists \mathbb{I}^{\mathcal{S}^{best}} \cap \mathbb{I}^{\mathcal{S}^i}$.

The *merge* generalisation entails the disjunctive conditions relies and comprises the union of two descriptions into one. We define the merged subgroup $\mathcal{S}^{merge}(\mathcal{S}^{best}, \mathcal{S}^i)$ over the set of items $\mathbb{I}^{\mathcal{S}^{merge}} = \mathbb{I}^{\mathcal{S}^{best}} \cup \mathbb{I}^{\mathcal{S}^i}$ if and only if the set of descriptive attributes $A_i \in \mathbb{A}$ encompassed by both $(\mathcal{S}^{best}, \mathcal{S}^i)$ are the same. Note that this *merge* operation is the procedure responsible for generating (and introducing to results) descriptions in $\mathcal{L}^{DS}$ language.

Hence, Figure 1c presents the fluxogram for handling the cases there do not comprise refinements. Basically, generalisations through *root* and *merge* are constructed and tested for inclusion in $\mathbb{S}$. The algorithm first tests if both root and merge operations are possible. If only one is, it is performed, and the generalisation is included in $\mathbb{S}$ (if it can). If the generalisation's model differs from $\mathcal{S}^{best}$ model, then the testing procedure to $\mathcal{S}^{best}$ inclusion in $\mathbb{S}$ follows to the next subgroup $\mathcal{S}^i \in \mathbb{S}$. If both root and merge are possible, we assess the model difference between them: if they are similar, only the root is tested to be added; otherwise, both are. Note that even when a more general subgroup (root or merge) is added to $\mathbb{S}$, if its model differs from the $\mathcal{S}^{best}$, the selection method to analyse the latter follows to the next subgroup in $\mathbb{S}$. If no generalisation exists (no root neither merge) between the two subgroups with similar models (and resemblance in description), the checking follows to the next $\mathcal{S}^i \in \mathbb{S}$.

4.4 EsmamDS: Parameters and Outputs

Finally, we overview the (input) parameters of EsmamDS algorithm and present the outputs it provides.

EsmamDS Parameters. The EsmamDS algorithms has following parameters:

1. $\mathcal{B}$: a baseline for subgroup exceptionality comparison (population or subgroup complement);
2. α: the level of significance for the logrank statistical test (and lower bound of subgroup quality);

3. $\mathcal{P}^{DS} = \{L_{TH}, W_{TH}\}$: the hyper-parameters of the EsmamDS *diverse search*, where L_{TH} is the x-value of the logistic sigmoid midpoint (for descriptive exploration) and W_{TH} is the weighted covering parameter (for covering exploration);
4. $\mathcal{P}^{ACO}$ – the hyper-parameters of the ACO meta-heuristic: (1) *ITS2STAGNATION* defines the maximum number of iterations without change in the dataset coverage so algorithm execution is terminated; (2) *N_ANTS* defines the number of ants (iterations) in the colony; (3) *N_CONVERG* defines the number of sequential identical constructed descriptions $\mathcal{D}^t$ in order to consider that the ants have converged to a single (best) solution (subgroup description); (4) *MIN_COV* is the minimum subgroup coverage (given in percentage of the dataset).

Although the algorithm presents eight configurable parameters, the only required configuration is the baseline $\mathcal{B}$ for EMM framework. The remainder (hyper-)parameters are related to the ACO meta-heuristic (the $\mathcal{P}^{DS}$ included). Although the proper configuration of such parameters is essential to the performance of the search, the advantage of employing a robust optimisation meta-heuristic is that a well tunned configuration of the algorithm is capable of assuring good results despite specifications of the problem in hand (although their consideration in the algorithm setting refines the performance).

Lastly, the EsmamDS algorithm provides three output files:

1. *subgroupSet.txt*: provides the discovered subgroup descriptions grouped by model equivalence and description similarity (groups of subgroups with similar models or some resemblance in description); also provide quantitative metrics such subgroup support and similarity measures.
2. *subgroupModel.csv*: provides the set of discovered subgroups ordered by quality; also provides some individual statistics of the subgroups.
3. *survivalModels.csv*: provides a table with the KM estimates of the population and each discovered subgroup.

5 Experimental Procedure

We conducted experiments to evaluate our proposed approach for mining exceptional survival behaviours. We aim at providing comprehensible and straightforward rule-based models that comprise characterisation of diverse (non-redundant) subgroups with unusual survival response. We assess the results with respect to the interpretability, generality and redundancy of the subgroups. We evaluate our algorithm against the state-of-the-art EMM and SD approaches - *beam-search* and *DSSD-CBSS* algorithm [114] (both presented in Section 2.2.3) - and the survival predictive covering algorithm *LR-Rules* [121] (reviewed in Section 3). The remainder of this section is as follows. In subsection 5.1, we present the experimental procedure setup: the datasets used in the experiments and the configuration of the algorithms' parameters, and in subsection 5.2 we report the results achieved.

5.1 Experimental Setup

To assess our approach's performance, we performed experiments on 14 (fourteen) real-world survival datasets obtained from online data repositories. Both the Es-

mamDS algorithm and its predecessor version, Esmam [83], were implemented in Python3, and they are publicly available on our website[2], along with all data and results of the empirical evaluation here presented.

Datasets Characterisation. Because we aim our efforts at providing better patient (individual) prognostic characterisation, we restrict the dataset scope to the medical domain. The datasets we investigated are presented in Table 1 and many of them are well investigated throughout the literature of survival methods, e.g. the German Breast Cancer Study Group 2 (gbsg2) and the Veteran's Administration Lung Cancer Trial (veteran) datasets. The online data sources, references and general data description of the datasets in Table 1 can be found in the Esmam [83] data repository.

We performed an initial processing of the data by removing observations containing missing values. A feature selection was employed on the datasets presented in Table 2. Once the EsmamDS algorithm only copes with categorical attributes, we discretised all numerical features using equal-frequency discretisation into five interval categories.

[2] ESMAM website: https://github.com/jbmattos/ESM-AM_bracis2020

Table 1: Characteristics of 14 datasets used in the experimental study: the number of observations ($|\Omega|$), the number of descriptive attributes ($|\mathbb{A}|$), the number items $\mathcal{I} = (A_i, V_i^j)$ in the dataset ($|\mathbb{I}^{\Omega}|$), the number of discretized attributes ($|\mathbb{A}^{disc}|$), the percentage of censored observations (%cens), the subject of research and the survival event description (Event)

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

Table 2: Selected descriptive attributes of the data sets with feature selection

Dataset	Set of Descriptive Attributes ($\mathbb{A}$)
actg320	tx, txgrp, strat2, sex, raceth, ivdrug, hemophil, karnof, cd4, priorzdv, age
mgus	age, sex, dxyr, pcdx, alb, creat, hgb, mspike
ptc	risk_group, histological_type, age, sex, path_t_stage, path_n_stage, path_m_stage, tumor_status, exome, extrathyroidal_extension, mrna_cluster, mirna_cluster, arm_scna_cluster, methylation_cluster, disease_stage, primary_exome, lowpass, wgs_status
whas500	age, gender, hr, sysbp, diasbp, bmi, cvd, afb, sho, chf, av3, miord, mitype, los

Table 3: Description of the algorithms compared in this work: the algorithms that search subgroups that are exceptional with relation to the population (*-pop* names); and the algorithms that provide subgroups that are exceptional with relation to their complement (*-cpm* names).

Algorithm	Search Strategy	DM Task	Target Concept	Quality Measure	Description Language
Baseline: Population					
EsmamDS-pop	ACO	EMM	KM model	ϕ_{EMM}	$\mathcal{L}^{DS}$
Esmam-pop	ACO	EMM	KM model	ϕ_{EMM}	$\mathcal{L}$
BS-EMM-pop	Beam Search	EMM	KM model	ϕ_{EMM}	$\mathcal{L}$
BS-SD-pop	Beam Search	SD	Survival time	ϕ_{SD}	$\mathcal{L}$
DSSD-CBSS	Cover Based Subgroup Selection	-	Survival time	t-Test	$\mathcal{L}$
Baseline: Complement					
EsmamDS-cpm	ACO	EMM	KM model	ϕ_{EMM}	$\mathcal{L}^{DS}$
Esmam-cpm	ACO	EMM	KM model	ϕ_{EMM}	$\mathcal{L}$
BS-EMM-cpm	Beam Search	EMM	KM model	ϕ_{EMM}	$\mathcal{L}$
BS-SD-cpm	Beam Search	SD	Survival time	ϕ_{SD}	$\mathcal{L}$
LR-Rules	Sequential covering	-	KM model	ϕ_{EMM}	$\mathcal{L}$

We executed the experiments considering both baselines for subgroup comparison - *population* and *complement* - and evaluated them separetly. For the *population*-baseline experiments, we compare the EsmamDS results with its early version Esmam, the beam-search algorithm (for both EMM and SD tasks) and the Diverse Subgroup Set Discovery variant - Cover Based Subgroup Selection (both heuristic subgroup search methods revised in Section 2.2.3). For the *complement*-baseline experiments, we also compare the EsmamDS against Esmam and the beam-search EMM/SD approaches, plus with the covering algorithm LR-Rules (revised in Section 3). Table 3 enunciate all comparable algorithms, for both baselines. We compare the results firstly regarding the characteristics of the final sets of discovered subgroups (complexity and generality), and then we assess the results regarding the redundancy in the subgroup sets. Table 4 summarises the description and definition of all evaluation metrics (introduced bellow).

Evaluation metrics of Subgroup Set. In order to evaluate the characteristics of the final sets, we first assess the exceptionality of the discovered subgroups and, then, we employ usual evaluation quality measures of Subgroup Discovery [51]. We use two measures of complexity to assess the interpretability of the subgroups (i.e. the simplicity of the knowledge provided): (i) the number of discovered subgroups (***#sg***); and (ii) the average subgroup description length (***lenght***), i.e. the average of the number of terms $\mathcal{T}_{ij} \in \mathcal{D}^{\mathcal{S}}$ in each subgroup $\mathcal{S} \in \mathbb{S}$. Furthermore, we assess the generality of the discoveries through the average percentual subgroup coverage (***sgCov***) – i.e. the average of the number of observations $o^i \in \Omega$ covered by the subgroups $\mathcal{S} \in \mathbb{S}$. We also assess the generalisation capacity of the compared methods through the evaluation of the dataset percentual coverage (***dbCov***) – i.e. the percentual of the total number of observations that is covered by any subgroup $\mathcal{S} \in \mathbb{S}$.

Evaluation metrics of redundancy in Subgroup Set. In order to assess the redundancy in the final subgroups' sets (for all three dimensions of redundancy), we rely on the redundancy metrics defined in Section 2.2.2: the *description redundancy* $\rho_{\mathcal{D}}$, the

Table 4: Description of the evaluation metrics employed in this work

Metrics	Description	Definition
Evaluation metrics of Subgroup Sets		
$\#sg$	Number of discovered subgroups	$\lvert\mathbb{S}\rvert$
$lenght$	Average subgroup description length	$\frac{\sum_{\forall \mathcal{S}\in\mathbb{S}} \lvert\mathcal{D}^{\mathcal{S}}\rvert}{\lvert\mathbb{S}\rvert}$
$sgCov$	Average percentual subgroup coverage	$\frac{\sum_{\forall \mathcal{S}\in\mathbb{S}} \lvert\mathcal{C}^{\mathcal{S}}\rvert}{\lvert\mathbb{S}\rvert}$
$dbCov$	Dataset percentual coverage	$\frac{\sum_{\forall o^i\in\Omega} \mathcal{D}(A^i_1,\ldots,A^i_{\lvert A\rvert})}{\lvert\Omega\rvert}$
Evaluation metrics of redundancy in Subgroup Sets		
$\rho_{\mathcal{D}}$	Description redundancy	Equation 7
$\rho_{\mathcal{C}}$	Coverage redundancy	Equation 7
CR	Cover Redundancy	Equation 7
$\rho_{\mathcal{M}}$	Model redundancy	Equation 8

coverage redundancy $\rho_{\mathcal{C}}$ and the *model redundancy* $\rho_{\mathcal{M}}$ – all three metrics defined in Equation 7; and the *Cover Redundancy* (***CR***) defined in Equation 8.

Finally, we define the configuration of the compared algorithms parameters. We first defined the (hyper-)parameters of EsmamDS algorithm through a randomised search for the following configurations:

- $N\_ANTS = \{100, 200, 500, 1000, 3000\}$
- $MIN\_COV = \{0.01, 0.02, 0.05, 0.1\}$
- $N\_CONVERG = \{5, 10, 30\}$
- $ITS2STAGNATION = \{20, 30, 40, 50\}$
- $L_{TH} = \{1, 3, 5, 10\}$
- $W_{TH} = 0.9$ (according to employed in the DSSD-CBSS in [114])
- $\alpha = 0.05$

We sampled 10% of the total number of possible combinations (an amount of 96 parameters configuration samples) and we executed the EsmamDS algorithm for three datasets (namely: actg320, breast-cancer and ptc). We selected the best configuration concatenating priorisations over the evaluation metrics, for each baseline ($\mathcal{B}$) parameter – population and complement. Therefore, we have the following EsmamDS setups – parameter (value):

EsmamDS-pop setup: $\mathcal{B}$ (population), α (0.05), N_ANTS (100), MIN_COV (0.1), $N\_CONVERG$ (5), $ITS2STAGNATION$ (40), W_{TH} (0.9), L_{TH} (5).

EsmamDS-cpm setup: $\mathcal{B}$ (complement), α (0.05), N_ANTS (100), MIN_COV (0.05), $N\_CONVERG$ (5), $ITS2STAGNATION$ (40), W_{TH} (0.9), L_{TH} (10).

For the compared approaches, we set the (user-defined) parameters accordingly to the settings and results of EsmamDS over 30 executions for each dataset (for each execution we used a different high prime number for the random generator, all provided in the repository). The following defined parameters settings are given for each dataset, and a table with the computed values for each parameter is also available in the repository.

- For the beam-size (or maximum number of discovered subgroups), we computed the average number of EsmamDS **#sg** metric.
- For the rule-depth (refinement/search depth), we averaged the maximum **length** achieved by EsmamDS in each execution.
- For the minimum subgroup coverage (or subgroup size), we employed the EsmamDS MIN_COV parameter (rounded up) – accordingly to the baseline employed by each approach.

Hence, we provide the specifics of each compared approach setup bellow.

Exceptional Survival Model AntMiner (Esmam):

- *Esmam-pop setup:* $\mathcal{B}$ (population), α (0.05), N_ANTS (100), MIN_COV (0.1), $N\_CONVERG$ (5), $ITS2STAGNATION$ (40).
- *Esmam-cpm setup:* $\mathcal{B}$ (complement), α (0.05), N_ANTS (100), MIN_COV (0.05), $N\_CONVERG$ (5), $ITS2STAGNATION$ (40).

Beam-Search (BS): The beam-search (*BS-*) algorithms were executed in Python3 using the PySubgroup package[3]. We implemented a beam-search framework to run both EMM and SD tasks, for both $\mathcal{B} = population$ (*BS-EMM-pop*/*BS-SD-pop*) and $\mathcal{B} = complement$ (*BS-EMM-cpm*/*BS-SD-cpm*). We employed the quality measures defined in Eq. 3 (for EMM) and in Eq. 2 (for SD). The implemented framework receives as parameters the definition of the task ($task = \{EMM, SD\}$ and the definition of baseline $\mathcal{B} = \{population, complement\}$. In addition, the *beam size*, the *rule depth* and the *minimum rule-coverage* were also defined.

Diverse Subgroup Set Discovery - Cover Based Subgroup Selection (DSSD-CBSS): This variant of the DSSD was executed using the Cortana[4] JAR file and selecting *cover-based beam selection* as the search strategy type. We defined the *single numeric* target type over the survival times and we defined the *t-Test* as the quality measure. The *refinement depth*, the *minimum coverage* and *maximum number of subgroups* were also defined. We set the maximum execution time to infinity, and the remaining user-defined parameters were kept as default.

LR-Rules: We executed the JAR file available in the LR-Rules repository[5], in training mode, according to the instructions in the available manual.

Finally, to assess the algorithms' performance in the proposed metrics, we employed the Friedman statistical test and the Nemenyi post-hoc test - performed in R language, using the *scmamp1* library. We executed the tests using EsmamDS (and Esmam) complete sample of 420 results for each metric - 30 experiments for each of the 14 datasets. For the remaining analysed exact algorithms, we repeated 30 times the results for each dataset. We assessed the tests using a level of significance of $\alpha = 5\%$. We display the results of the post-hoc test through the Critical-Distance (CD) plot. The plot's main horizontal line is the line where is

[3] PySubgroup website: https://github.com/flemmerich/pysubgroup
[4] Cortana website: https://datamining.liacs.nl/cortana.html
[5] LR-Rules website: https://github.com/adaa-polsl/LR-Rules/releases

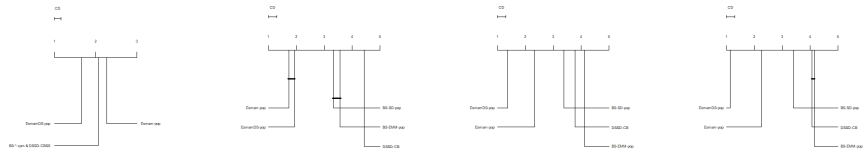
plotted the average ranks of the test. The lowest ranks are considered the best results. We consider both *sgCov* and *dbCov* maximisation metrics; for the others, we consider minimisation. The connected vertical lines are the groups of algorithms that can be considered statistically similar. The CD line on the up-left corner is the minimum distance between ranks to be considered the statistical difference. In the next section, we present and analyse the results achieved.

5.2 Results Analysis

Subgroups Exceptionality. We first assessed the exceptionality of the discovered subgroups regarding both the EMM and the SD tasks. In other words, for each subgroup in the final sets, we computed the logrank test and the t-Test between subgroups and baseline ($\mathcal{B}$). We could observe that the SD algorithms (the DSSD-CBSS included) do not assure the discovery of unusual survival models. In contrast, the EMM algorithms do not guarantee an unusual distribution of the survival time feature. This result is somehow expected since the different tasks search for exceptionalities of different natures. It also corroborates our premise that the EMM task is a more suitable approach when considering our goal in hand – which is to discover subgroups with unusual survival behaviour, i.e. subgroups with exceptional survival models.

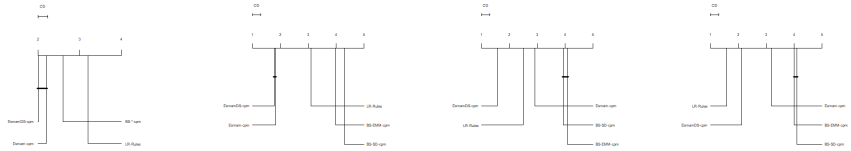
Subgroups Complexity and Generality. We evaluate the resultant sets of subgroups regarding their complexity and generality through the following metrics presented in Table 4: the number of discovered subgroups (*#sg*) and the average length of the descriptions (*length*) (for the first), and the average subgroup coverage (*sgCov*) and relative dataset coverage (*dbCov*) (for the latter). The (paired) Friedman test performed over 420 samples was rejected for all metrics, rejecting the null hypothesis that the compared algorithms present a similar performance. The Figures 2 (3) present the CD plot for *population* (*complement*) algorithms. For the *#sg* metric, we pinpoint that the *BS-* and *DSSD-CBSS* algorithms deliver a final set of predifined size - set as the EsmamDS average size. Therefore, although all approaches are represented in the graphical results, we evaluate the EsmamDS only against Esmam and LR-Rules.

We observe that our Diverse Search approach, the EsmamDS, returns more compact set of subgroups when compared to Esmam (although the *complement*-baseline presents no significant difference) and to the exhaustive LR-Rules algorithm (Figs 2a,3a). When analysing the descriptions' *length* (Figs 2b,3b), our approach returns shorter (simpler) patterns. When comparing specifically with its predecessor Esmam, no statistical difference is observed. However, one can remember that the EsmamDS employs a description language more capable of generalisation than the one employed by Esmam. Therefore, although both approaches present similar performances regarding the complexity of their descriptions, the analysis of the subgroups' coverage, *sgCov*, shows that the EsmamDS algorithms returns subgroups with larger coverage (Figs 2c,3c) – not only when compared to Esmam approach but also compared with the others. We observed that our EsmamDS approach returned subgroups presenting an (relative) average coverage of 28.0% (34.2%) for the baseline *population* (*complement*), comprising rules that neither cover the majority of the population nor very small subsets. When



(a) #sg (b) length (c) sgCov (d) dbCov

Fig. 2: Plot CD of Nemenyi post-hoc test over *population*-baseline algorithms for the following metrics: (a) the average ruleset size; (b) the average rule length; (c) the average rule coverage; (d) the average dataset coverage; and (e) the Cover Redundancy metric (CR)



(a) #sg (b) length (c) sgCov (d) dbCov

Fig. 3: Plot CD of Nemenyi post-hoc test over *complement*-baseline algorithms for the following metrics: (a) the average ruleset size; (b) the average rule length; (c) the average rule coverage; (d) the average dataset coverage; and (e) the Cover Redundancy metric (CR)

analysing the dataset coverage (*dbCov*), we observe that the EsmamDS present highest overall results - of 87.1% (97.8%) for the baseline *population* (*complement*) - being only behind of the exhaustive LR-Rules (complement baseline), with 99.8% of dataset coverage. When comparing to Esmam, it is possible to observe that the aggregation of search exploration resulted in an increase of over 20% of the dataset coverage and thus yields subgroups that better encompass all analysed cases. It is possible to observe that our approach outperforms the other approaches regarding the complexity (*#sg*/*length*) and generalisation (*sgCov*/*dbCov*) of the subgroups' sets, delivering simple models (i.e. small sets with small descriptions) that comprise relatively large subgroups.

Although simpler models are preferred for the sake of comprehension, it is also our goal to discover as many exceptional (and large) subgroups as possible. When analysing the set of subgroups, one can understand that smaller sets may not provide essential subgroups characterisation, while larger sets not necessarily yield a subgroup diversity (non-redundant subgroups). In the same sense, even though smaller descriptions are easier to interpret, a set of small-length subgroups not necessarily yields diverse characterisations; and a set of subgroups with high-coverage not necessarily encompass diversity in the representation of data observations. Therefore, the amount of diversity attained in the final models comprises the trade-off between simplicity and information. In the remainder of this section,

we assess diversity in the subgroup sets by analysing the three dimensions of redundancy *within* each approach's final subgroup set and *between* EsmamDS results against all other approaches.

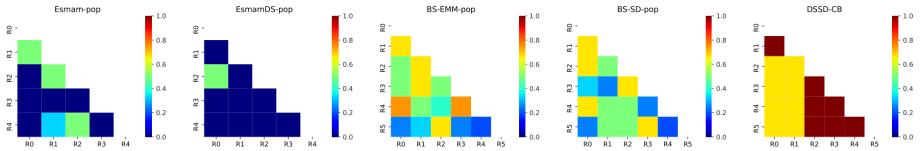
Redundancy within Subgroup Sets. We analyse diversity attained by each compared approach assessing the three dimensions of redundancy – descriptive, covering and model redundancies – *within* each resultant set of subgroups. For that matter, we first analyse the similarities between all combinations of subgroups within a set. Therefore, being $\mathbb{S} = \{\mathcal{S}^1, \ldots, \mathcal{S}^n\}$ a set of n subgroups $\mathcal{S}^i$, for each pair of subgroups $(\mathcal{S}^i, \mathcal{S}^j) \in \mathrm{C}_{\mathbb{S},2}$ we comput similarity regarding their description, coverage and model - $\varsigma_{\mathcal{D}}$ (Eq. 4), $\varsigma_{\mathcal{C}}$ (Eq. 5) and $\varsigma_{\mathcal{M}}$ (Eq. 6), respectively. We plotted such results in heatmaps, where the final subgroup set of each algorithm comprises a low-triangular matrix ($\mathbb{S} \times \mathbb{S}$) for each similarity measure. Note that for all measures, higher values represent higher redundancy.

Figure 4 shows the plots of *population*-baseline approaches on *whas500* dataset, and Figure 5 shows the plots of *complement*-baseline approaches on *breast-cancer* dataset. Note that the plots' granularity is related to the size ($\#sg$) of the sets. From the analysis of all 30 experiments over the 14 datasets, it is possible to observe that the EsmamDS algorithm consistently achieves lower similarity levels on rules' description and coverage while delivering a variety of exceptional survival models.

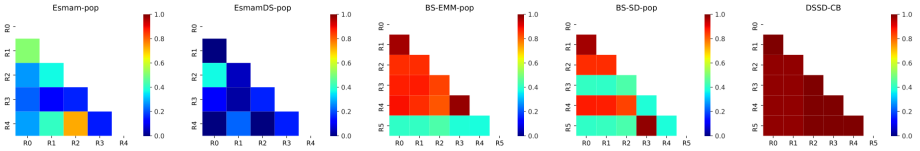
To evaluate the algorithms' performances regarding the diversity attained in their subgroups' sets, we employed the Friedman statistical test and the Nemenyi post-hoc test for the algorithms' performances over the $\rho_{\mathcal{D}}, \rho_{\mathcal{C}}$, $\rho_{\mathcal{M}}$ redundancy metrics (Eq. 7) and the CR metric (Eq. 8). For a level of significance of 5%, the Friedman test was rejected for all metrics, rejecting the assumption that all algorithms have similar performances. Figure 6 (7) present the CD plot of the redundancy metrics results for *population*(*complement*)-baseline algorithms. The plots show that the EsmamDS algorithm outperforms all other algorithms in attaining diversity on its subgroups sets – in all three dimensions of redundancy – being considered similar only with LR-Rules on coverage redundancy on both $\rho_{\mathcal{C}}$ and CR metrics.

It is possible to observe that the results of both covering redundancy metrics ($\rho_{\mathcal{C}}$ and CR) are consistent. When comparing the dataset coverage *dbCov* with such covering metrics, we observe that our approach not only better covers the data but also yields sets of subgroups with more diverse coverage. Also, by comparing the results of $\rho_{\mathcal{D}}$ with *length* (average description length), and the results of both $\rho_{\mathcal{C}}$ and CR with *sgCov* (average subgroup coverage), we also observe that the EsmamDS not only returns smaller descriptions and more general subgroup's coverage, but also achieves higher diversity on both dimensions.

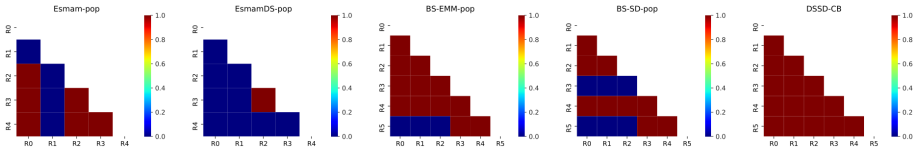
Additionally, from the model redundancy metric $\rho_{\mathcal{M}}$, we conclude that our approach attains higher diversity of the discovered survival models by achieving lower percentages of similar (pair of) models. Such a result is endorsed by the plots of the survival models of the discovered subgroups, presented in Figure 8(9) for the *population*(*complement*)-baseline algorithms on two datasets. Throughout the plots, it is possible to observe that EsmamDS survival models not only are more distant (distinct) from each other, but the EsmamDS complete set of subgroups also capture a wider range of survival responses (from lower to higher survival curves).



(a) Description Similarity

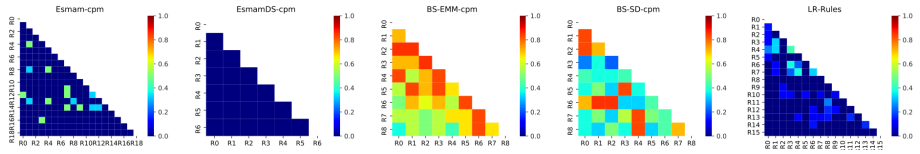


(b) Cover Similarity

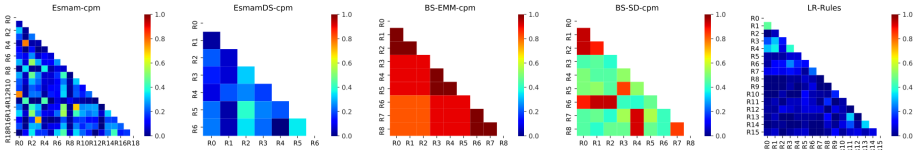


(c) Model Similarity

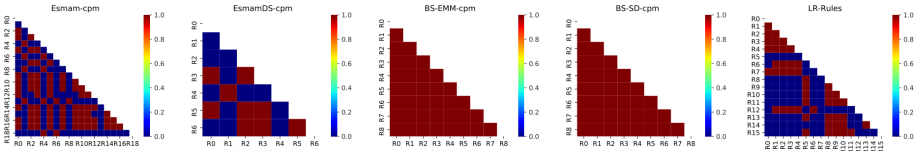
Fig. 4: Similarity measures within subgroup sets for *whas500* dataset (*exp0*) of *population*-baseline algorithms: (a) description similarity $\varsigma_{\mathcal{D}}$; (b) cover similarity $\varsigma_{\mathcal{C}}$; and (c) model similarity $\varsigma_{\mathcal{M}}$. (*From left to right: Esmam, EsmamDS, BS-EMM, BS-SD and DSSD-CBSS*)



(a) Description Similarity

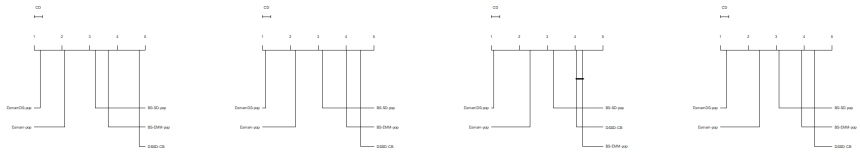


(b) Cover Similarity



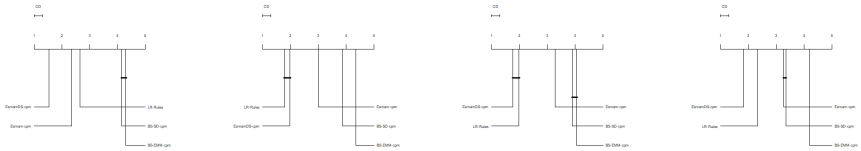
(c) Model Similarity

Fig. 5: Similarity measures within subgroup sets for *breast-cancer* dataset (*exp0*) of *complement*-baseline algorithms: (a) description similarity $\varsigma_{\mathcal{D}}$; (b) cover similarity $\varsigma_{\mathcal{C}}$; and (c) model similarity $\varsigma_{\mathcal{M}}$. (*From left to right: Esmam, EsmamDS, BS-EMM, BS-SD and LR-Rules*)



(a) ρ_d (b) ρ_c (c) CR (d) ρ_m

Fig. 6: Plot CD of Nemenyi post-hoc test over *population*-baseline algorithms for the following metrics: (a) description redundancy (ρ_d); (b) coverage redundancy (ρ_c); (c) Cover Redundancy measure (CR); and (d) model redundancy (ρ_m).



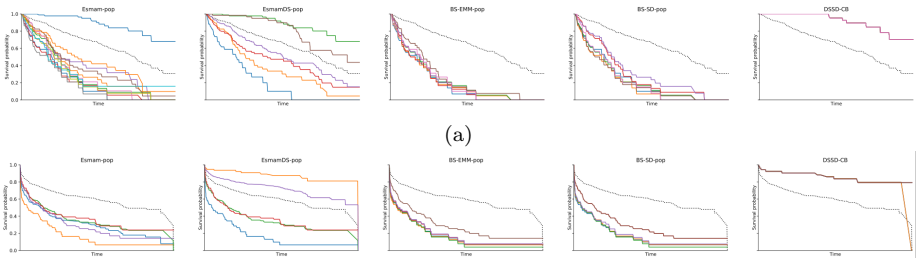
(a) ρ_d (b) ρ_c (c) CR (d) ρ_m

Fig. 7: Plot CD of Nemenyi post-hoc test over *complement*-baseline algorithms for the following metrics: (a) description redundancy (ρ_d); (b) coverage redundancy (ρ_c); (c) Cover Redundancy measure (CR); and (d) model redundancy (ρ_m).

Redundancy between Subgroup Sets. Finally, we assess the similarity between the sets of subgroups delivered by each approach. For that matter, we compare the EsmamDS subgroups' sets against the sets of each compared algorithm by computing the similarities $\varsigma_{\mathcal{D}}, \varsigma_{\mathcal{C}}$ and $\varsigma_{\mathcal{M}}$ for all combinations of subgroups discovered by both compared sets. Therefore, being $\mathbb{S}_{EsmamDS} = \{\mathcal{S}^1, \ldots, \mathcal{S}^n\}$ the resultant EsmamDS subgroup set and $\mathbb{S}_{Comp} = \{\mathcal{S}^1, \ldots, \mathcal{S}^k\}$ the set of unique subgroups delivered by the compared approach, we comput the similarities measures $\varsigma = \{\varsigma_d, \varsigma_d, \varsigma_m\}$ between all combinations of subgroups $(\mathcal{S}^i, \mathcal{S}^j)\ \forall\ i = \{1, n\}, j = \{1, k\}$.

Here is important to pinpoint that not all the compared algorithms deliver unique subgroups. In some cases, descriptions with the same terms but in different order are returned as two different subgroups. However, one can agree that descriptions with the same terms (conditions) delineate the same data subset and, therefore, comprise a single subgroup characterisation – therefore, we compare all combination of *unique* subgroups of EsmamDS and each other approach. We plotted such results in heatmaps matrix ($\mathbb{S}_{Comp} \times \mathbb{S}_{EsmamDS}$), for which the columns are all subgroups delivered by the EsmamDS algorithm, and each row is a unique subgroup discovered by the compared approach. Figure 10(11) presents the similarity between subgroup sets for *population*(*complement*)-baseline algorithms over XXX (XXX) datasets, in all three dimensions of redundancy.

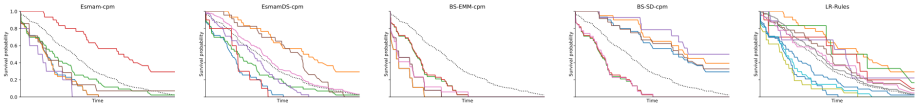
Note that higher values represent higher similarity, where maximum similarity is the same as equality. The different proportions between the plot dimensions are related to the difference between the number of unique subgroups in each



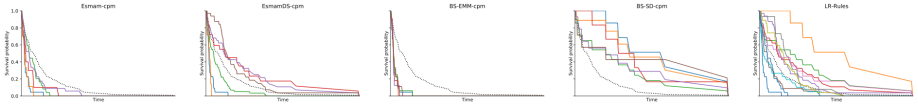
(a)

(b)

Fig. 8: Survival plots for *population*-baseline algorithms discovered subgroups on (a) *pbc (exp0)* dataset; and (b) *whas500 (exp0)* dataset. (*From left to right: Esmam, EsmamDS, BS-EMM, BS-SD and DSSD-CBSS*)



(a)



(b)

Fig. 9: Survival plots for *complement*-baseline algorithms discovered subgroups on (a) *lung (exp0)* dataset; and (b) *veteran (exp0)* dataset. (*From left to right: Esmam, EsmamDS, BS-EMM, BS-SD and LR-Rules*)

compared set. Here is important to remember that the granularity of a matrix dimension not necessarily corresponds to the set's size ($\#sg$), since we analyse the set of unique subgroups rather than all subgroups delivered by the algorithms. Also, note that the plots' vertical patterns reveal the similarity between a single EsmamDS subgroups and all unique subgroups discovered by the other algorithm. Similarly, horizontal patterns indicate the similarity between all subgroups in the EsmamDS set and a single subgroup of the compared approach. We observed that such vertical patterns are present in most experiments, indicating that the subgroups' sets delivered by EsmamDS somehow encompass the subgroups returned by the other algorithms.

To better analyse such results, we use the comparison between EsmamDS-pop and BS-EMM-pop – shown in the second-plot of Fig 10 – as an example. From the plot, we can observe that all subgroups in the BS-EMM-pop set are somehow similar to the first EsmamDS subgroup. Such subgroups' descriptions are as follows (the BS-EMM-pop subgroups were rearranged in increasing depth):

$$
\begin{array}{lll}
\mathcal{S}^0_{EsmamDS-pop}: & age = [74, 87] & \\
\mathcal{S}^1_{BS-EMM-pop}: & age = [74, 87] & \\
\mathcal{S}^3_{BS-EMM-pop}: & age = [74, 87] & \& \quad creat = 0 \\
\mathcal{S}^0_{BS-EMM-pop}: & age = [74, 87] & \& \quad pcdx = not\text{-}prog \\
\mathcal{S}^2_{BS-EMM-pop}: & age = [74, 87] & \& \quad pcdx = not\text{-}prog \quad \& \quad creat = 0
\end{array}
$$

We can observe that all subgroups discovered by BS-EMM-pop are refinements of a single EsmamDS-pop subgroup ($\mathcal{S}^0$). It is important to notice that although the description similarity measure (Figs 10a,11a) captures intersections in subgroups' descriptions, such measure decreases as a refinement becomes deeper (e.g. $\mathcal{S}^1, \mathcal{S}^3$ and $\mathcal{S}^2$ from BS-EMM-pop, respectively). Therefore, low levels of description similarity may reflect specialisations of a single pattern. From the cover similarity (Fig. 10b), we can observe that refinements of a subgroup may yield slightly variations in coverage, as is the case between $\mathcal{S}^0_{EsmamDS}$ and its refinement $\mathcal{S}^0_{BS-EMM-pop}$. Besides, when analysing the models' similarities (Fig 10c), we can observe that all refinements delivered by the BS-EMM-pop algorithm present survival models similar to their generalisation, $\mathcal{S}^0_{EsmamDS}$. That is because subsets of a subgroup's coverage, i.e. refinements of a subgroup's description, may present exceptional behaviour when compared to the baseline model but not necessarily yield a distinct distribution from their generalisation. There are also cases where we observe that different descriptions yield some degree of cover and model similarities.

Therefore, what we observe is that the EsmamDS algorithm not only yield more general subgroups (see *sgCov* results) but also better assures their generalisation. Our approach minimises the presence of subgroups' refinements by allowing their occurrence only if they present distinct model distribution. As a result, in most experiments, we observe that the EsmamDS somehow encompasses the subgroups delivered by the other approaches, but with more general and compact representations, while discovering subgroups that the other approaches do not uncover.

6 Conclusions

Acknowledgements This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 and by the National Council for Scientific and Technological Development – CNPq.

References

1. Attanasio, C.: Exceptional incidence distribution mining on a nationwide cancer registry: a descriptive approach. Master's thesis, Eindhoven University of Technology (2019)
2. Atzmueller, M.: Subgroup discovery. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **5**(1), 35–49 (2015)
3. Atzmueller, M., Lemmerich, F.: Fast subgroup discovery for continuous target concepts. In: International Symposium on Methodologies for Intelligent Systems, pp. 35–44. Springer (2009)
4. Atzmueller, M., Puppe, F.: Sd-map–a fast algorithm for exhaustive subgroup discovery. In: European Conference on Principles of Data Mining and Knowledge Discovery, pp. 6–17. Springer (2006)

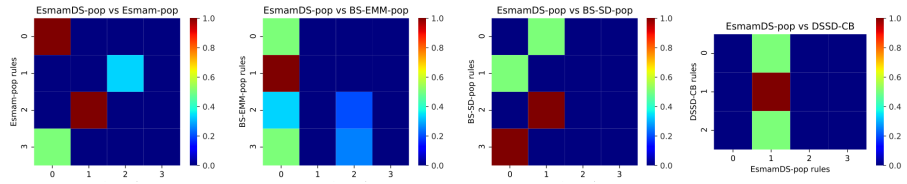
5. Bazan, J., Osmólski, A., Skowron, A., Ślçezak, D., Szczuka, M., Wroblewski, J.: Rough set approach to the survival analysis. In: International Conference on Rough Sets and Current Trends in Computing, pp. 522–529. Springer (2002)
6. Belfodil, A.: Exceptional model mining for behavioral data analysis. Ph.D. thesis, Institut National des Sciences Apliqueés de Lion (2019)
7. Belfodil, A., Cazalens, S., Lamarre, P., Plantevit, M.: Identifying exceptional (dis) agreement between groups. Data Mining and Knowledge Discovery **34**(2), 394–442 (2020)
8. Belfodil, A., Duivesteijn, W., Plantevit, M., Cazalens, S., Lamarre, P.: Deviant: Discovering significant exceptional (dis-) agreement within groups. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer (2019)
9. Berlanga, F., Del Jesus, M.J., González, P., Herrera, F., Mesonero, M.: Multiobjective evolutionary induction of subgroup discovery fuzzy rules: a case study in marketing. In: Industrial Conference on Data Mining, pp. 337–349. Springer (2006)
10. 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 **17**(10), 1169–1186 (1998)
11. 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, pp. 582–590 (2011)
12. 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 **32**(3), 604–650 (2018)
13. Bou-Hamad, I., Larocque, D., Ben-Ameur, H., et al.: A review of survival trees. Statistics Surveys **5**, 44–71 (2011)
14. 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 **89**(3), 431 (2003)
15. Bringmann, B., Zimmermann, A.: The chosen few: On identifying valuable patterns. In: Seventh IEEE International Conference on Data Mining (ICDM 2007), pp. 63–72. IEEE (2007)
16. Carmona, C.J., González, P., del Jesús, M.J., Herrera, F.: Non-dominated multi-objective evolutionary algorithm based on fuzzy rules extraction for subgroup discovery. In: International Conference on Hybrid Artificial Intelligence Systems, pp. 573–580. Springer (2009)
17. Carmona, C.J., González, P., del Jesus, M.J., Herrera, F.: Nmeef-sd: non-dominated multiobjective evolutionary algorithm for extracting fuzzy rules in subgroup discovery. IEEE Transactions on Fuzzy Systems **18**(5), 958–970 (2010)
18. Carmona, C.J., González, P., del Jesus, M.J., Herrera, F.: Overview on evolutionary subgroup discovery: analysis of the suitability and potential of the search performed by evolutionary algorithms. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **4**(2), 87–103 (2014)
19. Cavazzuti, M.: Optimization methods: from theory to design scientific and technological aspects in mechanics. Springer Science & Business Media (2012)
20. Davis, R.B., Anderson, J.R.: Exponential survival trees. Statistics in Medicine **8**(8), 947–961 (1989)
21. Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. Communications of the ACM **51**(1), 107–113 (2008)
22. Del Jesus, M.J., González, P., Herrera, F., Mesonero, M.: Evolutionary fuzzy rule induction process for subgroup discovery: a case study in marketing. IEEE Transactions on Fuzzy Systems **15**(4), 578–592 (2007)
23. Dorigo, M., Birattari, M., Stutzle, T.: Ant colony optimization. IEEE computational intelligence magazine **1**(4), 28–39 (2006)
24. Dorigo, M., Blum, C.: Ant colony optimization theory: A survey. Theoretical computer science **344**(2-3), 243–278 (2005)
25. Dorigo, M., Di Caro, G.: Ant colony optimization: a new meta-heuristic. In: Proceedings of the 1999 congress on evolutionary computation-CEC99 (Cat. No. 99TH8406), vol. 2, pp. 1470–1477. IEEE (1999)
26. Dorigo, M., Maniezzo, V., Colorni, A.: Positive feedback as a search strategy (1991)
27. 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) **26**(1), 29–41 (1996)

28. Dorigo, M., Stützle, T.: The ant colony optimization metaheuristic: Algorithms, applications, and advances. In: Handbook of metaheuristics, pp. 250–285. Springer (2003)
29. Dorigo, M., Stützle, T.: Ant colony optimization: overview and recent advances. In: Handbook of metaheuristics, pp. 311–351. Springer (2019)
30. Downar, L., Duivesteijn, W.: Exceptionally monotone models—the rank correlation model class for exceptional model mining. Knowledge and Information Systems **51**(2), 369–394 (2017)
31. Du, X., Duivesteijn, W., Klabbers, M., Pechenizkiy, M.: Elba: Exceptional learning behavior analysis. International Educational Data Mining Society (2018)
32. Du, X., Pei, Y., Duivesteijn, W., Pechenizkiy, M.: Exceptional spatio-temporal behavior mining through bayesian non-parametric modeling. Data Mining and Knowledge Discovery pp. 1–24 (2020)
33. Duivesteijn, W.: Exceptional model mining. Ph.D. thesis, Faculteit der Wiskunde en Natuurwetenschappen, Leiden Institute of Advanced Computer Science (LIACS), Faculty of Science, Leiden University (2013)
34. Duivesteijn, W., Farzami, T., Putman, T., Peer, E., Weerts, H.J., Adegeest, J.N., Foks, G., Pechenizkiy, M.: Have it both ways—from a/b testing to a&b testing with exceptional model mining. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 114–126. Springer (2017)
35. Duivesteijn, W., Feelders, A.J., Knobbe, A.: Exceptional model mining. Data Mining and Knowledge Discovery **30**(1), 47–98 (2016)
36. Duivesteijn, W., Knobbe, A., Feelders, A., van Leeuwen, M.: Subgroup discovery meets bayesian networks–an exceptional model mining approach. In: 2010 IEEE International Conference on Data Mining, pp. 158–167. IEEE (2010)
37. Evers, L., Messow, C.M.: Sparse kernel methods for high-dimensional survival data. Bioinformatics **24**(14), 1632–1638 (2008)
38. Faraggi, D., Simon, R.: A neural network model for survival data. Statistics in medicine **14**(1), 73–82 (1995)
39. 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 **28**(12), 3126–3139 (2016)
40. Fausto, F., Reyna-Orta, A., Cuevas, E., Andrade, Á.G., Perez-Cisneros, M.: From ants to whales: metaheuristics for all tastes. Artificial Intelligence Review **53**(1), 753–810 (2020)
41. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI magazine **17**(3), 37–37 (1996)
42. Floreano, D., Mattiussi, C.: Bio-inspired artificial intelligence. Ch **5**, 335–396 (2008)
43. Friedman, J.H., Fisher, N.I.: Bump hunting in high-dimensional data. Statistics and Computing **9**(2), 123–143 (1999)
44. Fürnkranz, J.: Separate-and-conquer rule learning. Artificial Intelligence Review **13**(1), 3–54 (1999)
45. Gamberger, D., Lavrac, N.: Expert-guided subgroup discovery: Methodology and application. Journal of Artificial Intelligence Research **17**, 501–527 (2002)
46. Gordon, L., Olshen, R.A.: Tree-structured survival analysis. Cancer treatment reports **69**(10), 1065–1069 (1985)
47. Goss, S., Aron, S., Deneubourg, J.L., Pasteels, J.M.: Self-organized shortcuts in the argentine ant. Naturwissenschaften **76**(12), 579–581 (1989)
48. Grosskreutz, H., Rüping, S.: On subgroup discovery in numerical domains. Data mining and knowledge discovery **19**(2), 210–226 (2009)
49. Grosskreutz, H., Rüping, S., Wrobel, S.: Tight optimistic estimates for fast subgroup discovery. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 440–456. Springer (2008)
50. Helal, S.: Subgroup discovery algorithms: a survey and empirical evaluation. Journal of Computer Science and Technology **31**(3), 561–576 (2016)
51. Herrera, F., Carmona, C.J., González, P., Del Jesus, M.J.: An overview on subgroup discovery: foundations and applications. Knowledge and information systems **29**(3), 495–525 (2011)
52. Hothorn, T., Lausen, B., Benner, A., Radespiel-Tröger, M.: Bagging survival trees. Statistics in medicine **23**(1), 77–91 (2004)
53. Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S., et al.: Random survival forests. The annals of applied statistics **2**(3), 841–860 (2008)

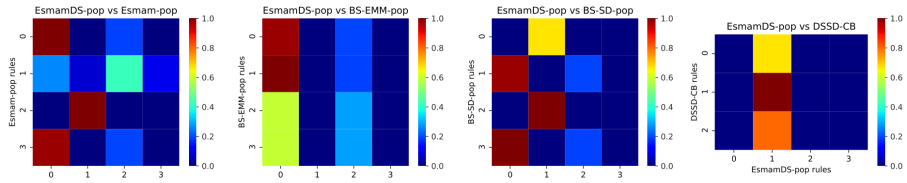
54. del Jesus, M.J., González, P., Herrera, F.: Multiobjective genetic algorithm for extracting subgroup discovery fuzzy rules. In: 2007 IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making, pp. 50–57. IEEE (2007)
55. Kaplan, E.L., Meier, P.: Nonparametric estimation from incomplete observations. Journal of the American statistical association **53**(282), 457–481 (1958)
56. Kavšek, B., Lavrač, N.: Apriori-sd: Adapting association rule learning to subgroup discovery. Applied Artificial Intelligence **20**(7), 543–583 (2006)
57. Kavšek, B., Lavrač, N., Jovanoski, V.: Apriori-sd: Adapting association rule learning to subgroup discovery. In: International Symposium on Intelligent Data Analysis, pp. 230–241. Springer (2003)
58. Khan, F.M., Zubek, V.B.: Support vector regression for censored data (svrc): a novel tool for survival analysis. In: 2008 Eighth IEEE International Conference on Data Mining, pp. 863–868. IEEE (2008)
59. Klösgen, W.: Explora: A multipattern and multistrategy discovery assistant. In: Advances in knowledge discovery and data mining, pp. 249–271 (1996)
60. Klösgen, W., May, M.: Spatio-temporal subgroup discovery. In: Mining Spatio-Temporal Information Systems, pp. 149–168. Springer (2002)
61. Knobbe, A.J., Ho, E.K.: Pattern teams. In: European Conference on Principles of Data Mining and Knowledge Discovery, pp. 577–584. Springer (2006)
62. Kononenko, I.: Inductive and bayesian learning in medical diagnosis. Applied Artificial Intelligence an International Journal **7**(4), 317–337 (1993)
63. Kononenko, I.: Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in medicine **23**(1), 89–109 (2001)
64. Krak, T.E., Feelders, A.: Exceptional model mining with tree-constrained gradient ascent. In: Proceedings of the 2015 SIAM International Conference on Data Mining, pp. 487–495. SIAM (2015)
65. Kronek, L.P., Reddy, A.: Logical analysis of survival data: prognostic survival models by detecting high-degree interactions in right-censored data. Bioinformatics **24**(16), i248–i253 (2008)
66. Lavrač, N., Kavšek, B., Flach, P., Todorovski, L.: Subgroup discovery with cn2-sd. Journal of Machine Learning Research **5**(Feb), 153–188 (2004)
67. Lavrač, N., Železnỳ, F., Flach, P.A.: Rsd: Relational subgroup discovery through first-order feature construction. In: International Conference on Inductive Logic Programming, pp. 149–165. Springer (2002)
68. LeBlanc, M., Crowley, J.: Relative risk trees for censored survival data. Biometrics pp. 411–425 (1992)
69. van Leeuwen, M.: Maximal exceptions with minimal descriptions. Data Mining and Knowledge Discovery **21**(2), 259–276 (2010)
70. Leman, D., Feelders, A., Knobbe, A.: Exceptional model mining. In: Joint European conference on machine learning and knowledge discovery in databases, pp. 1–16. Springer (2008)
71. Lemmerich, F., Becker, M., Atzmueller, M.: Generic pattern trees for exhaustive exceptional model mining. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 277–292. Springer (2012)
72. Lemmerich, F., Becker, M., Singer, P., Helic, D., Hotho, A., Strohmaier, M.: Mining subgroups with exceptional transition behavior. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 965–974 (2016)
73. Li, Y., Wang, J., Ye, J., Reddy, C.K.: A multi-task learning formulation for survival analysis. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1715–1724. ACM (2016)
74. 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: 2016 IEEE 16th International Conference on Data Mining (ICDM), pp. 231–240. IEEE (2016)
75. 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 **28**(1), 1–25 (2003)
76. Liu, X., Minin, V., Huang, Y., Seligson, D.B., Horvath, S.: Statistical methods for analyzing tissue microarray data. Journal of biopharmaceutical statistics **14**(3), 671–685 (2004)
77. Lowerre, B.T.: The harpy speech recognition system. Tech. rep., CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE (1976)

78. Lucas, P.J., Van der Gaag, L.C., Abu-Hanna, A.: Bayesian networks in biomedicine and health-care. Artificial intelligence in medicine **30**(3), 201–214 (2004)
79. 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 **59**, 487–499 (2017)
80. Lucas, T., Vimieiro, R., Ludermir, T.: Ssdp+: A diverse and more informative subgroup discovery approach for high dimensional data. In: 2018 IEEE Congress on Evolutionary Computation (CEC), pp. 1–8. IEEE (2018)
81. Luna, J.M., Romero, J.R., Romero, C., Ventura, S.: Discovering subgroups by means of genetic programming. In: European Conference on Genetic Programming, pp. 121–132. Springer (2013)
82. 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 **44**(12), 2329–2341 (2014)
83. Mattos, J.B., Silva, E.G., de Mattos Neto, P.S., Vimieiro, R.: Exceptional survival model mining. In: Brazilian Conference on Intelligent Systems, pp. 307–321. Springer (2020)
84. Meeng, M., Knobbe, A.: For real: a thorough look at numeric attributes in subgroup discovery. Data Mining and Knowledge Discovery **35**(1), 158–212 (2021)
85. Moens, S., Boley, M.: Instant exceptional model mining using weighted controlled pattern sampling. In: International Symposium on Intelligent Data Analysis, pp. 203–214. Springer (2014)
86. Neapolitan, R.E., et al.: Learning bayesian networks, vol. 38. Pearson Prentice Hall Upper Saddle River, NJ (2004)
87. 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 **10**(Feb), 377–403 (2009)
88. Pachón, V., Mata, J., Domínguez, J.L., Maña, M.J.: Multi-objective evolutionary approach for subgroup discovery. In: International Conference on Hybrid Artificial Intelligence Systems, pp. 271–278. Springer (2011)
89. Padillo, F., Luna, J.M., Ventura, S.: Subgroup discovery on big data: exhaustive methodologies using map-reduce. In: 2016 IEEE Trustcom/BigDataSE/ISPA, pp. 1684–1691. IEEE (2016)
90. Padillo, F., Luna, J.M., Ventura, S.: Exhaustive search algorithms to mine subgroups on big data using apache spark. Progress in Artificial Intelligence **6**(2), 145–158 (2017)
91. 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 **118**, 553–562 (2019)
92. Parpinelli, R.S., Lopes, H.S., Freitas, A.A.: Data mining with an ant colony optimization algorithm. IEEE transactions on evolutionary computation **6**(4), 321–332 (2002)
93. Pattaraintakorn, P., Cercone, N.: A foundation of rough sets theoretical and computational hybrid intelligent system for survival analysis. Computers & Mathematics with Applications **56**(7), 1699–1708 (2008)
94. Pawlak, Z.: Rough sets. International journal of computer & information sciences **11**(5), 341–356 (1982)
95. Pawlak, Z., Skowron, A.: Rudiments of rough sets. Information sciences **177**(1), 3–27 (2007)
96. Peto, R., Pike, M., Armitage, P., Breslow, N.E., Cox, D., Howard, S., Mantel, N., McPherson, K., Peto, J., Smith, P.: Design and analysis of randomized clinical trials requiring prolonged observation of each patient. ii. analysis and examples. British journal of cancer **35**(1), 1 (1977)
97. Pontes, T., Vimieiro, R., Ludermir, T.B.: Ssdp: A simple evolutionary approach for top-k discriminative patterns in high dimensional databases. In: 2016 5th Brazilian Conference on Intelligent Systems (BRACIS), pp. 361–366. IEEE (2016)
98. Proença, H.M., Bäck, T., van Leeuwen, M.: Robust subgroup discovery. arXiv preprint arXiv:2103.13686 (2021)
99. Pulgar-Rubio, F., Rivera-Rivas, A., Pérez-Godoy, M.D., González, P., Carmona, C.J., Del Jesus, M.: Mefasd-bd: multi-objective evolutionary fuzzy algorithm for subgroup discovery in big data environments-a mapreduce solution. Knowledge-Based Systems **117**, 70–78 (2017)
100. Radespiel-Tröger, M., Gefeller, O., Rabenstein, T., Hothorn, T.: Association between split selection instability and predictive error in survival trees. Methods of information in medicine **45**(05), 548–556 (2006)

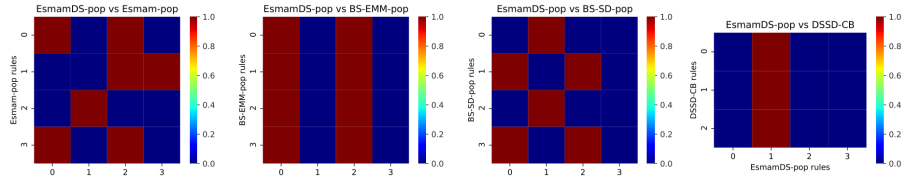
101. Raftery, A.E., Madigan, D., Volinsky, C.T.: Accounting for model uncertainty in survival analysis improves predictive performance. Bayesian statistics **5**, 323–349 (1996)
102. Riahi, F., Schulte, O.: Model-based exception mining for object-relational data. Data Mining and Knowledge Discovery pp. 1–42 (2020)
103. Rodríguez, D., Ruiz, R., Riquelme, J.C., Aguilar-Ruiz, J.S.: Searching for rules to detect defective modules: A subgroup discovery approach. Information Sciences **191**, 14–30 (2012)
104. Segal, M.R.: Regression trees for censored data. Biometrics pp. 35–47 (1988)
105. Shivaswamy, P.K., Chu, W., Jansche, M.: A support vector approach to censored targets. In: Seventh IEEE International Conference on Data Mining (ICDM 2007), pp. 655–660. IEEE (2007)
106. 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 **22**, 35–53 (2013)
107. Sikora, M., et al.: Censoring weighted separate-and-conquer rule induction from survival data. Methods of information in medicine **53**(02), 137–148 (2014)
108. Štajduhar, I., Dalbelo-Bašić, B.: Learning bayesian networks from survival data using weighting censored instances. Journal of biomedical informatics **43**(4), 613–622 (2010)
109. Štajduhar, I., Dalbelo-Bašić, B., Bogunović, N.: Impact of censoring on learning bayesian networks in survival modelling. Artificial intelligence in medicine **47**(3), 199–217 (2009)
110. Sutton, R.S., Barto, A.G., et al.: Reinforcement learning: An Introduction, vol. 135. MIT press Cambridge (1998)
111. Van Belle, V., Pelckmans, K., Suykens, J., Van Huffel, S.: Support vector machines for survival analysis. In: Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare (CIMED2007), pp. 1–8 (2007)
112. Van Belle, V., Pelckmans, K., Van Huffel, S., Suykens, J.A.: Support vector methods for survival analysis: a comparison between ranking and regression approaches. Artificial intelligence in medicine **53**(2), 107–118 (2011)
113. Van Leeuwen, M., Knobbe, A.: Non-redundant subgroup discovery in large and complex data. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 459–474. Springer (2011)
114. Van Leeuwen, M., Knobbe, A.: Diverse subgroup set discovery. Data Mining and Knowledge Discovery **25**(2), 208–242 (2012)
115. Ventura, S., Luna, J.M.: Supervised descriptive pattern mining. Springer (2018)
116. Ventura, S., Luna, J.M., et al.: Pattern mining with evolutionary algorithms. Springer (2016)
117. Vinzamuri, B., Li, Y., Reddy, C.K.: Active learning based survival regression for censored data. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp. 241–250. ACM (2014)
118. Wang, P., Li, Y., Reddy, C.K.: Machine learning for survival analysis: A survey. ACM Computing Surveys (CSUR) **51**(6), 110 (2019)
119. Widodo, A., Yang, B.S.: Application of relevance vector machine and survival probability to machine degradation assessment. Expert Systems with Applications **38**(3), 2592–2599 (2011)
120. Wróbel, Ł.: Tree-based induction of decision list from survival data. Journal of Medical Informatics & Technologies **20** (2012)
121. Wróbel, Ł., Gudyś, A., Sikora, M.: Learning rule sets from survival data. BMC bioinformatics **18**(1), 285 (2017)
122. Wrobel, S.: An algorithm for multi-relational discovery of subgroups. In: European Symposium on Principles of Data Mining and Knowledge Discovery, pp. 78–87. Springer (1997)
123. Wrobel, S.: Inductive logic programming for knowledge discovery in databases. In: Relational data mining, pp. 74–101. Springer (2001)
124. Zeev-Ben-Mordehai, O., Duivesteijn, W., Pechenizkiy, M.: Controversy rules-discovering regions where classifiers (dis-) agree exceptionally. arXiv preprint arXiv:1808.07243 (2018)
125. Železnỳ, F., Lavrač, N.: Propositionalization-based relational subgroup discovery with rsd. Machine Learning **62**(1-2), 33–63 (2006)
126. 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 **20**(1), 59–75 (2000)



(a) Description Similarity

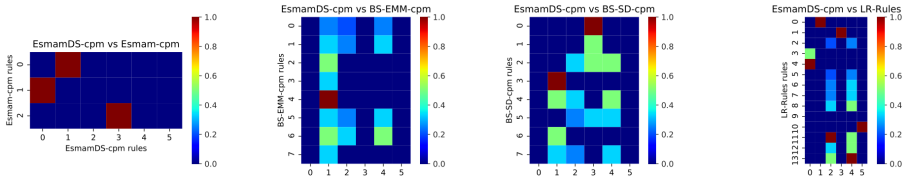


(b) Cover Similarity

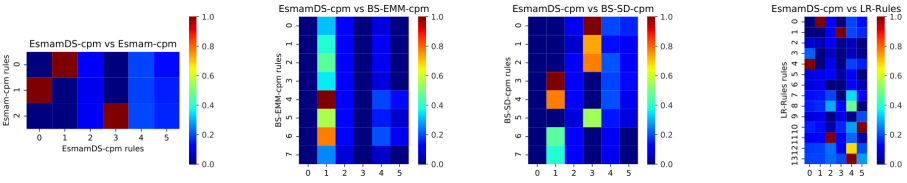


(c) Model Similarity

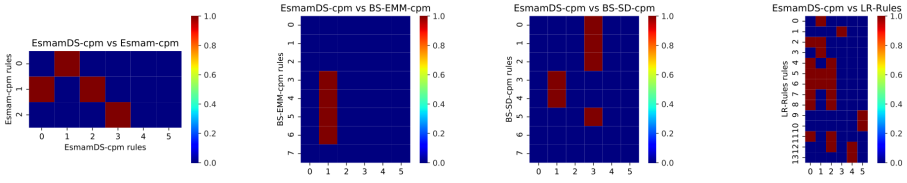
Fig. 10: Similarity measures between EsmamDS subgroups' set (columns) and the remaining *population*-baseline algorithms' set for *mgus* dataset (*exp0*): (a) description similarity $\varsigma_{\mathcal{D}}$; (b) cover similarity $\varsigma_{\mathcal{C}}$; and (c) model similarity $\varsigma_{\mathcal{M}}$. (*From left to right EsmamDS against Esmam, BS-EMM, BS-SD and DSSD-CBSS*)



(a) Description Similarity



(b) Cover Similarity



(c) Model Similarity

Fig. 11: Similarity measures between EsmamDS subgroups' set (columns) and the remaining *population*-baseline algorithms' set for *mgus* dataset (*exp0*): (a) description similarity $\varsigma_{\mathcal{D}}$; (b) cover similarity $\varsigma_{\mathcal{C}}$; and (c) model similarity $\varsigma_{\mathcal{M}}$. (*From left to right EsmamDS against: Esmam, BS-EMM, BS-SD and LR-Rules*)