
CONCLUSIVE LOCAL INTERPRETATION RULES FOR RANDOM FORESTS

A PREPRINT

Ioannis Mollas
Aristotle University of
Thessaloniki, 54636, Greece
iamollas@csd.auth.gr

Nick Bassiliades
Aristotle University of
Thessaloniki, 54636, Greece
nbassili@csd.auth.gr

Grigorios Tsoumakas
Aristotle University of
Thessaloniki, 54636, Greece
greg@csd.auth.gr

April 14, 2021

ABSTRACT

In critical situations involving discrimination, gender inequality, economic damage, and even the possibility of casualties, machine learning models must be able to provide clear interpretations for their decisions. Otherwise, their obscure decision-making processes can lead to socioethical issues as they interfere with people's lives. In the aforementioned sectors, random forest algorithms strive, thus their ability to explain themselves is an obvious requirement. In this paper, we present LionForests, which relies on a preliminary work of ours. LionForests is a random forest-specific interpretation technique, which provides rules as explanations. It is applicable from binary classification tasks to multi-class classification and regression tasks, and it is supported by a stable theoretical background. Experimentation, including sensitivity analysis and comparison with state-of-the-art techniques, is also performed to demonstrate the efficacy of our contribution. Finally, we highlight a unique property of LionForests, called conclusiveness, that provides interpretation validity and distinguishes it from previous techniques.

Keywords Explainable Artificial Intelligence · Interpretable Machine Learning · Local Interpretation · Model-Specific Interpretation · Random Forests

1 Introduction

It is apparent that machine learning (ML) models will be integrated into our society and daily life. However, in critical domains such as the aforementioned, where models can contain errors, or may suffer from biases, it is more than necessary to ensure their transparency. Gender inequality [1], inappropriate patient treatments [2] or easily tricked models [3] are only a few common problems when automated systems are used. Thereby, it is vital for secure, fair, and trustworthy intelligent systems to be able to explain how they work and why they predict a particular outcome. In addition, guaranteeing certain virtues on the behaviour of a model directly allows them to be compliant with legal frameworks, such as the General Data Protection Regulation (GDPR) [4] of the EU and the Equal Credit Opportunity Act of the US¹. These needs introduced a new area called explainable artificial intelligence (XAI) in the research community [5]. Interpretable machine learning (IML), a subfield of XAI, attempts to address these issues, proposing techniques to shed light into the inner workings of ML models [6, 7, 8].

Random forest (RF) is a highly accurate learning algorithm [9] that has been proven robust to overfitting [10], as well as to learning difficulties arising from phenomena such as class imbalance, or noisy and anomalous data [11]. RF excels in a lot of sectors. From applications related to security, such as intrusion detection [12], to the financial sector dealing with tasks including credit card fraud detection [13] and loan approval [14]. The presence of RF algorithms is also vivid in healthcare applications, like the patient safety culture [15] and the classification of different stages of Parkinson's disease [16]. Finally, in industry and law sectors, applications such as fault diagnosis in self-aligning conveyor idlers [17] and prediction of crime hotspots [18] are few examples among many.

¹EOCA 15 U.S. Code §1691 et seq.

The positive aspects of RFs in conjunction with their black box nature have drawn the attention of the IML research community, which introduced a variety of techniques to provide interpretation in the form of rules, trees, or feature importance. For RF models in particular, interpretation techniques follow two main directions. The first one is to create a surrogate model, intending to distil the knowledge of a complex RF model into a single tree [19, 20]. The other direction is to take full advantage of the inner structure of the RF and derive information from the individual trees that make it up [21, 22, 23]. However, these approaches have significant limitations. For example, most of them are only applicable to binary classification tasks [22, 23]. Another point to consider is that the interpretations produced are not always valid [21, 19]. This is particularly evident in techniques that seek to approximate the actual interpretations of a complex model.

This work presents LionForests (LF), an approach for interpreting individual predictions of RF models. LF performs path and feature reduction towards smaller interpretation rules with wider ranges, by using unsupervised techniques, such as association rules and k -medoids clustering, as well as a path-oriented dissimilarity metric. We extend our preliminary work [23] in multiple dimensions. First of all, we provide a stable theorem and a property, called conclusiveness, to support the validity of the produced rules. Besides that, we broaden LF's capabilities making it applicable to multi-class classification tasks, and we introduce a new concept, the “*allowed error*”, to render the technique applicable to regression tasks as well. In addition, we optimise LF's performance in terms of response time per produced interpretation. Further parameterisation options are also implemented in the technique, concerning the association rules and clustering algorithms. An improved and clearer method for dealing with categorical features is also provided. Moreover, a simple visualisation functionality is designed to increase the expressiveness and clarity of the generated interpretation rules. Experiments including sensitivity analysis, analysis of response time and comparison with other state-of-the-art (SOTA) interpretation techniques, as well as qualitative examples of actual interpretations, are provided to support LF's efficiency. Finally, we are investigating whether SOTA techniques have the same conclusiveness property as LF.

The remainder of this paper is structured as follows. Section 2 presents the related work, while Section 3 establishes the theoretical background. Section 4 describes the LF approach. Exhaustive experimentation from multiple perspectives, such as sensitivity analysis, time analysis, comparison with other methods, and qualitative assessments, is provided in Section 5. Section 6 presents an analysis of the experimental results. Finally, in Section 7 we discuss conclusions and future directions.

2 Related work

Over the last few years, IML has advanced so much, making the range of solutions provided wider than ever. Of the several dimensions of interpretability, those primarily used for categorisation and comparison of such approaches refer to the global-local aspect of the model and the agnostic-specific applicability of the technique to a model or architecture. Local-based techniques concern the interpretation of predictions for a single instance, while global-based techniques uncover the entire structure of a model. Both local and global based approaches include techniques that are either model-agnostic or model-specific. Model-agnostic approaches have the potential to interpret any machine learning model indifferently, while model-specific techniques interpret particular types of models, or even architectures.

Another interesting aspect of interpretation techniques is the form in which they deliver their results. It is possible for a technique to provide an explanation as a set of rules, a set of feature weights, images with skewed or highlighted sections, or prototype examples. In this section, we will present a few techniques related to these dimensions, in the context of ensemble models, such as random forests. These techniques share the same rule-based interpretation form. This means that the output of these techniques is a set of rules, or single rules, explaining a model or a specific prediction.

Surrogate models [24], based on the principle of model compression [25], are a global-based, model-agnostic interpretation technique that attempts to imitate the behaviour of more complex models. A decision tree (DT), for example, is a surrogate model of an RF model, when trained on training data, labelled by the RF model. Metrics such as fidelity emerged in order to evaluate the ability of these models to mimic the original models. High fidelity means the approximation is sufficient. However, it is still questionable how well the surrogate models should approximate black-box models in order to be trusted.

Anchors [26] and LORE [27], local-based model-agnostic techniques leveraging sub-spaces of the original data, create local surrogate models to explain particular instance predictions. Specifically, Anchors provides a single rule for interpreting the prediction of a single instance. This rule is the anchor that keeps the prediction (almost, as the authors acknowledge) always the same. Anchors is only applicable to datasets with discrete features. On the other hand, for each instance, LORE creates a local neighbourhood using genetic algorithms to construct a surrogate decision tree from which a single rule interpretation and counterfactual instances are extracted.

The single-tree approximation approaches like inTrees [20] and defragTrees [19] among others [28, 29], are global model-specific techniques that interpret tree ensembles by approximating the output of the model they attempt to explain through a single tree. However, this procedure is highly problematic because it is not feasible to summarise a complex model like tree ensembles to a single simple tree, as reported by other researchers [30].

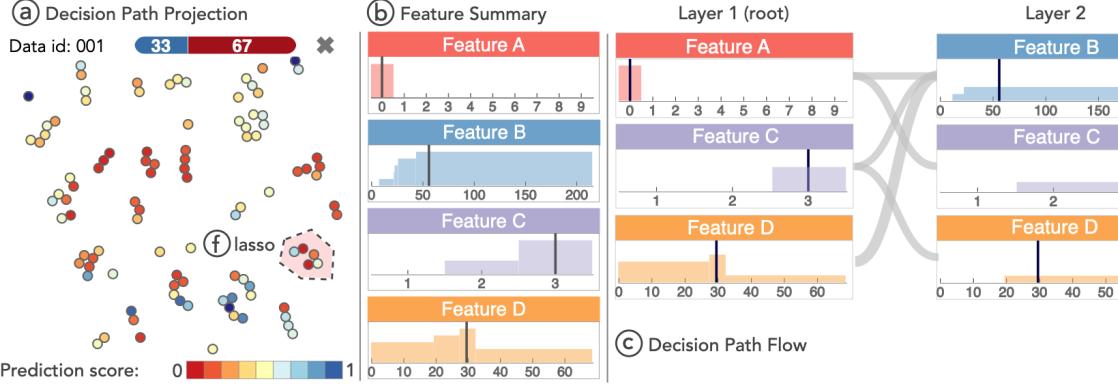


Figure 1: A sample example from the visualisation tool of iForest adapted from [31]

iForest [31] is a visualisation method that offers global and local interpretations of RF models. The most interesting aspect of this tool, though, concerns local explanations. Using a path distance metric, iForest projects the paths of an instance to a two-dimensional space (Figure 1a) using t-Distributed Stochastic Neighbour Embedding (t-SNE) [32], and it provides a feature summary (Figure 1b) and a decision path flow (Figure 1c), as well. iForest requires user input, in the form of drawing a lasso around a set of paths (Figure 1f), to produce the path flow. This is a disadvantage, because the user may give wrong input, leading to an incorrect feature summary and path flow, which will result in a flawed interpretation.

Class A (95.1%)				
Rank	Feature	Influence	Min	Max
1	Feature D	+74.9	7073.5	7074
2	Feature A	+22.0	12.5	∞
3	Feature C	+19.3	0.5	1
1	Feature E	-12.27	$-\infty$	0.5
2	Feature B	-0.73	0.25	0.5
3	Feature I	-0.60	$-\infty$	174

Figure 2: Template of explanation of Moore et al. [22]

Decision	Explanation	Contrast (%)	Confidence
Class A	Feature D $\leq 7074 \wedge$	-80.6	Covers 39.4% of historical
	Feature C $\neq 1 \wedge$	-24.9	Matches 97.3 of covered
	Feature A ≤ 20	-16.9%	Vote margin 40.0%

Figure 3: Template of explanation of CHIRPS [21]

Another local model-specific interpretation strategy [22] interprets RF models by providing a collection of features with their ranges, rated on the basis of their importance, as explanations (Figure 2). The process of interpretation consists of two steps. First, they measure the effect of a feature j on the prediction for a given instance x , and later this effect will be used for the ranking process. To accomplish this, a node per tree monitoring mechanism is used to find the aggregated effect of all the features for the prediction of a particular instance. The second step is to identify the narrowest range for each feature over all trees.

Finally, CHIRPS [21] proposes a technique for multi-class tasks, using frequent pattern (FP) mining on the paths of the majority class to identify the most influencing features and present them to the end user. CHIRPS promises interpretations (Figure 3) that will be minimally complete, providing information about counterfactual cases, and referring to real data, and not synthetic data. A drawback of this approach is that it lacks a strict restriction about the number of paths that will be covered through the generated rule, particularly in multi-class classification tasks. An extension of CHIRPS to gradient boosted tree ensembles is gbt-HIPS [33].

LF overcomes the issues of single-tree approximations, does not require user input like iForest, and comes with low computational cost in contrast to Anchors. Most importantly, LF provides rules that are always valid. Finally, it is applicable in a wider range of machine learning tasks compared to Anchors, LORE and CHIRPS, among other competitors.

3 Main concepts and notation

We here define the main concepts and notation concerning decision trees and random forests, which are necessary for the presentation of LF in the next section.

3.1 Decision trees

Decision Trees (DT) [34] is a classic machine learning algorithm that stimulates the growth of the Ensemble algorithms. A DT can be shown as an acyclic directed graph, as seen in Figure 4, containing a root node, decision nodes and leaf nodes, which are the prediction nodes. Regarding the learning algorithm, each node concerns a particular feature f_i and a condition relation. In the case of input instances, the decision tree traces the path to a leaf node containing a prediction. The prediction can be a class, in the case of a binary or multi-class classification, or a real number estimate, in the scenario of a regression task.

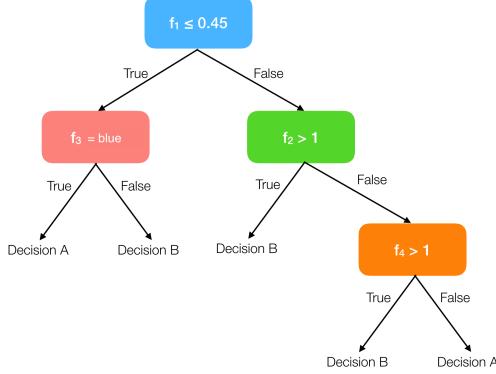


Figure 4: A simple decision tree classifier with 4 features

Each decision path p is a conjunction of conditions, and the conditions are features and values with relations \leq , $>$ and $=$. To provide an illustration, a path from the tree on Figure 4 would be: ‘if $f_1 > 0.45$ and $f_2 \leq 1$ and $f_3 = \text{blue}$ then *Decision A*’. Thus, each path p is expressed as a set:

$$p = \{f_i \boxtimes v_j | f_i \in F, ((\boxtimes \in \{\leq, >\} \rightarrow v_j \in \mathbb{R}) \wedge (\boxtimes \in \{=\} \rightarrow v_j \in S_j))\} \quad (1)$$

where v_j the instance’s value for the feature f_i , and S_j is the set of the categorical values if f_i is a discrete feature.

However, LionForests implementation relies on a library², which uses an optimised version of DTs, where the categorical features must be either encoded to numerical features, with encoding procedures like OneHot or Ordinal encoding. Hence, in the rest of the paper each path p is expressed as a set:

$$p = \{f_i \boxtimes v_j | f_i \in F, v_j \in \mathbb{R}, \boxtimes \in \{\leq, >\}\}. \quad (2)$$

3.2 Random forests

One of the first ensemble algorithms using DTs as foundation is the Random Forests (RF) [10] algorithm. RF is a collection of a specific number of trees, which are combined under an equal voting scheme. Abstractly, the inference process of an RF model can be seen as a voting procedure where the outcome (prediction) corresponds to the majority of the votes cast.

These trees are trained under different data and feature partitions, towards higher variance and lower bias, dealing with the overfitting problem. Then, for a specific prediction the trees are voting:

$$h(x_i) = \frac{1}{|T|} \sum_{t \in T} h_t(x_i) \quad (3)$$

where $h_t(x_i)$ is the vote cast from the tree $t \in T$ for the instance $x_i \in X$, representing the probability $P(C = c_j | X = x_i)$ of x_i to be assigned to class $c_j \in C$, and finally choosing the $\text{argmax}_{c \in C} P(C = c | X = x_i)$ for classification problems, and $h_t(x_i) \in \mathbb{R}$ in regression tasks.

²We use scikit-learn as core library (<https://scikit-learn.org>)

4 Our approach

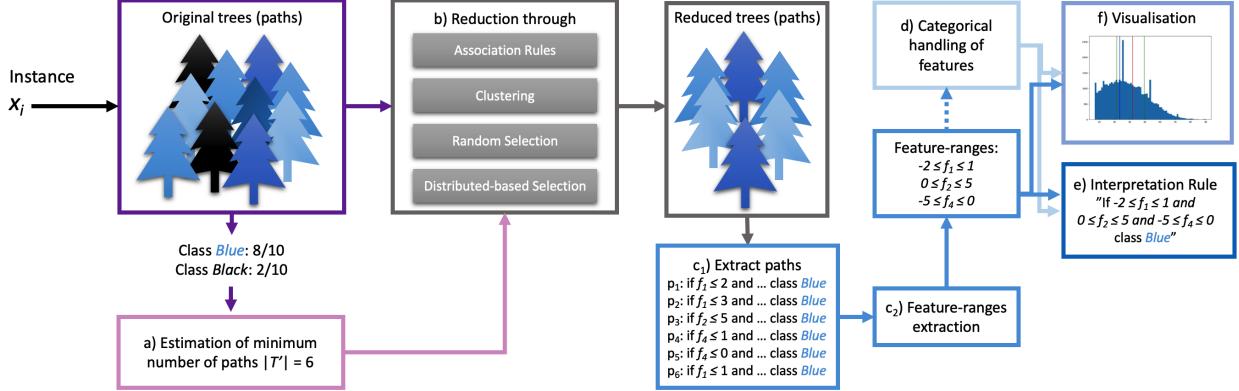


Figure 5: LionForests architecture

We present the core of LionForests (LF), along with its extensions, which concern primarily the theoretical grounding, the optimisation of the main algorithm, as well as the adaptation from binary classification to multi-class and regression tasks.

LF is a technique for local interpretation of binary, multi-class and regression RF models. This is achieved via the following sequence of actions: a) estimation of minimum number of paths, b) reduction through association rules, clustering, random selection, or distribution-based selection, c) extraction of feature-ranges, d) categorical handling of features and, ultimately, e) composition of interpretation and f) visualisation, as depicted in Figure 5. Before presenting to these actions, we introduce a new property called *conclusiveness* (in the following subsection). In order to produce rules with fewer features and broader ranges, LF first estimates the minimum number of paths that has to be maintained to produce *conclusive* rules. Then, it reduces the redundant paths and extracts the feature-ranges from the rest. In case of categorical features, a related procedure takes place. Finally, a single rule in natural language is provided, as well as a visualisation of the rule's details.

4.1 Conclusiveness property

conclusiveness - the quality of being final or definitely settled

We here introduce a property which concerns an aspect of a rule's quality, which we call *conclusiveness*. Thus, in a scenario with a dataset with e.g. 4 features ($F = [f_1, f_2, f_3, f_4]$), a predictive model h , an instance x with values for the corresponding features $[v_1, v_2, v_3, v_4]$, and the prediction c_k , $h(x) = c_k$, we have a rule “ $r = \text{if } 2 \leq f_1 \leq 10 \text{ and } 0.5 \leq f_3 \leq 1.2 \text{ then } c_k$ ” containing conditions with 2 of them ($F' = [f_1, f_3]$).

$$r = \{f_i \boxtimes v_j \rightarrow c_k | f_i \in F, v_j \in \mathbb{R}, \boxtimes \in \{\leq, >\}\}. \quad (4)$$

$$\text{conclusive}(r, c_k) = \begin{cases} \text{False}, & \exists v_l \in D(f_i), f_i = v_l \boxtimes v_j | f_i \boxtimes v_j \in r, h(x'_{f_i=v_l}) \neq c_k \\ \text{False}, & \exists v_l \in D(f_i), f_i = v_l | f_i \notin r, h(x'_{f_i=v_l}) \neq c_k \\ \text{True}, & \text{elsewhere} \end{cases}$$

, where $D(f_i)$ is the domain of feature f_i , and $x'_{f_i=v_l}$ is an alteration of instance x but with different value (v_l) for a specific feature f_i . Therefore, a rule will be conclusive if and only if the following restrictions are met:

1. if the instance's values for the excluded features are modified to any possible value the prediction will not be influenced,
2. if the value of one of the included features is modified within the specified range the prediction will not be influenced.

A technique that always produces conclusive rules owns the property of conclusiveness. As presented in the experiments of Section 5, however, a conclusive rule contains more features than an inconclusive one, increasing the rule length, and is specific to a particular instance, reducing the rule's coverage.

4.2 Estimation of minimum number of needed paths

LF attempts to identify and use only a subset of the paths that voted for the predicted class (C_M), in order to reduce the features appearing in a rule and to broaden the ranges of the features. However, to acquire the conclusiveness property, LF always has as many paths as needed to remain consistent with the original prediction. For each learning task (binary, multi-class and regression) it is necessary to define the theoretical foundations on which the estimation of the minimum number of needed paths will rely on.

We should mention here that in the implementation of LF, apart from the minimum number of paths, an additional restriction concerning the average probability of a class is added. This happens because unlike the original RF's majority voting schema (hard voting) [10], the implementation of RF we use [35] combines the probabilistic predictions (soft voting) of the trees.

4.2.1 Minimum number of paths in binary tasks

First, we will go through the theoretical foundations for estimating the minimum required paths on binary tasks. Based on Proposition 1, which states that we need at least $\frac{|T|}{2} + 1$ of the paths in order to maintain the same prediction C_M , LF will select at least a quorum. For example, if we had $|T| = 100$ trees in an RF model and 89 of them voted for class M for an instance x_i , it has to select at least 51 of the 89 paths extracted from the trees, in order to produce the interpretation rule.

Proposition 1 *An RF model, with T trees casting $|T|$ votes, predicts always class M (C_M) if and only if class M has at least a quorum of votes or more, where quorum = $\frac{|T|}{2} + 1$ out of $|T|$ votes.*

As a result, the optimisation problem (Eq. 5) is formulated to minimise the number of features ($|F'|$) that satisfy a reduced set of paths ($|T'|$). The constraint is to retain the same classification result as the original set of trees, making the number of the reduced paths equal to or greater than the quorum. Furthermore, the final interpretation rule will include any feature that appears in these paths, guaranteeing the property of conclusiveness.

$$\begin{aligned} & \underset{F' \subseteq F}{\text{minimise}} \quad |F'| \\ & \text{subject to} \quad p = \{f_i \boxtimes v_j \mid f_i \in F'\}, p \in P_t \quad \forall t \in T', \\ & \quad \lfloor \frac{1}{|T|} \sum_{t \in T'} h_t(x_i) + \frac{1}{2} \rfloor = \lfloor \frac{1}{|T|} \sum_{t \in T} h_t(x_i) + \frac{1}{2} \rfloor, \\ & \quad |T'| \geq \frac{|T|}{2} + 1 \end{aligned} \quad (5)$$

An example for the equation $\lfloor \frac{1}{|T|} \sum_{t \in T} h_t(x_i) + \frac{1}{2} \rfloor$ follows. When 70 out of $|T| = 100$ trees are voting for class 1, then we have $\lfloor \frac{1}{100} 70 + 0.5 \rfloor = \lfloor 1.2 \rfloor \rightarrow 1$. On the other hand, if 25 out of $|T| = 100$ trees are voting class 1 (the minority), then we have $\lfloor \frac{1}{100} 25 + 0.5 \rfloor = \lfloor 0.75 \rfloor \rightarrow 0$. Therefore, we are aiming to find the smallest $T' \subseteq T$, which will produce the same classification as the original T trees.

4.2.2 Minimum number of paths in multi-class tasks

In a multi-class classification task, the voting system is more complex than in a binary task, and thus Proposition 1 cannot cover every case. The following theorems are presented in order to apply the framework of LF to multi-class tasks. A running example will be given, along with the propositions and theorems, for better comprehension.

An RF model always predicts class M (C_M) if and only if class M has a majority of votes, while any other class J (C_J) has fewer votes. Let, the available classes for the RF model with $|T| = 100$ trees be C_1 (red), C_2 (blue) and C_3 (green). For a specific case, each tree voted and the result was $|C_1| = 45$, $|C_2| = 35$ and $|C_3| = 20$. Thus, the prediction was *red* with $|C_1| = |C_M| = 45$, and the second highest voting class was *blue* with $|C_2| = 35$. Proposition 1 cannot be used in this example because $45 = |C_1| = |C_M| < \text{quorum} = \frac{|T|}{2} + 1 = 51$.

Based on this, the result of the RF model is class M with the majority of votes. However, if we minimise the voting paths to the majority class votes, as we did in binary tasks (only if the number of votes is less than the quorum), and extract the interpretation rule, it is possible for another class to gain votes from other classes, because the paths voting for those classes will not be covered by the rule, and thus their vote (prediction) can change to the various paths of the

decision trees that might arise, and exceed the votes of the originally first class. As a consequence, the outcome of the prediction could change.

Lemma 1 *An RF model always predicts class $M (C_M)$ if and only if class M has a majority of votes, while any other class $J (C_J)$ can not exceed the votes of class M by obtaining at least $S = |C_M| - |C_J| + 1$ votes from the other classes.*

In our running example, if we apply Lemma 1, we can assume that RF will always predict the *red* class if the other classes do not exchange votes. For example, if $|C_1| - |C_2| + 1 = 45 - 35 + 1 = 11$ votes from the *green* class move to the *blue* class, the number of votes from the *blue* class would rise to $|C_2| = 35 + 11 = 46$, which are more than the *red* class. Or if $|C_1| - |C_3| + 1 = 45 - 20 + 1 = 26$ votes from the *blue* class move to the *green* class, the number of votes from the *green* class would rise to $|C_3| = 20 + 26 = 46$ more than the *red* class. As a result, if we had reduced the paths to the number of class *red* votes with LF, then the other votes could change and, therefore, the outcome of the prediction could also change, while Lemma 1 would be invalid.

Proposition 2 *An RF model always predicts class $M (C_M)$ if and only if the K number of votes from any other class remains stable, where $K = |T| + |C_L| - |C_M| + 1$, including the C_M and C_L , votes from the majority class and the second most voted class, respectively, out of the $|T|$ votes.*

Proposition 2 argues that in situations where the majority of the class has less than a quorum of votes, the number of votes can be reduced to K without affecting the outcome of the prediction. Maintaining only the votes of the two key classes, majority class and second most voted class, is not enough and hence we need to retain $K_{other} = K - |C_M| - |C_L|$ from the other classes, also at random.

Thus, in our running example, the minimum number of paths we can reduce to is $K = 100 + 35 - 45 + 1 = 91$. From these 91 paths, we will retain the 45 of class ‘red’ (C_1) and the 35 of class ‘blue’ (C_2) paths. Then we would have to hold $K_{other} = 91 - 45 - 35 = 11$ paths from the remaining classes, in this example from the ‘green’ class (C_3) which holds 20 paths. We use the LF’s techniques, which will be detailed later in this section, to select 11 out of 20 paths. Thus, holding 91 paths out of 100, the rest of the 9 paths of the class ‘green’, if they all switch from ‘green’ to ‘blue’, the class ‘blue’ will not exceed the votes of the class ‘red’, and the result will remain the same. This is presented in Theorem 1.

Theorem 1 *For $R = |T| - |C_M| - |C_L|$ as the remainder of the votes and $K_{other} = K - |C_M| - |C_L|$ as the remainder of the votes to be selected, S is always greater than $R - K_{other}$, which means that there is not a sufficient number of votes to be received by class L , or any other class, in order to surpass the votes of class M .*

In order to prove that the aforementioned Theorem holds, we will use contradiction to prove that $R - K_{other} < S$ is true.

Proof 1 *To prove Theorem 1 by contradiction we assume that the statement $R - K_{other} < S$ is false. We will prove that $R - K_{other} \geq S$ is true.*

$$\begin{aligned}
R - K_{other} \geq S &\iff \\
(|T| - |C_M| - |C_L|) - (K - |C_M| - |C_L|) &\geq |C_M| - |C_L| + 1 \iff \\
|T| - |C_M| - |C_L| - K + |C_M| + |C_L| &\geq |C_M| - |C_L| + 1 \iff \\
|T| - K &\geq |C_M| - |C_L| + 1 \iff \\
|T| - (|T| + |C_L| - |C_M| + 1) &\geq |C_M| - |C_L| + 1 \iff \\
|T| - |T| - |C_L| + |C_M| - 1 &\geq |C_M| - |C_L| + 1 \iff \\
|C_M| - |C_L| - 1 &\geq |C_M| - |C_L| + 1 \iff \\
-1 &\geq 1
\end{aligned}$$

Thus, we proved that the statement $R - K_{other} \geq S$ is not true. ■

Therefore, the number of paths we can keep in order to always maintain the same classification result, is provided by the following rules:

- If $|C_M| \geq quorum$ then apply LF reduction to features and paths as it happens to binary tasks, based on Proposition 1,

- If $|C_M| < quorum$ then identify the class with the second highest amount of votes, denoted as C_L . The number of paths we need to keep is equal to $quorum' = |C_L| - |C_M| + T + 1$. Then, we keep the paths from the C_M and C_L , and a random selection process is employed to collect $quorum'$ from the R remaining paths.

4.2.3 Minimum number of paths in regression tasks

The estimation of the required number of paths as defined in the binary and multi-class tasks, are not applicable for regression. Hence, we introduce an algorithm to adapt LF to regression models. In RF for regression, the prediction is determined by the average of the individual trees' predictions of the RF,

$$h(x_i) = \frac{1}{|T|} \sum_{t \in T} h_t(x_i) \quad (6)$$

where $h_t(x_i) \in \mathbb{R}$. On this basis, we cannot assume that by removing a few trees (for example, preserving a quorum) we can get the same outcome. To overcome this, we introduce the *allowed_error* definition.

The idea is to reduce the number of features and paths with regard to an *allowed_error*. We set the mean absolute error of the RF model as the default value for *allowed_error*. We also let the user set a preferred *allowed_error*. A sensitivity analysis between *allowed_error* and the feature and path reduction ratio is provided in Section 5.2 to guide users to choose the appropriate *allowed_error* for their application based on their needs. The value of *allowed_error* is strongly associated with the importance of each task. Saying we have a regression model that predicts the indication of blood sugar. The importance of the error has a mandatory sense, since it involves a health issue. On the other hand, when forecasting e.g. the quality of wine, the error might be less sensitive.

In order to measure the error of an interpretation and to use the following reduction techniques, the min and max predictions per tree of RF must be collected. Thus, for each tree of an RF model, we traverse to its leaves to identify the min and max predictions that the tree can provide. This kind of information will be used to estimate the worst-case error that may be added to the prediction given an interpretation. Therefore, the final interpretation will have the following form:

if $0.47 \leq f_1 \leq 0.6$ and $22 \leq f_3 \leq 54$ then Prediction: $24.15 \pm local_error$,
where $local_error \in [0, allowed_error]$

Algorithm 1: Calculation process of *local_error*

Input: *trees, allowed_error, tree_stats*
Output: *local_error*

```

1 L, K  $\leftarrow$  reduction_method(trees, allowed_error)
2 original_pred  $\leftarrow$  0, prediction  $\leftarrow$  0
3 for tree  $\in$  L do
4   original_pred  $\leftarrow$  original_pred + tree.predict(instance)
5   prediction  $\leftarrow$  prediction + tree.predict(instance)
6 end
7 for tree  $\in$  K do
8   tree_pred = tree.predict(instance)
9   min_pred = tree_stats[tree].min
10  max_pred = tree_stats[tree].max
11  extreme_pred  $\leftarrow$  max_pred
12  if  $|\text{tree\_pred} - \text{min\_pred}| > |\text{tree\_pred} - \text{max\_pred}|$  then
13    extreme_pred  $\leftarrow$  min_pred
14  end
15  original_pred  $\leftarrow$  original_pred + tree_pred
16  prediction  $\leftarrow$  prediction + extreme_pred
17 end
18 prediction =  $\frac{prediction}{|L+K|}$ 
19 original_pred =  $\frac{original\_pred}{|L+K|}$ 
20 local_error  $\leftarrow$   $|\text{original\_pred} - \text{prediction}|$ 
21 return local_error

```

The *local_error* is determined with regard to a selection of paths $|L|$. Three algorithms are introduced, presented in the following paragraphs, that will lead to a selection of $|L|$ trees from the initial $|T|$ trees. The $|K|$ remaining trees, which will be excluded from the final interpretation rule, can cause an error in the prediction. That is because the rule does not completely cover the decision of these trees, and thus, for a change in the instance to a feature that does not appear in the rule, those trees may produce different predictions, and the final prediction may change to:

$$h'(x) = \frac{1}{|T|} \left(\sum_{t \in T-K} h_t(x_i) + \sum_{t \in K} e_t(x_i) \right) \quad (7)$$

$$e_t(x_i) = \begin{cases} \min_pred_t, & |h_t(x_i) - \min_pred_t| > |h_t(x_i) - \max_pred_t| \\ \max_pred_t, & \text{elsewhere} \end{cases}$$

Consequently, for those $|K|$ trees, we replace their predictions with the min or max prediction that each tree can provide, choosing between them on the basis of which, min or max, is the most distant to the original prediction of the tree. The aforementioned process is depicted in the Algorithm 1, which is the base for Algorithm 6 presented later in Section 4.3.4.

4.3 Reduction techniques

We now introduce the reduction techniques applied after identifying how many paths we can reduce. We define four reduction techniques that we use to build rules with fewer features and larger ranges. Table 1 depicts an overview of the applicability of each algorithm presented in the following sections to the corresponding tasks. All of the techniques aim to reduce both the paths and the features, but with different degrees of effectiveness.

	Reduction through				
	Association Rules	Clustering	Random Selection	Distribution-based Selection	
				Inner	Outer
Binary	✓	✓	✓	-	-
Multi-class	✓	✓	✓	-	-
Regression	✓ <i>variant</i>	-	✓	✓	✓

Designed Towards	Feature Reduction	✓	✓	-	-	-
	Path Redution	-	-	✓	✓	✓

Table 1: Overview of algorithms we use for feature and path reduction for each task

4.3.1 Reduction through association rules

The first reduction process, reduction through association rules (AR), is applicable to all learning tasks with small variations, and it starts with the implementation of the association rules [36]. In association rules, the attributes are called items $I = \{i_1, i_2, \dots, i_n\}$. Each dataset contains sets of items, called itemsets $T = \{t_1, t_2, \dots, t_m\}$, where $t_i \subseteq I$. Using all possible items of a dataset, we can find all the rules $X \Rightarrow Y$, where $X, Y \subseteq I$. X is called antecedent, while Y is called consequent. The purpose of the association rules is to determine the support and confidence of every rule in order to find useful relations. A straightforward observation is that X is independent of Y when the confidence level is particularly low. What is more we can tell that X with high support implies it is probably very significant.

Algorithm 2 describes the reduction through association rules. The association rules will be implemented at the path level. The I items will contain the F features of our original dataset. The T dataset, which will be used to mine the association rules, will contain a collection of features that reflect each path $t_i = \{i_j | i_j = f_j, f_j \otimes v_k \in p_i, p_i \in P\}$. It is important to mention that we keep only the presence of a feature in the path, and we discard the value of v_j , in order to have itemsets only with features. It is then possible to apply association rules methods. To compute the association rules we use apriori [37] and fpgrowth [38], influenced by comparative studies of association rules algorithms [39, 40], presenting these as options for the reduction through association rules process.

The next step is to sort the association rules produced by the algorithm based on the ascending confidence score. For the rule $X \Rightarrow Y$ with the lowest confidence, we will take items of X and add them to an empty list of features F' .

Algorithm 2: Reduction through association rules

Input: $paths, features, trees, minimum_number_paths/allowed_error, task$

Output: $reduced_paths, reduced_feature_set$

```

1  $itemsets \leftarrow []$ ; // Empty list of itemsets
2 for  $path \in paths$  do
3    $itemset \leftarrow []$ 
4   for  $f_i \in path$  do
5     |  $itemset \leftarrow itemset + f_i$ 
6   end
7    $itemsets \leftarrow itemsets + itemset$ 
8 end
9  $frequent\_itemsets \leftarrow ar(itemsets)$ ; // ar can be apriori, fpgrowth
10  $antecedent\_features \leftarrow [], antecedents \leftarrow []$ 
11 for  $antecedent \in frequent\_itemsets[antecedents]$  do
12   if  $antecedent \notin antecedents$  then
13     |  $antecedents \leftarrow antecedents + antecedent$ 
14     for  $feature \in antecedent$  do
15       |   if  $feature \notin antecedent\_features$  then
16         |     |  $antecedent\_features \leftarrow antecedent\_features + feature$ ;
17     end
18   end
19  $reduced\_paths \leftarrow [], reduced\_feature\_set \leftarrow []$ 
20 if  $task \in \{regression\}$  then  $local\_error \leftarrow 2 * allowed\_error$ ;
21  $k \leftarrow 1$ 
22 while  $((task \in \{binary, multi-class\} \wedge |reduced\_paths| < minimum\_number\_paths) \vee (task \in \{regression\} \wedge local\_error > allowed\_error)) \wedge k < |antecedent\_features|$  do
23    $reduced\_feature\_set \leftarrow reduced\_features + antecedent\_features[k-1]$ 
24    $redundant\_features \leftarrow []$ 
25   for  $feature \in features$  do
26     | if  $feature \notin reduced\_feature\_set$  then  $redundant\_features \leftarrow redundant\_features + feature$ ;
27   end
28    $reduced\_paths \leftarrow []$ 
29   for  $path \in paths$  do
30     | if  $\forall feature \in path, feature \notin redundant\_features$  then  $reduced\_paths \leftarrow reduced\_paths + path$ 
31   end
32   if  $task \in \{regression\}$  then  $local\_error \leftarrow compute\_error(paths, reduced\_paths)$ ;
33    $k \leftarrow k + 1$ 
34 end
35 return  $reduced\_paths, reduced\_feature\_set$ 

```

After that, we determine the number of paths containing conjunctions that are satisfied with the new feature set F' . We stop when we acquire a number of paths either equal or more than the estimated required paths (as presented in Sections 4.2.1 and 4.2.2), or when we have a $local_error$ from the paths that cannot be valid with the new feature set, smaller or equal to the $allowed_error$ (Section 4.2.3).

Otherwise, we iterate and add more features taken from the next antecedent of the next in confidence score rule. By using this strategy, we keep the number of features low and smaller than the original $F' \subseteq F$. Reducing the features can also lead to a reduced collection of paths, as paths containing conjunctions with redundant features would no longer be valid. So, we have the following representation for every p path:

$$p = \{f_i \boxtimes v_j | f_i \in F', v_j \in \mathbb{R}, \boxtimes \in \{\leq, \geq\}\}. \quad (8)$$

It is evident that this reduction technique favours the feature reduction, as its main criteria are based on the association rules discovered in the feature sets of the paths. We should also mention here, that in contrast to our preliminary work, we optimised the feature set extracted from the frequent patterns (Algorithm 2, L.11-18) and now the reduction through association rules process is providing timely responses. This is also evident through our experiments in Section 5.3.

Algorithm 3: Path similarity metric

```

input :  $p_i, p_j, feature\_names, min\_max\_feature\_values$ 
return :  $similarity_{ij}$ 

1  $s_{ij} \leftarrow 0$ 
2 for  $f \in feature\_names$  do
3   if  $f \in p_i \wedge f \in p_j$  then
4     find  $l_i, u_i, l_j, u_j$  lower and upper bounds
5      $inter \leftarrow \min(u_i, u_j) - \max(l_i, l_j)$ ,  $union \leftarrow \max(u_i, u_j) - \min(l_i, l_j)$ 
6     if  $inter > 0 \wedge union \neq 0$  then  $s_{ij} \leftarrow s_{ij} + inter/union$ ;
7   else if  $f \notin p_i \wedge f \notin p_j$  then
8      $s_{ij} \leftarrow s_{ij} + 1$ 
9   end
10 end
11 return  $s_{ij}/|feature\_names|$ 

```

Algorithm 4: Reduction through clustering

```

Input:  $paths, minimum\_number\_paths, clusters$ 
Output:  $reduced\_paths, reduced\_feature\_set$ 

1  $dissimilarity \leftarrow []$ 
2 for  $i \in [0, |paths|]$  do
3    $vector \leftarrow []$ 
4   for  $j \in [0, |paths|]$  do
5     if  $i = j$  then
6        $vector \leftarrow vector + [0]$ 
7     else
8        $vector \leftarrow vector + [compute\_dissimilarity(path[i], path[j])]$ 
9     end
10   end
11    $dissimilarity \leftarrow dissimilarity + [vector]$ 
12 end
13  $clusters = cl(dissimilarity, clusters);$  // cr can be k-medoids, OPTICS, SC
14  $sorted\_clusters = sort\_by\_size(clusters)$ 
15  $reduced\_paths \leftarrow [], k \leftarrow 0$ 
16 while  $|reduced\_paths| < minimum\_number\_paths \wedge k < |antecedent\_features|$  do
17   for  $path \in sorted\_cluster[k]$  do
18      $reduced\_paths \leftarrow reduced\_paths + path$ 
19   end
20    $k \leftarrow k + 1$ 
21 end
22  $reduced\_feature\_set \leftarrow \emptyset;$  // Empty set of features
23 for  $path \in reduced\_paths$  do
24   for  $f_i \in path$  do
25      $reduced\_feature\_set \leftarrow reduced\_feature\_set + f_i$ 
26   end
27 end
28 return  $reduced\_paths, reduced\_feature\_set$ 

```

4.3.2 Reduction through clustering

A second reduction strategy based on clustering (CR) is utilised only on the binary and multi-class tasks. Reduction through clustering was not used in regression because reduction through association rules almost reaches the maximal local error allowed, and the overhead of clustering does not justify the effort³. Aside from k -medoids [41], which was used in our preliminary work, OPTICS [42] and Spectral Clustering [43] (SC) were added as additional choices. These

³Reduction through clustering was not used in regression because reduction through association rules almost reaches the maximal local error allowed, and the overhead of clustering does not justify the effort

algorithms are well-known clustering algorithms which need a distance or dissimilarity metric to find optimum clusters. Thus, the clustering of paths will require a distance or dissimilarity metric between two paths. Therefore, firstly, a similarity metric has been constructed (Algorithm 3), in order to transform it to a dissimilarity metric, in a way that favours the absence of a feature from both paths. Specifically, if a feature is missing from both paths, the similarity of these paths increases by 1. When a feature is present in both paths, the similarity increases by a value between 0 and 1, i.e. the intersection of the two ranges normalised by the union of the two ranges. Finally, we subtract by 1 to convert the similarity metric, which has a range [0, 1], where 0 means the paths are not similar at all and 1 means the paths are identical, to a dissimilarity metric.

Using one of the aforementioned clustering algorithms, by default we use k -medoids, the clusters are estimated using the dissimilarity metric discussed above. Subsequently, the ordering of the clusters is performed on the basis of the number of paths they cover. Then, paths from larger clusters are accumulated into a list, until at least the required number of paths (as estimated in Section 4.2) has been obtained.

By summing up larger clusters firstly, the probability of reducing the features is increasing, since the paths inside a cluster appear to be more similar among them. In addition, the biased dissimilarity metric will cluster paths with fewer insignificant features, leading to a subset of paths that are satisfied with a smaller set of features. As a result, we can assume that the reduction through clustering is also oriented toward feature reduction. The corresponding procedure is also represented in Algorithm 4.

4.3.3 Reduction through random selection

Random selection (Algorithm 5) (RS) of paths is used to achieve the minimum number of paths, in case of reduction through clustering, has not reached the quorum. For binary and multi-class tasks, RS removes randomly the unnecessary paths to only keep the minimum required number of paths. On the regression tasks, if the *local_error* after the AR is smaller than the *allowed_error* RS is applied in order to minimise the paths, while maintaining the *local_error* less or equal to *allowed_error*.

Algorithm 5: Reduction through random selection

Input: *paths*, *minimum_number_paths*/*allowed_error*, *task*
Output: *reduced_paths*, *reduced_feature_set*

```

1 reduced_paths  $\leftarrow \emptyset$ 
2 if task  $\in \{binary, multi-class\}$  then
3   for i  $\in [0, \text{minimum\_number\_paths}]$  do
4     j  $\leftarrow \text{random}(0, |\text{paths}|)$ 
5     reduced_paths  $\leftarrow \text{reduced\_paths} + \text{paths}[j]$ 
6     paths  $\leftarrow \text{paths} - \text{paths}[j]$ 
7   end
8 else
9   local_error  $\leftarrow 2 * \text{allowed_error}$ 
10  last_path  $\leftarrow \emptyset$ 
11  temp_paths  $\leftarrow \text{paths}$ 
12  while local_error  $< \text{allowed_error} \wedge |\text{paths}| > 1$  do
13    j  $\leftarrow \text{random}(0, |\text{temp_paths}|)$ 
14    last_path  $\leftarrow \text{temp_paths}[j]$ 
15    temp_paths  $\leftarrow \text{temp_paths} - \text{temp_paths}[j]$ 
16    local_error  $\leftarrow \text{compute_error}(\text{paths}, \text{temp_paths})$ 
17  end
18  reduced_paths  $\leftarrow \text{temp_paths} + \text{last_path}$ 
19 end
20 reduced_feature_set  $\leftarrow \emptyset$ ;                                // Empty set of features
21 for path  $\in \text{reduced_paths}$  do
22   for fi  $\in \text{path}$  do
23     | reduced_feature_set  $\leftarrow \text{reduced\_feature\_set} + f_i$ 
24   end
25 end
26 return reduced_paths, reduced_feature_set

```

Algorithm 6: Reduction through distribution-based selection

Input: $instance, trees, allowed_error, variation$

Output: $reduced_paths, reduced_feature_set, error$

```

1  $reduced\_paths \leftarrow [tree..decision\_path(instance), \forall tree \in trees]$ 
2  $error \leftarrow 0$ 
3  $predictions \leftarrow [tree.predict(instance), \forall tree \in trees]$ 
4  $mean \leftarrow \frac{1}{|predictions|} \sum_{prediction \in predictions} prediction$ 
5  $\sigma \leftarrow \sqrt{\frac{1}{|predictions|} \sum_{prediction \in predictions} (prediction - mean)^2}$ 
6 for  $s \in [.1, .2, .5, 1, 2, 4, 5, 6, 7, 8, 9, 10, 20, 50, 100]$  do
7    $normal\_distribution \leftarrow generate\_distribution(mean, \frac{\sigma}{s})$ 
8    $normal\_min = min(normal\_distribution)$ 
9    $normal\_max = max(normal\_distribution)$ 
10   $local\_reduced\_paths \leftarrow [], new\_prediction \leftarrow 0$ 
11  for  $tree \in trees$  do
12     $path \leftarrow tree.decision\_path(instance), prediction \leftarrow tree.predict(instance)$ 
13    if
14       $(variation \in \{DSi\} \wedge \neg(prediction < normal\_min \vee prediction > normal\_max)) \vee (variation \in \{DSo\} \wedge (prediction < normal\_min \wedge prediction > normal\_max))$  then
15         $local\_reduced\_paths \leftarrow local\_reduced\_paths + path$ 
16         $new\_prediction \leftarrow new\_prediction + prediction$ 
17    else
18       $distance\_with\_min \leftarrow |prediction - tree.min\_prediction|$ 
19       $distance\_with\_max \leftarrow |prediction - tree.max\_prediction|$ 
20      else if  $distance\_with\_min > distance\_with\_max$  then
21         $new\_prediction \leftarrow new\_prediction + tree.min\_prediction$ 
22      end
23       $new\_prediction \leftarrow new\_prediction + tree.max\_prediction$ 
24    end
25     $local\_error \leftarrow |mean - \frac{1}{|trees|} new\_prediction|$ 
26    if  $local\_error < allowed\_error \wedge |reduced\_paths| > |local\_reduced\_paths|$  then
27       $error \leftarrow local\_error, reduced\_paths \leftarrow local\_reduced\_paths$ 
28    end
29  end
30   $reduced\_feature\_set \leftarrow \emptyset;$  // Empty set of features
31  for  $path \in reduced\_paths$  do
32    for  $f_i \in path$  do
33       $reduced\_feature\_set \leftarrow reduced\_feature\_set + f_i$ 
34  end
35 return  $reduced\_paths, reduced\_feature\_set, error$ 

```

4.3.4 Reduction through distribution-based selection

The last technique, reduction through distribution-based selection (DS), is a reduction technique exclusively designed for regression tasks. This method utilises normal distributions, is not combined with any other method, and it attempts to reduce the features and the paths of a rule for a given instance. Each instance's prediction is the average of the individual predictions of the trees. For example, in an RF regression model with 100 trees, an instance's prediction will be the average of the 100 predictions by each one of trees. Those 100 predictions have their own distribution.

In Figure 6 the actual distribution of the RF's prediction is shown in blue, while the green distribution is a normal distribution generated with the mean and standard deviation (σ) values of the original distribution. We can change the way this artificial distribution is created by multiplying or dividing the σ value, generating wider or narrower distributions around the mean value. Therefore, for different variations of the σ value, we apply one of the following techniques, inner or outer selection, to choose the paths we will keep for the final rule. We select the distribution's σ value, which

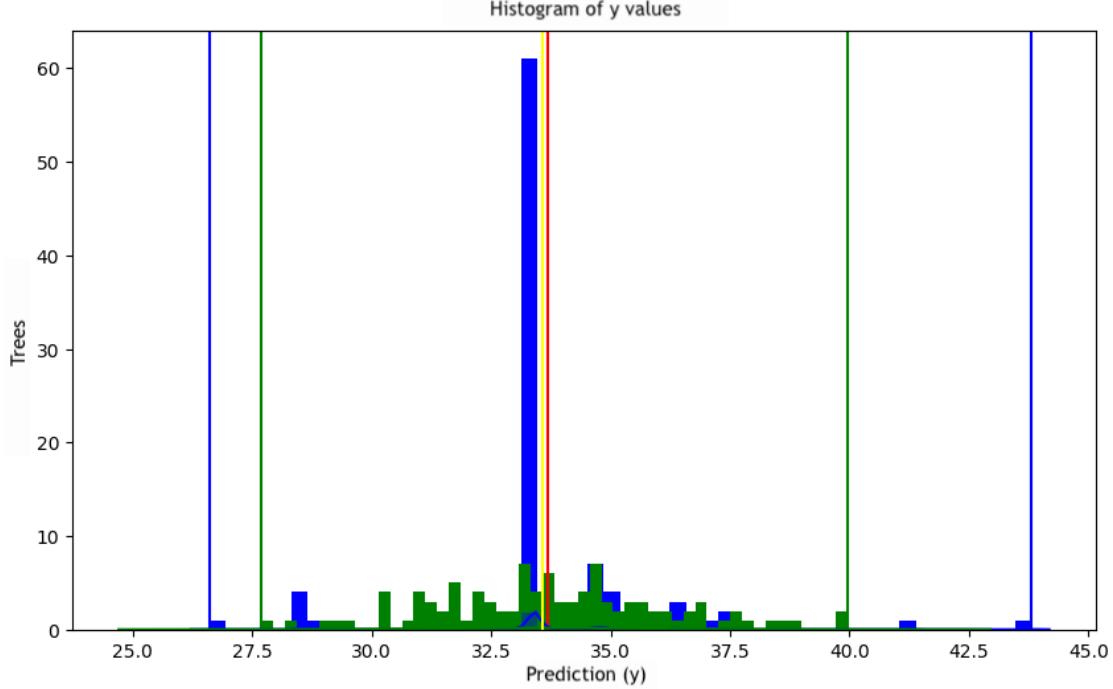


Figure 6: Distribution of predictions of a given instance

achieves the lowest number of paths with a $local_error \in [0, allowed_error]$ for an instance interpretation. This process is also presented in Algorithm 6, which is based on Algorithm 1.

Inner selection The distribution-based inner selection (DSi) technique, for all the different σ values, it selects the paths inside the range of the generated distribution to form the interpretation rule. For the paths outside the generated distribution, we compute the $local_error$. This approach is going to reduce the number of paths, when the majority of trees are providing similar predictions close to the mean value. Then, by isolating and removing distant predictions, even if they shift, the mean will not be greatly affected, and thus the $local_error$ will be relatively small.

Outer selection Similarly to the inner selection, the distribution-based outer selection (DSo) technique, for all the different σ values, it selects the paths outside the range of the generated distribution to form the interpretation rule. Again, for the paths outside the generated distribution, the $local_error$ is computed. This approach is providing better results when the original distribution is either like an inverse normal distribution or uniform. In this case the majority of the predictions will be distant to the mean value. Keeping those predictions, and leaving out from the interpretation rule the predictions closer to the mean, we will probably have a low $local_error$, because the predictions closer to the mean will be fewer. Moreover, those predictions are less likely to deviate a lot when small changes to the input will occur, thus the error most of the time will be smaller than the estimated $local_error$.

4.4 Feature-ranges formulation

In this section, we present how the formulation of feature-ranges are taking place in a binary classification scenario. However, in all tasks the formulation procedure is the same.

Consider an RF of T trees that predicts instance x as $c_j \in C = \{0, 1\}$. We concentrate on the $K \geq \frac{|T|}{2} + 1$ trees that classify x as c_j . For each feature we calculate the range of values imposed by the conditions in the root-to-leaf path of each of the K trees in which it appears.

For a concrete real-world example, we train an RF with $T = 50$ trees on the Banknote dataset [44]. We focus on the $skew$ feature of Banknote and a test instance x , whose $skew$ value is 0.25. The decision of the RF for this instance is supported by $K = 39$ of its trees. Figure 7 presents the value ranges for $skew$ in each of the 38 trees, whose decision paths contain the $skew$ feature. Figure 8 presents the corresponding stacked area plot, which shows the number of paths in the y axis whose value ranges include the corresponding value in the x axis.

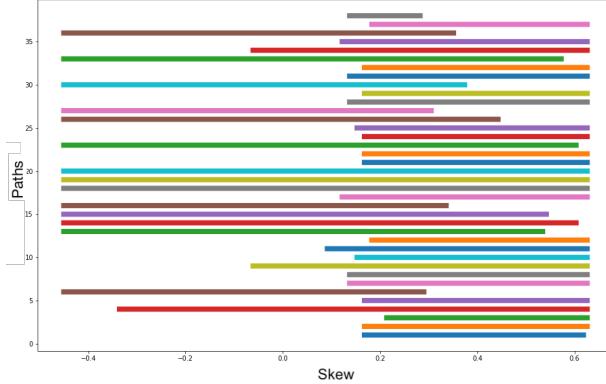


Figure 7: Bars plot of ‘skew’ feature

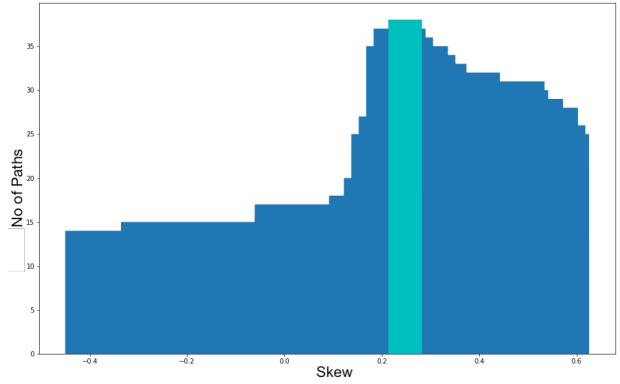


Figure 8: Stacked area plot of ‘skew’ feature

The cyan/light grey region of Figure 8 is the intersection between these 38 decision paths for *skew*. Then, we could say that the range of this feature will be: $0.21 \leq \text{skew} = 0.25 \leq 0.029$. This intersection will always contain the value (0.25) for the particular feature of the examined instance. Furthermore, no matter how much the value of the feature may shift, the decision paths will not change as long as it remains within this intersection. Thus, if the instance’s value for *skew* changes from 0.25 to 0.28 each tree will take its decision via the same path. Summing up the above the interpretation may have the following shape:

‘if $0.21 \leq \text{skew} \leq 0.29$ and . . . then *Fake Banknote*’

These paths will be used in order to form a final single rule. However, using all the available trees will produce interpretations with many features, thus less comprehensible to the end-user. Moreover, a large number of paths can lead to a small, strict, and very specific feature ranges. For example, the *skew* value of the instance x was 0.25 and the intersection range of all paths for this feature is $0.21 \leq \text{skew} \leq 0.29$, while the global range of *skew* is $[-1, 1]$. If the 10 first paths corresponding to the first 10 top rows (top to bottom) of the stacked area plot (Figure 8) could have been reduced, the range would get $0.15 \leq f_1 \leq 0.4$, thus a broader one. A narrow range, like the one mentioned above, would result in a negative perception of the model, which would be considered inconsistent and unreliable, whereas a wider range would be less rebuttable and preferable. Hence this is the reason why we compute the minimum number of paths required 4.2, and then we apply reduction techniques 4.3.

In this work, we also developed a much faster way for computing these intersections. This affected the responsiveness of any operation, including the extraction of an explanation without reduction. The following assumptions about the performance are confirmed by the time analysis studies in Section 5.3.

4.5 Categorical features

One more LF’s adaptation concerns the handling of categorical features. Focusing only in OneHot encoded cases, we designed a handling procedure, which is in accordance with this principle: if the OneHot encoded feature appears in the reduced paths, then the categorical feature where it originates is added to the rule. If this does not exist in the reduced paths, this means that the prediction has not been affected by this value of the categorical feature. However, we are looking for the other OneHot encoded features of the same categorical feature appearing in the reduced rules, but not in the instance (their values are equal to zero), and we present them to the user, along with the interpretation, as categorical values that may change the prediction if the category changes to one of them.

To give a brief example, let us have a feature *height* with values = [*short*, *normal*, *tall*]. After OneHot encoding, there are three features: $height_{\text{short}}$, $height_{\text{normal}}$ and $height_{\text{tall}}$. An instance with $height = \text{short}$ will therefore have $height_{\text{short}} = 1$, $height_{\text{normal}} = 0$, and $height_{\text{tall}} = 0$ after encoding. If the feature $height_{\text{short}}$ appears in the extracted paths, it means that it influences the prediction and will appear in the final rule as ‘ $height = \text{short}$ ’. Elsewhere, this indicates that this feature value does not affect the prediction. Then, we check whether the features $height_{\text{normal}}$ and $height_{\text{tall}}$ appear in the reduced paths. If any of them appears on the paths, this means that they influence the prediction if they change. Therefore, they are provided to the user as alternative values.

4.6 Interpretation composition

The last stage of LionForests combines the extracted ranges of features and the information derived from the categorical handling, if there was any categorical feature, in a single natural language rule. The order of appearance of the ranges of features in the rule is determined by a method of global interpretations, called permutation importance [10]. An example of an interpretation is as follows:

$$\text{'if } 0 \leq f_1 \leq 0.5 \text{ and } -0.5 \leq f_2 \leq 0.15 \text{ and } f_3 = \text{short then } Prediction' . \quad (9)$$

This rule can be interpreted in this way: “As long as the value of the f_1 is between the ranges 0 and 0.5, and the value of f_2 is between the ranges -0.5 and 0.15, and f_3 equals *sort*, the system will predict this instance as *Prediction*. If the value of f_1 , f_2 or both, surpass the limits of their ranges, and f_3 changes category then the prediction may change. Note that the features are ranked through their influence”. This type of interpretation is intuitive, human-readable and more comprehensible than other interpretations, like feature importance. Furthermore, this rule is conclusive since it will always include all of the feature ranges appearing in the paths produced the specific prediction, while these paths will be greater than or equal to the minimum number of paths required to maintain the prediction stable.

4.7 Visualisation

Apart from extending LionForests to multi-class and regression tasks, as well as optimising the core functionalities of LionForests, we are also presenting a proposed visualisation layout. Using this interface, the user will see the explanation rule for a prediction, and then the ability to choose to inspect each of the features that appear in the rule will be available. In the case of a feature with numerical values, the user will be supplied with a distribution plot with the values of the feature. Two vertical green lines are being used to represent the range that appears in the rule, while two blue vertical lines reflect the ranges of the original rule, without applying any reduction technique. On the other hand, if the selected feature is categorical, then if alternative values are available as described in Section 4.5, they will be visible to the user. An example is provided in Section 5.5 and Figures 30 and 31.

5 Evaluation

The performed experiments were related to response time, sensitivity analysis to inspect how RF parameters affect feature and path reduction, a comparative study with similar well-known techniques, as well as a qualitative assessment with examples. LionForests’ code and evaluation experiments are available at GitHub repository “LionLearn”⁴.

5.1 Setup and datasets

For the series of experiments, we have used the Scikit-learn’s [35] RandomForestClassifier and RandomForestRegressor. We used 8 different datasets, 3 for binary, 2 for multi-class, 2 for regression and 1 for multi-class and regression problems. The datasets for binary tasks are: Banknote [44], Heart (Statlog) [44] and Adult (Census) [45]. Glass [44] and Image Segmentation (Statlog) [44] for multi-class tasks. For regression, we used Wine Quality [46] and Boston Housing [47]. Finally, Abalone [48] was used for both multi-class and regression.

We applied a 10-fold cross-validation grid search for each dataset using the following set of parameters: $\max\ depth \in \{1, 5, 7, 10\}$ (max depth of each individual tree), $\max\ features \in \{\text{'sqrt'}, \text{'log2'}, 75\%, \text{None} = \text{all features}\}$ (max number of features per individual tree), $\min\ samples\ leaf \in \{1, 2, 5, 10\}$, $\text{bootstrap} \in \{\text{True}, \text{False}\}$ (bootstrap samples or the whole dataset is used to train each tree), $\text{estimators} \in \{10, 100, 500, 1000\}$ (number of trees in the ensemble). The scoring metric of the grid search was the weighted F_1 -score for the binary and multi-class datasets, and the mean absolute error (*mae*) for the regression datasets. Table 2 shows the number of instances and features of each dataset, the parameters’ values that achieved the best F_1 -score/*mae* for each dataset, along with the scores itself.

5.2 Sensitivity analysis

We have conducted the following sensitivity analysis experiments in order to compare the capability of LF to reduce features (feature reduction ratio - FR%) and paths (path reduction ratio - PR%) while tuning both the parameters of LF (LF) and the configuration of the Random Forests (RF) model. In order to simplify the information obtained from the analysis, we present each task (binary, multi-class, regression) separately. The parameters of the RF model under examination are the $\max\ depth \in \{1, 5, 7, 10\}$, $\max\ features \in \{\text{'sqrt'}, \text{'log2'}, 75\%, \text{None}\}$ and $\text{estimators} \in \{10, 100,$

⁴<https://git.io/JY0gF>

Task	Dataset	Classes	Instances	Features	Categorical	F_1 -score/ mae	Parameters				
							No. Estimators	Max Depth	Max Features	Min Samples Leaf	Bootstrap
Binary (B)	Banknote	2	1372	4	-	99.49%	500	10	0.75	1	True
	Heart (statlog)	2	270	13	1	84.60%	500	5	'sqrt'	5	False
	Adult Census	2	45167	12 (85)	7 (80)	84.84%	1000	10	0.75	2	True
Multi-Class (MC)	Glass	6	214	9	-	71.95%	1000	10	'sqrt'	2	True
	I. Segmentation	7	2310	19	-	96.98%	500	10	0.75	1	False
MC/R	Abalone	4/-	4027	8	-	68.76%/1.44	10/100	10	0.75	10/5	True
	Boston	-	506	13	-	2.95	1000	10	0.75	1	True
Regression (R)	Wine Quality	-	4898	12	-	0.55	1000	10	0.75	5	True

Table 2: Performance and best parameters for each dataset

500, 1000}. LF parameters tested in this sensitivity analysis are the *AR algorithm* ($AR \in \{\text{apriori}, \text{fpgrowth}\}$), *CL algorithm* ($CL \in \{\text{kmedoids}, \text{OPTICS}, \text{CS}\}$), and *method* $\in \{1, 2, 3, 12, 13, 23, 123\}$ for the binary and multi-class classification, where 1 represents the use of AR, 2 the use of CL and 3 the use of RS. Therefore, 123 means the utilisation of all 3 methods, in this order. For regression, we used *method* $\in \{\text{AR+RS}, \text{DSi}, \text{DSo}\}$, while an additional parameter of LF, *allowed_error* is also examined in the regression tasks.

5.2.1 Binary classification

The first sensitivity analysis we carried out applies to the Adult, Banknote and Heart (Statlog) datasets. Figure 9 presents the FR% for the parameters of the RF while Figure 10 refers to the parameters of LF. The parameter analysis reveals that when the RF's *max features* parameter is set to 75%, the reduction in features in all datasets is higher. With regard to *max depth* and *estimators*, LF achieves over 35% FR when the *max depth* is greater than or equal to 5 and the *estimators* are 100 or more.

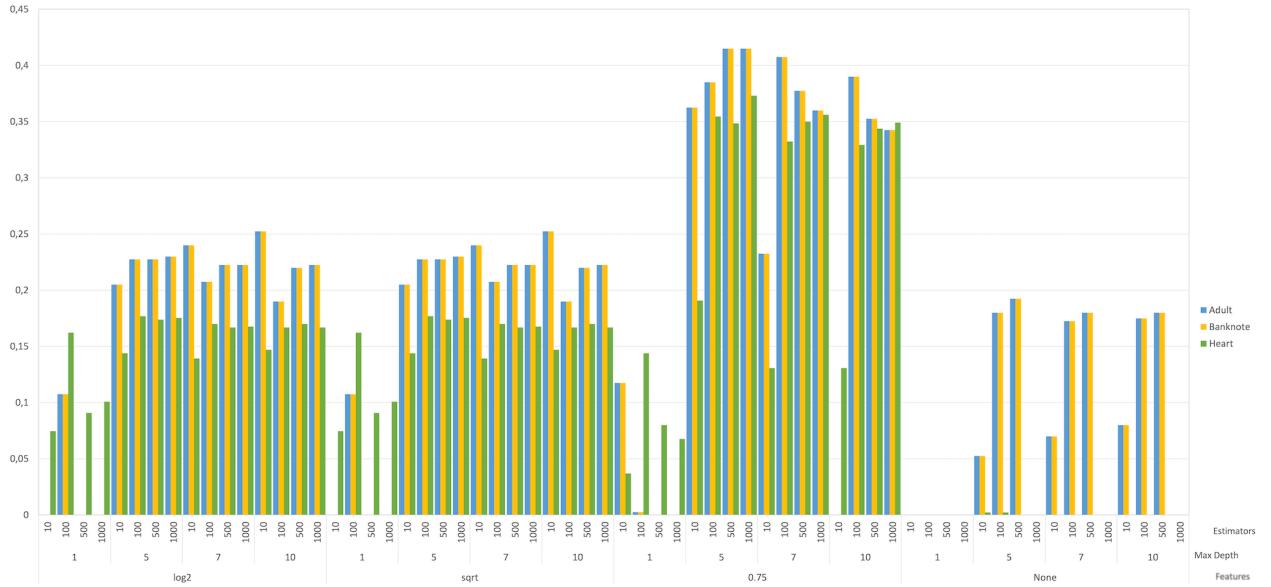


Figure 9: Binary classification: analysis of FR relation to RF's parameters

In Figure 10, we can see how the parameters of LF affect the FR% in these datasets. We initially observe that the two different AR (1) are performing identically in the FR%. In CR (2), the FR% seemed to diverge for the different algorithms. Specifically, k -medoids and SC manage to reduce the features by 20% or more in all datasets, while OPTICS could not manage to perform any reduction. RS (3) did not achieve any FR in Adult and Banknote, while it achieved low FR% in Heart (Statlog). Among the three approaches, AR seems necessary to achieve high FR%, while the combination of AR (1) with CL (2) seems to increase in all three datasets the FR%.

About the PR analysis, Figure 11 reveals that when the RF's *max features* parameter is set to 'None', the reduction in paths in all datasets is higher. With regards to *max depth* and *estimators*, LF achieves over 40% PR when the *max depth* is greater or equal to 5 and the *estimators* are 100 or over.

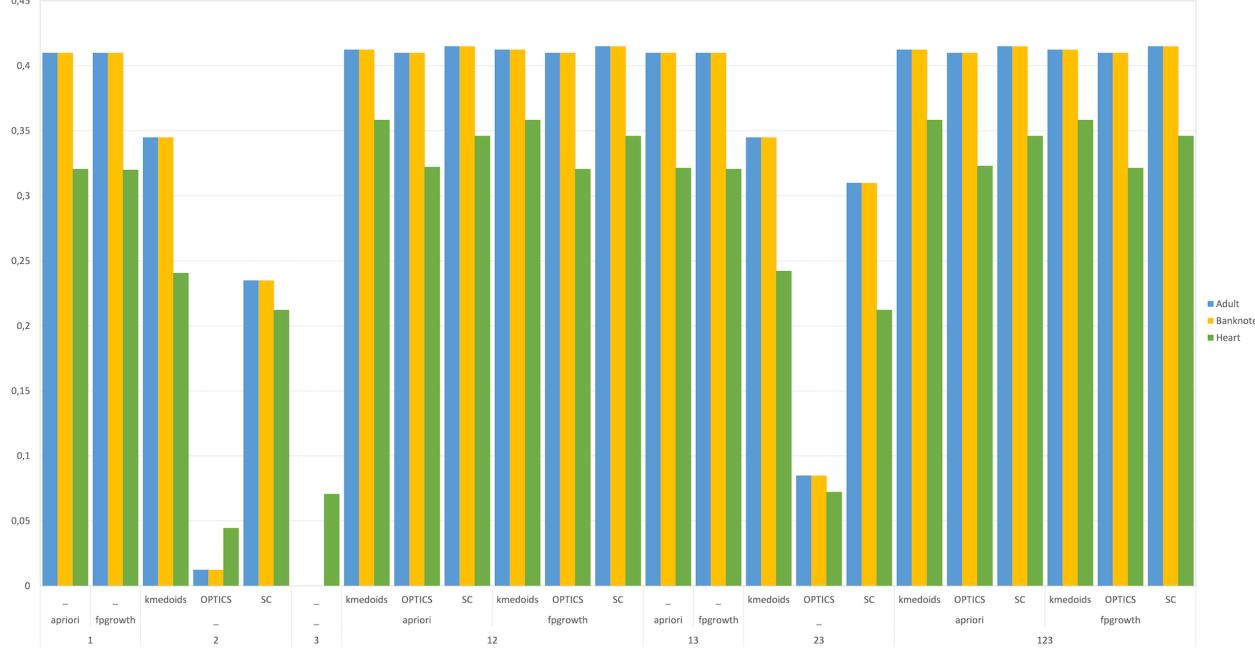


Figure 10: Binary classification: analysis of FR relation to LF's parameters

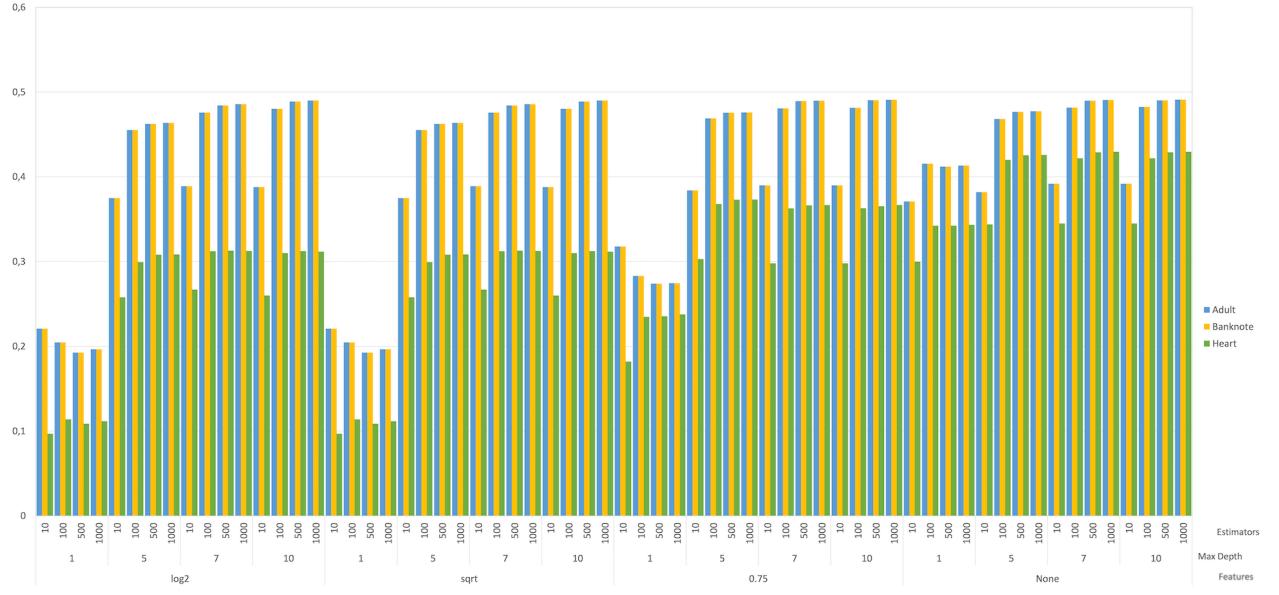


Figure 11: Binary classification: analysis of PR relation to RF's parameters

In Figure 12, we can see how the parameters of LF affect the PR% in these datasets. In contrast to FR, for PR, CR (2) and RS (3) are both maximising the PR%. Recall that we cannot reduce more than a quorum in a binary setup, thus these techniques achieving 49% PR are performing optimally. AR (1), on the other hand, cannot seem to be able to optimally reduce paths. Finally, we observe that when combining all three techniques (123), for every parameter setting, the PR% is higher than 40%.

5.2.2 Multi-class classification

We continue with the multi-class datasets Glass, I. Segmentation and Abalone. The tuning of RF's parameters and their impact to the FR% are visible in Figure 13. The analysis reveals that when the RF's *max features* parameter is set

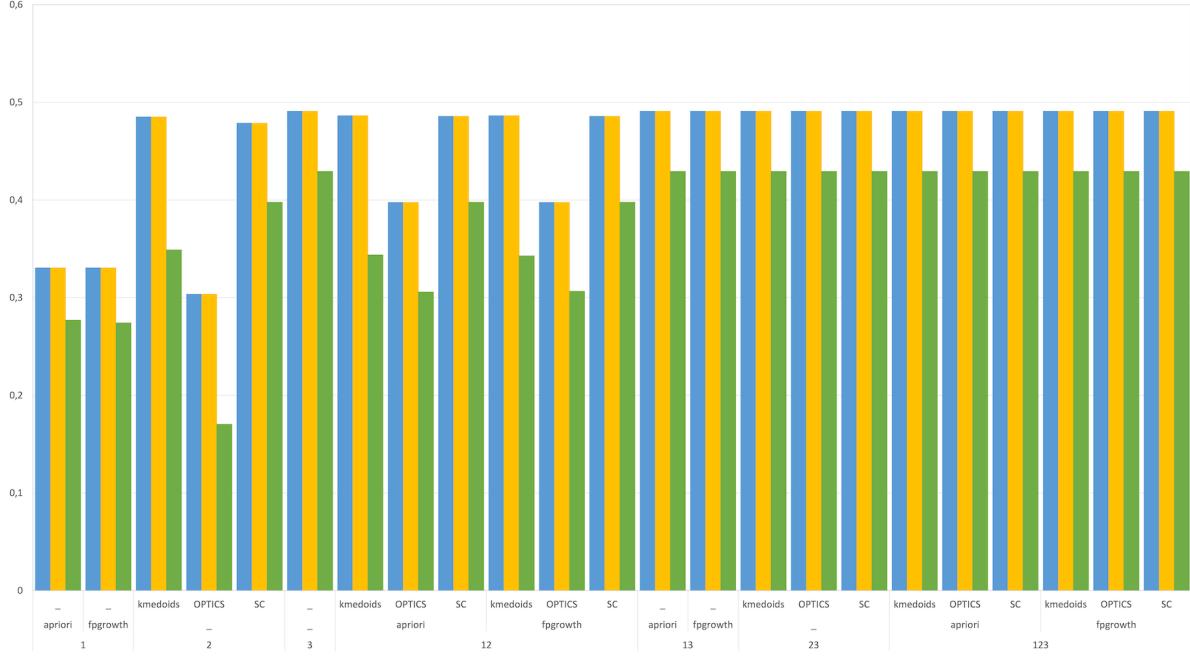


Figure 12: Binary classification: analysis of PR relation to LF's parameters

to 75%, the FR in all datasets is higher. LF achieves over 17% FR when the *max depth* is greater than or equal to 5 and *estimators* are 100 or over, while it achieves more than 25% and 34% for the individual datasets, Abalone and I. Segmentation, respectively.

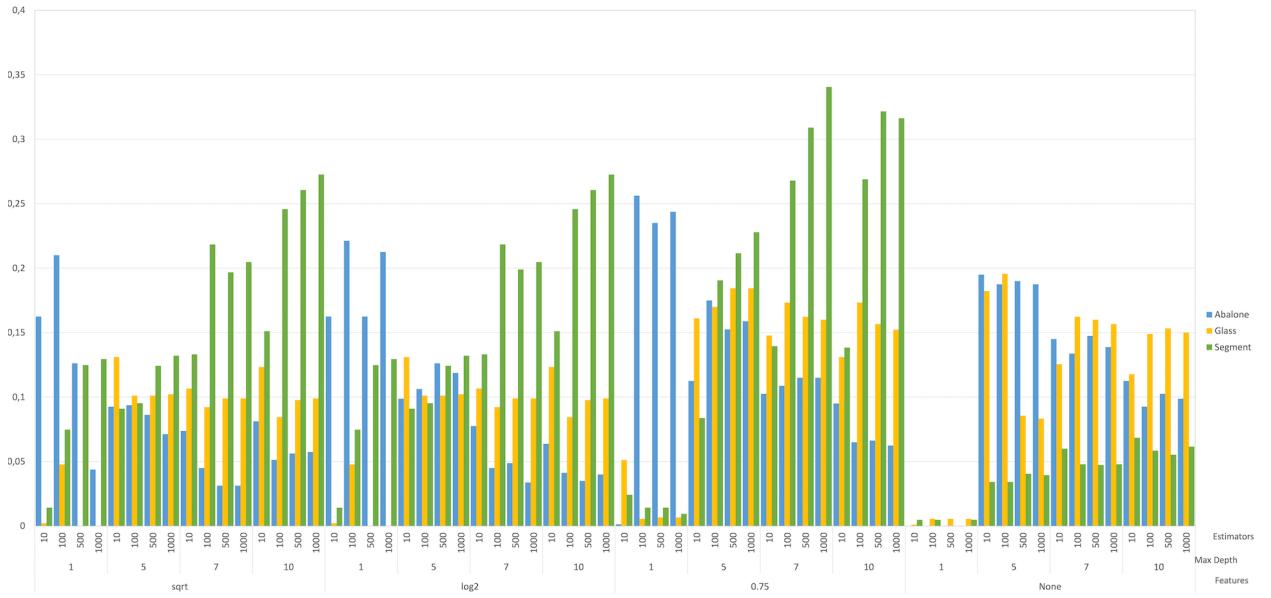


Figure 13: Multi-class classification: analysis FR relation to RF's parameters

Figure 14 presents how FR% is affected based on the different parameters of LF. As observed in the binary classification sensitivity analysis, here as well it is eminent that AR (1) is performing equally in the FR%. CR (2) in the Abalone dataset achieved a higher FR than AR, and when combined (123) with the other techniques, AR and RS, the FR% is not increasing. The analysis of Glass dataset revealed that rather than the FR achieved by the AR (1), no other method or combination managed to increase the FR%. Finally, on I. Segmentation seemed the combination of AR and CR (12), with specifically SC, to provide the highest FR%. Another interesting point is that the RS (3) managed to reduce the

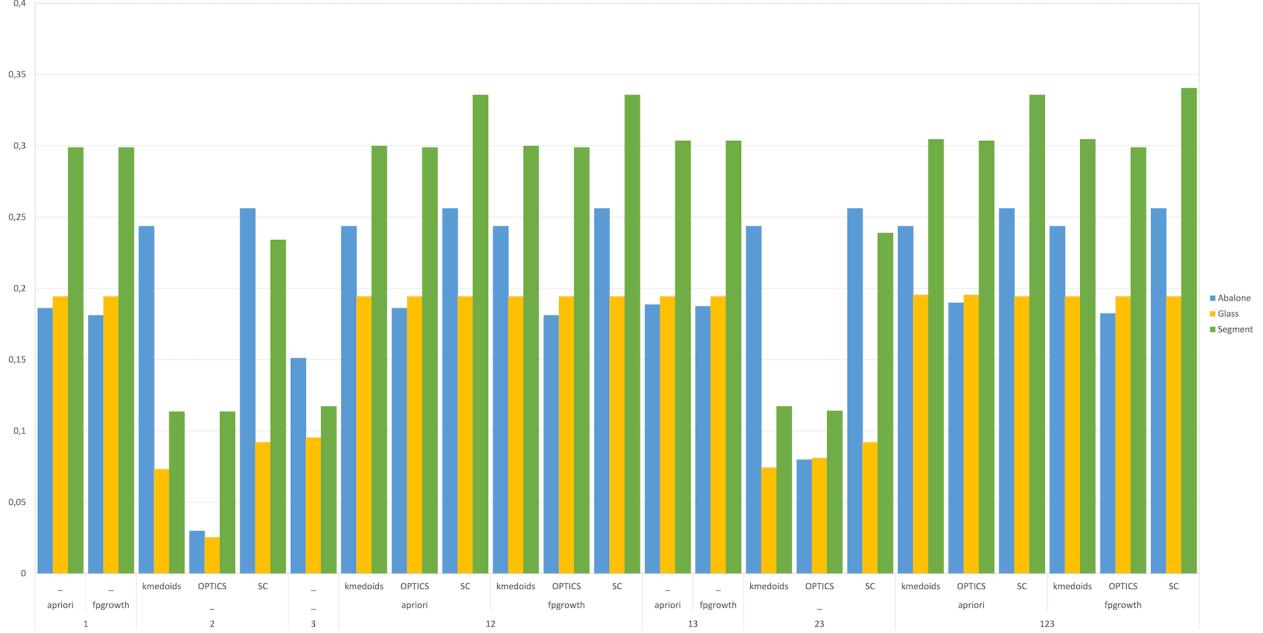


Figure 14: Multi-class classification: analysis of FR relation to LF's parameters

features of the rules in all three datasets in contrast to the RS's performance on the binary's classification sensitivity analysis.

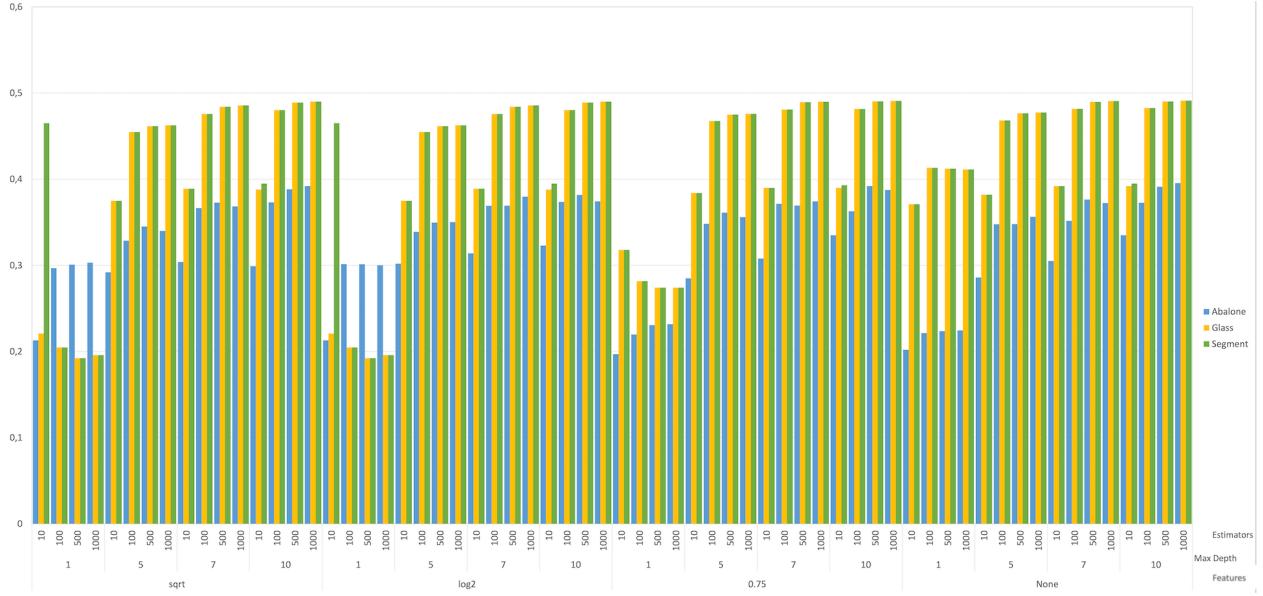


Figure 15: Multi-class classification: analysis of PR relation to RF's parameters

Through Figure 15, observing the PR while tuning the RF's parameters in these datasets, we can say that the *max features* parameters do not affect the PR%. We can not conclude the same for *max depth* and *estimators*, where we need 5 or higher and 100 or more, respectively, to achieve higher PR%. The highest PR% it is achieved when *max depth* equals 10 and *estimators* equals 1000.

In Figure 16, we can see how the parameters of LF affect the PR% in these datasets. RS (3) is maxing out the PR%. AR (1) cannot seem to achieve the desirable PR results, while CL (2) is performing well, but not as good as RS (3). Thus, RS or any combination with RS leads to a PR% of 38% or more.

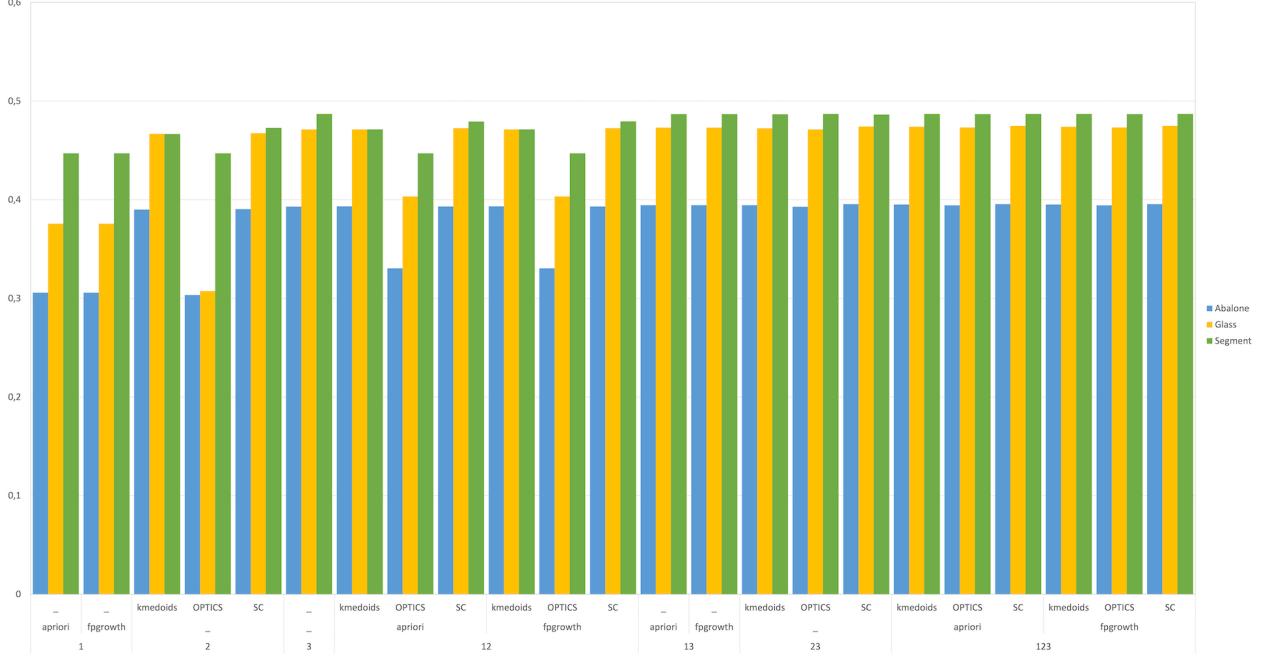


Figure 16: Multi-class classification: analysis of PR relation to LF's parameters

5.2.3 Regression

The last set of experiments is related to the analysis of the regression datasets, Abalone, Boston and Wine. In Figure 17, the relation of RF's parameters to the FR% is visible. We can say that the most influencing parameter is *estimators*. When *estimators* are equal or more than 500 and *max depth* is either 1 or 5, then the reduction is between 35% to 51%. Moreover, for Boston and Wine we observe that when *max features* is set to either ‘sqrt’ or ‘log2’, the FR% is higher. On the other hand, higher *max features* values like ‘0.75’ or ‘None’ seem to favour the FR% for Abalone.

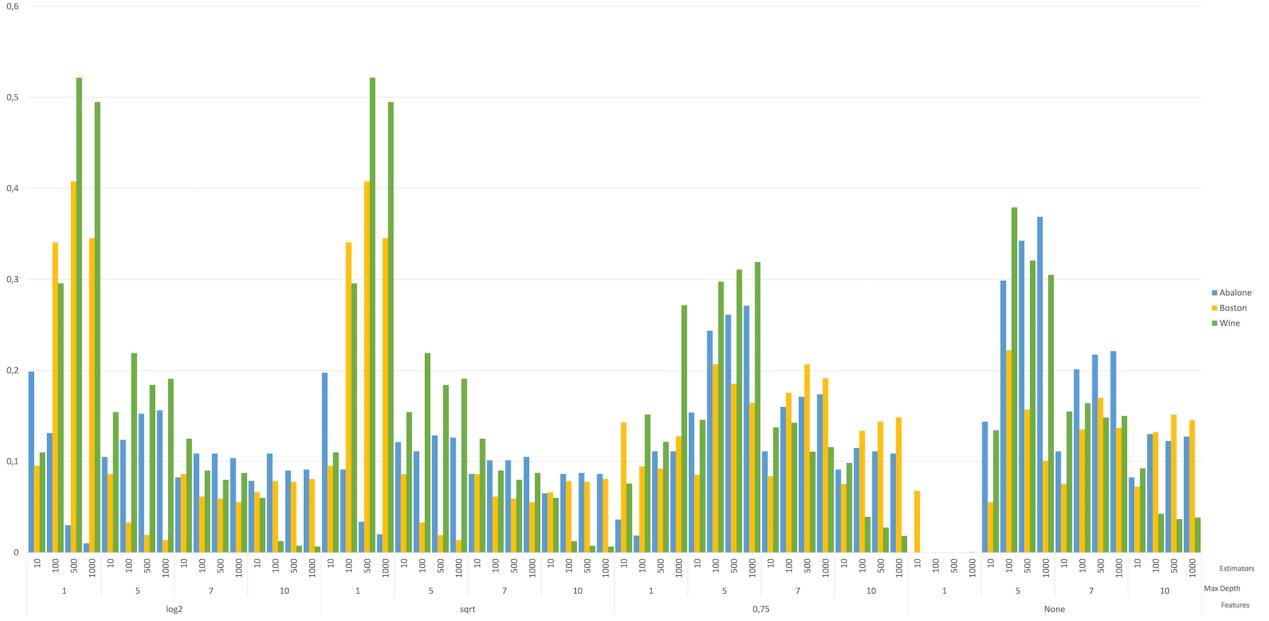


Figure 17: Regression: analysis FR relation to RF's parameters

Inspecting how the LF's parameters are affecting the FR%, in Figure 18, we can see that AR+RS method provides better results for Abalone, while DSi for Boston and Wine. However, DSo cannot reach desirable levels of FR% for any case.

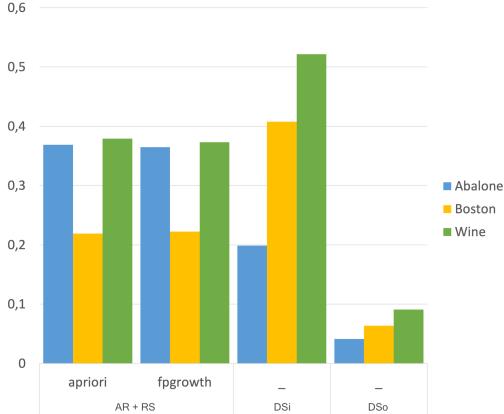


Figure 18: Regression: analysis of FR relation to LF's parameters

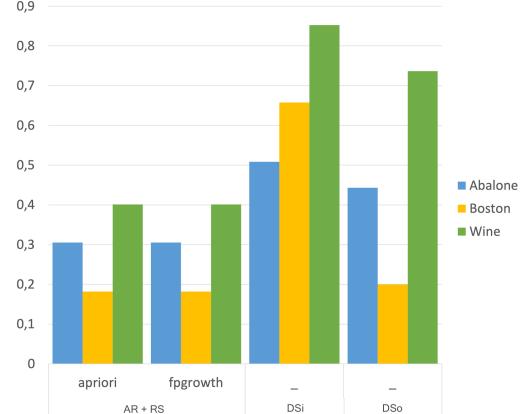


Figure 19: Regression: analysis of PR relation to LF's parameters

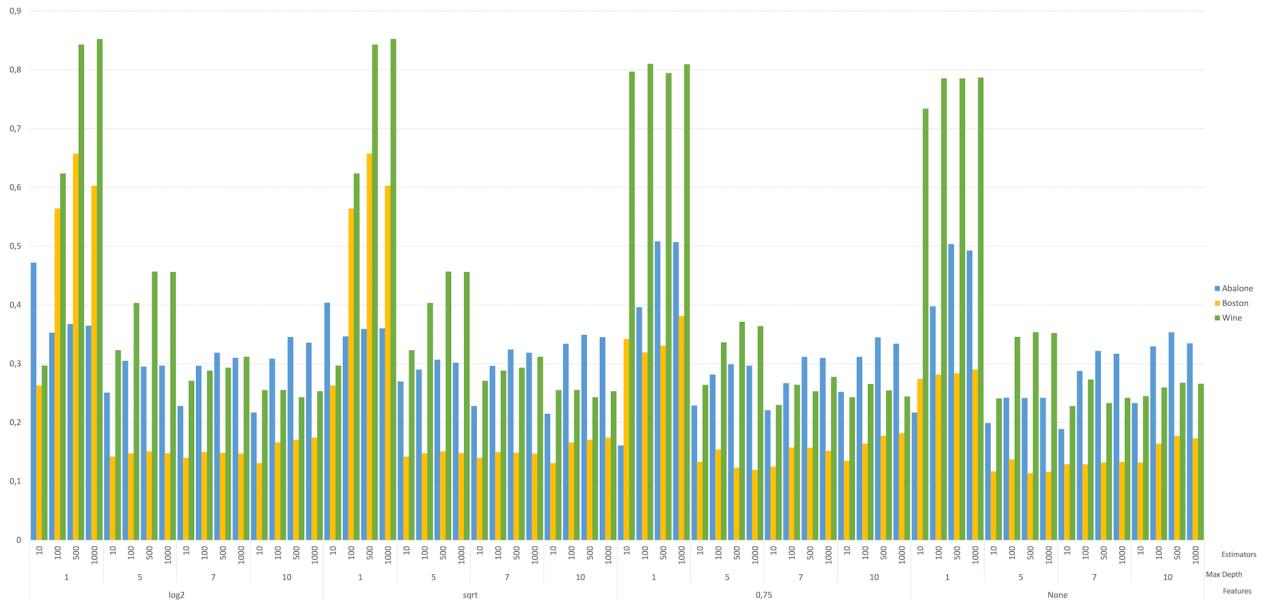


Figure 20: Regression: analysis of PR relation to RF's parameters

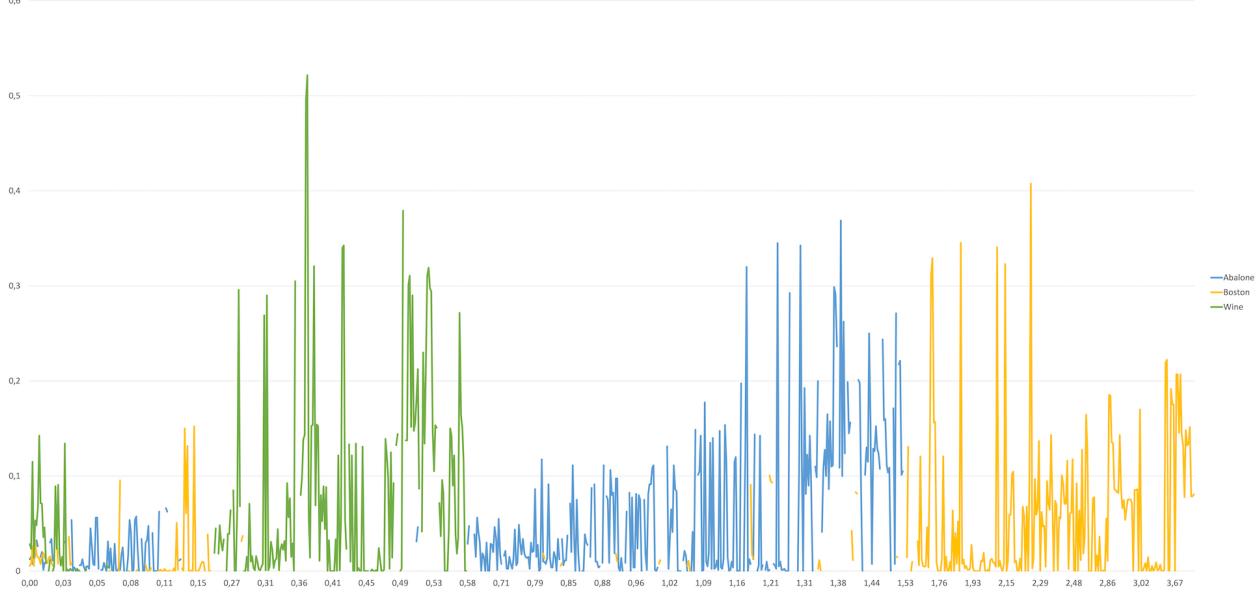
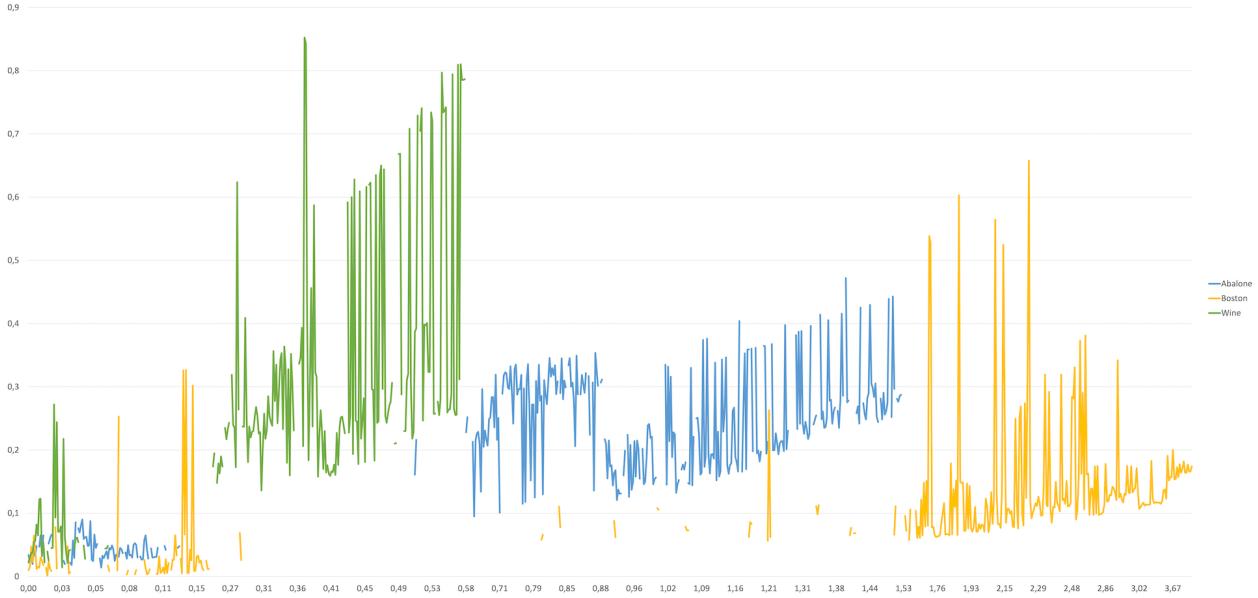
The same pattern we identified for the FR% relation to RF's parameters, it is apparent for the relation of PR% with the RF's parameters as well (Figure 20). Setting *estimators* between 500 or 1000 and *max depth* to either 1 or 5, the PR is between 50% to 85%. However, *max features* do not affect the PR%.

In Figure 19, we can see how the parameters of LF affect the PR% in these datasets. We observe the highest PR%, over 50%, with the DSi reduction method of LF. DSo is also better than AR+RS in terms of PR%.

Examining the relation of *local_error* with the FR% (Figure 21), we can say that for the Wine dataset we can achieve high FR%, over 50%, with a low *local_error* of around 0.36. For the Abalone dataset, we need a *local_error* with a value between [1.1, 1.4] in order to achieve approximately 35% of FR. Finally, for Boston in order to achieve FR% higher than 40% we need a *local_error* around 2.2.

	Min	Max	Mean	Std
Abalone	3.0	19.0	9.74	2.86
Boston	5.0	50.0	22.53	9.19
Wine	3.0	9.0	5.82	0.87

Table 3: Statistics of target variable of regression task's datasets

Figure 21: Regression: analysis of FR relation to *local_error*Figure 22: Regression: analysis of PR relation to *local_error*

Finally, about the relation of PR% with the *local_error*, we observe, in Figure 22, that we acquire higher PR% when we allow higher *local_error*, in every dataset. In order to let the reader understand better the relation of both the FR% and PR% with the *local_error*, we present the target variable statistics of each dataset in Table 3. This will help associating the *local_error* with the actual values of the target variable of each dataset.

5.3 Time analysis

Optimisation of LF implementation was an emerging issue to address, mainly due to the need for timely responses. The speed of providing explanations to a model's predictions in a variety of applications is indeed a very important factor, specifically, for on-line applications. The second direction of our experiments targets the analysis of response time. We are comparing LF to its preliminary version in order to back up the assumption of Sections 4.3.1 and 4.4. For this analysis we use only the binary datasets of Banknote and Heart (statlog) because the preliminary version of LF is only

applicable to binary tasks, and we do not use the Adult dataset, as the old version was unable to finish our exhaustive analysis within a relatively reasonable time.

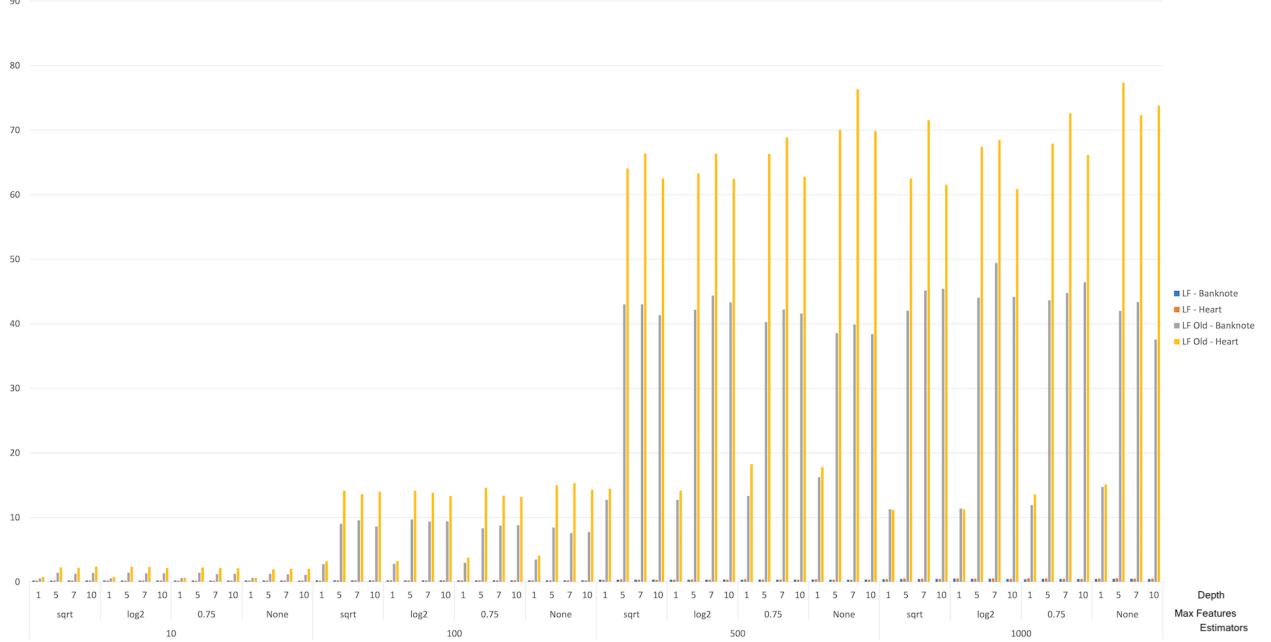


Figure 23: Comparison of preliminary LF and LF on features' ranges generation without reduction (y-axis in seconds)

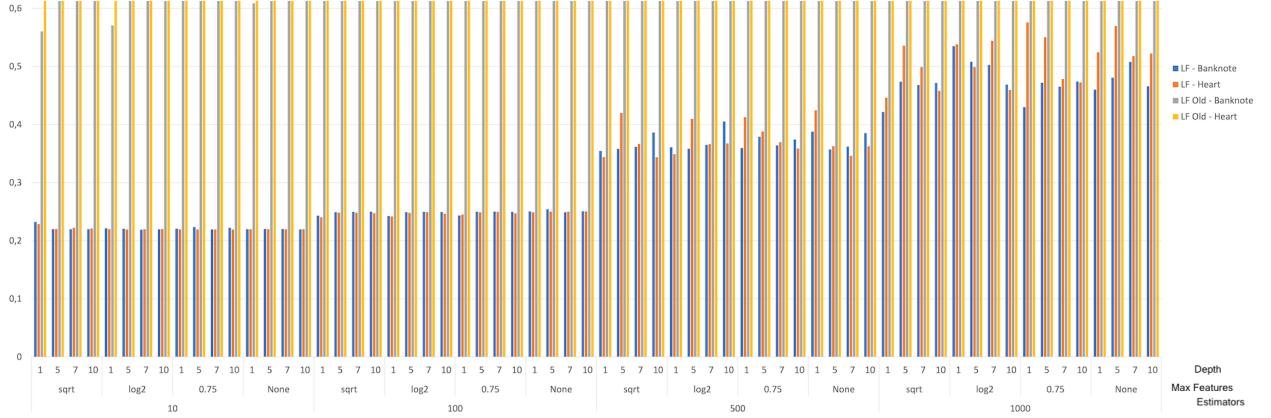


Figure 24: Comparison of preliminary LF and LF on features' ranges generation without reduction (y-axis in second - zoomed)

The first aspect of the algorithm we are going to inspect, concerns the identification of the ranges for each feature appearing in a rule without applying any reduction method. In Figure 23 and 24, we can see that LF is outperforming the preliminary version, with a huge decrease of 93% for Banknote and 95% for Heart datasets. In Figure 24 we are zooming the y-axis in order to make visible that LF runs approximately between 0.2 to 0.6 seconds per explanation, in contrast to the preliminary version which generates explanations from 0.2 to almost 80 seconds.

Then, we compare the two versions on the generation of rules applying reduction. We remind here that in the reduction pipeline we optimised the reduction through association rules, thus this experiment will inspect the effect of this optimisation. In Figure 25 and 26, we can see that LF is once again outperforming the preliminary version, with a huge decrease of 94% for Banknote and 96% for Heart datasets. In Figure 26 we are zooming the y-axis in order to make visible that the version of LF runs approximately between 0.2 to 11 seconds per explanation, in contrast to the preliminary version which generates explanations from 2 to 128 seconds, and even over 280 in few extreme cases.

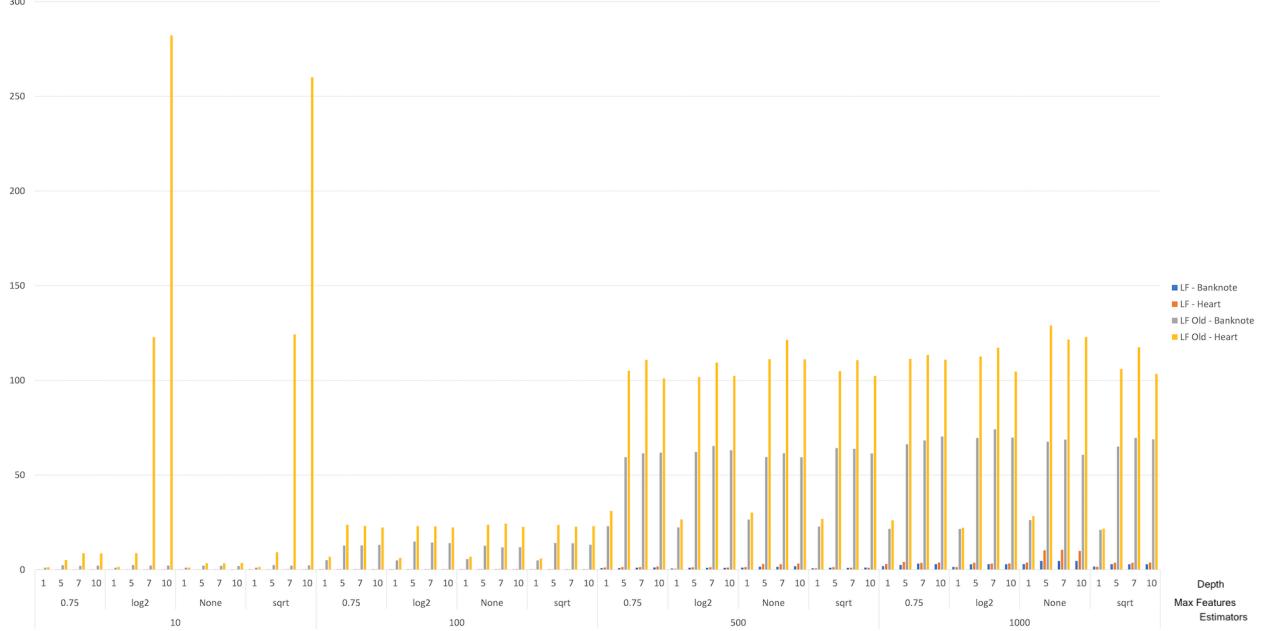


Figure 25: Comparison of preliminary LF and LF on features' ranges generation with reduction (y-axis in second)

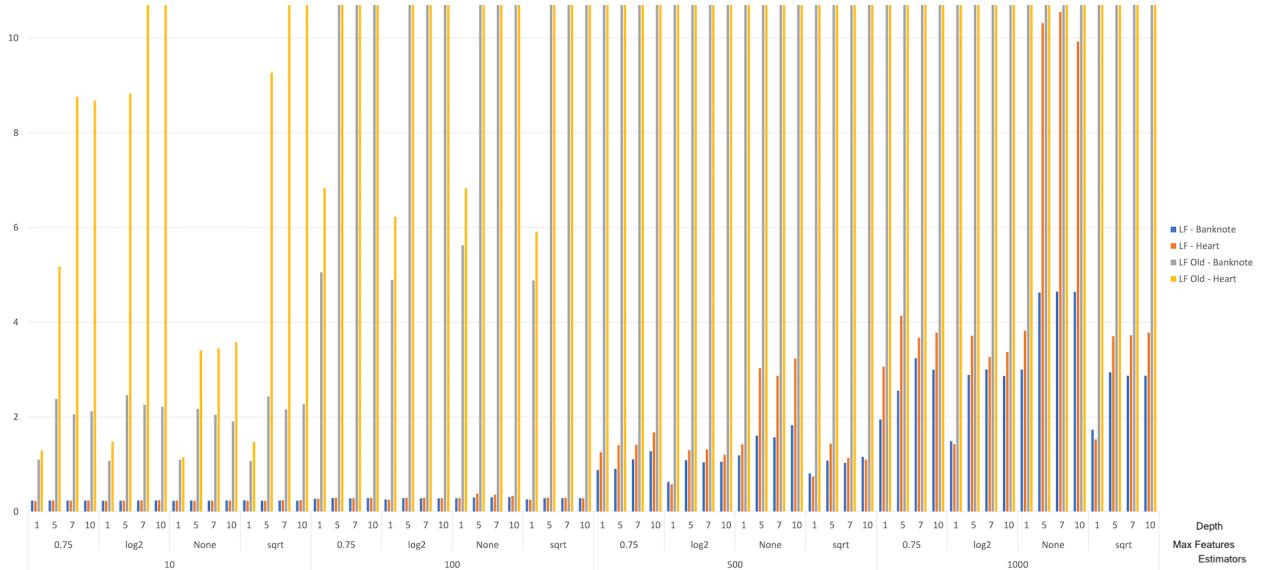


Figure 26: Comparison of preliminary LF and LF on features' ranges generation with reduction (y-axis in second - zoomed)

5.4 Comparison study with other techniques

In our experiments we include three techniques from the literature, Anchors [26] (AN), CHIRPS [21] (CH) and DefragTrees [19] (DF), as well as a simple surrogate model based on a decision tree in a global (GS) and a local (LS) fashion, to use it as a baseline. AN is a local, model-agnostic interpretation technique for binary and multi-class classification. CH is a local, model-specific technique for interpreting random forest models for binary and multi-class classification tasks. DF is a global, model-specific technique that interprets tree ensembles on binary and multi-class classification and regression. The applicability of each of these techniques to each dataset is presented in Table 4.

We use three different evaluation metrics. The first one computes the length of an explanation for an instance (*rule_length*). The second one measures the coverage of a rule, which describes the fraction of rules of a dataset that

	Banknote	Heart (Statlog)	Adult (Census)	Glass	Image Segmentation	Abalone (MC)	Abalone (R)	Boston	Wine Quality
Baseline Surrogate	✓	✓	✓	✓	✓	✓	✓	✓	✓
DefragTrees	✓	✓	✓	✓	✓	✓	✓	✓	✓
Anchors	✓	✓	✓	✓	✓	✓	-	-	-
CHIRPS	✓	✓	✓	✓	✓	✓	-	-	-

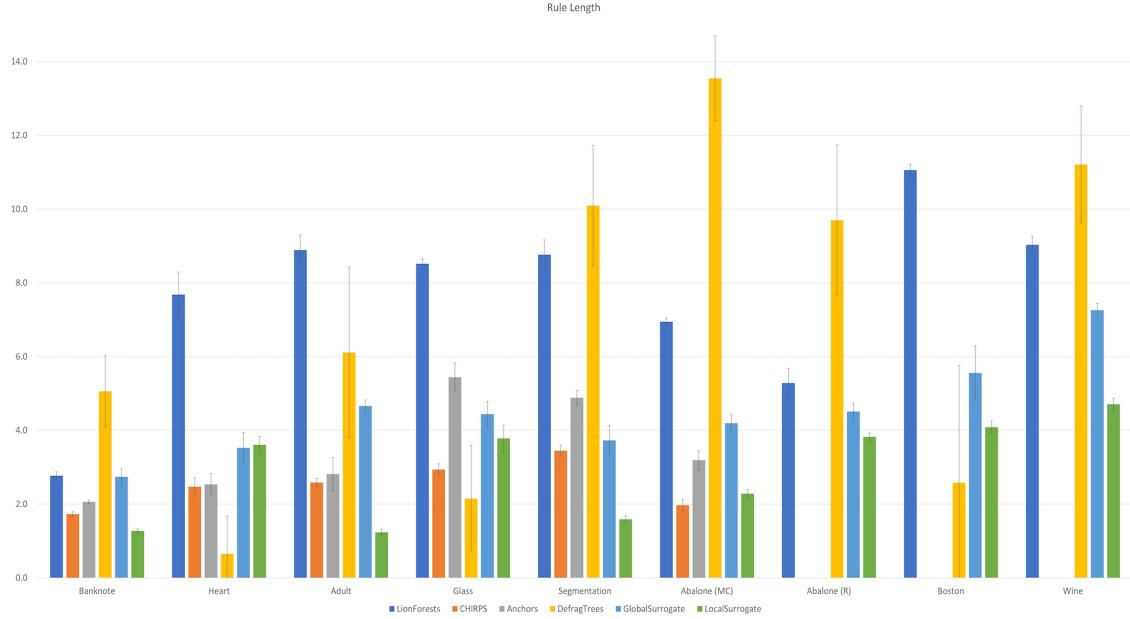
Table 4: Applicability of interpretation techniques to each dataset

satisfy the antecedent of that rule (*coverage*). The third one is the precision of a rule, which measures the performance of that rule on the covered instances (*precision*), measuring the actual *precision* in classification problems, and the *mean absolute error* (*mae*) in regression problems. We use the leave-one-out (LOO) cross-validation method in subsets of 10 random samples including the minimum between 10% or 100 instances for each dataset, to compute these metrics.

Finally, we propose a novel property called “conclusiveness”. A rule-based explanation method is defined as “conclusive” when the antecedent of a rule - explanation - is a definite proof for the consequent. LionForests by design is “conclusive” because each rule contains every necessary feature to cover at least the minimum amount of paths, which will always produce the same outcome. However, this automatically suggests that LionForests’ explanations will be more specific, and they will have low coverage. In the following, we experimentally prove, with at least one example per method, that all the other methods are not “conclusive”.

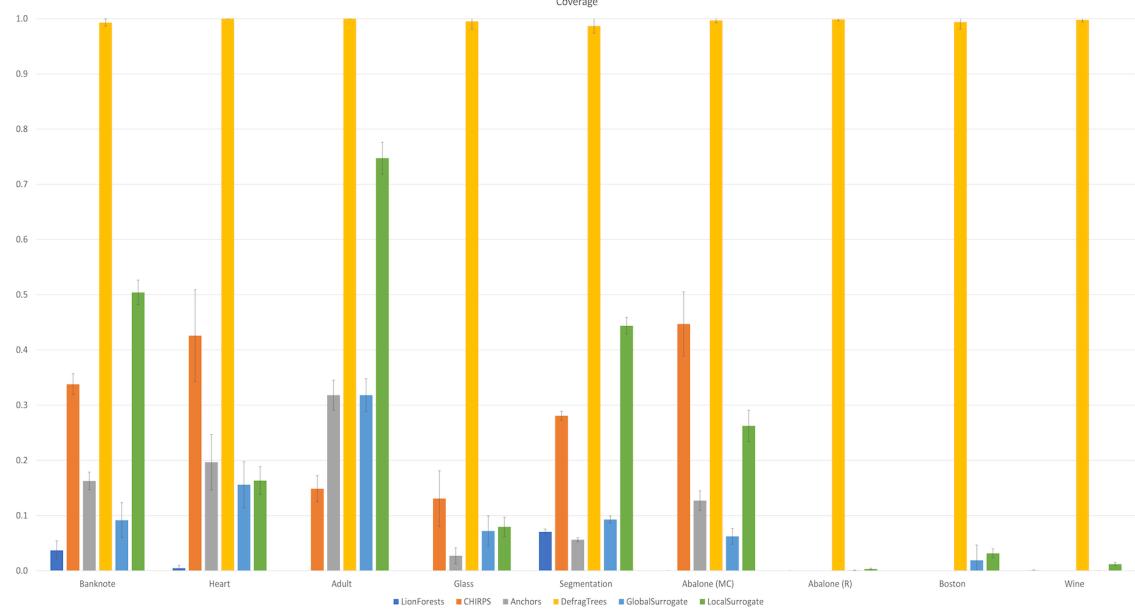
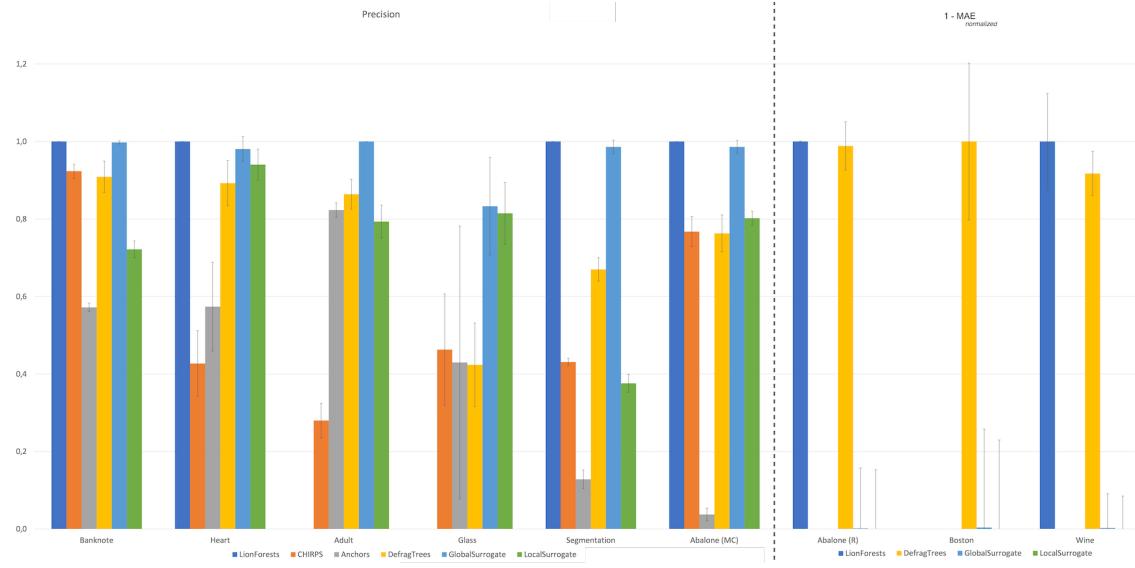
5.4.1 Comparison with evaluation metrics

The first metric to be examined is the *rule_length*. The results of the comparison are visible in Figure 27. In these results, it is clear that LF and DF are in almost every case the algorithms providing rules with a lot of conditions. Specifically, we can see that LF and DF are producing rules of around 7.66 ± 0.28 and 6.79 ± 1.7 length by average. The other techniques, CH= 2.53 ± 0.15 , AN= 3.49 ± 0.27 , GS= 4.51 ± 0.33 and LS= 2.93 ± 0.15 , seem to provide smaller rules. CH and AN are not present in the last three columns because these algorithms are not applicable in regression tasks.

Figure 27: Comparison of techniques in terms of the *rule_length* metric

Coverage is the second metric we investigate (Figure 28). The most interesting findings in this experiment are the extremely high *coverage* of DF ($99.59\% \pm 0.65$), and the low coverage of LF ($1.26\% \pm 0.31$). The performance of the remaining algorithms performed is much lower than the DF’s performance, but higher than the LF’s performance, CH= $29.52\% \pm 4.04$, AN= $14.79\% \pm 2.15$, GS= $9.02\% \pm 1.99$ and LS= $24.96\% \pm 1.67$.

Finally, we evaluated the techniques using the *precision* metric. As already mentioned, *precision* is measured differently across the classification and regression tasks. Therefore, in Figure 29 the first 6 columns are measured using

Figure 28: Comparison of techniques in terms of the *coverage* metricFigure 29: Comparison of techniques in terms of the *precision* metric and $1 - mae_{normalized}$

the actual *precision* of classification tasks, while in regression we used *mae*. Moreover, we scaled the *mae* error in each of the results of the regression datasets to $[0, 1]$ and subtracted by $1 (1 - mae_{normalized})$ for better comparison and visualisation. The very specific rules of LF achieved a high $100\% \pm 0.0$ *precision* and the highest 1 ± 0.01 *mae* in all tasks, except from Adult, Glass and Boston datasets, where the coverage was 0. In the classification tasks, GS had a high *precision* performance ($96.39\% \pm 3.28$), while DF ($75.37\% \pm 5.38$) and LS ($74.13\% \pm 3.74$) achieved similar to each other results. CH ($54.86\% \pm 5.67$) and AN ($42.74\% \pm 8.94$) had the worst performance. For the regression tasks, DF resulted in a high score ($0.97\% \pm 0.11$), while the performance of GS ($0\% \pm 0.17$) and LS ($1\% \pm 0.16$) was very low.

5.4.2 Conclusiveness property investigation

Even if LF does not always have the best performance in terms of the aforementioned metrics, it has the ability to provide “conclusive” interpretation rules, making it trustworthy and competitive. This occurs because LF always retains at least the number of paths that will always produce the same result for a particular instance, and every feature (and its

corresponding range) from those paths is included in the final rule (see Section 4.1 and 4.2). In fact, in this section, we demonstrate that all of the algorithms under consideration are not conclusive, by providing at least one example for each one of them.

We will use the Banknote dataset for this qualitative assessment. We retain 10 of the 1372 instances for manual inspection, and the rest are used to train the RF model. After normalisation, the distributions of Banknote's 4 features $F = [Variance, Skew, Curtosis, Entropy]$ are between $[-1, 1]$. Then, we choose a random instance among the 10, $x_1 = [0.53, -0.25, -0.24, 0.53]$, which was predicted to be a 'fake banknote'. The following rules were provided by the examined interpretation techniques:

- LF:** If $0.36 \leq Variance \leq 1$ and $-0.62 \leq Curtosis \leq 1$ then 'fake banknote'
- CH:** If $0.23 < Variance$ then 'fake banknote'
- AN:** If $0.42 < Variance$ then 'fake banknote'
- LS:** If {} then 'fake banknote'

LF indicates that the value of the feature *Curtosis* must be within the range $[-0.62, 1]$, while CH, AN, and LS provide no such requirement. When the instance's value for this feature is set to -1 , the prediction is changed from "false banknote" to "real banknote". As a result, the three rules, with the exception of LF, were all not conclusive. We proceed looking at another instance $x_2 = [0.52, 0.79, -0.91, -0.35]$. We produced the interpretation rules again after RF predicted the instance as "false banknote".

- LF:** If $0.43 \leq Variance \leq 1$ and $-0.92 \leq Curtosis \leq -0.89$ then 'fake banknote'
- DF:** If $0.34 < Skew$ and $-0.97 < Curtosis$ then 'fake banknote'

However, we can see that DF lacks a condition for *Variance*, whereas LF lacks a condition for *Skew*. We changed both of them, and LF's decision not to include *Skew* was correct because there was no change in the outcome when we altered *Skew* from 0.79 to either 1 or -1 . The prediction, on the other hand, changed when we changed *Variance* from 0.52 to -1 . As a result, DF is not conclusive. Finally, we examined another instance $x_3 = [-0.23, 0.24, -0.73, 0.2]$ to argue that GS is also not conclusive. The RF classified this instance as a "real banknote". We obtained the following two interpretation rules after requesting LF and GS to produce them:

- LF:** If $-0.23 \leq Variance \leq -0.12$ and $-1 \leq Skew \leq 0.32$ and $-1 \leq Curtosis \leq -0.57$ then 'real banknote'
- GS:** If $-0.26 < Variance \leq 0.06$ and $0.23 < Skew \leq 0.32$ then 'real banknote'

It is clear that the rules are not identical. GS omitted the *Curtosis* feature, while LF provides a condition about this feature. We proceeded to modify this feature's values from -0.23 to 1 and the prediction changed to 'fake banknote'. Hence, we can say that GS is also not conclusive.

5.5 Qualitative evaluation

In this final section of the experiments, we will present an example using an instance from the Adult dataset. Apart from the prediction and the interpretation rule, we will show the interpretation visually, and then perform a few manual tests to guarantee that the interpretation rule is conclusive. Initially, we obtain the following prediction and interpretation rule, with and without LF's reduction, for a random instance of Adult:

- Original:** if $Marital Status = Married$ and $5119.0 \leq Capital Gain \leq 5316.5$ and $0 \leq Capital Loss \leq 1782.5$ and $33.5 \leq Age \leq 49.5$ and $38.5 \leq Hours Per Week \leq 40.5$ and $Occupation = Exec Managerial$ and $108326.999 \leq fnlwgt \leq 379670.501$ and $Education = Bachelors$ and $Sex = Male$ and $Workclass = Private$ and $Native Country = United States$ then $Income > 50K$
- LF:** if $Marital Status = Married$ and $5119.0 \leq Capital Gain \leq 5316.5$ and $31.5 \leq Age \leq 49.5$ and $108326.999 \leq fnlwgt \leq 379670.501$ then $Income > 50K$

Based on the two explanations, we can draw the following conclusions. The capacity of LF to reduce the rules is obvious. The new interpretation is 63% shorter since it includes only four conditions rather than eleven. The ranges of the features *Age* and *fnlwgt* are also broader in the LF's rule, thus the rule is less sensitive to fluctuations. Finally, we can see that LF reduced nearly all of the categorical features, which is very positive.

We will analyse the features *United States* and *Age*, which are categorical and numerical, respectively, using the visualisation offered by LF. Someone may notice that the feature *United States* does not appear in the particular

interpretation rule. This means that the value for *United States* in this case, *United States*, has no effect on the prediction and has been diminished in this instance's interpretation. However, as the visualisation tool in Figure 30 indicates, the only potential values that can influence the prediction are *Mexico* and *South*.

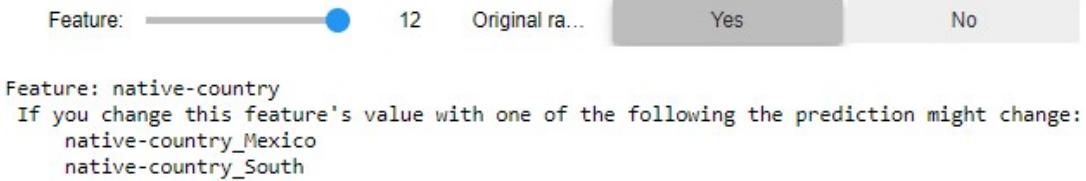


Figure 30: Visualisation of interpretation for feature *Native Country*

Then, we changed the value of *Native Country* to *Greece*, *United Kingdom*, *Japan* and another 36 countries not appearing in the list of values which may affect the prediction, and the outcome indeed remained the same. Our next modification concerns the *Age* feature, as presented in Figure 31. Since *Age* is a numerical feature, the distribution of this feature among the training data, as well as the range of the allowed values of this feature for the examined instance are presented in a plot.

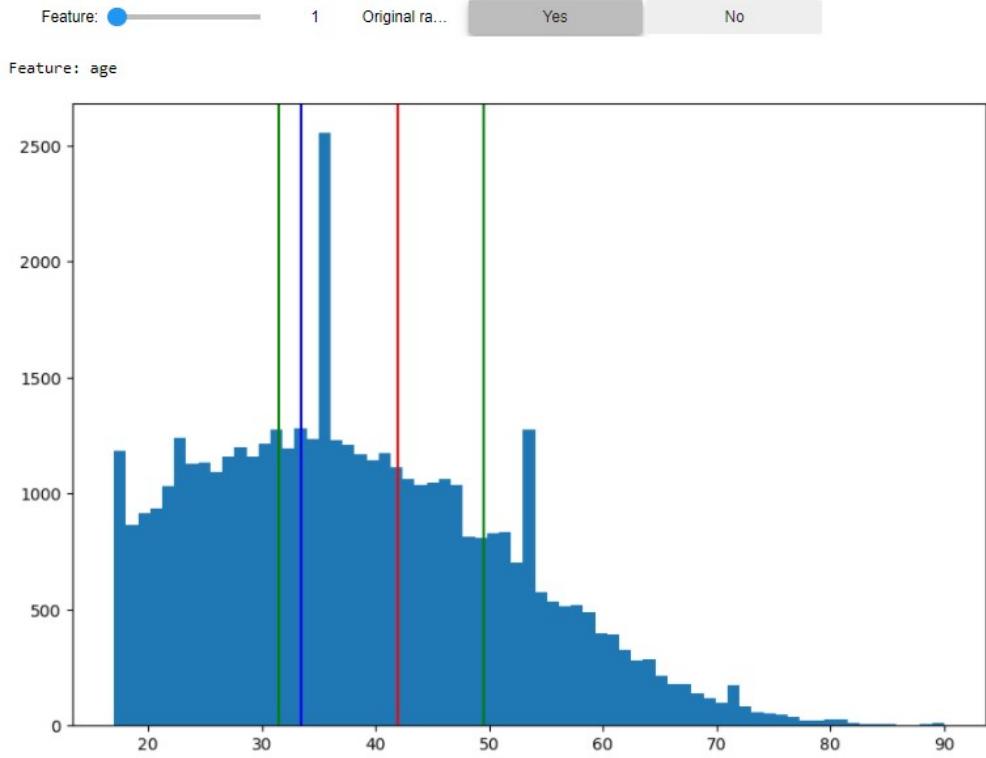


Figure 31: Visualisation of interpretation for feature *Age*

This plot firstly informs us about how the ranges of the feature's value were extended. The red vertical line reflects the instance's value for this feature (41), while the blue and green lines reflect the initial rule's and reduced rule's ranges, respectively (blue and green right lines are overlapping). We then considered modifying the *Age* with two values within the range, 32 and 48, and indeed the classification result remained the same as indicated by the extracted rule.

6 Discussion

We have used 8 different datasets from 3 different learning tasks in the above series of experiments to evaluate different aspects of LF and compare it to other techniques.

Via a sensitivity study, we initially investigated how the LF and RF parameters affect the reduction of features and paths (Section 5.2). The parameters ‘*max_features*’ and ‘*estimators*’ of the RF appear to be strongly associated with the reduction effect. Specifically, when estimators were 500 or greater, the feature reduction ration was always higher.

For the LF’s parameters, we discovered that reduction through association rules is required for high feature reduction in all classification scenarios, while reduction through random selection is always required for path reduction. Reduction through clustering seemed to assist in increasing feature reduction, but it was not as effective as reduction through association rules.

We can notice from the sensitivity analysis of the regression tasks that DSi and DSo are the most promising reduction techniques. We also investigated the impact of the *local_error* parameter, which confirmed the hypothesis that higher *local_error* would lead to higher feature and path reduction.

After the sensitivity analysis, we proceeded to the time analysis in Section 5.3. Those experiments proved how much the LF response time was improved in comparison to the preliminary version. Two core procedures were optimised, and the result was a 93% to 95% faster run-time on the first procedure, and a 94% to 96% on the second. This improvement renders the LF applicable even in online scenarios where immediate interpretations must be provided to users in real-time.

We then compared LF to SOTA algorithms using well-known metrics and a custom property. In terms of *rule_length* and *coverage*, LF did not outperform the other techniques, but it did achieve perfect *precision* in the classification experiments. The low performance of LF with respect to *coverage* is due to the extremely specific rules offered by LF. At the same time, the rules’ specificity renders them larger than the rules of the opposing algorithms. However, LF has the conclusiveness property, which renders it reliable and trustworthy, and by examining the other algorithms we proved that none of them has this property. Thus, despite their superior performance in the metrics favoured in the literature, they lack the “conclusiveness” property, which we believe is necessary for any interpretation algorithm that provides rules to the user towards more transparent, complete, and reliable explanations.

We finally presented a detailed use case from one dataset. We chose the Adult dataset which facilitates both numerical and categorical features. By taking a random instance we generated both the prediction by the RF model, the interpretation rule from the LF and the visual interface as well. We investigated both the ranges of a numerical value and the alternate values of a categorical feature using the user interface. In comparison to other algorithms that only provide the feature and the category in the rule, this tool allows the user to see alternate categorical values.

7 Conclusion

Random forest is one of the top performing machine learning algorithms in critical sectors such as health, industry, or retail. In the same time, its uninterpretable nature makes it an inappropriate solution, due to trustworthiness concerns, and even issues related to legal frameworks. Therefore, the need of injecting interpretability to such algorithms is evident. In a preliminary work, we introduced a random forest-specific local-based interpretation technique called LionForests. The interpretations are presented in the form of rules. Each rule is a conclusive set of conditions about the features that affected an instance’s prediction. LionForests implements a series of feature and path reduction approaches in order to provide smaller rules containing conditions with broader ranges. We are refining this methodology in a number of ways in this work. Providing a stable theoretic background, enhancing its core procedures towards timely responses, extending its applicability to a variety of learning tasks, LionForests technique undergoes a series of experiments. We investigated and evaluated how the parameters of a random forest model, as well as the parameters of LionForests, influence feature and path reduction in these experiments, which were focused on 8 separate datasets. A time analysis backs up our claim that LionForests responds in a timely manner, in contrast to its preliminary version. We compare LionForests using well-known metrics with cutting-edge techniques. However, while the performance of LionForests based on such metrics does not imply that our system is superior, its inherent property known as “conclusiveness” distinguishes it from the other techniques. We also demonstrate that all the other techniques are not conclusive. Finally, we give a detailed example of how to best exploit the LionForests technique, in a qualitative manner. Few of our next steps will be to evaluate LionForests in a human-oriented way, exploiting information from relevant works [49], to investigate its applicability to different types of data, as well as to examine different ensemble algorithms.

Acknowledgements

This paper is supported by the European Union’s Horizon 2020 research and innovation programme under grant agreement No 825619. AI4EU Project⁵.

⁵<https://www.ai4eu.eu>

References

- [1] Niall Firth. Apple card is being investigated over claims it gives women lower credit limits. <https://cutt.ly/oeGYCx5>, 2019. Accessed: 2019-11-18.
- [2] Angela Chen. Ibm's watson gave unsafe recommendations for treating cancer. <https://cutt.ly/keHQDma>, 2018. Accessed: 2019-11-18.
- [3] Samantha Cole. This trippy t-shirt makes you invisible to ai. <https://cutt.ly/FeHQHAA>, 2019. Accessed: 2019-11-18.
- [4] General Data Protection Regulation. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46. *Official Journal of the European Union (OJ)*, 59(1-88):294, 2016.
- [5] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 0210–0215. IEEE, 2018.
- [6] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.
- [7] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.
- [8] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. Benchmarking and survey of explanation methods for black box models. *arXiv preprint arXiv:2102.13076*, 2021.
- [9] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15(90):3133–3181, 2014.
- [10] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001.
- [11] Celine Vens and Fabrizio Costa. Random forest based feature induction. In Diane J. Cook, Jian Pei, Wei Wang, Osmar R. Zaïane, and Xindong Wu, editors, *11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11-14, 2011*, pages 744–753. IEEE Computer Society, 2011.
- [12] Paulo Angelo Alves Resende and André Costa Drummond. A survey of random forest based methods for intrusion detection systems. *ACM Comput. Surv.*, 51(3), May 2018.
- [13] T. Jemima Jebaseeli, R. Venkatesan, and K. Ramalakshmi. Fraud detection for credit card transactions using random forest algorithm. In J. Dinesh Peter, Steven L. Fernandes, and Amir H. Alavi, editors, *Intelligence in Big Data Technologies—Beyond the Hype*, pages 189–197, Singapore, 2021. Springer Singapore.
- [14] Huannan Zhang, Yilin Bi, Wangdong Jiang, Chuntian Luo, Shengjia Cao, Peng Guo, and Jianjun Zhang. Application of random forest classifier in loan default forecast. In Xingming Sun, Jinwei Wang, and Elisa Bertino, editors, *Artificial Intelligence and Security*, pages 410–420, Singapore, 2020. Springer Singapore.
- [15] Mecit Can Emre Simsekler, Abroon Qazi, Mohammad Amjad Alalami, Samer Ellahham, and Al Ozonoff. Evaluation of patient safety culture using a random forest algorithm. *Reliability Engineering & System Safety*, 204:107186, 2020.
- [16] Carlo Ricciardi, Marianna Amboni, Chiara De Santis, Gianluca Ricciardelli, Giovanni Improta, Luigi Iuppariello, Giovanni D'Addio, Paolo Barone, and Mario Cesarelli. Classifying different stages of parkinson's disease through random forests. In Jorge Henriques, Nuno Neves, and Paulo de Carvalho, editors, *XV Mediterranean Conference on Medical and Biological Engineering and Computing – MEDICON 2019*, pages 1155–1162, Cham, 2020. Springer International Publishing.
- [17] S. Ravikumar, V. Muralidharan, P. Ramesh, and Cheran Pandian. Fault diagnosis of self-aligning conveyor idler in coal handling belt conveyor system by statistical features using random forest algorithm. In Ning Zhou and S. Hemamalini, editors, *Advances in Smart Grid Technology*, pages 207–219, Singapore, 2021. Springer Singapore.
- [18] Shuyu Yao, Ming Wei, Lingyu Yan, Chunzhi Wang, Xinhua Dong, Fangrui Liu, and Ying Xiong. Prediction of crime hotspots based on spatial factors of random forest. In *15th International Conference on Computer Science & Education, ICCSE 2020, Delft, The Netherlands, August 18-22, 2020*, pages 811–815. IEEE, 2020.
- [19] Satoshi Hara and Kohei Hayashi. Making tree ensembles interpretable: A bayesian model selection approach. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 77–85, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

- [20] Houtao Deng. Interpreting tree ensembles with intrees. *International Journal of Data Science and Analytics*, 7(4):277–287, 2019.
- [21] Julian Hatwell, Mohamed Medhat Gaber, and R. Muhammad Atif Azad. CHIRPS: explaining random forest classification. *Artif. Intell. Rev.*, 53(8):5747–5788, 2020.
- [22] Alexander Moore, Vanessa Murdock, Yaxiong Cai, and Kristine Jones. Transparent tree ensembles. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1241–1244. ACM, 2018.
- [23] Mollas Ioannis, Bassiliades Nick, Vlahavas Ioannis, and Tsoumakas Grigorios. Lionforests: local interpretation of random forests. In Saffiotti Alessandro, Serafini Luciano, and Lukowicz Paul, editors, *First International Workshop on New Foundations for Human-Centered AI (NeHuAI 2020)*, number 2659 in CEUR Workshop Proceedings, page 17–24, Aachen, 2020.
- [24] Erik Strumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665, 2014.
- [25] Cristian Bucilua, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’06, page 535–541, New York, NY, USA, 2006. Association for Computing Machinery.
- [26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-Precision Model-Agnostic Explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [27] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. *CoRR*, abs/1805.10820, 2018.
- [28] Pedro Domingos. Knowledge discovery via multiple models. *Intelligent Data Analysis*, 2(1-4):187–202, 1998.
- [29] Yichen Zhou and Giles Hooker. Interpreting models via single tree approximation. *arXiv preprint arXiv:1610.09036*, 2016.
- [30] Stefan Th Gries. On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory*, 2019.
- [31] Xun Zhao, Yanhong Wu, Dik Lun Lee, and Weiwei Cui. iforest: Interpreting random forests via visual analytics. *IEEE transactions on visualization and computer graphics*, 25(1):407–416, 2018.
- [32] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.
- [33] Julian Hatwell, Mohamed Medhat Gaber, and R Azad. gbt-hips: Explaining the classifications of gradient boosted tree ensembles. *Applied Sciences*, 11(6):2511, 2021.
- [34] Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. Classification and regression trees. belmont, ca: Wadsworth. *International Group*, 432:151–166, 1984.
- [35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [36] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM, 1993.
- [37] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.
- [38] Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, 8(1):53–87, 2004.
- [39] B Nigam, A Nigam, and P Dalal. Comparative study of top 10 algorithms for association rule mining. *International Journal of Computer Sciences and Engineering*, 5(8), 2017.
- [40] Rufai Yusuf and Zaharaddeen Lawal. Performance analysis of apriori and fp-growth algorithms (association rule mining). *International Journal of Computer Applications in Technology*, 7:279–293, 03 2016.
- [41] Leonard Kaufman and Peter J Rousseeuw. Clustering by means of medoids. statistical data analysis based on the 11 norm. *Y. Dodge, Ed*, pages 405–416, 1987.
- [42] Erich Schubert and Michael Gertz. Improving the cluster structure extracted from OPTICS plots. In Rainer Gemulla, Simone Paolo Ponzetto, Christian Bizer, Margret Keuper, and Heiner Stuckenschmidt, editors, *Proceedings of the Conference "Lernen, Wissen, Daten, Analysen", LWDA 2018, Mannheim, Germany, August 22-24, 2018*, volume 2191 of *CEUR Workshop Proceedings*, pages 318–329. CEUR-WS.org, 2018.

- [43] Ulrike von Luxburg. A tutorial on spectral clustering. *Stat. Comput.*, 17(4):395–416, 2007.
- [44] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.
- [45] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, page to appear, 1996.
- [46] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4):547–553, 2009.
- [47] David Harrison Jr and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102, 1978.
- [48] D. Clark, Z. Schreter, A. Adams, R. C. Williamson, A. Burkitt, and P. Bartlett. A quantitative comparison of dystal and backpropagation, australian conference; 7th, neural networks. In *Neural networks, PROCEEDINGS OF THE SEVENTH AUSTRALIAN CONFERENCE ON NEURAL NETWORKS, Australian conference; 7th, Neural networks*, pages 132–137. Australian National University;, 1996.
- [49] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. Evaluating xai: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291:103404, 2021.