Machine Learning with Applications 6 (2021) 100094

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

A comparison among interpretative proposals for Random Forests

Massimo Aria [a,*], Corrado Cuccurullo [b], Agostino Gnasso [a]

[a] Department of Economics and Statistics, University of Naples Federico II, Italy
[b] Department of Economics, University of Campania Lugi Vanvitelli, Italy

ARTICLE INFO

Keywords:
Random Forest
Model interpretation
Rule extraction
inTrees
NodeHarvest

ABSTRACT

The growing success of Machine Learning (ML) is making significant improvements to predictive models, facilitating their integration in various application fields. Despite its growing success, there are some limitations and disadvantages: the most significant is the lack of interpretability that does not allow users to understand how particular decisions are made. Our study focus on one of the best performing and most used models in the Machine Learning framework, the Random Forest model. It is known as an efficient model of ensemble learning, as it ensures high predictive precision, flexibility, and immediacy; it is recognized as an intuitive and understandable approach to the construction process, but it is also considered a Black Box model due to the large number of deep decision trees produced within it.

The aim of this research is twofold. We present a survey about interpretative proposal for Random Forest and then we perform a machine learning experiment providing a comparison between two methodologies, inTrees, and NodeHarvest, that represent the main approaches in the rule extraction framework. The proposed experiment compares methods performance on six real datasets covering different data characteristics: n. of observations, balanced/unbalanced response, the presence of categorical and numerical predictors. This study contributes to picture a review of the methods and tools proposed for ensemble tree interpretation, and identify, in the class of rule extraction approaches, the best proposal.

1. Introduction

Machine learning is a data analysis method that automates the construction of analytical models. It is a branch of Artificial Intelligence and is based on the idea that systems can learn from data, identify patterns on their own and make decisions with minimal human intervention (Mitchell, 1997).

The use of Ensemble methods in Machine Learning is given by the fact that different predictive models produce different results for the same inputs.

Ensemble Learning refers to all approaches that increase predictive performance by combining the outputs of a set of induced hypotheses, also called base learners, into a single predictive model that has the purpose of decreasing variance, altering bias, and improving predictions.

An ensemble learner can match any machine learning algorithm, for example, it could be a decision tree, a neuronal network, or a linear regression model.

Classification and Regression Trees (CART) are supervised learning techniques that use a non-parametric approach (Breiman, Friedman, Olshen, & Stone, 1984).

The process of building trees is intuitive and simple for the human mind, which implies a simple and useful interpretation, but it is not competitive in terms of accuracy concerning other regression and classification approaches. However, by aggregating many decision trees, predictive performance can be substantially improved. For this purpose, Breiman proposed Random Forest as a non-linear approach that aims to achieve greater accuracy by averaging multiple decision trees, each of which is grown according to two random steps: the first step consist of the use of a bootstrap sample to train each tree, while the second is the use, at each internal node, of a random subset of variables to generate splits (Breiman, 2001).

Random Forest is an evolution of Bagging which aims to reduce the variance of a statistical model, simulates the variability of data through the random extraction of bootstrap samples from a single training set and aggregates predictions on a new record (see Breiman, 1996). It can improve predictions for many supervised methods, especially decision trees. These trees are grown in-depth without a pruning phase. The result is a set of classifiers characterized by low bias but also high variance. Then, the ensemble procedure reduces the variance by calculating predictions as the average of the generated trees (Hastie, Tibshirani, & Friedman, 2009).

Bagging produces high correlated trees when data are characterized by a subset of strong predictors. In this case the accuracy is not crucial with respect to a single tree (Hastie et al., 2009). Being an evolution

* Corresponding author.
E-mail addresses: massimo.aria@unina.it (M. Aria), corrado.cuccurullo@unicampania.it (C. Cuccurullo), agostino.gnasso@unina.it (A. Gnasso).

https://doi.org/10.1016/j.mlwa.2021.100094
Received 19 January 2021; Received in revised form 23 April 2021; Accepted 21 June 2021
Available online 27 June 2021

of Bagging, Random Forest aims to obtain even more different and unrelated trees.

In order to achieve this goal, when the trees grow, a subset of features is selected as candidates split from the complete set of predictors. The splitting phase can only use one of these features.

Consequently, at each split the algorithm will not consider the majority of the available predictors. All this can be translated into the goal of obtaining a less variable and more reliable ensemble learner, therefore as a process of decorrelation of trees.

The use of the Random Forest models is spreading in all research domains, from the hard sciences to the humanities, even if in some specific contexts, such as economics and medicine, their use is limited by the impossibility to obtain an interpretation of causal links between predictors and response. The main feature of RF models is the high predictive accuracy obtained through non-parametric approaches based on iterative algorithms which however generate so-called "black-box" models. These models are not interpretable through parameters and functional forms, as is the case with classical regression analysis. This characteristic is completely unsatisfactory in application domains where the decision-maker needs to understand the causal mechanisms that generate a model output and therefore justify the choices.

Hence the need to develop new techniques able to reconstruct the causal relationships and interactions between predictors and response used in an RF model.

The purpose of this study, therefore, is represented by a twofold objective:

- Introduce a series of approaches aimed at providing the interpretation necessary to understand the Random Forest model;
- Identify the best approach to provide a model interpretation for Random Forest models, through a comparative experiment.

2. Related work

In machine learning, interpretability is defined as the ability to explain or provide meaning in terms that a human being can understand (Doshi-Velez & Kim, 2017). To undertake a decision-making process, having confidence in a machine learning model is essential in order to feel reassured when analyzing and using it.

Ribeiro, Singh, and Guestrin (2016) identify two different but related definitions of trust: trust in a prediction and trust in a model.

Trusting a prediction implies that the user will take a certain action based on it; it is important to determine this trust since the model will be used to make decisions. It is a striking example, the use of a decision-making process in the medical or financial field and the consequence of acting with absolute confidence in the predictions obtained.

While having confidence in a model is equivalent to evaluating the model as a whole and testing its generalization capacity with appropriate evaluation metrics. A problem that recurs in using data from real contexts is that they are often significantly different and the chosen metric may not be adequate, therefore an inspection procedure of individual predictions and their interpretations may be the best choice.

Following the common goal of interpretability, in recent years there has been a considerable interest that has developed a large body of research on Interpretable Machine Learning (see e.g. Adadi & Berrada, 2018; Došilović, Brčić, & Hlupić, 2018; Du, Liu, & Hu, 2019; Guidotti et al., 2018; Haddouchi & Berrado, 2019; Lipton, 2018).

All these papers compared the different proposals analyzing the methodological aspects but neither of these performed an accuracy comparison identifying the best interpretative approach for Random Forest. For example, Haddouchi & Berrado proposed a taxonomy of RF interpretative methods, that we adopt in this work, but they did not consider the accuracy aspects.

In today's society, machine learning models are recurrent in many application contexts, which makes it necessary to research methods of interpreting such systems.

In some application contexts such as medicine or economics, machine learning models must be interpretable as we are more interested in the interpretation of the causal relationships between predictors and response rather than in the accuracy of predictions.

It is important to understand how a prediction is reached to control the functioning of the model. Consider, for example, the financial loan: to decide the assignment of a loan, a machine learning model is implemented, the outcome of which can be negative or positive; through interpretability, it could be possible to check for errors and/or satisfy the customer's requests regarding the exact reasons why the loan was rejected.

We will present some methodologies and tools proposed to interpret the Random Forest.

In order to have an organized and clear vision of the studies, the classification of techniques proposed by Haddouchi and Berrado (2019) will be used (Fig. 1).

Proposals are divided into two categories, those that use insights inherent to the internal processing approaches and those that use methods based on RF post-processing (Post-Hoc approaches).

2.1. Internal processing approaches

These regards all tools and methods as well as techniques whose purpose is to provide a global overview of the model. In general, they do not provide a structure but a series of measures useful for interpreting the results obtained.

2.1.1. Random forest extra information

The variable importance (VI) of a feature X_m (where m indicates the number of features or variables) for the prediction of Y can be evaluated in two ways: a measure that evaluates the improvement made by the split criterion for each split, called *Mean Decrease Impurity (MDI)* is a measure that is aimed at calculating the permutations of OOB observations, called *Mean Decrease Accuracy (MDA)* (see Breiman, 2001; Genuer, Poggi, & Tuleau-Malot, 2010; Louppe, Wehenkel, Sutera, & Geurts, 2013).

In addition to Variable Importance, another approach to analyzing the relationships between features and predictions is partial dependence through the *Partial Dependence Plot (PDP)* (Friedman, 2001).

It describes the marginal effect that the features assume concerning predictions. Friedman (2001) elaborates this approach through the display of functions with a subset selected by input variables as it is simpler than a function having a large argument.

Let z_l be a target subset of size l of the input variable x, $z_l = \{z_1, \dots, z_l\} \subset \{x_1, \dots, x_n\}$, and let z_* be the complement subset $z_* \cup z_l = x$.

The approximation $\hat{F}(x)$ depends on the variables included in the subset $\hat{F}(x) = \hat{F}(z_l, z_*)$.

If there is a dependence on a specific value for the variables in z_*, then $\hat{F}(x)$ can be considered as a function of variables chosen in the subset z_l[1]: in particular $\hat{F}_{z_*}(z_l) = \hat{F}(z_l|z_*)$.

If this dependence turns out to be not too strong, then the mean $\bar{F}_l(z_l)$ can be understood as the partial dependence of $\hat{F}(x)$ on the chosen subset of variables z_l: the mean would then correspond to

$$\bar{F}_l(z_l) = E_{z_*}\left[\hat{F}(x)\right] = \int \hat{F}(z_l, z_*) p_*(z_*) dz_* \tag{1}$$

where $p_*(z_*)$ is the marginal probability density of z_*. If this is estimated on the training set then $\bar{F}_l(z_l)$ becomes

$$\bar{F}_l(z_l) = \frac{1}{N} \sum_{i=1}^{N} \hat{F}(z_l, z_{i,*}) \tag{2}$$

[1] In general, the functional form of $\hat{F}_{z_*}(z_l)$ will depend on the particular values chosen for z_*.

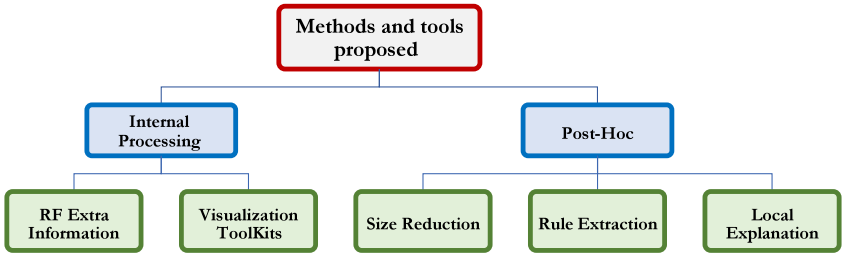

Fig. 1. Classification of the methods aiming the random forest interpretability.

In conclusion, $\bar{F}_l(z_l)$ allows obtaining a complete description of the nature of the variation of $\hat{F}(x)$ on the chosen subset of input variables z_l.

Finally, the last approach that allows the analysis of additional information is the *Proximity Matrix*. It represents the fraction of trees in which elements i and j fall into the same terminal node.

The *random forest proximity* of a pair of points is an unweighted average of the number of trees in the random forest model where the points end in the same leaf [6]:

$$proximity_{RF}(x, x') = \frac{1}{T}\sum_{i=1}^{T}\sum_{j=1}^{\tau_i} f(x \in R_{j,i}) f(x' \in R_{j,i}) \tag{3}$$

where T denotes the number of trees, with τ_i the leaf node and with $R_{j,i}$ the region of the feature space.

It results that the distance between a pair of points is $d^{RF}(x, x') = 1 - proximity_{RF}(x, x')$. Since the regions $\{R_{j,i}\}_{j=1}^{\tau_i}$ divide the feature space, each point $x \in X$, can be at most in one region and therefore the internal sum of *RF proximity* takes values 0 and 1 for each tree.

This matrix is used to obtain a *Multidimensional scaling plot (MDS)*[2] which is useful to identify observations that the Random Forest perceives as similar since they often end up in the same terminal nodes.

The basic idea is that similar observations should be found more frequently in the same terminal nodes than dissimilar observations.

2.1.2. Visualization toolkits

The visualization toolkits (Haddouchi & Berrado, 2019) allow the analysis of the aforementioned approaches from a graphic point of view. Liaw, Wiener, et al. (2002) propose the *randomForest* tool implemented in R which allows obtaining both the RF classifier and the internal measures of variable importance, proximity matrix, and partial dependence.

Ehrlinger (2016) proposes the package in R *ggRandomForests* where it uses measures for the accuracy of the predictions such as the Variable Importance and the Minimal Depth, measures of association of variables such as the Partial Variable Dependence and measures of interactions of the variables such as the Pairwise Minimal Depth Interactions. The variable importance is calculated using the Mean Decrease Accuracy, evaluating the difference between the OOB prediction error before and after the permutation.[3] That allows to show the importance that each variable assumes on the respective class of the response. A high value of VI indicates that the omission or incorrect specification of that particular variable greatly reduces the predictive accuracy.

As an alternative to the variable importance, the Minimal Depth inspects the construction of the forest to classify the variables. It assumes that the high importance variables are those that most frequently divide the nodes in the initial levels of a tree, that is, those that divide large samples of the population. The Minimal Depth averages the depth of the first division for each variable considering all trees in the forest; corresponds to the threshold θ where smaller values of it $\theta < v$ indicate greater importance assumed by that particular variable. Since VI and Minimal Depth are two different measures, the instrument compares them to have definitive confirmation of which variable contributes or not to the prediction.

Partial Dependence Plots are constructed by taking equidistant points along the distribution of the variable of interest X; for each value $(X = x)$, the average of the RF predictions over all the covariates in X is calculated.

Conditioning Plots are also implemented that allow you to analyze the dependence of the prediction on two or more variables. By characterizing the Conditional Plots as Stratified Variable Dependence Plots, it is also possible to obtain the Conditional Partial Dependence Plot, that is, before determining the conditional membership of a group, the partial dependence estimates on each subgroup are calculated. Furthermore, with the use of different variable dependence measures, it is possible to calculate the pairwise interaction between variables. In particular, the calculation of the Minimal Depth is altered by recalculating it for the jth variable with respect to the maximum subtree of the ith variable. Finding the interaction consists of calculating all the interactions of minimum depth in pairs, thus obtaining a pxp interaction matrix. The values on the diagonal are normalized and correlated with the Minimal Depth with respect to the root node.

Multiple authors have focused on the interpretation of RF through a visualization system. Zhao, Wu, Lee, and Cui (2018) recently developed iForest, a tool that allows interactive RF visualization in Python. In short, this tool develops a Feature View capable of illustrating the relationship between input features and predictions, it is able to show the underlying mechanism of RF by summarizing different decisional paths, thus making comprehension and exploration accessible to users. of the partition logics of the analyzed paths. It can be guessed that the literature has had a greater focus on techniques that describe how features affect predictions. It may happen that an increase in the number of input variables greatly increases the complexity of these techniques and also the understanding. To this end (Tan, Hooker, & Wells, 2016) prepare their research on distance metrics, introducing four prototype selection methods, representative points that provide a condensed view of a set of data with the intent of explaining a prediction through the use of these.

2.2. Post-Hoc approaches

These approaches aim to identify a relationship structure among response and predictors. An example could be represented by a surrogate model that approximates the original RF model with the aim of making it simpler and easier to interpret for the intended user.

2.2.1. Size reduction

An essential premise is that Tree ensembles, like Random Forests, have internal characteristics that can be used to approximate an understandable global model: Chipman, George, and McCulloh (1998) provide an interesting observation: although a large number of different trees are identified by the RF, many of them differ only for a few nodes,

[2] The Multidimensional scaling plot is the equivalent of the Principal Components.

[3] Unlike the tool proposed by Liaw, where in addition to the MDA it also measures the total decrease of the impurities of the node for each split averaged over all the trees.

or only in architecture, but use the same division of space as features X.

Wang and Zhang (2009) contribute to the search for an optimal small forest by setting themselves two objectives: to maintain a level of prediction accuracy similar to, or even better than, the original forest; reduce the number of trees to a level that is considered manageable. In order to reach the minimum forest size, they consider three measures that determine the importance of a tree in terms of prediction performance: the first focuses on prediction and determines whether a tree can be removed considering the impact of this removal on overall accuracy; another measure is based on the similarity between two trees, a tree will be removed if it is similar to others in the forest; the last measure is an altered version of the previous one, which considers a narrow similarity, i.e. the correlation between two trees is evaluated in terms of similarity of the predictions.

After evaluating the three methods, to select the optimal forest that has a reduced number of trees compared to the original one, the performance of the subforests $h(i)$, where $i = 1, \dots, N_{RF} - 1$, is analyzed in order to identify the size that maximizes its performance. The last step, on the other hand, consists of choosing the smallest sub-forest so that the corresponding mean $\bar{h}$ is within a standard error of $\bar{h}(i_m)$, where i_m is the dimension that maximizes the average, which is equivalent to the optimal size of the sub-forest.

Wang and Zhang state that in terms of heuristics, in this study some sub-forests have come to outperform the original RF in terms of prediction accuracy, it has therefore been shown that it is possible to obtain a highly accurate forest with a small and manageable number of trees to allow an in-depth analysis of the same.

Gibbons et al. (2013) generated a very large artificial dataset using Random Forest predictions. To understand the latter, they grow a decision tree with the artificial dataset. Finally, they increase the comprehensibility of the tree obtained by carrying out a pruning phase in which an understandable human depth is considered.[4] Subsequently, Zhou, Zhou, and Hooker (2018) propose an Approximation Tree procedure, an extension of the previous method (Gibbons et al., 2013) as it allows the construction of the decision tree using hypothesis tests to understand which are the best splits, through the analysis of Gini indices on the trees of the RF. Their goal is to study the asymptotic behavior of splits and introduce a better methodology based on splits, which stabilizes the structure of the trees.

A further study that is provided by Zhou et al. (2018) is the development of a stopping rule, which indicates the depth of a tree with the use of the variability of the RF when this is used as an instructor. The question that arises is whether the RF predictions within a node are statistically different from the constant ones.[5]

In conclusion, through the use of the statistical properties of RF, stopping rules are created, which ensure that the results deriving from approximated trees reflect a coherent signal, rather than a background noise of the sample.

2.2.2. Rule extraction

Some studies aim to solve interpretative problems through the use of understandable rule-based predictors; for example, an understandable global predictor provides global explanatory power with the use of the entire set of rules or a set of reduced dimensions, which leads to every possible decision. In most studies, a rule means a decision path, from the root node to a leaf node.

Node Harvest (Meinshausen, 2010) is proposed with the aim of selecting the set of rules by assigning weights, on the basis of quadratic programming with linear inequality constraints. Further it also tries to reconcile two objectives, such as interpretability and accuracy in predictions, by combining the positive aspects of trees and tree ensembles.

The work takes place in two phases: In the first phase the shallow part of the trees is extracted, while the remaining ones are removed. Next, the shallow trees are combined so that they can work well on the training set.

Advantage of Node Harvest that the combination phase can be formulated as a convex quadratic programming, which allows to obtaining an optimal global solution in an effective way and in addition it simplifies this combination using the shallow parts of the trees. Further advantages are the management of missing values without the explicit use of imputations or surrogate split,[6] the management of mixed data and is finally not sensitive to monotonous data transformations.

Despite this, the limit of Node Harvest is that the resulting model is still an ensemble of trees and it is therefore often difficult to interpret this simplified ensemble except through the use of additional methodologies. Node Harvest treats binary response as a particular case of a regression analysis. *inTrees* (Deng, 2019), also called *Simplified tree ensemble learner (STEL)*, aims to obtain interpretable information through the extraction, measurement and processing of rules deriving from an ensemble of trees such as the RF. In addition, these rules are used to create a learner for future predictions. In particular, this method consists of a series of algorithms that allow first of all to extract the rules and classify them; then to carry out a pruning phase on each rule, that is to cut rules that produce background noise or that are irrelevant. Then a compact set of relevant and non-redundant rules is selected, frequent interactions are extracted and finally, everything is summarized in a learner that can be used to make predictions on new data. The inTrees tool can be applied to both classification and regression problems, though it is designed for classification; in regression analysis, the output is discretized and then transformed into a classification problem. Furthermore, the number of levels obtained with the discretization remains as a tuning parameter and affects the resulting rules.

STEL aims obtaining a list of rules $\mathfrak{R}$ sorted by priority, which will be applied from top to bottom to a new instance until a condition is will be satisfied. The right side of the rule then becomes the prediction for the new instance. We indicate the *Default Rule* as $\{\} = t^*$, where t^* is the class with the highest frequency in the training set for classification, while it is the mean for regression. By default, the rule assigns t^* to the result of an instance ignoring the values of predictive variables. Let S be the set of rules that includes the Default Rule and the rules extracted from RF, while let D be the set that includes all the training instances at the beginning of the construction of algorithm process. Rules having a low frequency, for example below 0.01, are removed from S to avoid overfitting problems. The algorithm has multiple iterations and, to each of them, the best rule (BR) in S, evaluated by D, is selected and added to $\mathfrak{R}$. The BR corresponds to the rule with minimum error, where the latter is defined as the ratio between the number of incorrectly classified instances and the number of instances that satisfy the condition of the rule for classification problems. For regression problems, the error is calculated using the Mean Square Error (MSE). Instances that satisfy the Best Rule are removed from D and the Default Rule is updated with the metrics related to the rules depending on the instances left. If the Default Rule is selected as the best rule in an iteration, then the algorithm returns $\mathfrak{R}$ as output.

2.2.3. Local explanation

Through a local decomposition (Haddouchi & Berrado, 2019) it is possible to check and analyze the predictions made by a Random Forest. In recent years, approaches defined as agnostics (see Guidotti et al., 2018) have been developed with respect to the Black Box model that we want to interpret. By definition, these approaches are

[4] Approximately between 6 and 11 knots in depth.

[5] It is equivalent to ask whether the further division of the node is simply modeling the background noise.

[6] If for an observation the value of the split variable is not detected, it is possible to refer to a surrogate split variable by identifying among the explanatory variables the one most linked to the split variable, chosen through a study of the relationships between couples of variables.

generalizable and return an understandable local predictor: a function is implemented that allows the interpretation of any Black Box model. Probably the first study attempting to provide an agnostic solution is the one proposed by Lou, Caruana, and Gehrke (2012): a method is implemented which, through the use of Generalized Additive Models (GAM), is able to interpret Black Box, such as tree ensembles and, therefore, the Random Forest.

Recently, Ribeiro et al. (2016) proposed the Local Interpretable Model-Agnostic Explanations (LIME) approach which does not depend on the type of Black Box to be interpreted, the particular type of local understandable predictor, and the nature of data. In this way, it can be defined as a general technique that explains the predictions of any classifier or regressor in an interpretable and faithful way.

The essential question they try to answer is, for example, how a prediction was made or which variables were key to obtaining it. Therefore, with LIME we try to understand how the model predictions vary through the perturbations of the input data,[7] modifying the values of the variables and observing the impact caused. The obtainable output is consequently a list of explanations, which reflects the contribution made by each variable to the prediction in order to provide local interpretability.

A single explanation is created by approximating the underlying model locally with an interpretable model.[8] Although LIME is based on simple and understandable ideas, it also has some limitations: only linear models are used to approximate local behavior, which is justified when looking at a very small region around the data sample; if instead, an expansion of the region is carried out, it is likely that a linear model is not powerful enough to explain the behavior of the original model. A second point could be represented by the type of perturbation that must be performed on the data: to obtain correct explanations, often, the perturbations must be specific to each individual use case, since the predefined ones are generally not sufficient. Theoretically, the perturbations should be driven by the variation observed in the data set.

In other words, LIME aims to generate dummy records by perturbing an instance, then approximating the local behavior of the original model in proximity to the perturbed instance. LIME does not take into account the distribution of the values of the variables. In fact, it approaches a model based on randomly perturbed element values that may never appear in a record in the dataset.

Unlike LIME, an interpretation method called Confident Itemsets Explanation (CIE) (Moradi & Samwald, 2020) aims to approximate the correlations between features and a result based on the real values that appeared in the data set. This method only considers real samples and predictions made by a Black Box learner. This solves the problem of Black Box approximations based on unrealistic perturbations. Like LIME, the CIE method is model-independent and can provide explanations for various Black Box models.

CIE is based on confident itemsets or sets of elements that accurately represent the local behavior of a model in different parts of the input space. This strategy addresses the difficulties of approximating more complex correlations, such as multi-class text data sets.

Compared to frequent itemsets which only consider the values of the most frequent features to extract associations between features and classes, confident itemsets precisely approximate the relationships between features and class labels. The feature space is discretized into small subspaces, so that in each subspace the confident itemsets specify the decision boundaries of a class. Confident itemsets can also reveal relationships between multiple feature values and a target class label. Compared to decision set methods that use frequent itemsets and optimize accuracy with respect to overall interpretability, the CIE method can produce concise and easily understandable explanations without diminishing descriptive accuracy.

From the experimental study conducted by Moradi and Samwald (2020), the authors concluded that: confident itemsets provide an effective way to acquire local relationships between variable values and class labels in various subspaces of a Black Box learner's decision space. Furthermore, the optimization of fidelity, interpretability, and coverage measures on extracted local relationships can lead to the production of high-quality class interpretations. The perturbation and decision set-based explanation methods impose some limitations that decrease the accuracy and descriptive interpretation, especially when producing class explanations, to this end the CIE method addresses these limitations by relying on real feature values (instead of perturbed features), using confidence measure to capture all important correlations in subspaces, providing a confidence measure that quantifies the strength of correlations, revealing ambiguity and uncertainty in the model's decision space or data, and optimizing essential topical measures, interpretability, and coverage.

Table 1
Main characteristics of the selected datasets.

Datasets	Obs.	Qual. Feat.	Quant. Feat.	0/1 response Rate	Unbalanced response
Bank Marketing	4119	11	10	1072/35	True
Churn Modeling	10 000	5	6	2301/281	True
Banknote Authentication	1372	1	4	222/184	False
Cardiovascular Disease	70 000	7	5	785/736	False
Pima Indians Diabetes	768	8	1	125/56	False
Carseat sales	400	4	7	61/34	False

3. Comparison study

In this section, a machine learning experiment is conducted among the rule extraction approaches previously described. First, we illustrate the methodological assumptions and the design criteria underlying the experiment. In particular, we explain the research strategy and the main characteristics of the real datasets used. Successively, we show the experiment results discussing the main findings by comparing the approaches considered.

3.1. Experimental design

The aim is to compare rule extraction approaches through the use of several datasets. We focus on Node Harvest (Meinshausen, 2010) and inTrees (Deng, 2019): these methodologies provide an interpretative structure that synthesizes and interprets the forest of trees produced by the Random Forest. This type of interpretation is crucial in many practical context.

To perform the comparative analysis we selected six binary classification datasets with different characteristics in term of number of observations, number of features, and different ratio of balanced/unbalanced response (Table 1). The datasets are published on the UCI Machine Learning repository.[9] For Cardiovascular Disease data, 10% of the total observations were selected for computational simplicity.

At first, a random forest is grown for each of the six datasets in order to predict the target variable. Subsequently, the set of rules is extracted for the investigation of the paths taken by each individual observation, also illustrating the most important and frequent rules of the same set. Finally, we proceed with the comparison of the various sets of rules deriving from the two methodologies investigated, specially paying attention to the validity of the experiment in terms of performance and generalization. Performance evaluation is carried

[7] For example, for continuous variables, background noise is entered in order to perturb the data.

[8] The interpretable models can be linear models with a strong regularization such as decision trees, formed on small perturbations of the original instance, which provide a good local approximation.

[9] https://archive.ics.uci.edu/ml/index.php.

out using nine metrics: Accuracy, Precision, Sensitivity, Specificity, G-Mean, F1 Score, Youden's Index, Balanced Accuracy, Kappa (Akosa, 2017; García, Mollineda, & Sánchez, 2009; Sokolova, Japkowicz, & Szpakowicz, 2006).

All considered metrics are calculated starting from a confusion matrix. It is a special kind of 2 × 2 contingency table in which each row represents the instances in a predicted class, while each column represents the instances in an actual class. Matrix cells are labeled as follow: true positive (TP), false positive (FP), true negative (TN), and false negative (FN).

Accuracy measure aims to assess the overall efficiency of a model, but requires a balanced distribution of the response variable:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Precision identifies the proportion of positive observations that are correctly classified:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

Sensitivity detects the percentage of positive observations identified correctly, while Specificity returns the proportion of negative cases that are correctly classified:

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$Specificity = \frac{TN}{TN + FP} \tag{7}$$

The Geometric Mean (G-Mean) is a metric that measures the balance between classification performances on both the majority and minority classes:

$$G - mean = \sqrt{Sensitivity * Specificity} \tag{8}$$

The F1-Score index is defined as the harmonic average between Sensitivity and Precision and measures the robustness of the model:

$$F1\ Score = \frac{2 * Sensitivity * Precision}{Sensitivity + Precision} = \frac{2TP}{2TP + FP + FN} \tag{9}$$

Youden's index evaluates whether the classifier is able to avoid classification errors and can be described as a synthetic measure of accuracy:

$$Youden's Index = Sensitivity - (1 - Specificity) \tag{10}$$

Balanced Accuracy measures the average accuracy achieved by both the minority and majority category, thus avoiding the problem related to unbalanced responses:

$$Balanced\ Accuracy = \frac{1}{2}(Sensitivity * Specificity) \tag{11}$$

Finally, Cohen's Kappa coefficient considers the precision that would only be generated by chance:

$$Kappa = \frac{Total\ Accuracy - Random\ Accuracy}{1 - Random\ Accuracy} \tag{12}$$

The two Rule extraction approaches are compared with each other and with the Random Forest reference standard; in particular, considerable attention is paid to the presence of balanced and unbalanced data. The accuracy metrics have been used to measure how results of an interpretative approach are similar to the performance of a classical Random Forest applied on the same data. Closer results mean a better ability to reconstruct the original "black-box" supervised model with an interpretable structure.

The machine learning experiment was conducted in the R environment.

Model validation has been performed using the holdout technique that provides an impartial estimate of the generalization performance. The original dataset is randomly divided in two parts following Guyon (1997) which suggests to use about 70% for the training set and the rest for the test set.

Table 2
Random forest confusion matrices.

(a) Bank marketing			(b) Churn modeling			(c) Banknote authentication		
Predicted	Actual		Predicted	Actual		Predicted	Actual	
	0	1		0	1		0	1
0	1072	98	0	2301	333	0	222	2
1	31	35	1	85	281	1	4	184
(d) Cardiovascular disease			(e) Pima indians diabetes			(f) Carseat sales		
Predicted	Actual		Predicted	Actual		Predicted	Actual	
	0	1		0	1		0	1
0	785	319	0	125	22	0	61	20
1	260	736	1	27	56	1	5	34

Table 3
Node harvest confusion matrices.

(a) Bank marketing			(b) Churn modeling			(c) Banknote Authentication		
Predicted	Actual		Predicted	Actual		Predicted	Actual	
	0	1		0	1		0	1
0	1095	116	0	2328	392	0	207	13
1	8	17	1	58	222	1	19	173
(d) Cardiovascular disease			(e) Pima indians diabetes			(f) Carseat sales		
Predicted	Actual		Predicted	Actual		Predicted	Actual	
	0	1		0	1		0	1
0	812	367	0	136	35	0	59	31
1	226	695	1	16	43	1	7	23

3.2. Analysis

In this section, we show results regarding the prediction performance of classical Random Forest, Node Harvest, and inTrees. In particular, for each approach, we report confusion matrices and performance measures. Random Forest is used as benchmark of the two rule extraction methods.

- *Random Forest* (benchmark model): the model was created using the randomForest package (Liaw et al., 2002). Hyperparameters are set by default. Results are reported in Table 2.

- *Node Harvest*: the estimator is obtained with the package nodeHarvest (Meinshausen, 2010), setting as hyperparameter the number of nodes in the initial set equal to 500 and the minimum number of samples in each node equal to 10. It is observed that increasing the number of initial nodes, performance and computation time increase.

- *inTrees*: it was implemented with the homonymous package in R (Deng, 2019). The hyperparameters coincide with those chosen for the randomForest algorithm. The produced object is transformed into a list of trees from which the rules are extracted and then processed.

Analyzing confusion matrices, inTrees and nodeHarvest show a quite similar accuracy. However, inTrees appears to be slightly better probably because it works on the same set of trees that is produced by randomForest. Node-Harvest, on the contrary, builds up a new forest, different from the original one, to extract partition rules (see Table 3 and Table 4).

The results are shown in Table 5, where we report all performance metrics calculated on each dataset. The best value is marked in bold, and for sake of clarity, the performance of randomForest is reported,

Table 4
InTrees confusion matrices.

(a) Bank marketing			(b) Churn modeling			(c) Banknote Authentication		
Predicted	Actual		Predicted	Actual		Predicted	Actual	
	0	1		0	1		0	1
0	1096	119	0	2306	368	0	221	8
1	7	14	1	80	246	1	5	178
(d) Cardiovascular disease			(e) Pima indians diabetes			(f) Carseat sales		
Predicted	Actual		Predicted	Actual		Predicted	Actual	
	0	1		0	1		0	1
0	694	253	0	128	24	0	58	19
1	344	809	1	24	54	1	8	35

but not highlighted. For the Bank Marketing and Churn Modeling datasets, characterized by unbalanced responses, accuracy is not a relevant metric and then it values are not discussed (Branco, Torgo, & Ribeiro, 2016).

For the unbalanced data, there is no prevalence between the two methods analyzed as they both appear to work similarly (Table 5a and b). On the contrary, considering the analysis of balanced data, the inTrees method achieves better performance than nodeHarvest in 3 out of 4 datasets, Banknote Authentication, Pima Indians Diabetes and Carseat sales respectively. NodeHarvest wins only for Cardiovasular Datasets but not in all metrics. In fact inTrees shows again best Precision and Specificity values. The sample size seems not playing a significant role in identifying the best method. inTrees works better both for small and bigger samples.

In conclusion, inTrees appears to be the best rule extraction algorithm to interpret Random Forest structures and, more in general, any other algorithm implementing a forest of trees.

The inTrees method provides an excellent choice for deriving simple, rule-based learners from complex ensembles of trees. The predictive results obtained are just as good even though they probably have a greater variance than the randomForest benchmark prediction method. Therefore, it achieves the set goal by offering an intuitive way to understand a complex Black Box model such as the Random Forest, in order to provide a means that is able to help, for example, in the detection of financial fraud or disease.

4. Conclusion

In addition to having high predictive performance, the Random Forest has the advantage of being an intuitive algorithm as regards its construction and, therefore, accessible for use by even less experienced users. Therefore, by using the right approach it is also possible to gain internal interpretation, in order to have a powerful predictive tool that is accurate and interpretable.

Here, the Rule Extraction approach is considered as the key to an effective natural understanding and able to provide an intuitive interpretation of the model produced by the RF. Through the use of this approach, a set of representative rules is obtained, which allow the discovery of the structure that synthesizes and interprets the entire RF model, while offering significant information such as the relationships and importance of the variables. This type of information is fundamental in the practical contexts: deriving representative rules can be useful to experts who can analyze each possible scenario in more detail and, therefore, create treatment options suitable for each of these scenarios.

For this reason, the implementation of the Machine Learning experiment aims to highlight the two main approaches proposed in the

Table 5
Performance metrics. An higher value means a better performance. The best value is marked in bold. *For unbalanced datasets, Accuracy has not been considered for comparison.

(a) Bank marketing				(b) Churn modeling			
	randomForest	nodeHarvest	inTrees		randomForest	nodeHarvest	inTrees
Accuracy*	0.8956	0.8997	0.8981	Accuracy*	0.8607	0.8500	0.8507
Balanced Accuracy	0.6175	**0.5603**	0.5495	Balanced Accuracy	0.7110	0.6686	**0.6836**
Kappa	0.3019	**0.1875**	0.1571	Kappa	0.4965	0.4226	**0.4446**
Specificity	0.2632	**0.1278**	0.1053	Specificity	0.4577	0.3616	**0.4007**
Sensitivity	0.9719	0.9927	**0.9937**	Sensitivity	0.9644	**0.9757**	0.9665
Precision	0.9162	**0.9042**	0.9021	Precision	0.8736	0.8559	**0.8624**
G-mean	0.5057	**0.3562**	0.3234	G-mean	0.6643	0.5939	**0.6223**
F1	0.9432	**0.9464**	0.9456	F1	0.9167	**0.9119**	0.9115
Youden's index	0.2350	**0.1206**	0.0989	Youden's index	0.4220	0.3373	**0.3671**
(c) Banknote Authentication				(d) Cardiovascular disease			
	randomForest	nodeHarvest	inTrees		randomForest	nodeHarvest	inTrees
Accuracy	0.9854	0.9223	**0.9684**	Accuracy	0.7243	**0.7176**	0.7157
Balanced Accuracy	0.9858	0.9230	**0.9674**	Balanced Accuracy	0.7244	**0.7183**	0.7152
Kappa	0.9706	0.8436	**0.9362**	Kappa	0.4487	**0.4360**	0.4308
Specificity	0.9892	0.9301	**0.9570**	Specificity	0.6976	0.6544	**0.7618**
Sensitivity	0.9823	0.9159	**0.9779**	Sensitivity	0.7512	**0.7823**	0.6686
Precision	0.9911	0.9409	**0.9651**	Precision	0.7111	0.6887	**0.7328**
G-mean	0.9858	0.9230	**0.9674**	G-mean	0.7239	**0.7155**	0.7137
F1	0.9867	0.9283	**0.9714**	F1	0.7306	**0.7325**	0.6992
Youden's index	0.9715	0.8460	**0.9349**	Youden's index	0.4488	**0.4367**	0.4304
(e) Pima indians diabetes				(f) Carseat sales			
	randomForest	nodeHarvest	inTrees		randomForest	nodeHarvest	inTrees
Accuracy	0.7870	0.7783	**0.7913**	Accuracy	0.7917	0.6833	**0.7750**
Balanced Accuracy	0.7702	0.7230	**0.7672**	Balanced Accuracy	0.7769	0.6599	**0.7635**
Kappa	0.5320	0.4741	**0.5344**	Kappa	0.5682	0.3333	**0.5369**
Specificity	0.7179	0.5513	**0.6923**	Specificity	0.6296	0.4259	**0.6481**
Sensitivity	0.8224	**0.8947**	0.8421	Sensitivity	0.9242	**0.8939**	0.8788
Precision	0.8503	0.7953	**0.8421**	Precision	0.7531	0.6556	**0.7532**
G-mean	0.7684	0.7023	**0.7635**	G-mean	0.7628	0.6171	**0.7547**
F1	0.8361	**0.8421**	**0.8421**	F1	0.8299	0.7564	**0.8112**
Youden's index	0.5403	0.4460	**0.5344**	Youden's index	0.5539	0.3199	**0.5269**

literature and identify which of the two is more performing in terms of predictive performance. The findings are in favor of the inTrees approach when datasets are characterized by balanced responses. Table 5 shows how inTrees outperforms nodeHarvest in 3 out of 4 datasets considering all metrics. On the contrary, when datasets have unbalanced responses, the experiment does not provide a leading approach, probably also because we analyzed only two datasets. This is a limitation of this study.

However, the current study is a first step in comparing these methodologies. Future researches should focus on an extensive approach considering more deeply the analysis of data characterized by unbalanced response and the relevance of the extracted rules in real world applications, especially in health and economics

CRediT authorship contribution statement

Massimo Aria: Conceptualization, Methodology, Formal analysis, Investigation, validation, Writing - review & editing. **Corrado Cuccurullo:** Conceptualization, Writing - review & editing. **Agostino Gnasso:** Conceptualization, Methodology, Formal analysis, Investigation, Supervision, Writing - original draft.

Acknowledgment

This work has been partially supported by the Italian Research Programme “V:ALERE 2019”. Project “Leading change in the Italian public Academic Medical Centers”, Grant Agreement Number 403.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Akosa, J. (2017). Predictive accuracy: A misleading performance measure for highly imbalanced data. In *Proceedings of the SAS Global Forum* (pp. 2–5).

Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys*, *49*(2), 1–50.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Belmont, CA: Wadsworth. *International Group*, *432*, 151–166.

Chipman, H., George, E., & McCulloh, R. (1998). Making sense of a forest of trees. *Computing Science and Statistics*, 84–92.

Deng, H. (2019). Interpreting tree ensembles with intrees. *International Journal of Data Science and Analytics*, *7*(4), 277–287.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. ArXiv Preprint arXiv:1702.08608.

Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 0210–0215). IEEE.

Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, *63*(1), 68–77.

Ehrlinger, J. (2016). Ggrandomforests: random forests for regression. ArXiv Preprint arXiv:1501.07196.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 1189–1232.

García, V., Mollineda, R. A., & Sánchez, J. S. (2009). Index of balanced accuracy: A performance measure for skewed class distributions. In *Iberian Conference on Pattern Recognition and Image Analysis* (pp. 441–448). Springer.

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236.

Gibbons, R. D., Hooker, G., Finkelman, M. D., Weiss, D. J., Pilkonis, P. A., Frank, E., et al. (2013). The CAD-MDD: a computerized adaptive diagnostic screening tool for depression. *The Journal of Clinical Psychiatry*, *74*(7), 669.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1–42.

Guyon, I. (1997). A scaling law for the validation-set training-set size ratio. *AT&T Bell Laboratories*, *1*(11).

Haddouchi, M., & Berrado, A. (2019). A survey of methods and tools used for interpreting random forest. In *2019 1st International Conference on Smart Systems and Data Science (ICSSD)* (pp. 1–6). IEEE.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R News*, *2*(3), 18–22.

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, *16*(3), 31–57.

Lou, Y., Caruana, R., & Gehrke, J. (2012). Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 150–158).

Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. In *Advances in Neural Information Processing Systems* (pp. 431–439).

Meinshausen, N. (2010). Node harvest. *The Annals of Applied Statistics*, 2049–2072.

Mitchell, T. (1997). *Machine Learning*. McGraw Hill.

Moradi, M., & Samwald, M. (2020). Post-hoc explanation of black-box classifiers using confident itemsets. ArXiv Preprint arXiv:2005.01992.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In *Australasian Joint Conference on Artificial Intelligence* (pp. 1015–1021). Springer.

Tan, H. F., Hooker, G., & Wells, M. T. (2016). Tree space prototypes: Another look at making tree ensembles interpretable. ArXiv Preprint arXiv:1611.07115.

Wang, M., & Zhang, H. (2009). Search for the smallest random forest. *Statistics and its Interface*, *2*(3), 381–388.

Zhao, X., Wu, Y., Lee, D. L., & Cui, W. (2018). Iforest: Interpreting random forests via visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 407–416.

Zhou, Y., Zhou, Z., & Hooker, G. (2018). Approximation trees: Statistical stability in model distillation. ArXiv Preprint arXiv:1808.07573.