Simple explanations to summarise Subgroup Discovery outcomes: a case study concerning patient phenotyping

Enrique Valero-Leal[1,2] [0000−0002−5797−360X], Manuel Campos[2,3] [0000−0002−5233−3769], and Jose M. Juarez[2] [0000−0003−1776−1992]

[1] Technical University of Madrid
[2] AIKE group (INTICO), University of Murcia
[3] Murcian Bio-Health Institute (IMIB-Arrixaca)

Abstract. Phenotyping is essential in medical research, as it provides a better understanding of healthcare problems owing to the fact that clinical phenotypes identify subsets of patients with common characteristics. Subgroup discovery (SD) appears to be a promising machine learning approach because it provides a framework with which to search for interesting subgroups according to the relations between the individual characteristics and a target value. Each single pattern extracted by SD algorithms is human-readable. However, its complexity (the number of attributes involved) and the high number of subgroups obtained make the overall model difficult to understand. In this work, we propose a method with which to explain SD, designed for the clinical context. We have employed a two-step process in order to obtain SD model-agnostic explanations based on a decision tree surrogate model. The complexity involved in evaluating explainable methods led us to adopt a multiple strategy. We first show how explanations are built, and test a selection of state-of-the-art SD algorithms and gold-standard datasets. We then illustrate the suitability of the method in a clinical use case for an antimicrobial resistance problem. Finally, we study the utility of the method by surveying a small group in order to validate it from a human-centric perspective.

Keywords: explainable artificial intelligence · subgroup discovery · biomedical informatics

1 Introduction

Although explainability is a term that predates this century, there is now an increasing interest in explainable artificial intelligence (XAI). Much of the work in this field revolves around classification and Deep Learning techniques, while some areas – such as unsupervised and semi-supervised learning - are barely explored.

This is the case of subgroup discovery (SD) [15,36], a family of descriptive induction algorithms that find subgroups of interesting members of a particular population with regard to a certain characteristic (target attribute). SD techniques are, in practice, particularly helpful in biomedical science for the purpose of patient phenotyping, i.e. the characterisation of groups of patients given their traits and clinical evidence [9,12,25]. However, from the XAI point of view, the main limitations as regards making SD outcomes understandable are the volume of subgroups obtained and their complexity (the

number of descriptors involved in order to define each pattern discovered). In this work, we tackle both aspects of the problem with the objective of providing a more compact and understandable representation of the whole SD model. The main contributions of this paper are the following:

- A new SD model-agnostic explanation method based on a global surrogate model.
- The evaluation of the explanatory SD capacity of our proposal: (1) using gold-standard datasets, (2) illustrating the utility of the proposal with a real clinical phenotype problem concerning infectious diseases, and (3) carrying out an empirical survey analysis to study subjective human satisfaction.

The paper is organised as follows. In Section 2, we introduce the background knowledge and the notation used. The SD explainer is described in Section 3, while Section 4 shows the preliminary experimental results obtained with synthetic data, a proof of concept in the antimicrobial infection domain and a study of the usefulness of our proposal by means of a survey. Finally, Sections 5 and 6 respectively provide a discussion of the results obtained and our conclusions.

2 Background

2.1 XAI methods and healthcare

In XAI, it is possible to distinguish between model-specific and model-agnostic explanations. The objective of the former is to explain the model itself and can be understood as explainability by design, whereas model-agnostic explanations are independent of the model. Model agnostic methods are post-hoc, i.e. we first train a machine learning (ML) model and we then attempt to explain it by considering the outputs of that model.

An example of a model-agnostic technique is the global surrogate model, which consists of approximating the results obtained by a black box using a simple and intrinsically interpretable model [24]. We first train the original (black box) model, and the outputs obtained are then used to train the interpretable model, after which it is necessary only to study the interpretable model. Similar works have already been carried out using this approach in the context of explaining neural networks [6].

Given the importance of transparency in the healthcare domain, there has been a growing interest in improving the explainability of opaque yet powerful models, such as random forests [20] and different types of artificial neural networks [13,22]. The use of model-agnostic methods, such as local explanations provided by LIME [30] or SHAP [21], makes it possible to both maintain high accuracy in the system and provide approximate explanations of the reasoning.

Although the aforementioned methods have yielded relatively good results, they are designed to explain classification tasks. We are, therefore, of the opinion that the current state of the art lacks an exploratory analysis of descriptive methods, and believe that a promising approach would be to adapt the philosophy of the global explanations to these methods. The use of a global surrogate approach might, for example, make it possible to gain further insights into the relations between the data and the output, which is exactly what doctors require when SD is explained to them.

Decision trees have been adopted by the healthcare community as a graphical method with which to express most of the medical decisions described in clinical guidelines and protocols [27]. The later success of decision tree algorithms in the 1990s contributed to these structures becoming gold standards for clinical knowledge extraction using ML [29]. Clinicians' familiarity with decision trees helps to answer questions about the importance of features, data distribution and the output (subgroups) obtained [8].

Other models that could be considered as surrogates are the state-of-the-art algorithms designed with the philosophy of being interpretable and transparent. Generalized linear rule models [35] generate a linear combination of interpretable rules for classification and regression using column generation to deal with the explosion of possible conjunctions/disjunctions of the rules. Similarly, in [7], an algorithm with which to construct rules in conjunctive and disjunctive normal forms, denominated as boolean decision rules via column generation, is presented. However, both methods generate a set of rules rather than compacting all the information into a single structure.

2.2 Subgroup discovery

SD can be defined [15,36] as a ML task at the intersection of predictive and descriptive modelling [26] whose objective is to discover the most interesting subgroups within a given population according to a certain set of characteristics of interest and to a target variable.

In this work we provide the following conventions based on [3]:

Given a dataset $D = (I, A)$ where I defines the set of individuals and $A = \{a_1, \ldots, a_m\}$ is the set of attributes, a **selector condition** sc describes a constraint on the values of an attribute $a_i \in A$ of the dataset.

A **selector** is a function $s_{sc} : I \rightarrow Boolean$ that returns $True$ if an individual $i \in I$ of the dataset has the characteristics described by the selector condition sc.

A **pattern** P is a finite set of selectors $P = \{s_1, \ldots, s_l\}$, interpreted as a conjunctive or disjunctive form. For the sake of simplicity, in this work we restrict the patterns to their conjunctive form.

A **subgroup** is a 2-tuple $SG = (P, s_t)$, where P is a pattern and s_t is a selector. The subgroup can be interpreted as an IF-THEN rule. For example, given the selector condition $s_t = (susceptibility = Resistant)$ and the pattern $P = \{age > 35, culture = EnterococcusFaecium\}$ the following subgroup can be defined: $IF\ (age > 35 \wedge culture = EnterococcusFaecium)\ THEN\ susceptibility = Resistant$. Given a dataset $D = (I, A)$ and the subgroup $SG = (P, s_t)$, the instances of a subgroup are formalised as $SG(\cdot) = \{\forall i \in I | s_{sc}(i) = True, \forall sc \in P\}$. It is worth mentioning that $SG(\cdot)$ also includes false positive instances, that is $i \in I$ where $s_t(i) = False$.

The interest of a subgroup is computed by employing a **quality function** $qf(P, D) \in \mathbb{R}^+$ that maps every pattern P in the search space onto a real number that reflects the quality of the subgroup. For example, in Equation (1) we formalise the weighted relative accuracy (WRAcc) [16], a widely used quality measure that provides a trade-off between the generality and distributional unusualness of the subgroup [5,17]. The first part of the product refers to the support of the subgroup, whereas the second refers to

the relative accuracy, i.e., the accuracy minus the proportion of (true and false) positives instances.

$$WRAcc(SG) = \frac{|SG(\cdot)|}{|I|} \cdot \left(\frac{|\{\forall i \in SG(\cdot) | s_t(i) = True\}|}{SG(\cdot)} - \frac{|\{\forall i \in I | s_t(i) = True\}|}{|I|} \right). \tag{1}$$

2.3 SD algorithms and explainability

SD algorithms generally follow a top-down search approach in order to find subgroups of interest. The algorithm starts from the most basic subgroup descriptions, and explore the search space by specialising the most promising subgroup descriptions. According to [31], it is possible to distinguish between exhaustive and beam search approaches. For reduced data sets, the whole space is often traversed using adapted versions of frequent pattern mining algorithms, such as $SD\text{-}MAP$ [4], $Dp\text{-}Subgroup$ [11] or BSD [18]. If an exhaustive search is not possible (owing to, e.g., certain medical problems), beam search is frequently adopted in order to explore only a portion of the vast space of solutions by relying on heuristics such as SD [9], $CN2\text{-}SD$ [17] or $SD4TS$ [25] algorithms. Finally, other local-search strategies, such as evolutionary techniques [34], have also been considered.

To the best of our knowledge, little attention is paid to SD algorithms and explainability principles. In [19], the utility of employing SD strategies to provide model-agnostic local (or even global) explanations for recommending systems is discussed. However, rather than use SD to explain other more complex systems, we are interested in studying how explainable is SD itself.

Some proposals use ontologies to generate Subgroup explanations [32,33]. In these works, a SD algorithm is first used to obtain subgroups, after which the dataset is labelled with the subgroups that cover each example, an algorithm ranks the attributes according to their capability to discriminate between subgroups and, finally, a semantic SD algorithm is applied, taking more generic ontology terms as characteristics and the induced subgroups as the target. Other than this approach, as far as we know, there is no literature regarding SD explainability that does not rely on additional knowledge.

3 Methods

In this section, we propose an approach with which to explain SD: the SubgroupExplainer methodology, which can be summarized as follows: (0) We apply a SD algorithm. Any algorithm can be used without restrictions, since our proposal is model agnostic; (1) the dataset is automatically labelled, according to the subgroup coverage of the examples; (2) the surrogate model is learnt from the labelled dataset taking the newly created label as the target attribute. This will build an explainable model for the dataset that has the objective of helping to interpret the SD task. The key elements of the SubgroupExplainer methodology are shown in Figure 1.

The **first step** in our methodology with which to explain SD is that of labelling each instance of the dataset according to the subgroups to which the instance belongs. The

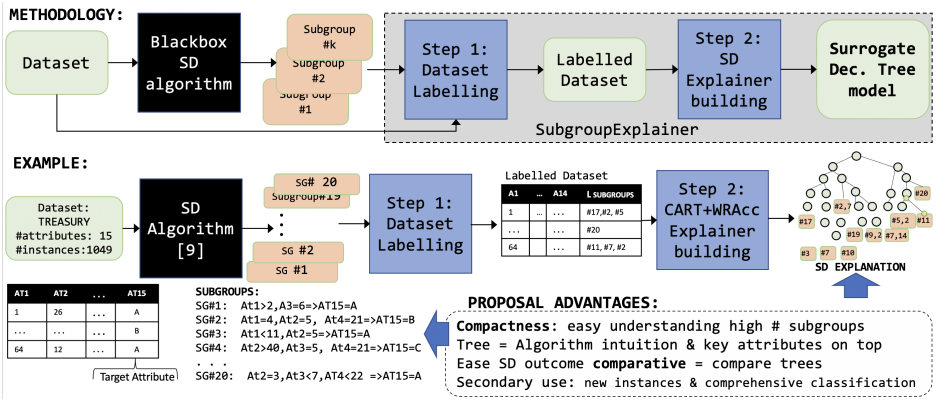

Fig. 1. Intuitions of SubgroupExplainer methodology

fact that an instance belongs to each of the k induced subgroups can be expressed as a k-dimensional boolean label , where $true$ in i signifies that the instance belongs to the i-th subgroup and $false$ is the opposite, with $i \in 1..k$.

Formally, let $S_{SG} = \{SG_1, \ldots, SG_k\}$ be the set of subgroups discovered by a SD algorithm from a given dataset $D(I, A)$. We temporally label the dataset by adding to A a new k-dimensional binary attribute $L' = (l'_1, \ldots, l'_k)$ which expresses the fact that the individual $i \in I$ belongs to the subgroup $SG_j \in S_{SG}$ (Equation (2)). We then transform the vector of labels L' into a single label L, using a label powerset. Working with binary numbers, this equates to transform each label l'_j of the vector L' into one-hot encoding with a 1 bit in the position j if the instance belongs to subgroup j and then sum all of them, getting then a single binary string L in which, consequently, a 1 in the position j means that the instance belongs to subgroup j and a 0 means the opposite. Finally, we label the dataset, $D' = (I, A \cup L)$ (Equation (3)).

$$l'_j(i) = \begin{cases} 1 \text{ if } i \in SG_j(\cdot) \\ 0 \; otherwise \end{cases} \tag{2}$$

$$L(i) = \sum_{j=1}^{k} l'_j(i)^j \tag{3}$$

The **second step** consists of building a global surrogate model in order to explain SD outcomes. This methodology specifically employs a decision tree, which is a human-interpretable graphical model that has been widely used by clinicians to represent medical decision knowledge and can potentially be interactive and improved by using visual tools to make it more user-friendly. Moreover, the tree will make it possible to visualise the overlapping between subgroups, as the instances that belong to multiple subgroups will be represented by their own branches on the tree.

Although the SubgroupExplainer methodology states that any decision tree algorithm can be used, we illustrate the suitability of our approach by using the CART al-

gorithm [10]. This decision was motivated by the binary nature of the selectors, which allows them to be used as *splits* for a binary tree.

We propose a new strategy with which to simplify and accelerate the construction of the tree described in Algorithm 1 using the information obtained from the subgroups. The algorithm considers only as possible splits for the nodes the selectors present in the subgroups, and we have prioritised the use of splits (selectors) that are present in generic subgroups (the one with fewer selectors). We additionally decided to split each node using the selector with the highest WRAcc rather than computing the Gini impurity or entropy of the possible splits, thus giving the splits a semantic sense closer to that of the SD framework. When Gini is employed, the branching maximises the classification accuracy of a random instance, whereas our proposal distinguishes between those examples that (according to the WRAcc) belong to the best one-selector subgroup and those that do not.

Algorithm 1 Selection of the SD-split attribute

Input: D', qf, $Output$ ▷ Labeled dataset D', a quality function qf, the set $Output$ of subgroups founded by the black-box algorithm
Output: $split$ ▷ Split attribute

$S_{SG} \leftarrow \emptyset$ ▷ List of all possible one-selector-pattern subgroups of D
$S \leftarrow \emptyset$
$L \leftarrow$ Set of all feasible labeled-classes ▷ Target-value tuples
$i \leftarrow 1$
repeat
  $S \leftarrow$ selectors present in patterns of $Output$ whose $length = i$
  if $S \neq \emptyset$ **then**
    $\forall s \in S,\ \forall l \in L$, add $SG = (s, l)$ to S_{SG} ▷ "IF s THEN l"
    $(s, l) \leftarrow \arg\max_{SG \in S_{SG}} qf(SG, D')$ ▷ "best pattern subgroup"
    $split \leftarrow s$
    return $split$
  else
    $i \leftarrow i + 1$
until $length(out) < i, \forall out \in Output$
return $\varnothing$ ▷ No possible split was found

4 Experiments

Explainable AI is a multi-disciplinary field that involves computer science, human-computer interaction and social sciences [23], signifying that a more complex evaluation that is not limited to numerical experiments is required. Our evaluation method, therefore, comprised three stages: (1) the scalability and computational properties of the proposal were studied using various SD algorithms and gold-standard datasets commonly used in the SD domain, as shown in section 4.1; (2) the suitability of the method was studied by employing it in a clinical use case for an antimicrobial resistance problem, as described in section 4.2, and (3), since the ultimate goal of XAI is for it to be

understood by actual people, the utility of the method was studied by surveying a small group in order to validate it from a human-centric perspective, as explained in section 4.3. The results obtained are preliminary and will be discussed in Section 5.

Table 1. Dataset description

Dataset	$\|D\|$	categorical	numerical	Dataset	$\|D\|$	categorical	numerical
autoMPG8	392	0	6	abalone	4177	0	8
dee	365	0	6	puma32h	8192	0	32
ele-1	495	0	2	elevators	16599	0	18
forestFires	517	0	12	bikesharing	17379	2	10
concrete	1030	0	8	california	20640	0	8
treasury	1049	0	15	house	22784	0	16

With regard to SD algorithms, we selected a local search strategy, since these strategies perform better than exhaustive search algorithms in large databases, such as those used in biomedical and healthcare domains. We specifically selected SD, $CN2\text{-}SD$ and $SD4TS$. The last two algorithms are implemented in the Pyhton3 library `Subgroups`[4], whereas the already the SD algorithm already implemented was improved in terms of execution time efficiency and the quality of the subgroups found. The other experiments were implemented outside `Subgroups`, in a separate repository[5].

4.1 Performance and scalability

The aim of these preliminary experiments was to understand the performance and scalability of our proposal in terms of the complexity of the explanation model - studying the size, branching and complexity of the tree – and its relation to: (1) the SD Algorithm selected, and (2) the characteristics of the subgroups obtained. The study was essential in order to attain a first impression of the usefulness of the tree from a computational perspective. A further evaluation will be carried out in the following sections.

The experiments were carried out using 12 gold-standard datasets for SD analysis [28], which are available in the Keel repository [1] and have a wide range of examples and features. The description of the datasets is summarised in Table 1. While most of them are not related to the healthcare domain, these datasets allowed us to study the viability of our algorithms in numerical terms (number and size of subgroups, size of the trees...). The datasets selected are intended to be used in regression tasks, and the target attribute is, therefore, numerical (discrete or real). [**?**] However, we are interested in classification tasks with a discrete target for our use case and thus we grouped all the possible values of the target into five equal-sized bins that would be used as the new target attribute.

As stated previously, all of the three algorithms selected (SD, $CN2\text{-}SD$, $SD4TS$) use a beam search to traverse the space of solutions, but they require different parameters and use diverse quality measures that shape how the algorithms behave in different

[4] Available at PyPI, `https://pypi.org/project/subgroups/`

[5] Available at GitHub, `https://github.com/Enrique-Val/SubgroupExplainer`

manners. The most noteworthy divergence concerns the quality function. SD uses the Q_g quality function, whose parameter g allows a trade-off between general and specific subgroups, while $CN2$-SD uses a modified version of the WRAcc that considers example weighting (referred in this work as WRAcc'). In [25], the algorithm $SD4TS$ use a domain-specific quality measure, but in our experiments we selected the WRAcc. The selection of parameters is summarised in Table 2.

Since the size of the tree can sometimes grow rapidly, we decided to add an input parameter min_split to the tree. This parameter specifies the minimum number of samples of the total dataset that a node should contain in order to be split. If $min\_split = 0$, we will split each node regardless of the number of examples that it contains and, as a result, the leaf nodes of the tree will be *completely pure*, whereas if $min\_split \geq 0$ some leaf nodes might be *impure* (contain instances with a different label L), but the number of nodes might decrease. This provides an interesting trade-off between accurate trees and small trees that will be discussed in the following sections.

Table 2. Algorithms and parameters used

Algorithm	Quality measure	Beam width	Min. support	Weighting scheme
SD	Q_g ($g = 10$)	20	15%	$None$
$CN2$-SD	WRAcc'	3	$None$ (0%)	Multiplicative ($\gamma = 0.3$)
$SD4TS$	WRAcc	20	15%	$None$

The results of the experiments are depicted in Tables 3 and 4. The study parameters regarding SD will be the number of induced subgroups $|S_{SG}|$, the total number of selectors $|S|$ of the set S_{SG}, the number of non-repeated selectors of S, namely $|S_u|$, and the mean cardinality of the subgroups $card = \frac{|S|}{|S_{SG}|}$. The study parameters of the tree, will be the number of nodes (T), the depth of the tree ($Depth$), the depth of the shortest branch of the tree (Min_depth) and a purity ratio ($Purity$), which is a measure that we defined as the proportion of examples of the training set that have been perfectly classified, i.e. the number of instances that attain a pure leaf node that label them correctly divided by the total number of instances. Two trees will be generated for each SD algorithm: one with the parameter min_split set to 0 and the other with a min_split value of 0.05, thus allowing for some degree of impurity in the leaf nodes.

4.2 Use case: patient phenotype

The objective of this use case is to identify potential patient phenotypes of antimicrobial resistance. In this problem, we analyse the increase in its Minimum Inhibitory Concentration (MIC). The MIC is the lowest concentration of a chemical that will inhibit the growth of a microorganism, and they are considered highly important when determining the susceptibility of bacteria to an antibiotic [2].

For the sake of reproducibility, in this research we used a dataset obtained from the public database MIMIC-III [14], which integrates information from the health records of over 60,000 admissions. The dataset used contains 1280 samples that represent medical episodes. These contain clinical information about the episode registered, such as

Table 3. Results (1)

Dataset	SG alg.	SG metrics				min split	SGExplainer metrics			
		$\|S_{SG}\|$	$\|S\|$	$\|S_u\|$	$card$		$\|T\|$	$Depth$	Min_depth	$Purity$
autoMPG8	SD	20	117	9	5.85	0	9	5	2	1.0
						0.05	9	5	2	1.0
	CN2-SD	24	63	33	2.62	0	413	17	5	1.0
						0.05	73	13	4	0.16
	SD4TS	20	**67**	**13**	**3.35**	0	**5**	**3**	**2**	1.0
						0.05	5	3	2	1.0
dee	SD	20	112	10	5.6	0	9	5	2	1.0
						0.05	9	5	2	1.0
	CN2-SD	24	68	33	2.83	0	457	17	5	1.0
						0.05	79	12	4	0.09
	SD4TS	20	68	11	3.4	0	59	8	4	1.0
						0.05	43	8	3	0.81
ele-1	SD	11	20	4	1.82	0	17	5	3	1.0
						0.05	17	5	3	1.0
	CN2-SD	**29**	**59**	**15**	2.03	0	**49**	**9**	**3**	1.0
						0.05	49	9	3	1.0
	SD4TS	20	55	7	2.75	0	13	5	2	1.0
						0.05	13	5	2	1.0
forestFires	SD	20	25	16	1.25	0	455	15	5	1.0
						0.05	91	15	3	0.09
	CN2-SD	**29**	**84**	**56**	2.9	0	953	22	6	1.0
						0.05	93	12	4	0.01
	SD4TS	20	27	19	1.35	0	671	14	6	1.0
						0.05	65	8	4	0.0
concrete	SD	20	88	8	4.4	0	49	8	3	1.0
						0.05	39	7	3	0.9
	CN2-SD	38	107	40	2.82	0	945	19	5	1.0
						0.05	73	9	4	0.0
	SD4TS	20	26	17	1.3	0	711	14	6	1.0
						0.05	65	8	4	0.0
treasury	SD	**20**	**117**	**24**	**5.85**	0	**21**	**8**	**2**	**1.0**
						0.05	**15**	**8**	**2**	**0.97**
	CN2-SD	19	51	39	2.68	0	273	21	4	1.0
						0.05	97	16	3	0.34
	SD4TS	20	25	20	1.25	0	127	16	4	1.0
						0.05	**57**	16	3	**0.82**

the sex and age of the patient, the month, year and season of the medical episode, the episode ID (the same patients that have the same episode ID) or the duration of the episode. Key attributes of the dataset are: (1) Microorganism: the bacteria observed in the study of the MIC; (2) Susceptibility: the microbiological study of the reaction of bacteria to an antibiotic; (3) MIC Increases: whether or not the Minimum Inhibitory Concentration increases. If it increases, this means that the susceptibility was lower in a previous observation, and (4) Culture service: the hospital service (ICU, cardiology, etc) that requested the bacterial culture used in the study. The problem falls in the category of binary classification, as we are interested to study if the MIC increases or not.

Table 5 provides a summary of the dataset used for the experiments. The EiD column shows the duration of the patient episode in days.

We are interested in studying the subgroups of a population in which there is a high chance of a microbiological resistance being developed. We use the $CN2\text{-}SD$ algorithm instead of the less sophisticated SD and $SD4TS$. According to our experi-

Table 4. Results (2)

Dataset	SG alg.	SG metrics				min split	SGExplainer metrics			
		$\|S_{SG}\|$	$\|S\|$	$\|S_u\|$	$card$		$\|T\|$	$Depth$	Min_depth	$Purity$
abalone	SD	**20**	**96**	**14**	**4.8**	0	**9**	**5**	**2**	1.0
						0.05	9	5	2	1.0
	CN2-SD	**35**	**95**	**40**	**2.71**	0	**945**	22	5	**1.0**
						0.05	**135**	18	3	**0.32**
	SD4TS	20	20	20	1	0	425	13	5	1.0
						0.05	51	13	3	0.61
puma32h	SD	20	61	5	3.05	0	21	6	3	1.0
						0.05	21	6	3	1.0
	CN2-SD	40	84	37	2.1	0	12097	26	6	1.0
						0.05	49	9	3	0.0
	SD4TS	20	21	12	1.05	0	351	10	6	1.0
						0.05	69	10	4	0.28
elevators	SD	20	111	8	5.55	0	9	5	2	1.0
						0.05	9	5	2	1.0
	CN2-SD	53	150	63	2.83	0	20757	30	6	1.0
						0.05	89	13	3	0.0
	SD4TS	20	20	20	1.0	0	135	9	5	1.0
						0.05	49	9	3	0.84
bikesharing	SD	20	72	8	3.6	0	5	3	2	1.0
						0.05	5	3	2	1.0
	CN2-SD	20	48	22	2.4	0	309	13	3	1.0
						0.05	63	10	3	0.58
	SD4TS	20	57	7	2.85	0	7	4	2	1.0
						0.05	7	4	2	1.0
california	SD	20	35	11	1.75	0	471	12	6	1.0
						0.05	77	12	3	0.4
	CN2-SD	47	115	45	2.45	0	5075	23	6	1.0
						0.05	79	10	4	0.0
	SD4TS	20	51	8	2.55	0	65	7	3	1.0
						0.05	25	7	3	0.79
house	SD	20	83	9	4.15	0	169	10	3	1.0
						0.05	39	8	3	0.63
	CN2-SD	62	164	71	2.65	0	40335	33	6	1.0
						0.05	85	13	4	0.0
	SD4TS	20	51	8	2.55	0	65	7	3	1.0
						0.05	25	7	3	0.79

ments (see Section 4.1), $CN2\text{-}SD$ tends to induce subgroups with a higher variety of non-repeated selectors. We used the same parameters as in the previous section (see Table 2). In addition, we focused the search ton only look for subgroups with the target $MIC\ Increases = yes$ rather than subgroups with any target value, since we were specifically interested in defining the population in which the resistance is prone to appear. This can be easily done in $CN2\text{-}SD$, since it launches an individual beam search procedure for every target value. As shown in Tables 3 and 4, the use of $CN2\text{-}SD$ algorithm results also in larger tree sizes, and we accordingly placed the threshold of instances on a node for it to be split by 5%. The numeric results and subgroups obtained can be seen in Table 6, while a visual representation of the tree is shown in Figure 2.

4.3 Human subjective study

Interpretations are social [23] and thus have a strong cognitive and subjective factor. The usefulness of an interpretation can depend on many variables, such as the users'

Table 5. Dataset: Minimum inhibitory concentration (only 2 rows shown)

sex	age	season	month	year	episode Id	microorganism	EiD	susceptibility	MIC Increases	culture service
F	ELDERLY	SPRING	6	2015	10119	S. EPIDERMIDIS	1	SENSIBLE	Yes	TRA
M	ADULT	SUMMER	8	2016	12731	S. COAGULASA NEG.	1	SENSIBLE	No	ORL
.	...	...	.	...	...	...	.	...	.	...

Table 6. MIC detection results

SG metrics				min split	SGExplainer metrics			
$\|S_{SG}\|$	$\|S\|$	$\|S_u\|$	$card$		$\|T\|$	$Depth$	Min_depth	$Purity$
4	12	7	4.95	0	17	7	3	1.0
				0.05	15	7	3	0.98

IF microorg not E. FAECALIS, microorg not E. FAECIUM AND microorg not MARSA THEN MIC Increases
IF age not NEWBORN, microorg not S. AUREUS, microorg not S. COAGULASE NEG. THEN MIC Increases
IF microorg not E. FAECALIS, microorg not MARSA, microorg not S. EPIDERMIDIS THEN MIC Increases
IF microorg not E. FAECIUM, microorg not S. AUREUS, microorg not S. EPIDERMIDIS THEN MIC Increases

academic background, their cognitive abilities or their knowledge of ML and statistics. It is insufficient to study the characteristics of an explanation by simply looking at the numbers obtained. We, therefore, decided to carry out a survey that would allow us to study the efficacy of a decision tree with which to explain SD.

The study was carried out with a total of 18 participants, most of whom (90%) had University studies and were between 18 and 25 years old. Furthermore, 70% of the respondents had a background in Computer Science and 55% of them (a third of the total population) had knowledge of Artificial Intelligence and ML. However, SD is a highly specific task that is not usually taught at universities, and is, therefore, known only by ML practitioners.

In the survey, we first provided an intuitive explanation of what a subgroup and SD are, and the idea of SubgroupExplainer, after which we presented a set of five subgroups and posed a control question in order to validate whether the respondents had understood what the key concepts of SD are. Two thirds of the population understood the intuitive concept of subgroup, while the other third assumed that the members of a subgroup always had to have the value of the target specified by the subgroup description (which, although desirable, does not always hold true).

Three SubgroupExplainer trees were then presented, each with a different threshold with which to divide a node. The first had no threshold, in the second, a node had to have at least 15% of the examples and in the third, 30%. There were two questions for each tree regarding whether an individual belonged to the subgroups presented, showing descriptions of both the subgroup and the tree. Domain-specific and more subjective questions, such as identifying key features, were not included, since we were focusing on the actual comprehension of the subgroups. After this objective question, we then asked the users whether they found the tree helpful, complete (the tree is sufficient to be able to visualise and classify the subgroup) and simple, rating each tree along these axes on a scale of 1 to 5, where 1 was “Not at all” and 5 was “Very”.

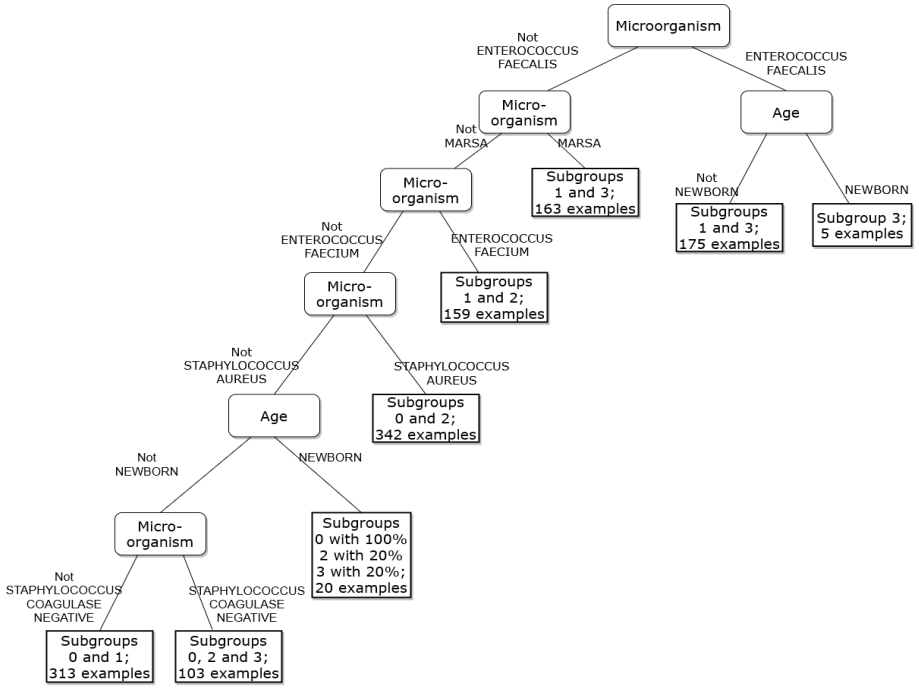

Fig. 2. SD Explanation tree: global surrogate model

With regard to the objective question, the percentage of correct answers varied depending on the tree: (1) in the case of the first, 86% of the answers given were correct once the average for the two questions had been obtained; (2) in that of the second, 94% of the respondents answered correctly, and (3), in that of the third, 17% answered the first question correctly and 94% answered the second correctly (an average of 56%).

With regard to the subjective evaluation of the explanation, the results are shown in Table 7, in which the overall results and the answers given by AI/ML practitioners are separated. As will be noted, the average user prefers a tree with a certain trade-off between node purity and tree size (second tree). In contrast, AI students prefer a precise tree, even if it has more nodes. They actually rate the precise (first) tree as being less simple than the second, despite the fact that the second has more nodes. Both groups rate the highly-pruned tree (the third) as being the least helpful, complete and simple, showing that, overall, a more accurate tree is preferred to a small one.

5 Discussion

5.1 Scalability of the SD algorithms

With regard to the the results of the scalability analysis described in Tables 3 and 4 , it is worth noting that the $CN2$-SD discovers more subgroups because it is not limited to

Table 7. Summary about the mean user experience with the SubgroupExplainer trees.

Polled	Tree	Helpful	Complete	Simple	Mean Help.-Simple	Mean Compl.-Simple
All	no min. samples	4.94	4.56	4.06	4.5	4.32
	min. samples = 15%	4.89	4.61	4.28	4.58	4.44
	min. samples = 30%	4.06	3.89	4	4.03	3.94
AI/ML knowledge	no min. samples	4.83	4.33	4.33	4.58	4.33
	min. samples = 15%	4.83	4.33	4	4.42	4.17
	min. samples = 30%	4.33	3.83	3.83	4.08	3.83

20 subgroups, as opposed to SD and $SD4TS$. While the algorithm that generates more or less selectors is dependent on the dataset, it will be noted that the variety of selectors induced is always higher in $CN2$-SD, and that the SD algorithm discovers subgroups with a higher selector cardinality in 9 out of 12 experiments.

5.2 Decision trees as explainers

The trees built with the subgroups obtained with $CN2$-SD are the largest, while those obtained with SD have a slight tendency towards being the smallest. In the cases, the trees tend to be very imbalanced - that is, there is a considerable difference between the longest and shortest branch. As expected, the purity ratio when min_split is set to 0 is maximum, since all the nodes are pure and the tree perfectly classifies the instances into subgroups. When min_split is set to 0.05 the trees that have a higher number of nodes are those that are reduced the most, possibly owing to an inefficient branching that occurs when attempting to classify very specific and unique instances. The reduction in the size of the tree comes at the cost in its $Purity$, which tends to be higher in the trees obtained with the $CN2$-SD subgroups.

In the case of the relation between the characteristics of each set of subgroups obtained and the trees, it is possible to observe that the size and depth of the tree is independent of the number of subgroups induced and the cardinality. Although we cannot state that there is a clear dependence between the number of unique selectors and the size of the tree, the probability of the tree growing larger seems to be higher when the number of unique selectors is also large.

Our use case (Section 4.2) proves the potential of compacting all the subgroups are compacted into a single tree. It helps to highlight the attributes that make it possible to better discriminate between subgroups, which are the attributes that are in a higher position in the tree and that are selected more frequently as $split$. This specific use case shows us that if the microorganism is an Enterococcus Faecalis or MARSA, the individual will not belong to subgroups 0 and 2.

Visualising the tree makes it possible to see the imbalance previously identified in Section 4.1. The tree contains only 15 nodes, a quantity of information that is usually easy to handle, but the imbalance makes the tree look large and complex, thus potentially limiting the explanation.

5.3 Understanding of the subgroups by humans

Concerning the opinions gathered from the users in Section 4.3, we highlight that most people find it difficult to understand the fact that an individual may belong to two subgroups. The tree can be helpful as regards solving this problem, as it can explicitly show that an individual with certain characteristics can belong to multiple subgroups.

The AI/ML practitioners' preference for the larger (although more accurate) decision tree can be explained by their greater familiarity with ML models. Analogously, non ML students have favoured a smaller tree for the sake of simplicity, even if it was not as accurate because it is a completely new concept for them.

6 Conclusions

The objective of this paper is to provide clinicians with tools that will allow them to better understand SD algorithms and their outcomes for patient phenotyping. We propose SubgroupExplainer, a methodology that provides SD model-agnostic explanations. This method is based on the hypothesis that decision trees are an effective approach by which to provide global surrogate explanations for medical problems. We have evaluated the suitability of our proposal by studying the ML pipeline and by providing a clinical use case and a human-centric analysis.

Unlike the state-of-the-art approaches [32,33], SubgroupExplainer does not require additional knowledge (ontologies) in order to generate explanations. As a result, the subgroup explanations might be less compact than those obtained using higher level concepts of an ontology, but structuring the partition of the space in a tree-like form is still helpful as regards understanding the data and the subgroups.

While interviewing clinicians would have been more helpful for our study, the results obtained with our current sample are still helpful in order to validate the usefulness of the explanations of SD.

In future research it will be necessary to analyse both the explanatory potential of n-ary decision trees, as well as looking for correlations between the characteristics of the subgroups and the size of the tree. Even if subgroup discovery is not strictly designed for classification, a comparison between the accuracy of the subgroups and the surrogate tree would be another method to examine its fidelity. A baseline comparison with a tree whose split are found using the Gini impurity would also be a valuable addition. The trees could be further improved by using visual and interactive keys, such as colouring the nodes and branches and allowing user interaction. From a practical perspective, we plan to extend the study to clinicians working in MIC detection or related problems.

Acknowledgement

This work was partially funded by the CONFAINCE project (Ref: PID2021-122194OB-I00), supported by the Spanish Ministry of Science and Innovation, the Spanish Agency for Research and the IMPACT-T2D project (PMP21/00092) supported by the Spanish Health Institute Carlos III (ISCIII).

References

1. Jesús Alcalá-Fdez, Alberto Fernández, Julián Luengo, Joaquín Derrac, Salvador García, Luciano Sánchez, and Francisco Herrera. KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing*, 17(2-3):255–287, 2011.
2. Jennifer M Andrews. Determination of minimum inhibitory concentrations. *Journal of antimicrobial Chemotherapy*, 48(Suppl. 1):5–16, 2001.
3. Martin Atzmueller. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(1):35–49, 2015.
4. Martin Atzmueller and Frank Puppe. SD-map – a fast algorithm for exhaustive subgroup discovery. In *Lecture Notes in Computer Science*, pages 6–17. Springer Berlin Heidelberg, 2006.
5. Cristóbal J Carmona, María José del Jesus, and Francisco Herrera. A unifying analysis for the supervised descriptive rule discovery via the weighted relative accuracy. *Knowledge-Based Systems*, 139:89–100, 2018.
6. Mark Craven and Jude Shavlik. Extracting tree-structured representations of trained networks. *Advances in Neural Information Processing Systems*, 8:24–30, 1995.
7. Sanjeeb Dash, Oktay Gunluk, and Dennis Wei. Boolean decision rules via column generation. *Advances in Neural Information Processing Systems*, 31:4655–4665, 2018.
8. Federica Di Castro and Enrico Bertini. Surrogate decision tree visualization interpreting and visualizing black-box classification models with surrogate decision tree. In *Joint Proceedings of the ACM IUI 2019 Workshops co-located with the 24th ACM Conference on Intelligent User Interfaces*, volume 2327 of *CEUR Workshop Proceedings*. CEUR-WS, 2019.
9. Dragan Gamberger and Nada Lavrac. Expert-guided subgroup discovery: Methodology and application. *Journal of Artificial Intelligence Research*, 17:501–527, 2002.
10. A. D. Gordon, L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and regression trees. *Biometrics*, 40(3):874, 1984.
11. Henrik Grosskreutz, Stefan Rüping, and Stefan Wrobel. Tight optimistic estimates for fast subgroup discovery. In *Machine Learning and Knowledge Discovery in Databases*, pages 440–456. Springer Berlin Heidelberg, 2008.
12. Sumyea Helal. Subgroup discovery algorithms: a survey and empirical evaluation. *Knowledge and Information Systems*, 3(29):495—525, 2011.
13. Lujain Ibrahim, Munib Mesinovic, Kai-Wen Yang, and Mohamad A. Eid. Explainable Prediction of Acute Myocardial Infarction Using Machine Learning and Shapley Values. *IEEE Access*, 8:210410–210417, 2020.
14. Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):1–9, 2016.
15. Willi Klösgen. Explora: A multipattern and multistrategy discovery assistant. In *Advances in Knowledge Discovery and Data Mining*, pages 249–271. AAAI/MIT Press, 1996.
16. Nada Lavrač, Peter Flach, and Blaz Zupan. Rule evaluation measures: A unifying view. In *International Conference on Inductive Logic Programming*, pages 174–185. Springer, 1999.
17. Nada Lavrac, Branko Kavsek, Peter Flach, and Ljupco Todorovski. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5(2):153–188, 2004.
18. Florian Lemmerich, M. Rohlfs, and M. Atzmueller. Fast discovery of relevant subgroup patterns. In *Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pages 428–433. AAAI Press, 2010.
19. C. Lonjarret, C. Robardet, M. Plantevit, R. Auburtin, and M. Atzmueller. Why should I trust this item? Explaining the recommendations of any model. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 526–535, 2020.

20. Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020.
21. Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:4765–4774, 2017.
22. Pavan Rajkumar Magesh, Richard Delwin Myloth, and Rijo Jackson Tom. An explainable machine learning model for early detection of Parkinson's disease using LIME on DaTSCAN imagery. *Computers in Biology and Medicine*, 126:104041, 2020.
23. Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.
24. Christoph Molnar. *Interpretable Machine Learning*. Lulu.com, 2019.
25. Marianne Mueller, Rómer Rosales, Harald Steck, Sriram Krishnan, Bharat Rao, and Stefan Kramer. Subgroup discovery for test selection: a novel approach and its application to breast cancer diagnosis. In *International Symposium on Intelligent Data Analysis*, pages 119–130. Springer, 2009.
26. Petra Kralj Novak, Nada Lavrač, and Geoffrey I Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10(2):377–410, 2009.
27. Vili Podgorelec, Peter Kokol, Bruno Stiglic, and Ivan Rozman. Decision trees: An overview and their use in medicine. *Journal of Medical Systems*, 26(5):445–463, 2002.
28. Hugo M. Proença, Peter Grünwald, Thomas Bäck, and Matthijs van Leeuwen. Discovering outstanding subgroup lists for numeric targets using MDL. In *Machine Learning and Knowledge Discovery in Databases*, pages 19–35. Springer International Publishing, 2021.
29. J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.
30. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.
31. Matthijs van Leeuwen and Arno Knobbe. Non-redundant subgroup discovery in large and complex data. In *Machine Learning and Knowledge Discovery in Databases*, pages 459–474. Springer Berlin Heidelberg, 2011.
32. Anže Vavpetič, Vid Podpečan, and Nada Lavrač. Semantic subgroup explanations. *Journal of Intelligent Information Systems*, 42(2):233–254, 2013.
33. Anže Vavpetič, Vid Podpečan, Stijn Meganck, and Nada Lavrač. Explaining subgroups through ontologies. In *Lecture Notes in Computer Science*, pages 625–636. Springer Berlin Heidelberg, 2012.
34. Sebastián Ventura, José María Luna, et al. *Supervised Descriptive Pattern Mining*. Springer, 2018.
35. Dennis Wei, Sanjeeb Dash, Tian Gao, and Oktay Gunluk. Generalized linear rule models. In *International Conference on Machine Learning*, pages 6687–6696. Proceedings of Machine Learning Research, 2019.
36. Stefan Wrobel. An algorithm for multi-relational discovery of subgroups. In *European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 78–87. Springer, 1997.