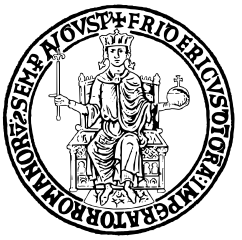
UNIVERSITY OF NAPLES FEDERICO II



DOCTORAL THESIS IN ECONOMICS

E2Tree:
A methodological proposal to explain tree-based ensemble decision rules

---

Candidate
Dr. Agostino Gnasso

Supervisor
Prof. Massimo Aria

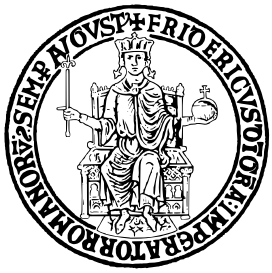
Ph.D. Director
Prof. Marco Pagnozzi

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy in Economics - XXXVII Cycle
at the Naples School of Economics
Department of Economics and Statistics

The important thing is not to stop questioning.

Curiosity has its own reason for existing.

(Albert Einstein, Life magazine, May 2, 1955)



Candidate's declaration

I hereby declare that this thesis submitted to obtain the academic degree of Philosophiæ Doctor (Ph.D.) in Economics - Department of Economics and Statistics, University of Naples Federico II - is my own unaided work, that I have not used other than the sources indicated, and that all direct and indirect sources are acknowledged as references.

Parts of this dissertation have been published in international journals and/or conference articles (see list of the author's publications at the end of the thesis).

Naples, December 2, 2025

_____

Agostino Gnasso

Abstract

The aim of this work is to provide a methodological contribution on ensemble methods.
Ensemble learning methods constitute a cornerstone of modern supervised learning, providing state-of-the-art predictive performance in both classification and regression tasks. Among them, Random Forests have emerged as one of the most effective algorithms, combining robustness, flexibility, and high accuracy. Yet, the very mechanism that grants Random Forests its predictive power (aggregating the output of numerous decision trees) renders the model largely opaque. This opacity hampers interpretability and explainability, turning Random Forest into a *black box* model and limiting its applicability in domains where transparency is essential, such as finance, medicine, and the social sciences. This dissertation presents a novel methodology, *Explainable Ensemble Trees (E2Tree)*, which is a significant advancement for the field of explainable artificial intelligence (XAI). E2Tree merges the intuitive structure of decision trees with the predictive strength of Random Forests. By exploiting a dissimilarity matrix derived from co-occurrence patterns of observations within terminal nodes, the proposed methodology reconstructs a unified tree-like structure that effectively summarizes the ensemble process. In doing so, E2Tree preserves the predictive accuracy of the Random Forest model while providing a transparent, explainable, interpretable, and graphical representation of the relationships between predictors and outcomes. The dissertation advances this methodology through five main directions. First, it introduces the fundamental concepts necessary for comprehending the methodological and theoretical framework underlying the proposal. Second, a critical evaluation of existing interpretative strategies for Random Forests is presented, with the shortcomings of these strategies in capturing both local and global model behaviour being identified. Third, it formalizes the E2Tree frame-

work in classification settings. Fourth, it demonstrates the explanatory capacity of the method through a benchmark application in the classification domain. Finally, the dissertation extends the framework to regression tasks by introducing novel node-level measures of goodness of fit and validating the approach through appropriate statistical tests, including the Mantel test. By offering a unified and explainable framework for ensemble tree models, E2Tree has the potential to bridge the long-standing gap between predictive performance and explainability. Its intuitive structure can enhance decision-making in domains where accountability and transparency are critical, foster greater trust in machine learning applications, and serve as a versatile tool for researchers to understand complex interactions within data-driven analyses.

Keywords: Explainable Artificial Intelligence; Explainability; Interpretability; Random Forest; Ensemble Methods; Explainable Ensemble Trees

Contents

Introduction

Machine Learning (ML) has become a cornerstone of modern data analysis, providing powerful tools to uncover patterns and generate accurate predictions across a wide range of fields, including finance and healthcare. Among supervised learning algorithms, ensemble methods – especially Random Forests (RF) – have been shown to be highly effective due to their robustness, versatility and outstanding predictive accuracy. [1]. By aggregating multiple decision trees, RF reduces variance and enhances stability, thereby outperforming single-tree models such as CART [2]. Despite these advantages, the widespread adoption of RF has raised concerns regarding its opacity: the internal logic of ensemble models is often inaccessible to users, relegating them to the status of "black boxes." This lack of transparency undermines trust in predictions, complicates their validation, and hinders their integration into high-stakes decision-making contexts.

The debate around interpretability and explainability in ML has grown substantially over the past decade. While interpretability is generally refers to the ability to intuitively link inputs to outputs, explainability focuses on uncovering and reconstructing the internal mechanisms that guide the model's decisions [3, 4]. Both dimensions are crucial in domains where the costs of errors are high and where accountability is required, such as medicine, finance, and public policy. Several post-hoc methods have been developed to address these needs, including Partial Dependence Plots (PDP) [5], Local Interpretable Model-Agnostic Explanations (LIME) [6], and SHapley Additive exPlanations (SHAP) [7]. Despite the extensive implementation of these methodologies, they are characterised by significant limitations. It has been demonstrated that PDPs have the capacity to

obscure variable interactions. Furthermore, LIME, despite their intuitive nature, frequently produce unstable results. Finally, although SHAP are theoretically robust, they are computationally demanding and depend on assumptions of feature independence. Collectively, these approaches offer fragmented insights, either concentrating on local explanations of individual predictions or generating global attributions that fail to capture the comprehensive decision-making architecture of an ensemble model.

The trade-off between predictive accuracy and transparency represents one of the most pressing challenges in the field of Explainable Artificial Intelligence (XAI). RF is an example of this issue. They are consistently ranked amongst the most accurate methods for problems in both low and high dimensions [8], but their complexity resists conventional interpretability. Existing strategies for RF explanation, including variable importance measures, rule extraction techniques, and surrogate models, provide partial solutions but often compromising accuracy, stability, or interpretability [9].

To address this, E2Tree has been introduced as a novel methodological framework to balance accuracy with interpretability [10]. E2Tree builds on the insight that the relational structure embedded within a RF can be reconstructed in the form of a single tree-like model. The E2Tree methodology employs a dissimilarity matrix derived from the co-occurrence patterns of observations in RF terminal nodes, thereby producing a global representation that maintains the ensemble's predictive fidelity while explicitly revealing hierarchical interactions among predictors. This method combines the clarity of decision trees with the robustness of Random Forests, providing both graphical visualizations and formal "if–then" decision rules.

E2Tree has been further developed to include regression tasks, creating a coherent explanatory framework for continuous outcomes. [11]. The methodology has been successfully applied in policy-relevant fields, such as public health, where the opacity of black-box models can otherwise preclude actionable insights. For instance, in a study investigating depression risk factors, E2Tree facilitated the identification of significant predictor interactions, including the complex relationships between chronic illness, family structure, and psychological symptoms, thereby supporting evidence-based healthcare recommendations [12]. However, the E2Tree framework is not devoid of limitations. Constructing the dissimilarity matrix is computationally intensive, raising concerns of scalability for very

large datasets. Additionally, its dependence on RF as the base learner limits its applicability in contexts where alternative ensemble methods, such as Gradient Boosting Machines or XGBoost, are more appropriate. In cases where predictor interactions are weak or data are sparse, the added complexity of E2Tree may provide only marginal interpretative benefits over to simpler models. These challenges highlight the need for further methodological refinement, including the development of approximation strategies to reduce computational burden, validation metrics to assess explanatory fidelity, and extensions to other ensemble-tree algorithms.

This dissertation contributes to the evolving discourse on XAI by positioning E2Tree as a methodology that has the potential to offer comprehensive explanations for ML models. The study undertakes a critical examination of the theoretical foundations of ensemble methods and their challenges to interpretability. It also evaluates E2Tree's methodological innovations in comparison to existing explanation techniques and explores its practical applications in socially and scientifically significant domains. The objective of this work is twofold: firstly, to highlight the strengths and limitations of E2Tree; and secondly, to contribute to the broader goal of reconciling predictive accuracy with transparency in ML.

This work consists of four papers written during my Ph.D. period. The thesis is structured into five chapters.

Chapter 1 provides the theoretical background on ensemble learning methods, with a particular focus on RF. It discusses their predictive advantages, the trade-off between accuracy and interpretability, and sets the stage for the need of more transparent methodologies.

Chapter 2 reviews the main interpretative approaches for RF. It surveys the existing literature, classifies the proposals according to established taxonomies, and critically discusses their strengths and limitations. This review sets the stage for developing a novel methodology that bridges the gap between predictive accuracy and interpretability.

Chapter 3 introduces the E2Tree methodology. The chapter presents the theoretical framework, the construction of the dissimilarity matrix, the algorithmic steps, and the stopping criteria that define the proposed approach.

Chapter 4 applies the E2Tree methodology to a case study of public health relevance: predicting depression in the Italian population using the

European Health Interview Survey (EHIS). The analysis highlights how E2Tree complements the predictive power of RF with a transparent and interpretable representation of the decision process.

Chapter 5 extends the E2Tree framework to regression problems, introducing specific adaptations to handle continuous response variables, including new definition of local goodness-of-fit and additional stopping rules. This extension broadens the scope of the methodology and demonstrates its versatility.

Finally, the conclusion summarizes the main contributions of the thesis, reflects on its limitations, and outlines directions for future research.

Chapter 1

Preliminary tools and concepts

1.1 Supervised machine learning methods

Supervised machine learning constitutes the backbone of modern predictive analytics, providing algorithms capable of learning from labeled data to infer outcomes for unseen observations. Within this paradigm, tree-based approaches occupy a prominent role, owing to their intuitive structure and adaptability. Decision Trees offer an interpretable baseline but often suffer from instability and limited predictive accuracy. To address these limitations, Ensemble Methods combine multiple learners to enhance robustness and generalization. Bagging, for instance, reduces variance by aggregating predictions across bootstrap samples, while Boosting sequentially re-weights observations to iteratively correct errors. Building on these principles, RF introduce randomness in variable selection at each split, producing decorrelated trees and yielding highly accurate models with wide applicability across disciplines. In the following sections, these methods will be examined in greater detail, providing the background necessary for the work carried out in this dissertation.

1.1.1 Tree-based Methods

Tree-based methods represent a family of non-parametric supervised learning techniques that recursively partition the predictor space into a set of disjoint regions and fit a simple model, typically a constant, within each region [13]. Their conceptual simplicity and intuitive graphical representation make them particularly appealing, both for classification and regression tasks. The general idea is to split the dataset according to cut-off values of the features, thereby creating subsets that are progressively more homogeneous with respect to the outcome variable [14]. The recursive partitioning results in a hierarchical structure of nodes: *internal nodes* (or split nodes) represent decision rules, while *terminal nodes* (or leaves) contain the final predictions. For regression problems, the prediction at each leaf corresponds to the average outcome of the training observations in that node, whereas in classification tasks the prediction is the most frequent class. Formally, the predictive function of a tree with M terminal nodes can be written as:

$$\hat{f}(x) = \sum_{m=1}^{M} c_m \, I\{x \in R_m\},$$

where R_m denotes the region of the feature space corresponding to the m-th terminal node, c_m is the predicted value associated with that region, and $I\{\cdot\}$ is the indicator function [15]. A crucial step in tree construction is the choice of the splitting rule. In the CART (Classification and Regression Trees) algorithm, introduced by Breiman [13], the split is chosen by evaluating all possible cutoffs across all features and selecting the one that maximizes the homogeneity of the resulting subsets. For regression tasks, this is achieved by minimizing the variance of the response within each node, while for classification the most common criterion is the Gini index, which measures node impurity. This recursive process continues until a stopping criterion is met, such as a minimum number of observations in a node or a maximum tree depth. The interpretability of decision trees arises from their hierarchical structure: starting at the root, an observation follows a unique path defined by a sequence of splitting rules until it reaches a leaf. Each path can be expressed as a conjunction of conditions of the type *"if feature x_j is greater/less than threshold c, AND ..."*, which

directly maps to a predicted outcome in the corresponding leaf. This property makes trees highly transparent compared to most other ML models. Figure 1.1 provides a graphical example of the decision tree representation, which allows a straightforward visualization of the recursive partitioning process.
An additional interpretative tool is the assessment of feature importance [15]. The relevance of a variable can be quantified by aggregating its contribution to impurity reduction across all nodes where it is used for splitting. Impurity is typically measured through the variance of the response (for regression) or the Gini index (for classification). After normalization, these contributions can be interpreted as relative shares of the overall explanatory power of the model. Beyond global importance, decision trees also allow decomposition of individual predictions. Starting from the root, each split contributes positively or negatively relative to the average prediction at the root node. The prediction for an observation x can be represented as:

$$\hat{f}(x) = \bar{y} + \sum_{d=1}^{D} \text{split.contrib}(d, x) = \bar{y} + \sum_{j=1}^{p} \text{feat.contrib}(j, x),$$

where $\bar{y}$ is the mean of the training outcome, and the contributions are aggregated either by split or by feature. This decomposition enables a feature-level explanation of how each variable influenced a specific decision.

Decision Rules. Within the family of tree-based methods, decision rules can be regarded as the elementary building blocks that summarize the splitting logic of a tree [16]. Each rule takes the form of an IF–THEN statement, where the antecedent specifies conditions on the predictor variables and the consequent provides the corresponding prediction. Their semantic structure closely resembles natural reasoning, which makes them particularly suitable for interpretation and communication [13]. For example, in a credit risk setting a rule might be: *IF income < 20,00 € AND credit history = negative THEN loan status = high risk.* The usefulness of a rule can be quantified through two complementary indicators [15]. The *support* (or coverage) measures the proportion of instances for which the antecedent holds. Suppose a rule states: *IF firm size > 250 employees*

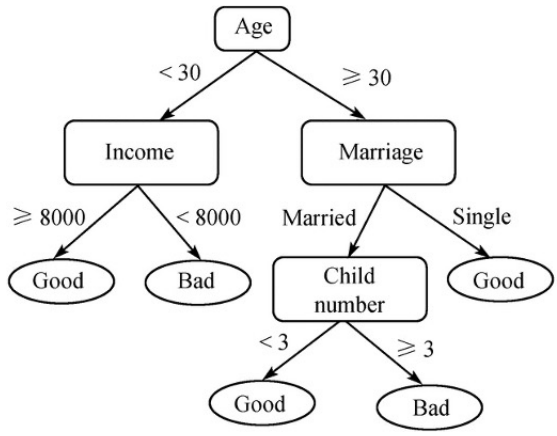


Figure 1.1. Example of a decision tree representation.

AND sector = manufacturing THEN probability of exporting = high. If 150 out of 1,000 firms satisfy these conditions, the support of the rule is:

$$\text{Support} = \frac{150}{1000} = 0.15 \ (15\%).$$

The *accuracy* (or confidence) indicates the proportion of correct predictions among the cases covered by the rule. If 120 of those 150 firms are indeed exporters, the accuracy is:

$$\text{Accuracy} = \frac{120}{150} = 0.80 \ (80\%).$$

Typically, rules with many conditions achieve higher accuracy but lower support, as they apply to fewer observations. When multiple rules are extracted from a tree or from an ensemble of trees, two organizational strategies are commonly adopted. A *decision list* orders the rules sequentially: the first rule that matches an instance provides the prediction [14]. For instance, in consumer choice analysis, the first rule might be: *IF income > 60,000 € THEN probability of purchasing luxury goods = high*; if this does not apply, the algorithm proceeds to the next rule, such as *IF age < 25 AND education = university THEN probability = medium.* Alternatively, a *decision set* treats rules as unordered, allowing multiple rules to

apply simultaneously and resolving conflicts through a voting mechanism, possibly weighted by rule quality. To guarantee completeness, models often include a *default rule*, which assigns a prediction when no other rule applies [16]. This default is usually chosen as the most frequent outcome in the dataset; for example, in a credit scoring model, if no specific condition matches, the prediction might default to *loan = approved*. In this way, decision rules provide an intuitive lens through which to understand the logic of tree-based methods. They make explicit the conditions driving predictions, thereby enhancing interpretability and offering practical insights in applied contexts such as economics and finance, where transparency and accountability are crucial.

Strengths and limitations of decision trees. Decision trees naturally capture non-linearities and interactions between predictors, often producing rules that are easier to interpret than coefficients in a multidimensional regression [17]. Their graphical representation is straightforward, providing human-friendly explanations of both global decision boundaries and local predictions. Furthermore, tree paths allow contrastive reasoning: for any given observation, one can ask "what if" questions by changing the branch taken at a split, thus obtaining counterfactual predictions. When the tree is shallow, explanations are also highly selective, as only a few features are required to justify a prediction [13]. Despite their interpretability, decision trees have several well-known weaknesses. First, they approximate linear relationships poorly, since such patterns are represented through multiple stepwise splits, leading to discontinuous prediction functions [18]. As a result, small changes in the input may cause disproportionate shifts in the output. Second, trees are highly unstable: minor perturbations in the training data may lead to different splits at the root, and consequently to entirely different structures [13]. Finally, interpretability declines rapidly as depth increases: while a tree of depth one has only two leaves, depth three already produces up to eight terminal nodes, and in general the number of leaves grows exponentially as 2^{depth} [15]. Deep trees may thus become overly complex and less transparent, undermining the main advantage of this method.
While decision trees remain among the most interpretable and conceptually straightforward learning algorithms, their instability and limited

predictive accuracy restrict their standalone use in complex tasks. These limitations have driven the development of ensemble methodologies (most notably Bagging, Boosting, and Random Forests) which exploit collections of trees to enhance stability and performance.

1.1.2 Ensemble Methods

The motivation for adopting ensemble methods in supervised learning arises from the observation that different predictive models may yield different outputs for the same input. *Ensemble learning* refers to the class of approaches that improve predictive performance by combining the outputs of multiple base learners, denoted as $\{h_1, \ldots, h_K\} \in H$, into a single predictive model [19]. The overarching goal is to reduce variance, control bias, and enhance generalization. A base learner is any supervised algorithm capable of mapping labeled examples into a predictive model. Depending on the context, this learner can be a decision tree, a neural network, a linear regression model, or another supervised learning algorithm. Ensemble methods can be broadly categorized into two groups [9]. *Sequential ensembles* (e.g., Boosting) construct learners in sequence, exploiting dependencies between them by assigning higher weights to previously misclassified observations. *Parallel ensembles* (e.g., Bagging and Random Forests) train learners independently on perturbed versions of the data and aggregate their predictions, relying on diversity among classifiers to reduce error through averaging. The intuition underlying ensemble learning is straightforward [20]: by aggregating multiple models, errors of individual classifiers are likely to be offset by others, leading to improved predictive accuracy. A necessary condition, however, is that the base learners perform better than random guessing and are sufficiently diverse. This principle echoes the *Condorcet Jury Theorem* [21], which states that if independent voters are more likely to be correct than incorrect, then majority voting increases the probability of reaching the correct decision as the number of voters grows. In ML, this translates into the bias–variance trade-off [22]: predictive error can be decomposed into systematic error (bias) and sensitivity to data fluctuations (variance). Ensembles achieve better performance by reducing one or both of these components [18]. Formally, in regression with squared error loss, the expected prediction error

can be decomposed as:

$$\mathbb{E}_{\mathscr{D},\varepsilon}\Big[(Y-\hat{f}(X))^2\Big] = \underbrace{\sigma^2}_{\text{irreducible error}} + \underbrace{\big(\mathbb{E}_{\mathscr{D}}[\hat{f}(X)] - f^*(X)\big)^2}_{\text{bias}^2} + \underbrace{\mathbb{V}_{\mathscr{D}}[\hat{f}(X)]}_{\text{variance}},$$

where σ^2 is the irreducible error, f^* is the true underlying function, and $\hat{f}$ is the predictor learned from training data $\mathscr{D}$. Averaging over multiple, decorrelated base learners reduces the variance term, while sequential methods like Boosting can also reduce bias [23]. A complementary intuition comes from majority voting: if base classifiers are independent and each has accuracy $p > 1/2$, the ensemble's accuracy is

$$\Pr(\text{correct}) = \sum_{j=\lceil K/2 \rceil}^{K} \binom{K}{j} p^j (1-p)^{K-j},$$

which increases with K, provided that diversity among classifiers is maintained. Dietterich [19] further formalized three reasons why ensembles tend to generalize better than individual classifiers. The first is *statistical*: a single learner trained on a dataset may be suboptimal, while averaging across several learners reduces the risk of selecting a poorly performing hypothesis. The second is *computational*: learning algorithms may get trapped in local minima; ensembles mitigate this by combining multiple local solutions. The third is *representational*: the true target function may lie outside the hypothesis space of any individual learner, but a collection of learners can approximate it more closely. Decision trees provide a clear example: their decision boundaries are axis-aligned and piecewise linear. When combined in an ensemble, such as a RF, the resulting model can approximate more complex, diagonal, or nonlinear boundaries. From a structural perspective, an ensemble system requires four components [24]: (i) a training set; (ii) a base learning algorithm (e.g., a decision tree inducer); (iii) a mechanism for diversifying the learners (e.g., bootstrap resampling, feature subsampling, or reweighting of hard cases); and (iv) a combiner function to aggregate predictions. The aggregation may take the form of *weighted voting*, where each classifier contributes to the final decision according to its weight, or *meta-learning*, where a secondary learner integrates the outputs of the base models (as in Stacking) [19]. For tree en-

sembles (often called *Additive Tree Models, ATM* [25]), the final predictor can be expressed as:

$$F(x) = \sum_{t=1}^{T} w_t \, f_t(x),$$

where f_t is the prediction of the t-th tree and w_t its weight (uniform in Bagging/Random Forests; learned in Boosting). Tree ensembles routinely yield state-of-the-art accuracy in applications such as credit risk assessment and clinical prediction. However, the aggregation of hundreds of trees produces thousands of decision rules, which makes direct inspection infeasible [9]. This accuracy–interpretability tension motivates the explainability techniques and surrogate structures discussed in the following sections.

Bagging

Bootstrap Aggregating, commonly known as *Bagging*, is an ensemble method introduced to reduce the variance of statistical learning models [26]. The central idea is to simulate data variability by repeatedly drawing B bootstrap samples (with replacement) from the original training set and training a base learner on each resampled dataset. The predictions of the individual learners are then aggregated to obtain a final, more stable predictor. Formally, for a new observation x_i, the Bagging estimator is defined as:

$$\hat{M}_{\text{bag}}(x_i) = \frac{1}{B} \sum_{b=1}^{B} \hat{M}_b^*(x_i), \tag{1.1}$$

where $\hat{M}_b^*$ denotes the model trained on the b-th bootstrap sample. Bagging is particularly effective when applied to unstable learning algorithms, such as decision trees, whose predictions can vary substantially in response to small perturbations of the training data [27]. In regression tasks, Bagging averages the outputs of multiple regression trees, thus smoothing fluctuations caused by the instability of individual learners. In classification tasks, each tree casts a "vote" for a class label, and the final decision is obtained through a *majority voting* scheme. A distinctive feature of Bagging applied to decision trees is that the trees are usually grown to full depth without pruning [26]. This construction yields base

learners with low bias but high variance. Averaging across many such learners substantially reduces variance, leading to more accurate and robust predictions. Figure 1.2 illustrates the Bagging procedure, showing how bootstrap samples are used to generate diverse classifiers combined into an ensemble prediction. Intuitively, although a single deep tree tends to overfit the training data, an ensemble of many overfitted trees can generalize well once their outputs are aggregated. The bootstrap resampling mechanism is the cornerstone of Bagging. Since each resample contains roughly two-thirds of the unique instances from the original training set, each tree is trained on a slightly different subset of data. This randomness enhances the diversity among base learners, which is crucial for variance reduction: the greater the independence of individual predictors, the stronger the variance reduction effect. As the number of bootstrap replications B increases, the Bagging predictor converges towards a stable solution with low variability. In summary, Bagging is especially suitable for combining high-variance, low-bias classifiers such as decision trees [18]. By exploiting bootstrap resampling and aggregation, it transforms unstable predictors into a stable ensemble with improved generalization performance.

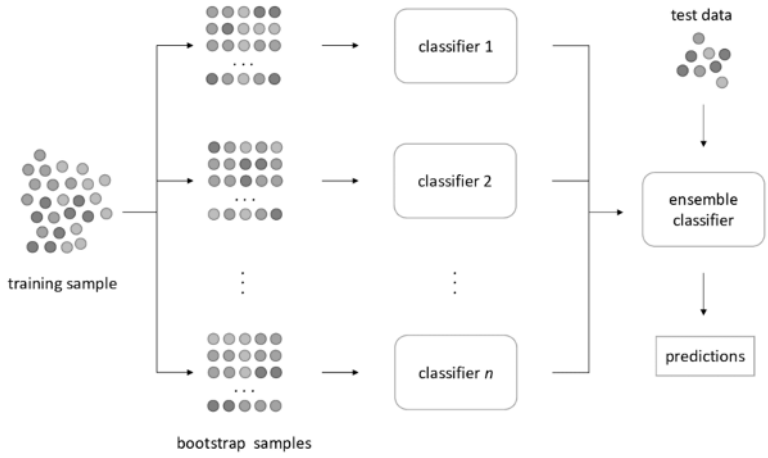


Figure 1.2. Schematic representation of the Bagging procedure, where multiple bootstrap samples are used to train base classifiers and their predictions are aggregated into an ensemble model.

Boosting

Boosting is a family of ensemble methods designed to convert a collection of weak learners into a strong predictor. Unlike Bagging, which builds learners in parallel and aggregates their predictions, Boosting constructs base learners sequentially, with each learner attempting to correct the errors of its predecessors [28]. This adaptive reweighting of training instances allows Boosting to reduce both bias and variance, often achieving state-of-the-art predictive accuracy across a wide variety of applications. Formally, suppose we have a set of n labeled training examples $(x_1, y_1), \ldots, (x_n, y_n)$, and a base learner $\hat{h}(x)$ that performs slightly better than random guessing [23]. Boosting maintains a distribution of weights $\{w_1, \ldots, w_n\}$ over the training instances. Initially, all weights are uniform. At each iteration $m = 1, \ldots, M$, a new base learner $h_m(x)$ is trained on the weighted dataset, with greater emphasis placed on instances that were misclassified by the ensemble so far. The final predictor is obtained as a weighted sum (or majority vote) of the M base learners:

$$F_M(x) = \sum_{m=1}^{M} \alpha_m h_m(x),$$

where α_m is the weight assigned to the m-th learner, typically reflecting its predictive accuracy. The most widely known Boosting algorithm is *AdaBoost* (Adaptive Boosting) introduced by Freund and Schapire [29]. In AdaBoost, after training each weak classifier h_m, its weighted error rate ε_m is computed. The coefficient α_m is then determined as

$$\alpha_m = \frac{1}{2} \log \left(\frac{1 - \varepsilon_m}{\varepsilon_m} \right),$$

which ensures that learners with lower error rates receive higher weights in the final ensemble. The sample weights are updated so that misclassified observations receive higher importance in the next iteration, forcing subsequent classifiers to focus on harder cases. Figure 1.3 illustrates the Boosting procedure, highlighting how misclassified instances are reweighted to iteratively improve the ensemble performance. Boosting is not limited to classification tasks. Extensions such as *Gradient Boosting* [5] generalize the

framework to arbitrary differentiable loss functions, including regression and probabilistic modeling. In gradient boosting, the model is built additively by fitting each base learner to the negative gradient of the loss function with respect to the current ensemble's predictions. This perspective reveals Boosting as a stage-wise functional gradient descent procedure, a connection that has been critical to the development of powerful algorithms such as XGBoost [30] and LightGBM [31]. From a statistical perspective, Boosting reduces bias by iteratively refining the decision boundaries of the model, while its variance tends to remain moderate compared to Bagging because the learners are not built independently. However, Boosting is sensitive to noise and outliers: since misclassified points receive increasing weight, the algorithm may overfit in the presence of mislabeled data [5]. To mitigate this issue, regularization techniques such as shrinkage (using a learning rate to scale α_m), subsampling of the training data (stochastic gradient boosting), and constraints on tree depth are commonly employed [14].

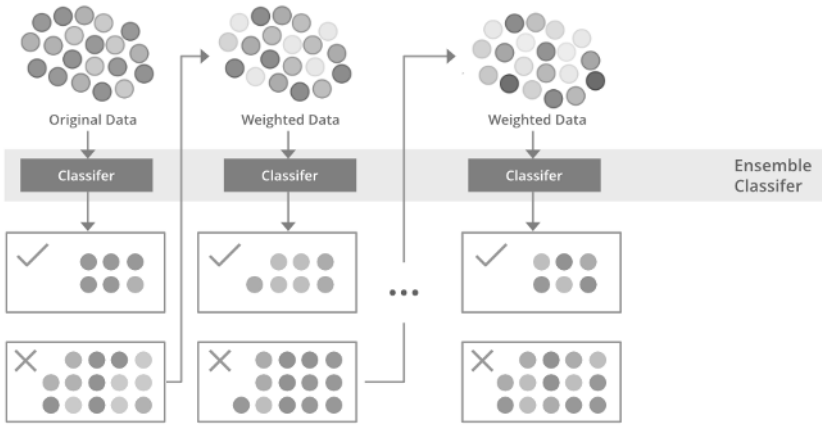


Figure 1.3. Schematic representation of the Boosting procedure, where sequential classifiers are trained on reweighted data, emphasizing previously misclassified observations, and combined into a final ensemble model.

Random Forest

RF are ensemble classifiers/regressors that aggregate the predictions of many decision trees grown on randomized versions of the training data [1]. Let $\{h(x;\Theta_k)\}_{k=1}^{K}$ denote the trees in the forest, where the Θ_k's are i.i.d. randomization vectors (bootstrap sample and feature subsampling). For a new input x, the RF predictor is obtained by averaging (regression) or by majority voting (classification):

$$\hat{f}_{\mathrm{RF}}(x) = \begin{cases} \dfrac{1}{K}\displaystyle\sum_{k=1}^{K} h(x;\Theta_k), & \text{regression,} \\ \operatorname{mode}\left\{h(x;\Theta_k)\right\}_{k=1}^{K}, & \text{classification.} \end{cases} \tag{1.2}$$

Construction and sources of randomness. RF extends Bagging by injecting two layers of randomness during tree growth: (i) *bootstrap sampling* of the training set for each tree; (ii) *feature subsampling* at each split: instead of considering all p predictors, a random subset of size $m_{\mathrm{try}} = v$ is drawn and only these features are eligible for splitting [1]. In practice, $v \approx \sqrt{p}$ is a common default for classification, while $p/3$ used for regression. Trees are commonly grown to full depth without pruning, yielding low-bias but high-variance base learners whose correlations are reduced by the feature randomization.

Variance reduction and decorrelation. Let $Z_k(x) = h(x;\Theta_k)$ for the k-th tree prediction, with mean μ and variance σ^2, and let ρ denote the average pairwise correlation among the $\{Z_k\}$. The variance of the forest average in Equation (1.2) (regression case) is

$$\mathbb{V}\left[\frac{1}{K}\sum_{k=1}^{K} Z_k\right] = \rho\,\sigma^2 + \frac{1-\rho}{K}\,\sigma^2, \tag{1.3}$$

showing that increasing K reduces the second term, while reducing ρ (via feature subsampling) decreases the first term [14]. Thus, RF improves upon Bagging by explicitly targeting low inter-tree correlation without inflating the variance of individual trees.

Prediction mechanism. Each tree maps x to a terminal leaf and outputs either a numeric value (node mean for regression) or a class vote (most frequent class in the leaf). The forest aggregates these leaf-level predictions as in Equation (1.2). Equivalently, RF can be viewed as an Additive Tree Model

$$F(x) = \sum_{t=1}^{K} w_t f_t(x), \qquad w_t = \frac{1}{K} \text{ (Bagging/RF)}, \tag{1.4}$$

with uniform weights for standard RF and node-wise impurity criteria (e.g., Gini for classification, variance reduction for regression) guiding each split [1].

Out-of-Bag (OOB) estimation. Because each bootstrap sample leaves out on average a fraction $e^{-1} \approx 0.368$ of the training observations, one can obtain *out-of-bag* predictions by evaluating, for each training case x_i, only the trees for which x_i was not used for fitting [32]. Let OOB_i denote this subset of trees; the OOB predictor is

$$\hat{f}_{\mathrm{OOB}}(x_i) = \begin{cases} \dfrac{1}{|\mathrm{OOB}_i|} \displaystyle\sum_{k \in \mathrm{OOB}_i} h(x_i; \Theta_k), & \text{regression}, \\ \mathrm{mode}\left\{h(x_i; \Theta_k)\right\}_{k \in \mathrm{OOB}_i}, & \text{classification}. \end{cases} \tag{1.5}$$

Aggregating (1.5) over $i = 1, \ldots, n$ yields an internal estimate of generalization error that is asymptotically comparable to leave-one-out cross-validation, while avoiding explicit data splitting.

Practical properties and hyperparameters. RFs are robust across a wide range of applications and data types (binary, categorical, continuous). They typically exhibit: (i) strong resistance to overfitting as K increases; (ii) stability to moderate perturbations/noise in the response; (iii) insensitivity to inclusion of irrelevant predictors due to randomized split selection; (iv) built-in diagnostics such as OOB error, proximity measures, and variable-importance scores [1]. Key hyperparameters are the number of trees K and the number of candidate features per split m_{try}; default values (large K, moderate m_{try}) often perform well with minimal

tuning.

Rationale for feature subsampling. If a few predictors are much stronger than the rest, Bagging (which considers all p features at each split) tends to select the same strong variables near the root across trees, producing highly correlated trees and limited variance reduction. By sampling only v candidates at each split, RF probabilistically excludes the strongest variables with frequency $(p - v)/p$, enabling alternative predictors to participate in early splits and thereby lowering ρ in Equation (1.3) [33]. The resulting ensemble acts as a "committee" of weakly correlated trees whose aggregated decision is more accurate and stable than that of any single tree.

1.2 Interpretability vs Explainability

The concepts of interpretability and explainability have become central to the discussion on transparency in ML. Although these terms are often used interchangeably, they capture different dimensions of how models can be understood and evaluated. The purpose of this chapter is to disentangle these notions, clarify their theoretical underpinnings, and highlight their complementary roles. By distinguishing between interpretability, which focuses on the direct comprehensibility of a model's structure, and explainability, which aims to make complex or opaque models intelligible, this chapter establishes the conceptual foundation for subsequent methodological discussions within the broader framework of Explainable Artificial Intelligence (XAI).
To provide an initial overview, Table 1.1 summarizes the main distinctions between interpretability and explainability. These concepts will then be introduced and examined in greater depth in the following sections of the chapter.

1.2.1 Interpretability

Interpretability is commonly defined as the degree to which a human can understand the cause-and-effect relationships within a model and trace how individual inputs contribute to a given output [34] [15]. It reflects the

Aspect	Interpretability	Explainability
Focus	Why the model produces specific outputs	How the model produces specific outputs
Scope	Causal relationships between variables	Internal processes and decision pathways
Model Dependency	Typically model-specific	Can be model-agnostic
Complexity	Relatively simple	Often requires additional tools or frameworks

Table 1.1. Key Differences Between Interpretability and Explainability.

intuitive transparency of a model: an interpretable system is one whose mechanisms can be grasped without the need for additional layers of abstraction or external explanatory tools. In this sense, interpretability is concerned not only with the accuracy of predictions but also with their legibility to human reasoning. As argued by Lipton [35], interpretability should not be reduced to a single dimension. It encompasses several overlapping notions, including simulatability (the ability of a human to mentally simulate the entire model) decomposability, referring to the clarity of each component such as coefficients or rules, and algorithmic transparency, which indicates the ease of describing the training procedure and assumptions. These dimensions highlight that interpretability is not an inherent binary property but rather a spectrum that varies across models and contexts. The most interpretable models are those whose structure directly maps onto human cognitive processes. Linear and logistic regression models, for example, provide coefficients that offer immediate insights into the direction and magnitude of relationships between predictors and outcome variables. In econometrics, regression coefficients allow analysts to quantify the marginal effect of unemployment on inflation or the elasticity of demand with respect to price. Similarly, decision trees of limited depth enable predictions through a clear sequence of if–then rules, making them particularly useful in social sciences and policy evaluation, where they

mimic decision-making processes familiar to human reasoning. Finally, rule-based models assign each prediction to an explicit and understandable rule, such as "if income $<$ X and education $=$ low, then probability of poverty $=$ 0.65." These examples illustrate that interpretability arises when the model structure itself is sufficiently simple for humans to comprehend without mediation.

Advantages and limitations of interpretability

The advantages of interpretability extend beyond academic elegance and are particularly relevant in practical applications. Interpretable models provide stakeholders (such as policy-makers, clinicians, or economists) with a clear rationale for decisions, thereby fostering trust and accountability. This is crucial in domains where opaque predictions may be legally or ethically unacceptable. In addition, transparency enables the integration of statistical results with substantive domain knowledge. Economists, for instance, can compare estimated coefficients with established theoretical expectations, ensuring that predictions are not only accurate but also consistent with economic reasoning. Another important advantage lies in communication: interpretable models facilitate the dissemination of findings to non-technical audiences, an essential feature in policy contexts where results must inform public debate and decision-making. Despite these strengths, interpretability comes at a cost. Human cognitive constraints limit the number of relationships that can be meaningfully understood. A regression model with dozens of variables, though formally interpretable, may be cognitively opaque to a practitioner [15]. Moreover, interpretable models often lack the flexibility to capture highly nonlinear or high-dimensional relationships, leading to reduced predictive accuracy compared to more complex black-box methods [36]. This trade-off raises a fundamental dilemma: privileging interpretability may undermine predictive performance, while maximizing accuracy may sacrifice transparency. Another limitation concerns the risk of apparent interpretability. Simple models may appear transparent but can still be misleading. For example, a decision tree might suggest that a single variable dominates the prediction process, but this apparent clarity may mask hidden interactions or data biases. As noted by Doshi et al. [3], interpretability must therefore be critically evaluated, since not all forms of "simplicity" correspond to

genuine understanding.

Interpretability in critical domain

The importance of interpretability is particularly pronounced in economics and the social sciences, where transparency, causal reasoning, and communicability are paramount. In these fields, statistical models are not only predictive devices but also instruments for theory building and policy design. Econometricians, for instance, have long relied on interpretable models to investigate the mechanisms underlying economic phenomena. A regression coefficient, for example, is not merely a numerical estimate; it embodies an interpretable statement about the relationship between economic variables, such as the marginal effect of unemployment on inflation or the elasticity of demand with respect to price. These coefficients provide direct evidence for or against theoretical assumptions and allow researchers to frame results within established conceptual frameworks. Survey-based research provides another important context in which interpretability becomes essential. Studies on health inequalities, unemployment risk, or educational attainment often involve diverse stakeholders, including policymakers, social workers, and the general public. In such cases, the ability to clearly interpret model results outweighs small gains in predictive accuracy, because the ultimate goal is to support evidence-based interventions that are both rigorous and understandable. Transparent models facilitate communication across disciplines, ensuring that complex statistical results can be translated into practical recommendations. For example, an interpretable regression or decision tree can reveal how socio-economic status, education, and health conditions interact to influence labor market participation, offering insights that are actionable by policymakers. From a broader epistemological perspective, interpretability aligns with the goals of the social sciences, which extend beyond prediction to explanation and understanding of human behavior. While opaque models may deliver marginally higher accuracy, they often fall short of offering meaningful explanations that can guide institutional change or social policy. Interpretable models, by contrast, bridge the gap between quantitative analysis and theoretical reasoning, ensuring that statistical findings contribute not only to forecasting but also to knowledge accumulation and societal decision-making. In this sense, interpretability remains indispens-

able, even in an era increasingly dominated by high-performing but opaque algorithms, because it preserves the explanatory and normative functions that are central to the mission of economics and the social sciences.

1.2.2 Explainability

Explainability refers to the degree to which the internal logic of a machine learning (ML) model can be articulated in a way that is comprehensible to humans. Unlike interpretability, which emphasizes the direct transparency of a model's structure, explainability deals with the challenge of making complex or opaque models intelligible. It addresses the question of how predictions are produced: how a model transforms input features into outputs and through which internal processes or pathways those transformations occur [34]. This distinction is crucial because many modern ML models, such as ensemble methods or deep neural networks, lack inherent interpretability due to their structural complexity. They involve numerous parameters, nonlinear transformations, and high-dimensional interactions that make it impossible for a human to fully grasp their mechanics without mediation. Explainability thus emerges as a post hoc strategy, whereby an additional layer of analysis or representation is applied to reveal insights about a model's decision-making process. A defining feature of explainability is its broad applicability. It is not restricted to a specific class of models: it can be employed to deepen the understanding of inherently interpretable structures and, more importantly, to render black-box systems intelligible. In the literature, three main characteristics of explainability are typically emphasized. First, it is model-agnostic, since explanation methods can often be applied independently of the underlying algorithmic architecture, making them suitable for diverse approaches ranging from decision trees to neural networks. Second, explainability operates at multiple levels of granularity: global explanations aim to clarify the overall decision logic of the model, while local explanations focus on individual predictions or subsets of the data, thus allowing users to reconcile general patterns with case-specific reasoning. Third, explainability is human-centered, as its ultimate purpose is not merely to document algorithmic mechanics but to provide narratives that foster trust, support accountability, and inform human action [3].

Advantages and limitations of explainability

Explainability serves several key functions that extend beyond the mere documentation of a model's internal mechanisms. It bridges the gap between predictive power and transparency, making it possible to deploy highly accurate models in sensitive domains such as healthcare, finance, or criminal justice. By reconstructing the rationale behind predictions, explainability enhances stakeholder trust and facilitates adoption in contexts where accountability is crucial. Furthermore, it enables error analysis and model debugging: by clarifying how predictions are generated, practitioners can identify biases, spurious correlations, or unintended behaviors embedded in the model. Explainability also contributes to ethical and regulatory compliance, as many emerging legal frameworks, including the European Union's General Data Protection Regulation (GDPR), explicitly require that algorithmic decisions be explainable to affected individuals. Beyond these practical advantages, explainability has an epistemic value, since it allows researchers to generate scientific insights from complex models. In medical research, for example, it can help uncover previously unknown associations between patient characteristics and disease outcomes, thereby advancing both predictive practice and domain knowledge. Despite these benefits, explainability faces several challenges. A central issue is the faithfulness–simplicity trade-off: explanatory models or representations often simplify the complexity of the original model, raising doubts about whether they accurately reflect the true decision logic. This can create an "illusion of transparency," where stakeholders are reassured by plausible but ultimately misleading explanations. Another limitation concerns computational costs, as many explanation methods are resource-intensive, particularly when applied to large-scale models, which restricts their feasibility in real-time decision-making contexts. Moreover, stability and consistency are not always guaranteed, since explanations for the same model may vary across runs or depend heavily on arbitrary methodological choices, undermining their reliability. Finally, explainability introduces epistemological ambiguity. While it aims to provide human-understandable accounts of algorithmic reasoning, it does not necessarily imply causality. Explanations may describe correlations or approximations rather than genuine causal pathways, which can mislead policymakers or practitioners if not critically assessed.

Explainability in critical domains

In socio-economic applications, explainability plays a complementary role to interpretability. While policymakers may prefer interpretable models for their directness, complex socio-economic systems often require high-capacity models that cannot be directly understood. In such cases, explainability becomes indispensable, as it offers a structured account of the relationships learned by the model without sacrificing predictive performance. For example, in the context of public health or labor market analysis, explainability enables researchers to uncover non-linear interactions (such as how education, income, and regional disparities jointly influence unemployment risk) that would remain hidden in simpler models. In this way, explainability provides a more nuanced evidence base for interventions, while still satisfying the demand for transparency and accountability in policy-making. The literature, however, remains divided on the role of explainability. Some scholars argue that reliance on post hoc explanations may obscure the importance of developing inherently interpretable models, which should be preferred in high-stakes applications [36]. Others contend that explainability is indispensable, since complex systems cannot realistically be reduced to fully interpretable forms without significant losses in predictive capacity. This debate underscores the necessity of a balanced approach: explainability should not be viewed as a substitute for interpretability, but as a complementary strategy that extends transparency to models that would otherwise remain opaque. Ultimately, explainability represents both a methodological tool and a normative requirement. Its effectiveness depends on the careful calibration of accuracy, transparency, and ethical accountability. By providing structured accounts of complex models, explainability empowers ML to operate not only as a predictive technology but also as a trustworthy instrument for decision-making in socially consequential domains.

Chapter 2

Random Forest interpretability: methods and comparison

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.[a]
Keywords: *Random Forest, Model interpretation, Rule extraction, inTrees, NodeHarvest.*

[a]This chapter has been published as: Aria, M., Cuccurullo, C., & Gnasso, A. (2021). A comparison among interpretative proposals for Random Forests. Machine Learning with Applications, 6, 100094.

In the previous chapter, the concepts of interpretability and explainability were introduced as key dimensions of transparency in machine learning. Building on this foundation, the present chapter focuses on interpretability in the specific context of RF models. A growing body of research has proposed various methodologies to open the "black box", aiming to provide meaningful insights into their decision processes (see, e.g., [37, 38, 39, 35]). While these contributions have compared interpretative techniques from a methodological perspective, less attention has been devoted to systematically assessing their relative accuracy and practical effectiveness. To structure this review, we adopt the taxonomy proposed by Haddouchi & Berrado [40], which distinguishes between approaches that exploit the internal mechanisms of RF and post-hoc methods applied after model training (Figure 2.1). Within this framework, the following sections present the main interpretative proposals, highlighting their theoretical underpinnings, practical implementation, and comparative strengths and weaknesses.

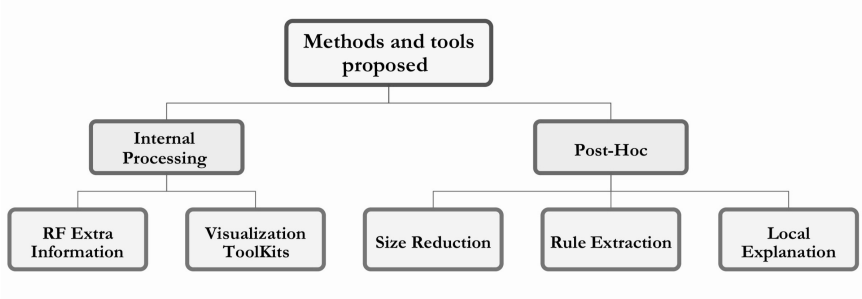


Figure 2.1. Classification of the methods aiming the RF interpretability.

2.0.1 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.

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 [1, 41, 42]). In addition to Variable Importance, another approach to analyzing the relationships between features and predictions is partial dependence through the *Partial Dependence Plot (PDP)* [5]. It describes the marginal effect that the features assume concerning predictions. [5] 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, $z_l = \{z_1, ..., z_l\} \subset \{x_1, ..., x_n\}$, and let $z_\prime$ be the complement subset $z_\prime \cup z_l = x$. The approximation $\hat{F}(x)$ depends on the variables included in the subset $\hat{F}(x) = \hat{F}(z_l, z_\prime)$. If there is a dependence on a specific value for the variables in $z_\prime$, then $\hat{F}(x)$ can be considered as a function of variables chosen in the subset z_l[1]: in particular $\hat{F}_{z_\prime}(z_l) = \hat{F}(z_l|z_\prime)$. 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

$$\bar{F}_l(z_l) = E_{z_\prime}\left[\hat{F}(x)\right] = \int \hat{F}(z_l, z_\prime) p_\prime(z_\prime) dz_\prime \tag{2.1}$$

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

$$\bar{F}_l(z_l) = \frac{1}{N} \sum_{i=1}^{N} \hat{F}(z_l, z_{i,\prime}) \tag{2.2}$$

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

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

Proximity Matrix. It represents the fraction of trees in which elements i and j fall into the same terminal node. The *RF proximity* of a pair of points is an unweighted average of the number of trees in the RF model where the points end in the same leaf [6]:

$$proximity_{RF}(x, x^{'}) = \frac{1}{T} \sum_{i=1}^{T} \sum_{j=1}^{\tau_i} f(x \in R_{j,i}) f(x^{'} \in R_{j,i}) \tag{2.3}$$

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 RF 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.

Visualization ToolKits

The visualization toolkits [43] allow the analysis of the aforementioned approaches from a graphic point of view. Liaw [44] 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 [45] 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

[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.

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 $j - th$ variable with respect to the maximum subtree of the $i - th$ 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 [46] 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 tech-

niques 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 et al. [47] 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.

2.0.2 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.

Size Reduction

An essential premise is that Tree ensembles, like RF, have internal characteristics that can be used to approximate an understandable global model: Chipman et al. [48] 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, or only in architecture, but use the same division of space as features X. Wang et al. [49] 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 $i = 1, ..., N_{RF} - 1$, is analyzed in order to iden-

tify 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. [50] 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 et al. [51] propose an Approximation Tree procedure, an extension of the previous method [50] 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. [51] 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.

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

[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.

possible decision. In most studies, a rule means a decision path, from the root node to a leaf node. *Node Harvest* [52] 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* [53], 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

[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.

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 $\Re$ 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 $\Re$. 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 $\Re$ as output.

Local Explanation

Through a local decomposition [43] it is possible to check and analyze the predictions made by a RF. In recent years, approaches defined as agnostics [38] have been developed with respect to the Black Box model that we want to interpret. By definition, these approaches are 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 et al. [54]: 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 RF. Recently, Ribeiro et al. [55] 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) [56] 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

[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.

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 et al. [56], 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.

2.1 Comparison Study

In this section, a ML experiment is conducted among the rule extraction approaches previously described. First, we illustrate the methodolog-

ical 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.

2.1.1 Experimental Design

The aim is to compare rule extraction approaches through the use of several datasets. We focus on Node Harvest [52] and inTrees [53]: these methodologies provide an interpretative structure that synthesizes and interprets the forest of trees produced by the RF. 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 2.1). 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 RF is grown for each of the six datasets in order to predict the

Datasets	Obs.	Qual. Feat.	Quant. Feat.	0/1 Response Rate	Unbalanced Response
Bank Marketing	4119	11	10	1072/35	True
Churn Modelling	10000	5	6	2301/281	True
Banknote Authentication	1372	1	4	222/184	False
Cardiovascular Disease	70000	7	5	785/736	False
Pima Indians Diabetes	768	8	1	125/56	False
Carseat Sales	400	4	7	61/34	False

Table 2.1. Main characteristics of the selected datasets.

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 out using nine metrics: Accuracy, Precision, Sensitivity, Specificity, G-Mean, F1 Score, Youden's Index, Balanced Accuracy, Kappa ([57], [58], [59]). All considered metrics are calculated starting from

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

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:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.4}$$

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

$$Precision = \frac{TP}{TP + FP} \tag{2.5}$$

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

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.6}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2.7}$$

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

$$G - mean = \sqrt{Sensitivity * Specificity} \tag{2.8}$$

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

$$F1\,Score = \frac{2 * Sensitivity * Precision}{Sensitivity + Precision} = \frac{2TP}{2TP + FP + FN} \tag{2.9}$$

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

$$Youden's Index = Sensitivity - (1 - Specificity) \tag{2.10}$$

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

$$Balanced\,Accuracy = \frac{1}{2}(Sensitivity * Specificity) \tag{2.11}$$

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

$$Kappa = \frac{Total\,Accuracy - Random\,Accuracy}{1 - Random\,Accuracy} \tag{2.12}$$

The two Rule extraction approaches are compared with each other and with the RF 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 RF applied on the same data. Closer results mean a better ability to reconstruct the original "black-box" supervised model with an interpretable structure. The ML 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 [60] which suggests to use about 70% for the training set and the rest for the test set.

2.1.2 Analysis

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

- *Random Forest* (benchmark model): the model was created using the randomForest package [44]. Hyperparameters are set by default. Results are reported in table 2.2.

Bank Marketing

	Actual	
Predicted	0	1
0	1072	98
1	31	35

Churn Modelling

	Actual	
Predicted	0	1
0	2301	333
1	85	281

Banknote Authentication

	Actual	
Predicted	0	1
0	222	2
1	4	184

Cardiovascular Disease

	Actual	
Predicted	0	1
0	785	319
1	260	736

Pima Indians Diabetes

	Actual	
Predicted	0	1
0	125	22
1	27	56

Carseat Sales

	Actual	
Predicted	0	1
0	61	20
1	5	34

Table 2.2. RF confusion matrices

- *Node Harvest*: the estimator is obtained with the package nodeHarvest [52], 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 [53]. 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.

Bank Marketing

	Actual	
Predicted	0	1
0	1095	116
1	8	17

Churn Modelling

	Actual	
Predicted	0	1
0	2328	392
1	58	222

Banknote Authentication

	Actual	
Predicted	0	1
0	207	13
1	19	173

Cardiovascular Disease

	Actual	
Predicted	0	1
0	812	367
1	226	695

Pima Indians Diabetes

	Actual	
Predicted	0	1
0	136	35
1	16	43

Carseat Sales

	Actual	
Predicted	0	1
0	59	31
1	7	23

Table 2.3. Node Harvest confusion matrices

Bank Marketing

	Actual	
Predicted	0	1
0	1096	119
1	7	14

Churn Modelling

	Actual	
Predicted	0	1
0	2306	368
1	80	246

Banknote Authentication

	Actual	
Predicted	0	1
0	221	8
1	5	178

Cardiovascular Disease

	Actual	
Predicted	0	1
0	694	253
1	344	809

Pima Indians Diabetes

	Actual	
Predicted	0	1
0	128	24
1	24	54

Carseat Sales

	Actual	
Predicted	0	1
0	58	19
1	8	35

Table 2.4. inTrees confusion matrices

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. The results are shown in the table 2.5, 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, 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 [61]. For the unbalanced data, there is no prevalence between the two methods analyzed as they both appear to work similarly (Tables 2.5a and 2.5b). 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 RF 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 RF, in order to provide a means that is able to help, for example, in the detection of financial fraud or disease.

	randomForest	*nodeHarvest*	*inTrees*
*Accuracy**	0.8956	0.8997	0.8981
Balanced Accuracy	0.6175	**0.5603**	0.5495
Kappa	0.3019	**0.1875**	0.1571
Specificity	0.2632	**0.1278**	0.1053
Sensitivity	0.9719	0.9927	**0.9937**
Precision	0.9162	**0.9042**	0.9021
G-mean	0.5057	**0.3562**	0.3234
F1	0.9432	**0.9464**	0.9456
Youden's Index	0.2350	**0.1206**	0.0989

(a) *Bank Marketing*

	randomForest	*nodeHarvest*	*inTrees*
*Accuracy**	0.8607	0.8500	0.8507
Balanced Accuracy	0.7110	0.6686	**0.6836**
Kappa	0.4965	0.4226	**0.4446**
Specificity	0.4577	0.3616	**0.4007**
Sensitivity	0.9644	**0.9757**	0.9665
Precision	0.8736	0.8559	**0.8624**
G-mean	0.6643	0.5939	**0.6223**
F1	0.9167	**0.9119**	0.9115
Youden's Index	0.4220	0.3373	**0.3671**

(b) *Churn Modelling*

	randomForest	*nodeHarvest*	*inTrees*
Accuracy	0.9854	0.9223	**0.9684**
Balanced Accuracy	0.9858	0.9230	**0.9674**
Kappa	0.9706	0.8436	**0.9362**
Specificity	0.9892	0.9301	**0.9570**
Sensitivity	0.9823	0.9159	**0.9779**
Precision	0.9911	0.9409	**0.9651**
G-mean	0.9858	0.9230	**0.9674**
F1	0.9867	0.9283	**0.9714**
Youden's Index	0.9715	0.8460	**0.9349**

(c) *Banknote Authentication*

	randomForest	*nodeHarvest*	*inTrees*
Accuracy	0.7243	**0.7176**	0.7157
Balanced Accuracy	0.7244	**0.7183**	0.7152
Kappa	0.4487	**0.4360**	0.4308
Specificity	0.6976	0.6544	**0.7618**
Sensitivity	0.7512	**0.7823**	0.6686
Precision	0.7111	0.6887	**0.7328**
G-mean	0.7239	**0.7155**	0.7137
F1	0.7306	**0.7325**	0.6992
Youden's Index	0.4488	**0.4367**	0.4304

(d) *Cardiovascular Disease*

	randomForest	*nodeHarvest*	*inTrees*
Accuracy	0.7870	0.7783	**0.7913**
Balanced Accuracy	0.7702	0.7230	**0.7672**
Kappa	0.5320	0.4741	**0.5344**
Specificity	0.7179	0.5513	**0.6923**
Sensitivity	0.8224	**0.8947**	0.8421
Precision	0.8503	0.7953	**0.8421**
G-mean	0.7684	0.7023	**0.7635**
F1	0.8361	**0.8421**	**0.8421**
Youden's Index	0.5403	0.4460	**0.5344**

(e) *Pima Indians Diabetes*

	randomForest	*nodeHarvest*	*inTrees*
Accuracy	0.7917	0.6833	**0.7750**
Balanced Accuracy	0.7769	0.6599	**0.7635**
Kappa	0.5682	0.3333	**0.5369**
Specificity	0.6296	0.4259	**0.6481**
Sensitivity	0.9242	**0.8939**	0.8788
Precision	0.7531	0.6556	**0.7532**
G-mean	0.7628	0.6171	**0.7547**
F1	0.8299	0.7564	**0.8112**
Youden's Index	0.5539	0.3199	**0.5269**

(f) *Carseat Sales*

Table 2.5. 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.

2.2 Concluding remarks

In addition to having high predictive performance, the RF 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 ML experiment aims to highlight the two main approaches proposed in the 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 2.5 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

Chapter 3

Explainable Ensemble Trees

Ensemble methods are supervised learning algorithms that provide highly accurate solutions by training many models. Random forest is probably the most widely used in regression and classification problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. However, such an algorithm suffers from a lack of explainability and thus does not allow users to understand how particular decisions are made. To improve on that, we propose a new way of interpreting an ensemble tree structure. Starting from a random forest model, our approach is able to explain graphically the relationship structure between the response variable and predictors. The proposed method appears to be useful in all real-world cases where model interpretation for predictive purposes is crucial. The proposal is evaluated by means of real data sets.[a]
Keywords: *Machine Learning, Random Forest, Classification, Explainability.*

[a]This chapter has been published as: Aria, M., Gnasso, A., Iorio, C., & Pandolfo, G. (2024). Explainable ensemble trees. Computational Statistics, 39(1), 3-19.

In the previous chapter, existing interpretative approaches for RF were reviewed and critically assessed, highlighting their advantages and limitations. Although these methods provide valuable insights into the functioning of ensemble models, they often remain fragmented, either focusing on feature-level attributions or relying on surrogate simplifications that do not fully capture the global decision structure. This chapter introduces the *E2Tree*, a novel methodology designed to combine the predictive accuracy of RF with the interpretability and explainability of decision trees. E2Tree reconstructs the logic of an ensemble into a single tree-like structure, derived from a dissimilarity matrix that summarizes co-occurrence patterns among observations across trees. In doing so, the method provides a coherent and transparent representation of the ensemble's decision-making process, bridging the gap between accuracy, interpretability and explainability.

3.1 Explainable Ensemble Trees: the key idea

The *E2Tree* key idea consists of the definition of an algorithm to represent an RF model using a single tree-like structure. The goal is to explain the results from the RF algorithm while preserving its level of accuracy, which always outperforms those provided by a decision tree. We want to emphasize that our contribution is new in the field of model explainability for ensemble tree algorithms. The proposed method is based on identifying the relationship tree-like structure explaining the classification or regression paths summarizing the whole ensemble process. There are two main advantages of E2Tree: *(i)* building an explainable tree that ensures the predictive performance of an RF model and *(ii)* allowing the decision-maker to manage with an intuitive structure (such as a tree-like structure).
In this work, we focus on RF but the algorithm can be generalized to every ensemble approach based on decision trees.
Let H be an RF model composed by B weak learners

$$H = \{H_1, H_2, \ldots, H_b, \ldots, H_B\}.$$

Let $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ be a sample of size n used to train the RF.
Let $U_i = \{Y_i, X_i\}$ be the vector associated to the i-th observation.
The algorithm starts from the definition of a dissimilarity matrix D:

$$D = \begin{array}{c} \\ u_1 \\ u_2 \\ u_3 \\ \\ u_i \\ u_j \\ \\ u_n \end{array} \begin{array}{c} \begin{array}{cccccccc} u_1 & u_2 & u_3 & & u_i & u_j & & u_n \end{array} \\ \begin{bmatrix} 0 & d_{12} & d_{13} & \cdots & d_{1i} & d_{1j} & \cdots & d_{1n} \\ d_{21} & 0 & d_{23} & \cdots & d_{2i} & d_{2j} & \cdots & d_{2n} \\ d_{31} & d_{32} & 0 & \cdots & d_{3i} & d_{3j} & \cdots & d_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ d_{i1} & d_{i2} & d_{i3} & \cdots & 0 & d_{ij} & \cdots & d_{in} \\ d_{j1} & d_{j2} & d_{j3} & \cdots & d_{ij} & 0 & \cdots & d_{jn} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & d_{n3} & \cdots & \cdots & d_{nj} & \cdots & 0 \end{bmatrix} \end{array} = 1 - \begin{bmatrix} \\ O_{ij} \\ \\ \end{bmatrix} \tag{3.1}$$

where O_{ij} is a weighted co-occurrence measure between u_i and u_j along the sequence $H_1, \ldots, H_B$ weak learnes. Hence, d_{ij} measures the dissimilarity (or the "discordance") between two observations with respect to a given classifier H. In other words, O_{ij} counts how many times u_i and u_j co-occur in a terminal node along the B weak learners sequence as:

$$O_{ij} = \frac{\sum_{b=1}^{B} I(u_i \wedge u_j) \cdot W_{t_{ij}|b}}{max(\sum_b u_i W_{t_i|b}, \sum_b u_j W_{t_j|b})} \tag{3.2}$$

where $W_{t_{ij}|b}$ is a local goodness of fit measure of the node $t_{ij}|b$.
If $I(u_i \wedge u_j) = 1$, then $t_{ij}|b$ exists and it represents the terminal node t of the weak learner b that contains i and j. Each weak learner b can contain only one terminal node $t_{ij}|b$ so the summation with respect to b indicates that, along the ensemble sequence for each tree b, only $t_{ij}|b$ nodes will be considered when they exist.
By definition $0 \leq O_{ij} \leq 1$ and $0 \leq W_{t_{ij}|b} \leq 1$.
Here, we set $W_{t_{ij}|b} = R_t$, where R_t is the correct classification rate at a generic node t, defined as:

$$R_t = \sum_{i=1}^{n_t} \frac{I(y_i = \hat{y}_t)}{n_t},$$

n_t is the size of node t and $\hat{y}_t$ is the classification returned by the RF algorithm in the same node. The goal is to build a single tree that explains

the information contained in D through the p predictors. The innovative contribution of our proposal refers to the learning of a tree representing, in a single, explainable, logical, and graphical structure, the aggregated classification obtained along with the ensemble procedure. The matrix D summarizes the aggregated classification in a new way. Indeed, at each node, the values of O measure the strength with which a set of observations must be forced to fall into the same node along with the construction of the tree, minimizing the values of O (or maximizing the values of D).
To build E2Tree, we define a set of stopping rules that retrace the classical rules of tree construction. In particular, we consider thresholds to the size and impurity of the terminal nodes and the complexity of the structure.
It is well-know that RF tend to have many non-informative splits because of the random selection of split variable. We might expect that with more nodes in a tree, the d_{ij} dissimilarity values between two observations (i, j) with respect to a given classifier tend more and more strongly to one. To overcome this problem, in the proposed algorithm we set an additional stopping rule. Once we have identified the best split concerning dissimilarity, we check that both child nodes have a different response class from each other and a well-classified rate below a certain threshold (e.g., 95% or 99%). In this way, we are able to avoid non-informative branches that result in overfitting.
More formally, let $s \in S$ be a candidate split for a variable X that splits t into left and right children nodes t_L and t_R according to whether $X \leq s$ or $X > s$. These two nodes will be denoted by $t_L = U \in t : X \leq s$ and $t_R = U \in t : X > s$.
In order to maximize the difference in discordance among the parent and children nodes, we define the following measures:

1. $\bar{d}_t$ is the impurity measure at node t: $\bar{d}_t = \sum_i \sum_j \frac{d_{ij}}{n_t \cdot (n_t - 1)}$

2. $\bar{d}_{t|s}$ is the pooled impurity measure of the children nodes generated

by s at node t

$$\bar{d}_{t|s} = \left\{ \left[\left(\frac{\sum_i^{n_{t_L}} \sum_j^{n_{t_L}} d_{ij}}{n_{t_L} \cdot (n_{t_L} - 1)} \cdot \frac{n_{t_L}}{n_t} \right) + \left(\frac{\sum_i^{n_{t_R}} \sum_j^{n_{t_R}} d_{ij}}{n_{t_R} \cdot (n_{t_R} - 1)} \cdot \frac{n_{t_R}}{n_t} \right) \right] \middle| s \right\}$$
$$= \left\{ \left[\frac{1}{n_t} \cdot \left(\frac{\sum_i^{n_{t_L}} \sum_j^{n_{t_L}} d_{ij}}{(n_{t_L} - 1)} + \frac{\sum_i^{n_{t_R}} \sum_j^{n_{t_R}} d_{ij}}{(n_{t_R} - 1)} \right) \right] \middle| s \right\}.$$

3. $\Delta\bar{d}_{t|s}$ is the decrease of impurity at node t splitted by s

$$\Delta\bar{d}_{t|s} = \{\bar{d}_t - [(\bar{d}_{t_L} \cdot p_{t_L} + \bar{d}_{t_R} \cdot p_{t_R})|s]\},$$

where $s \in S$. Note that $[(\bar{d}_{t_L} \cdot p_{t_L} + \bar{d}_{t_R} \cdot p_{t_R})|s] = \bar{d}_{t|s}$. Then, it is straightforward that

$$\max_{s \in S} \Delta\bar{d}_{t|s} \equiv \min_{s \in S} \{[(\bar{d}_{t_L} \cdot p_{t_L} + \bar{d}_{t_R} \cdot p_{t_R})|s]\}.$$

The best split selected at node t, denoted by s^*, maximizes the decrease of impurity $\Delta\bar{d}_{t|s^*}$.

Algorithm Explainable Ensemble Trees

```
1: procedure
2:     System Initialization:
3:       - Calculate the dissimilarity matrix D from the RF model
4:       - Set Stopping Rules
5:       - Generate Matrix S of all possible splits
6:     repeat
7:           - Select node t
8:           - Identify best split s* : max_{s∈S} Δd̄_{t|s}
9:           - Generate the two children nodes
10:            t_L = t · 2
11:            t_R = (t · 2) + 1
12:          - Calculate statistics for t_L and t_R
13:    until At least one of the stopping rules is satisfied
14: end procedure
```

3.2 Illustrative examples

To evaluate the performance of our proposal, we performed applications on freely available data sets. The source of data is the UCI Machine Learning repository (https://archive.ics.uci.edu/ml/index.php).
Data analysis was performed using the `R` software with our routines, which are available at the `GitHub` repository *https://github.com/massimoaria/eTree*.
Following Guyon [60], for each data set, we used 70% for the training sample and 30% for the testing sample. We set the number of trees in the forest equal to 1000 and the number of randomly selected features/variables to evaluate at each tree node equal to 2. These were selected to exhibit different levels of class imbalance and to be a mixture of discrete and continuous features.
In all the tables, t denotes the number of the node following the approach proposed by the authors of the statistical software SPAD (Cisia Institute, France), n_t is the number of the instances in each node t, where the category predicted is assigned to the "Class" $\hat{Y}_t$ which presents the highest proportion of observations, R_t is the correct classification rate for each node t, $\bar{d}_t$ is the impurity measure at node t as defined in the above Section, s^* is the best split selected at node t, and the entire path of the tree is denoted by "Path".
In all the figures depicting the "Explainable Ensemble Tree", the terminal nodes are denoted by square boxes, while non-terminal nodes are denoted by circles.

3.2.1 Iris data

Iris data [62] is a standard benchmark set and have frequently been used for illustrating new multivariate techniques. It has some well-known structure and is thus useful to see how well this structure is recovered.
In this data set, there are four measurements on 50 plants from each of three species of Iris: *Iris setosa*, *Iris versicolor*, and *Iris virginica*. The *setosa* plants are well separated from the others, but there is some overlap between the *versicolor* and *virginica* species. Each iris has four measurements *sepal length*, *sepal width*, *petal length*, and *petal width*. Table 3.1 and Table 3.2 show the results of our proposal. As can be noticed by looking

at both the tables, our proposal highlights the entire path, also offering, for each node, the most important information for predictive purposes. Table 3.1 indicates the importance of the split showing the contribution of the nodes with the lower (higher) impurity measure to the higher (lower) correct classification rate. Table 3.2 indicates the path of each prediction made by the RF algorithm. The main advantage of the proposed E2Tree is depicted in Figure 3.1 which offers, at first sight, the importance of the splits, showing the relational structure between the response variable and the predictors and their interactions. This represents an important advantage since no visual information is provided by an RF model.

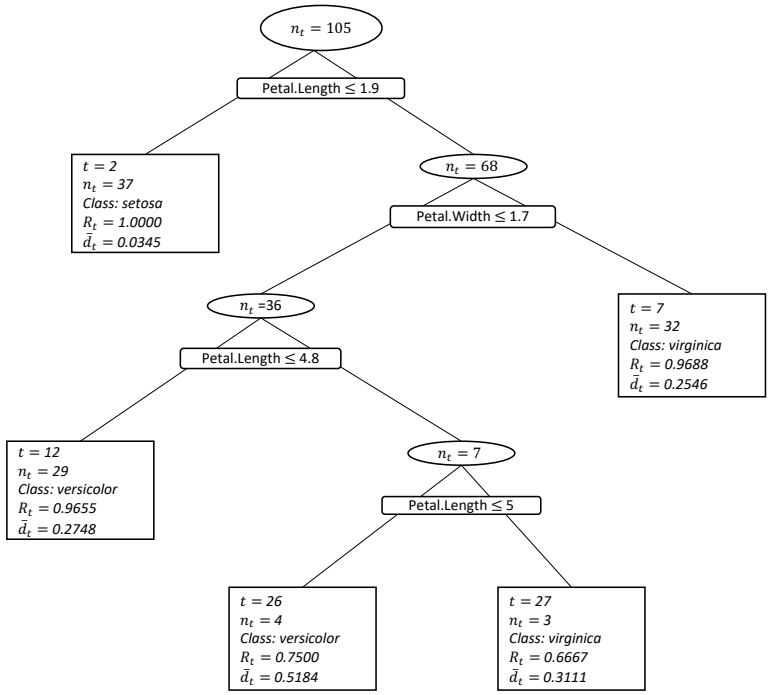


Figure 3.1. Iris data set: path visualization of the E2Tree.

t	n_t	Class	R_t	$\bar{d}_t$	s^*	Node Type
1	105	setosa	0.3524	0.7342	Petal.Length $\leq$ 1.9	Non-terminal
2	37	setosa	1.0000	0.0345	-	Terminal
3	68	virginica	0.5147	0.6476	Petal.Width $\leq$ 1.7	Non-terminal
6	36	versicolor	0.8889	0.4232	Petal.Length $\leq$ 4.8	Non-terminal
7	32	virginica	0.9688	0.2546	-	Terminal
12	29	versicolor	0.9655	0.2748	-	Terminal
13	7	versicolor	0.5714	0.5506	Petal.Length $\leq$ 5	Non-terminal
26	4	versicolor	0.7500	0.5184	-	Terminal
27	3	virginica	0.6667	0.3111	-	Terminal

Table 3.1. Iris data set: results of the E2Tree.

t	Class	Path
2	setosa	Petal.Length $\leq$ 1.9
12	versicolor	Petal.Length $>$ 1.9 & Petal.Width $\leq$ 1.7 & Petal.Length $\leq$ 4.8
26	versicolor	Petal.Length $>$ 1.9 & Petal.Width $\leq$ 1.7 & Petal.Length $>$ 4.8 & Petal.Length $\leq$ 5
27	virginica	Petal.Length $>$ 1.9 & Petal.Width $\leq$ 1.7 & Petal.Length $>$ 4.8 & Petal.Length $>$ 5

Table 3.2. Iris data set: paths of the E2Tree.

3.2.2 Breast Cancer Wisconsin data

This data set is used to classify a set of breast cancer patients. The characteristics are calculated from a digitized image of a needle aspiration of a breast mass taken from 569 patients with or without cancer. The features describe the cell nuclei present in the image. Ten real-valued features are computed for each cell nucleus. The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. The goal is to classify these cell nuclei as benign or malignant. One of the most important features of E2Tree is the possibility to provide a clear interpretation of the pathways "If-Then" that identify how interactions among predictors define the final classification in each terminal node. This is possible both in a graphical and tabular representation. Figure 3.2 shows a graphical representation of the E2Tree structure while Table 3.3 and Table 3.4 show the results for this data set.

By looking at the Figure 3.2, individuals who have a lesion characterized by a "perimeter" worst greater than 110.3 (t=3), a "standard error area"

less than or equal to 43.4 (t=6), and a "concavity mean" greater than 0.101 (t=13) are classified with a malignant status with a correct classification rate of 98%. Following the same pathway, in the last split when the "concavity mean" is less than or equal to 0.101 the malignant status is diagnosed with a correct classification rate of 60%, while the benign status is diagnosed with 40%. In this case, an expert in the research domain should carefully examine the status of the patient.

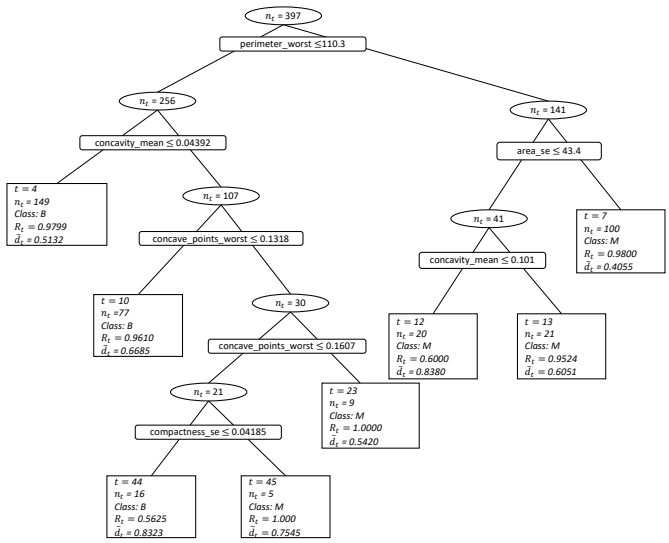


Figure 3.2. Breast data set: path visualization of the E2Tree.

t	n_t	Class	R_t	$\tilde{d}_t$	s^*	Node Type
1	397	B	0.6121	0.7991	perimeter_worst $\leq$ 110.3	Non-terminal
2	256	B	0.9062	0.6706	concavity_mean $\leq$ 0.04392	Non-terminal
3	141	M	0.9220	0.5748	area_se $\leq$ 43.4	Non-terminal
4	149	B	0.9799	0.5132	-	Terminal
5	107	B	0.8037	0.7715	concave_points_worst $\leq$ 0.1318	Non-terminal
6	41	M	0.7805	0.7988	concavity_mean $\leq$ 0.101	Non-terminal
7	100	M	0.9800	0.4055	-	Terminal
10	77	B	0.9610	0.6685	-	Terminal
11	30	M	0.6000	0.8292	concave_points_worst $\leq$ 0.1607	Non-terminal
12	20	M	0.6000	0.8380	-	Terminal
13	21	M	0.9524	0.6051	-	Terminal
22	21	B	0.5714	0.8505	compactness_se $\leq$ 0.04185	Non-terminal
23	9	M	1.0000	0.5420	-	Terminal
44	16	M	0.5625	0.8323	-	Terminal
45	5	B	1.0000	0.7545	-	Terminal

Table 3.3. Breast data set: results of the E2Tree.

t	Class	Path
4	B	perimeter_worst $\leq$ 110.3 & concavity_mean $\leq$ 0.04392
7	M	perimeter_worst $>$ 110.3 & area_se $>$ 43.4
10	B	perimeter_worst $\leq$ 110.3 & concavity_mean $>$ 0.04392 & concave_points_worst $\leq$ 0.1318
12	M	perimeter_worst $>$ 110.3 & area_se $\leq$ 43.4 & concavity_mean $\leq$ 0.101
13	M	perimeter_worst $>$ 110.3 & area_se $\leq$ 43.4 & concavity_mean $>$ 0.101
23	M	perimeter_worst $\leq$ 110.3 & concavity_mean $>$ 0.04392 & concave_points_worst $>$ 0.1318 & concave_points_worst $>$ 0.1607
44	M	perimeter_worst $\leq$ 110.3 & concavity_mean $>$ 0.04392 & concave_points_worst $>$ 0.1318 & concave_points_worst $\leq$ 0.1607 & compactness_se $\leq$ 0.04185
45	B	perimeter_worst $\leq$ 110.3 & concavity_mean $>$ 0.04392 & concave_points_worst $>$ 0.1318 & concave_points_worst $\leq$ 0.1607 & compactness_se $>$ 0.04185

Table 3.4. Breast data set: path of the E2Tree.

3.2.3 Indian diabetes data

The goal of the data set is to predict diagnostically whether a patient has diabetes or not, through various diagnostic measurements. Considering the constraints placed on the selection of the instances, we note that all patients are women of at least 21 years of Pima Indian origin. Among the features used there are the number of pregnancies the patient has had, her body mass index, insulin level, age, and so on. It contains 768 subjects with eight variables relevant to subjects.
Table 3.5, Table 3.6, Figure 3.3 show the results for this data set. As can be noticed, by leveraging the advantages of RF algorithms that handle large datasets efficiently, we offer a greater explanation of their results.
As for the previous dataset, here we try to describe some pathways to provide an example of how our proposal helps the decision-maker to understand the relationships among the predictor interactions and the final classification in each terminal node. Patients with a BMI greater than 27.8(t=3), glucose less than or equal to 113(t=6), age less than or equal to 27 (t=12), BMI less than or equal to 31.6 (t=24) are classified as non-diabetic with a correct classification rate of 93%.

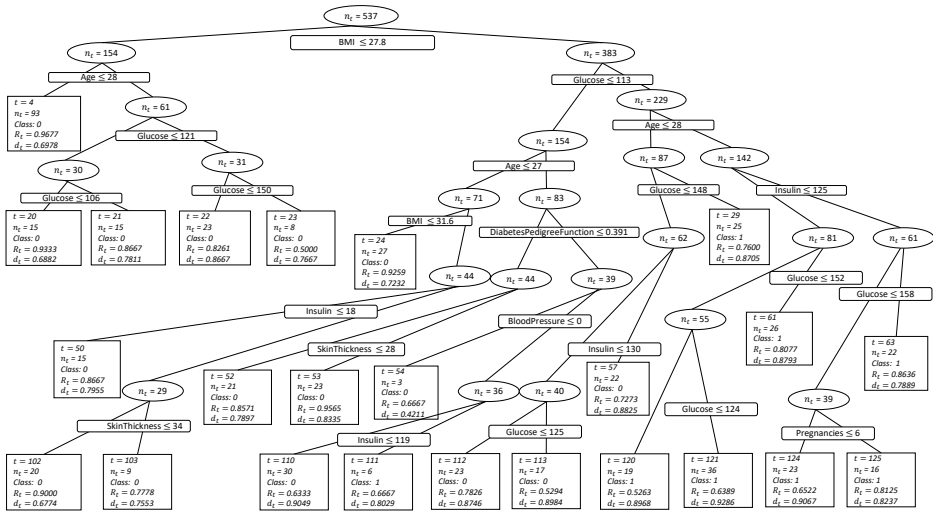


Figure 3.3. Diabetes data set: path visualization of the E2Tree.

3.3 Concluding remarks

We have proposed E2Tree aimed at explaining and graphically representing the relationships and interactions between the variables used to perform an ensemble method, i.e. random forest algorithm. The latter is a well-known supervised ML algorithm that creates many decision trees and combines the output of all trees. Compared to a decision tree, the main advantage of a random forest is the accuracy of prediction, but the main disadvantage is the lack of explainability of the final result. Nowadays, there are a lot of domains characterized by making critical decisions demanding explainability (e.g., medicine, finance, and marketing). To preserve the same advantages of an ensemble method while also ensuring its greater explainability, our proposal starts by defining a dissimilarity matrix measuring the discordance between two observations concerning a given classifier of a random forest model. Hence, we provide a dendrogram-like structure capable of explaining all the information contained in the output of a RF algorithm. We want to stress that our proposal has no predictive purposes, since we aim to bridge the lack of interpretability typical of ensemble methods like a RF model. We also want to point out that we refer to the accuracy in order to evaluate how our proposal reflects the accuracy of any benchmarking RF. The proposed method ensures the following advantages: represents the ensemble process through a single tree-like structure; saves the accuracy performance of the ensemble process; and provides a powerful tool to explain iterations among predictors determining the final node classifications. The explainable power and the performances of our proposal are evaluated by real data sets from the UCI Machine Learning Repository. The results showed that by using E2Tree we provide a global explanation of the model and provide a powerful predictive tool. Our further aim is to extend this proposal to a regression framework and generalize the approach to other ensemble-tree methodologies.

t	n_t	Class	R_t	$\widetilde{d}_t$	s^*	Node Type
1	537	0	0.6629	0.9578	BMI $\leq$ 27.8	Non-terminal
2	154	0	0.9091	0.8377	Age $\leq$ 28	Non-terminal
3	383	0	0.5640	0.9652	Glucose $\leq$ 113	Non-terminal
4	93	0	0.9677	0.6978	-	Terminal
5	61	0	0.8197	0.8761	Glucose $\leq$ 121	Non-terminal
6	154	0	0.8182	0.9071	Age $\leq$ 27	Non-terminal
7	229	1	0.6070	0.9644	Age $\leq$ 28	Non-terminal
10	30	0	0.9000	0.7850	Glucose $\leq$ 106	Non-terminal
11	31	0	0.7419	0.8727	Glucose $\leq$ 150	Non-terminal
12	71	0	0.8873	0.7976	BMI $\leq$ 31.6	Non-terminal
13	83	0	0.7590	0.9085	DiabetesPedigreeFunction $\leq$ 0.391	Non-terminal
14	87	0	0.5632	0.9407	Glucose $\leq$ 148	Non-terminal
15	142	1	0.7113	0.9507	Insulin $\leq$ 125	Non-terminal
20	15	0	0.9333	0.6882	-	Terminal
21	15	0	0.8667	0.7811	-	Terminal
22	23	0	0.8261	0.8667	-	Terminal
23	8	0	0.5000	0.7667	-	Terminal
24	27	0	0.9259	0.7232	-	Terminal
25	44	0	0.8636	0.7943	Insulin $\leq$ 18	Non-terminal
26	44	0	0.9091	0.8505	SkinThickness $\leq$ 28	Non-terminal
27	39	0	0.5897	0.9163	BloodPressure $\leq$ 0	Non-terminal
28	62	0	0.6935	0.9303	Insulin $\leq$ 130	Non-terminal
29	25	1	0.7600	0.8705	-	Terminal
30	81	1	0.6667	0.9449	Glucose $\leq$ 152	Non-terminal
31	61	1	0.7705	0.9016	Glucose $\leq$ 158	Non-terminal
50	15	0	0.8667	0.7955	-	Terminal
51	29	0	0.8621	0.7462	SkinThickness $\leq$ 34	Non-terminal
52	21	0	0.8571	0.7897	-	Terminal
53	23	0	0.9565	0.8335	-	Terminal
54	3	0	0.6667	0.4211	-	Terminal
55	36	0	0.5833	0.9159	Insulin $\leq$ 119	Non-terminal
56	40	0	0.6750	0.9170	Glucose $\leq$ 125	Non-terminal
57	22	0	0.7273	0.8825	-	Terminal
60	55	1	0.6000	0.9401	Glucose $\leq$ 124	Non-terminal
61	26	1	0.8077	0.8793	-	Terminal
62	39	1	0.7179	0.9087	Pregnancies $\leq$ 6	Non-terminal
63	22	1	0.8636	0.7889	-	Terminal
102	20	0	0.9000	0.6774	-	Terminal
103	9	0	0.7778	0.7553	-	Terminal
110	30	0	0.6333	0.9049	-	Terminal
111	6	1	0.6667	0.8029	-	Terminal
112	23	0	0.7826	0.8746	-	Terminal
113	17	0	0.5294	0.8984	-	Terminal
120	19	1	0.5263	0.8968	-	Terminal
121	36	1	0.6389	0.9286	-	Terminal
124	23	1	0.6522	0.9067	-	Terminal
125	16	1	0.8125	0.8237	-	Terminal

Table 3.5. Diabetes data set: results of the E2Tree.

t	Class	Path
4	0	BMI $\leq$ 27.8 & Age $\leq$ 28
20	0	BMI $\leq$ 27.8 & Age $>$ 28 & Glucose $\leq$ 121 & Glucose $\leq$ 106
21	0	BMI $\leq$ 27.8 & Age $>$ 28 & Glucose $\leq$ 121 & Glucose $>$ 106
22	0	BMI $\leq$ 27.8 & Age $>$ 28 & Glucose $>$ 121 & Glucose $\leq$ 150
23	0	BMI $\leq$ 27.8 & Age $>$ 28 & Glucose $>$ 121 & Glucose $>$ 150
24	0	BMI $>$ 27.8 & Glucose $\leq$ 113 & Age $\leq$ 27 & BMI $\leq$ 31.6
29	1	BMI $>$ 27.8 & Glucose $>$ 113 & Age $\leq$ 28 & Glucose $>$ 148
50	0	BMI $>$ 27.8 & Glucose $\leq$ 113 & Age $\leq$ 27 & BMI $>$ 31.6 & Insulin $\leq$ 18
52	0	BMI $>$ 27.8 & Glucose $\leq$ 113 & Age $>$ 27 & DiabetesPedigreeFunction $\leq$ 0.391 & SkinThickness $\leq$ 28
53	0	BMI $>$ 27.8 & Glucose $\leq$ 113 & Age $>$ 27 & DiabetesPedigreeFunction $\leq$ 0.391 & SkinThickness $>$ 28
54	0	BMI $>$ 27.8 & Glucose $\leq$ 113 & Age $>$ 27 & DiabetesPedigreeFunction $>$ 0.391 & BloodPressure $\leq$ 0
57	0	BMI $>$ 27.8 & Glucose $>$ 113 & Age $\leq$ 28 & Glucose $\leq$ 148 & Insulin $>$ 130
61	1	BMI $>$ 27.8 & Glucose $>$ 113 & Age $>$ 28 & Insulin $\leq$ 125 & Glucose $>$ 152
63	1	BMI $>$ 27.8 & Glucose $>$ 113 & Age $>$ 28 & Insulin $>$ 125 & Glucose $>$ 158
102	0	BMI $>$ 27.8 & Glucose $\leq$ 113 & Age $\leq$ 27 & BMI $>$ 31.6 & Insulin $>$ 18 & SkinThickness $\leq$ 34
103	0	BMI $>$ 27.8 & Glucose $\leq$ 113 & Age $\leq$ 27 & BMI $>$ 31.6 & Insulin $>$ 18 & SkinThickness $>$ 34
110	0	BMI $>$ 27.8 & Glucose $\leq$ 113 & Age $>$ 27 & DiabetesPedigreeFunction $>$ 0.391 & BloodPressure $>$ 0 & Insulin $\leq$ 119
111	1	BMI $>$ 27.8 & Glucose $\leq$ 113 & Age $>$ 27 & DiabetesPedigreeFunction $>$ 0.391 & BloodPressure $>$ 0 & Insulin $>$ 119
112	0	BMI $>$ 27.8 & Glucose $>$ 113 & Age $\leq$ 28 & Glucose $\leq$ 148 & Insulin $\leq$ 130 & Glucose $\leq$ 125
113	0	BMI $>$ 27.8 & Glucose $>$ 113 & Age $\leq$ 28 & Glucose $\leq$ 148 & Insulin $\leq$ 130 & Glucose $>$ 125
120	1	BMI $>$ 27.8 & Glucose $>$ 113 & Age $>$ 28 & Insulin $\leq$ 125 & Glucose $\leq$ 152 & Glucose $\leq$ 124
121	1	BMI $>$ 27.8 & Glucose $>$ 113 & Age $>$ 28 & Insulin $\leq$ 125 & Glucose $\leq$ 152 & Glucose $>$ 124
124	1	BMI $>$ 27.8 & Glucose $>$ 113 & Age $>$ 28 & Insulin $>$ 125 & Glucose $\leq$ 158 & Pregnancies $\leq$ 6
125	1	BMI $>$ 27.8 & Glucose $>$ 113 & Age $>$ 28 & Insulin $>$ 125 & Glucose $\leq$ 158 & Pregnancies $>$ 6

Table 3.6. Diabetes data set: path of the E2Tree.

Chapter 4

Explaining depression risk in Italy with E2Tree

Machine Learning techniques are celebrated for their predictive accuracy, uncovering subtle patterns beyond human perception. Among these, Random Forest is a widely adopted ensemble method, valued for its robust performance and ease of use, especially in scenarios where the cost of errors is significant. However, the inherent opacity of such models raises concerns about their interpretability and trustworthiness. In this study, we apply the Random Forest algorithm to analyze data from the Italian National Institute of Statistics (Istituto Nazionale di Statistica, Istat), specifically the European Health Interview Survey (EHIS), to identify risk factors associated with depression in Italy. To enhance transparency and interpretability, we subsequently employ the Explainable Ensemble Trees methodology to explore and understand the decision-making processes of the Random Forest model. The goal is to demonstrate how E2Tree can provide actionable insights for targeted interventions, supporting public health efforts in addressing mental health challenges.[a]
Keywords: *Explainability, Interpretability, Machine Learning, EHIS, Mental Health.*

[a]This chapter has been published as: Aria, M., Gnasso, A., Rivieccio, R., & Siciliano, R. (2025). Predicting depression in Italy using random forest through the E2Tree methodology. Annals of Operations Research, 1-18.

In this chapter, the E2Tree methodology is applied to a case study of significant social relevance: the prediction of depression in the Italian population using microdata from the European Health Interview Survey (EHIS), conducted by the Italian National Institute of Statistics (ISTAT). Mental health, and depression in particular, has become a crucial issue in contemporary societies, with profound consequences at both the individual and collective level. Identifying the main risk factors associated with depression is therefore essential for designing effective public health policies. Traditional statistical approaches often struggle to capture the complex, non-linear relationships underlying this phenomenon, which makes ML methods particularly suitable. In the following sections, RF is first employed to model the relationship between socio-demographic, health, and lifestyle variables and the likelihood of depression. Building on this, the E2Tree methodology is used to reconstruct the ensemble's decision structure into a transparent and interpretable form. This allows not only for accurate predictions but also for a clear understanding of how different factors interact in shaping the risk of depression, thus providing insights that are directly relevant for policy and practice.

4.1 Depression prediction among Italians

Typically, ML models are applied to prevent some risks in several policy domains like health, education, crime, and finance and this is called an early warning system (EWS) [63]. Explanation approaches can be implemented to recognize the risk factors to intervene. This section describes the dataset and presents the results of the RF model, which are further understood through the E2Tree explanation method.

4.1.1 Data description

The European Health Interview Survey (EHIS)[1] was implemented across all EU Member States to facilitate a comparative analysis of health conditions and healthcare utilization. The survey's topics covered three main areas: health status, determinants of health, and use of healthcare ser-

[1]For further details: https://www.istat.it/microdati/european-health-interview-survey-ehis-public-use-micro-stat-files/

vices. These are investigated alongside the socio-demographic context of each individual in the interviewed households. The results are highly relevant to society, informing national healthcare planning and contributing to European policy development. The survey involves approximately 30,000 families from around 840 Italian municipalities of different sizes. The term "family" refers to a *de facto family*, a group of people cohabiting and connected by ties of marriage, kinship, affinity, adoption, guardianship, or affection. The sampling design is a two-stage stratified design with municipalities as the primary sampling units. The second-stage units are households, selected through a random sampling procedure from the population register for municipalities with less than 1000 inhabitants, and from the list of households selected for the 2018 Permanent Census for municipalities with 1000 inhabitants or more, to ensure a statistically representative sample. Data for all individuals in the sample households is collected through two methods: face-to-face interviews using paper-and-pencil questionnaires (PAPI) conducted by interviewers, and self-administered questionnaires completed by the respondents. The most recent EHIS Italian public health dataset, sourced from the National Institute of Statistics (Istat), was employed to predict depression. The dependent variable measures the presence of depression in individuals within the 12 months preceding the data collection. The 22 selected variables, shown in Table 4.1, encompass a range of socio-demographic, educational, familial, and health-related factors. Specifically, to account for the potential impact of chronic illnesses on the phenomenon of interest, a variable was included to identify individuals who had experienced at least one long-term illness listed in the questionnaire within the past year. The dataset captures personal perceptions of economic status, housing conditions, and social support. It also assesses various psychological symptoms, including apathy, fatigue, attention issues, eating disorders, and self-esteem problems. To improve output visualization, ordinal variables, which represent characteristics with a natural order, are coded numerically. This application domain is characterized by two key challenges: class imbalance in the response variable and the difficulty of accurately detecting depression using the available predictors. Considering units with complete responses to the questionnaire, which represent 92%, a stratified sample was drawn from the original dataset in order to address class imbalance and computational constraints. All

depressed individuals were included, along with a random sample of 7000 non-depressed observations. Thus, the final dataset includes 9463 individuals, 26% of which are depressed. The decision to analyze only the available information is motivated by the context in which the data were collected. Since the questionnaire captures personal perceptions of the respondents, imputing missing answers based on observed responses could introduce bias and distort the underlying patterns. Additionally, the Istat survey techniques manual [64] outlines a scenario where, during an interview, information about the respondent may be provided by another family member. This can occur for sensitive questions, like those related to emotional state, or when the respondent is absent throughout the interview. The interviewer may decide not to accept such responses, which are then labeled as proxies. While accepting these answers may affect the reliability of the collected data, on the other hand, it reduces the costs and time of data collection, also avoiding a sampling error due to the failure to interview a portion of the designated units. Consequently, this study includes feedback accepted by the interviewer even when they were not given directly by the respondent, assuming that close units possess reliable information about the subject. This is especially true in the context of depression, where individuals may struggle to discuss certain sensitive topics.

4.1.2 Analysis

The application was fully implemented in R enviroment, employing the *randomForest* package for model building and the *e2tree* package for model explanation. The RF model was constructed by dividing the dataset into a training set and a testing set. In line with the methodology proposed by Guyon [65], 70% of the dataset was used for training, while the remaining 30% was set aside for testing. The RF model was configured with 500 trees, and the number of variables randomly selected as candidates at each split was set to $\sqrt{p} = 5$, where p represents the total number of variables. Table 4.2 displays the confusion matrix for the test set. Balanced Accuracy, F1-Score, and Cohen's Kappa were selected as performance metrics due to their robustness in handling imbalanced datasets. Unlike Accuracy, these metrics provide a more reliable assessment of the model's predictive capabilities, particularly when the classes are not equally distributed [59].

Variable Name	Variable Encoding/Range	Variable Type
SEX	Female; Male	Categorical
AGE	1 = 15–17; 2 = 18–24; 3 = 25–34; 4 = 35–44; 5 = 45–49; 6 = 50–54; 7 = 55–59; 8 = 60–64; 9 = 65–69; 10 = 70–74; 11 = $\geq$ 75	Ordinal
GEO	C = Center; I = Island; N-E = Northeast; N-O = Northwest; S = South	Categorical
JOB	Worker; Unemployed; Retired; Other	Categorical
MARITAL.STATUS	Single; Married; Widowed; Divorced	Categorical
INCOME	1 = I; 2 = II; 3 = III; 4 = IV; 5 = V	Ordinal
EDU	1 = Elementary; 2 = Middle School; 3 = Diploma; 4 = Post-secondary and Tertiary	Ordinal
FAM.MEMBERS	1–7	Numerical
FAMILY.TYPE	Single Person; Single Parent; Couple with Children; Couple without Children; Other	Categorical
APATHY	1–4	Ordinal
SLEEPING.DIS	1–4	Ordinal
TIRED	1–4	Ordinal
EATING.DIS	1–4	Ordinal
SELF.EST.DIS	1–4	Ordinal
ATTENTION.DIS	1–4	Ordinal
SPORT	0–7	Numerical
BODY.MASS	1 = Underweight; 2 = Normal Weight; 3 = Overweight; 4 = Obese	Ordinal
DISEASE	Yes; No	Categorical
CARE	1–5	Ordinal
SMALL.HOUSE	Yes; No	Categorical
BAD.HOUSE	Yes; No	Categorical
ECO.STATUS	1 = Very Insufficient; 2 = Limited; 3 = Good; 4 = Excellent	Ordinal

Table 4.1. Description and encoding of variables.

Table 4.3 presents the evaluation metrics derived from the test set confusion matrix. Balanced Accuracy, which is the average of *sensitivity* (true depressed rate) and *specificity* (true non-depressed rate), is 0.776, indicating a reasonable balance between the model's ability to identify both depressed and non-depressed cases correctly. F1-Score index is the harmonic mean between specificity and *precision* (proportion of depressed predic-

Observed	Predicted	
	Yes	**No**
Yes	472	265
No	187	1915

Table 4.2. Test set confusion matrix.

tions that are correctly classified). In this implementation, it assumes a moderate value of 0.676. Cohen's Kappa, a measure for chance agreement, is 0.571, suggesting discrete concordance between the predicted and actual classifications. Despite the class imbalance in the response variable, the reported metrics show satisfactory performance in predicting the presence of depression in individuals. However, the performance of the predictive model is further influenced by the discriminative power of the selected predictors in accurately identifying individuals' mental health status. To investigate the role of the independent variables in the decision-making process, the E2Tree explanation model is implemented and presented in the following section.

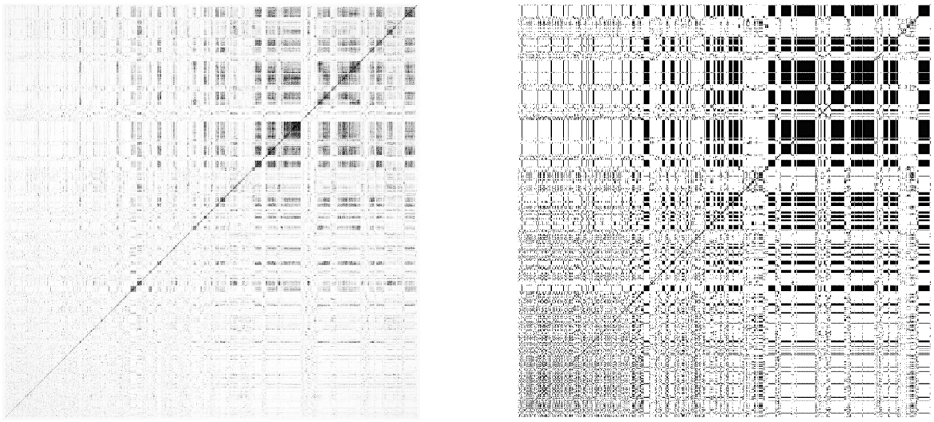
Metrics	Value (95% CI)
Accuracy	0.841 (0.827, 0.854)
Sensitivity	0.640 (0.604, 0.676)
Specificity	0.911 (0.898, 0.923)
Precision	0.716 (0.680, 0.751)
Balanced Accuracy	**0.776** (0.756, 0.793)
F1-Score	**0.676** (0.645, 0.705)
Kappa	**0.571** (0.533, 0.606)

Table 4.3. Evaluation metrics.

4.1.3 Results

The generation of the E2Tree begins with the construction of the matrix O_{ij}, as defined in Equation 3.2, derived from the outputs of the RF

algorithm. The heatmap in *panel a* of the Figure 4.1 visually represents O_{ij}, summarizing the frequency with which pairs of observations co-occur within the nodes of the RF. This provides an insightful depiction of the relationships identified by the RF algorithm during its decision-making process. Similarly, in the *panel b* of the Figure 4.1 illustrates the heatmap of the matrix $\hat{O}_{ij}$, estimated by the E2Tree methodology. This visualization highlights the co-occurrence frequencies of observation pairs in the nodes of the E2Tree structure, effectively demonstrating its ability to replicate and refine the relational patterns captured by the RF algorithm. A side-by-side comparison of *panel a*, depicting the RF-derived O_{ij}, and *panel b*, representing the E2Tree-estimated $\hat{O}_{ij}$, reveals how E2Tree preserves the relational structure of RF. Beyond this, it significantly enhances interpretability and explainability, providing a transparent and comprehensible framework to approximate the internal mechanisms of the RF model. Hav-



(a) (b)

Figure 4.1. Heatmap of the matrix O_{ij} (a) and heatmap of the matrix $\hat{O}_{ij}$ estimated by E2Tree (b). The higher the values in the matrix cells, the darker the colour of the heatmap cells.

ing established E2Tree's capacity to reconstruct relationships within the matrices, attention can now be directed to its primary output, a single decision tree. This decision tree, presented in Figure 4.3, offers a simplified yet comprehensive representation of the RF s decision-making process.

The single decision tree produced by E2Tree is not intended for predictive purposes but it has to be used as an interpretative tool to understand the inner workings of the RF model. Using the complexity of an RF ensemble into a single tree structure, E2Tree enables both interpretability and explainability, allowing users to trace the model's decisions with clarity and ease. The decision tree itself encapsulates the hierarchy of variables and the conditions under which RF model makes predictions. Each node represents a splitting criterion based on specific variables, while the terminal nodes (leaves) indicate the final classification outcomes along with their associated probabilities. This structure not only preserves the predictive insights of the RF but also renders the decision process transparent and readily understandable, bridging the gap between model performance and interoperability. To further validate the alignment between the RF and E2Tree methodologies, a Mantel test [66] was conducted to assess the correlation between the pairwise similarity matrices generated by both approaches. The results, depicted in the density plot in Figure 4.2, visually convey the statistical significance of the observed correlation. The Mantel test yielded a p-value of 0.001, confirming a statistically significant correlation and allowing the null hypothesis of no correlation to be rejected with 99.9% confidence. Furthermore, the high z-statistic of 118316.3 reflects the strong relationship between the two matrices. The alternative two-sided hypothesis further confirms the strength of this correlation. These findings underscore the ability of E2Tree to accurately approximate the structure and relationships encoded in the RF model. At the same time, it introduces much-needed interpretability and transparency, validating E2Tree's utility as a powerful tool for enhancing the explainability of ensemble methods without compromising predictive fidelity. Figure 4.3 visually represents the explanation tree. Instead, Table 4.4 and Table 4.5 is the single table split into two to illustrate the E2Tree and provide insight into the path leading to each terminal node. The initial split in the tree partitions the data according to the presence or absence of chronic diseases, excluding depression. Individuals without long-term illnesses are classified as non-depressed. Individuals who have experienced one or more chronic illnesses follow a more complex decision path, considering additional factors and interactions between variables to determine the presence of depression. The model proceeds to create two distinct subtrees based on the type of fam-

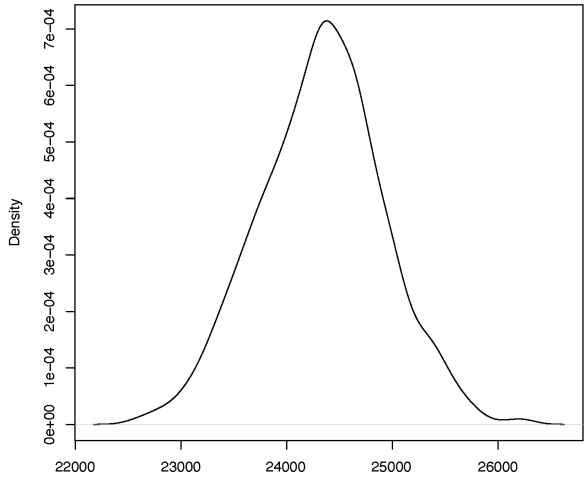


Figure 4.2. Density plot of Mantel test permutations ($N = 999$, $Bandwidth = 140$).

ily. One subtree includes lonely families, such as couples without children, single parents, and single individuals. The remaining subtree considers observations belonging to households with children under 25 years of age. Subsequently, the decision-making process primarily relies on the occurrence of symptoms such as fatigue, apathy, and self-esteem disorders. For families with children, the presence of depression is linked only to individuals who feel tired, with a lack of interest in life activities, and have negative self-perception. For the units belonging to families of different structures, even those without self-esteem issues are labeled as depressed if they display disinterest in life. Then, the explanation tree is predominantly composed of leaf nodes that classify individuals as non-depressed. Based on the decision-making process, workers with chronic illnesses, belonging to single-parent households, and who never feel tired, have not experienced depression. Therefore, according to E2Tree, the RF model predictive performance is strictly connected with the presence of long-term illnesses and the sensation of apathy, tiredness, and low self-worth, proving that the

decision-making process is reasonable for these paths.

4.2 Concluding remarks

This study employs the E2Tree methodology to enhance the interpretability and explainability of RF models, focusing on predicting depression in Italy. By synthesizing the RF model into a tree-like structure, the E2Tree methodology provides a framework that is both interpretable and explainable, thus facilitating the understanding of the relationships captured by the model. The analysis yielded several significant findings. The comparison of the heatmaps revealed a high degree of similarity between the RF model and E2Tree, as evidenced by a statistically significant Mantel test result. This suggests that E2Tree has successfully captured the relational structure of the RF model. Regarding the provided insights, the primary factor associated with the experience of depression in Italy is strongly linked to the presence of other chronic illnesses, highlighting a potential public health concern. This may suggest that the current healthcare system is not sufficiently effective in ensuring the mental well-being of the Italian population. Family structure also plays a crucial role in individuals' psychological well-being. Lonely families are particularly at risk of experiencing depression. Moreover, the presence of this mental state manifests through symptoms triggered by difficulties in coping with daily life. Indeed, fatigue, low self-esteem, and apathy emerge as key contributing factors. In light of these findings, the Italian government should implement policies aimed at improving both the treatment of chronic illnesses and the psychological support available to affected individuals. For instance, tailored programs could help patients manage the physical and emotional burden of chronic conditions, including the establishment of dedicated support groups to foster connection, reduce feelings of loneliness and abandonment, and help individuals regain a sense of vitality and connection to life. Future research could aim to improve the performance of the predictive approach by first designing a more targeted questionnaire focused explicitly on individuals' psychological well-being. Such an instrument should be capable of capturing depression in all its facets. Additionally, in addressing the natural imbalance between response categories, oversampling techniques could be implemented that take into account both the nature of

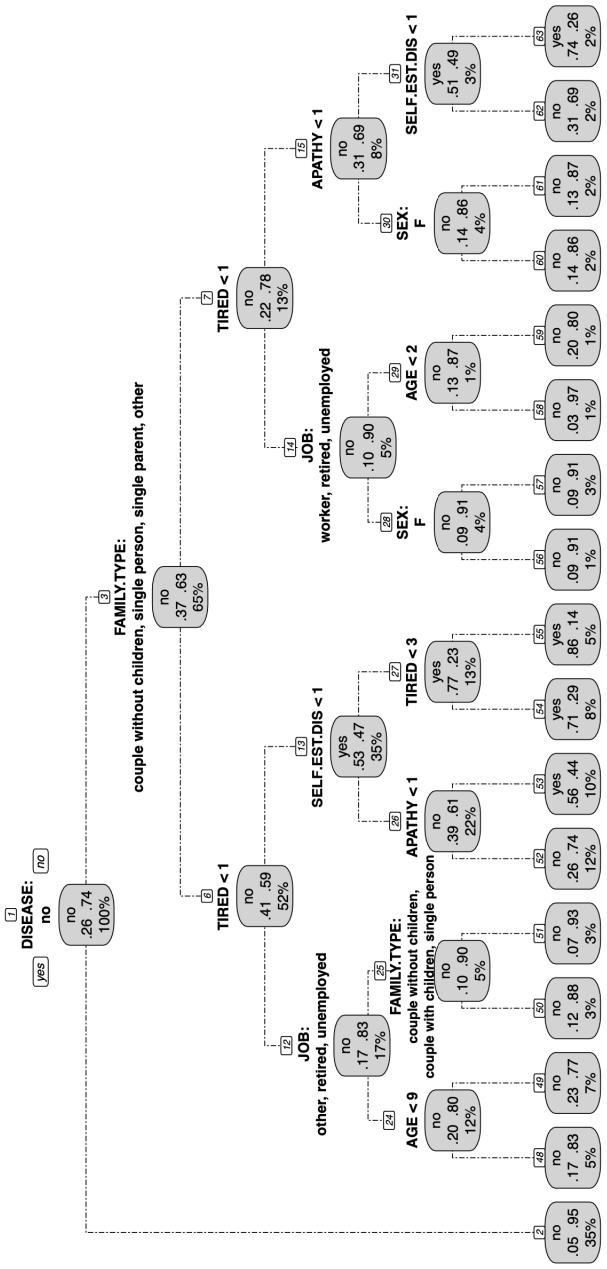


Figure 4.3. E2Tree output. Each node displays the proportion of observations belonging to each class of the target variable, as well as the percentage of the total dataset represented by the node.

t	n_t	$Class$	R_t	s^*	$NodeType$	$Path$
1	6624	no	0.74	DISEASE = no	Non-Terminal	
2	2299	no	0.95	-	Terminal	DISEASE= no
3	4325	no	0.63	FAMILY.TYPE = (couple without children, other, single parent, single person	Non-Terminal	DISEASE = yes
6	3470	no	0.59	TIRED $\leq$ 1	Non-Terminal	DISEASE = yes & FAMILY.TYPE = (couple without children, other, single parent, single person)
12	1140	no	0.83	JOB = other, retired, unemployed)	Non-Terminal	DISEASE = yes & FAMILY.TYPE = (couple without children, other, single parent, single person) & TIRED $\leq$ 1
24	786	no	0.80	AGE $\leq$ 9	Non-Terminal	DISEASE = yes & FAMILY.TYPE = (couple without children, other, single parent, single person) & TIRED $\leq$ 1 & JOB = (other, retired, unemployed)
48	355	no	0.83	-	Terminal	DISEASE = yes & FAMILY.TYPE = (couple without children, other, single parent, single person) & TIRED $\leq$ 1 & JOB = (other, retired, unemployed) & AGE $\leq$ 9
49	431	no	0.77	-	Terminal	DISEASE = yes & FAMILY.TYPE = (couple without children, other, single parent, single person) & TIRED $\leq$ 1 & JOB = (other, retired, unemployed) & AGE $\leq$ 9
25	354	no	0.90	FAMILY.TYPE = (couple with children, couple without children, single person)	Non-Terminal	DISEASE = yes & FAMILY.TYPE = (couple without children, other, single parent, single person) & TIRED $\leq$ 1 & JOB = worker

Table 4.4. Results of E2Tree (a).

t	n_t	$Class$	R_t	s^*	$NodeType$	$Path$
50	184	no	0.88	-	Terminal	DISEASE = yes & FAMILY.TYPE = (couple without children, other, single parent, single person) & TIRED $\leq$ 1 & JOB = worker & FAMILY.TYPE = (couple with children, couple without children, single person)
51	170	no	0.93	-	Terminal	DISEASE = yes & FAMILY.TYPE = (couple without children, other, single parent, single person) & TIRED $\leq$ 1 & JOB = worker & FAMILY.TYPE = (single parent, other)
13	2330	yes	0.53	SELF.EST.DIS $\leq$ 1	Non-Terminal	DISEASE = yes & FAMILY.TYPE = (couple without children, other, single parent, single person) & TIRED $\geq$ 1
26	1485	no	0.61	APATHY $\leq$ 1	Non-Terminal	DISEASE = yes & FAMILY.TYPE = (couple without children, other, single parent, single person) & TIRED $\geq$ 1 & SELF.EST.DIS $\leq$ 1
52	826	no	0.74	-	Terminal	DISEASE = yes & FAMILY.TYPE = (couple without children, other, single parent, single person) & TIRED $\geq$ 1 & SELF.EST.DIS $\leq$ 1 & APATHY $\leq$ 1
53	659	yes	0.56	-	Terminal	DISEASE = yes & FAMILY.TYPE = (couple without children, other, single parent, single person) & TIRED $\geq$ 1 & SELF.EST.DIS $\leq$ 1 & APATHY $\geq$ 1
27	845	yes	0.77	TIRED $\leq$ 3	Non-Terminal	DISEASE = yes & FAMILY.TYPE = (couple without children, other, single parent, single person) & TIRED $\geq$ 1 & SELF.EST.DIS $\geq$ 1

Table 4.5. Results of E2Tree (b).

t	n_t	$Class$	R_t	s^*	$NodeType$	$Path$
54	511	yes	0.71	-	Terminal	DISEASE = yes & FAMILY.TYPE = (couple without children, other, single parent, single person) & TIRED > 1 & SELF.EST.DIS > 1 & TIRED *leq* 3
55	334	yes	0.86	-	Terminal	DISEASE = yes & FAMILY.TYPE = (couple without children, other, single parent, single person) & TIRED $\geq$ 1 & SELF.EST.DIS $\geq$ 1 & TIRED > 3
7	855	no	0.78	TIRED <=1	Non-Terminal	DISEASE = yes & FAMILY.TYPE = couple with children
14	357	no	0.90	JOB = (retired, unemployed , worker)	Non-Terminal	DISEASE = yes & FAMILY.TYPE = couple with children & TIRED $\leq$ 1
28	271	no	0.91	SEX = F	Non-Terminal	DISEASE = yes & FAMILY.TYPE = couple with children & TIRED $\leq$ 1 & JOB = (retired, unemployed, worker)
56	76	no	0.91	-	Terminal	DISEASE = yes & FAMILY.TYPE = couple with children & TIRED <=1 & JOB = (retired, unemployed, worker) & SEX = F
57	195	no	0.91	-	Terminal	DISEASE = yes & FAMILY.TYPE = couple with children & TIRED $\leq$ 1 & JOB = (retired, unemployed, worker) & SEX = M
29	86	no	0.87	AGE <=2	Non-Terminal	DISEASE = yes & FAMILY.TYPE = couple with children & TIRED $\leq$ 1 & JOB = other

Table 4.6. Results of E2Tree (c).

t	n_t	$Class$	R_t	s^*	$NodeType$	$Path$
58	36	no	0.97	-	Terminal	DISEASE = yes & FAMILY.TYPE = couple with children & TIRED <=1 & JOB = other & AGE $\leq$ 2
59	50	no	0.80	-	Terminal	DISEASE = yes & FAMILY.TYPE = couple with children & TIRED $\leq$ 1 & JOB = other & AGE $>$ 2
15	498	no	0.69	APATHY $\leq$ 1	Non-Terminal	DISEASE = yes & FAMILY.TYPE = couple with children & TIRED $>$ 1
30	272	no	0.86	SEX = F	Non-Terminal	DISEASE = yes & FAMILY.TYPE = couple with children & TIRED $\geq$ 1 & APATHY $\leq$ 1
60	147	no	0.86	-	Terminal	DISEASE = yes & FAMILY.TYPE = (couple with children, other) & TIRED $\geq$ 1 & APATHY $\leq$ 1 & SEX = F
61	125	no	0.87	-	Terminal	DISEASE = yes & FAMILY.TYPE = (couple with children, other) & TIRED $\geq$ 1 & APATHY $<$ 1 & SEX = M
31	226	yes	0.51	SELF.EST.DIS $\leq$ 1	Non-Terminal	DISEASE = yes & FAMILY.TYPE = couple with children & TIRED $>$ 1 & APATHY $\geq$ 1
62	119	no	0.69	-	Terminal	DISEASE = yes & FAMILY.TYPE = couple with children & TIRED $\geq$ 1 & APATHY $\geq$ 1 & SELF.EST.DIS $\leq$ 1
63	107	yes	0.74	-	Terminal	DISEASE = yes & FAMILY.TYPE = couple with children & TIRED $\geq$ 1 & APATHY $\geq$ 1 & SELF.EST.DIS $\geq$ 1

Table 4.7. Results of E2Tree (d).

the variables included in the questionnaire and the need to generate realistic observations. Moreover, future studies could further refine the analysis by exploring the relationship between depression and various chronic diseases, in order to more precisely identify the areas in which the healthcare system may be inadequate and where governmental intervention is most urgently required.

Chapter 5

Extending E2Tree to regression contexts

The advent of ensemble methods, such as Random Forest, has led to a paradigm shift in supervised learning. These methods have achieve remarkable levels of prediction accuracy by aggregating multiple weak learners. However, a drawback of these methods is their lack of transparency, which often prevents users from understanding their prediction processes. In light of these challenges, Explainable Ensemble Trees (E2Tree) has recently been proposed, that provides a graphical representation of the relationships between response variables and predictors in Random Forests for classification. E2Tree constructs a single decision tree based on (dis)similarities between observations. By summarizing both distances in terms of predictors, and a forest as a single decision tree, E2Tree merges the strengths of both decision trees and decision tree ensembles. In this paper, we propose to extend the E2Tree methodology to regression contexts. We investigate the performance of E2Tree for regression using real-world datasets. We use the Mantel test to test the correlation between similarities of the Random Forest and E2Tree.[a]
Keywords: *E2Tree, Machine Learning, Random Forest, Explainability, Interpretability*

[a]This chapter has been accepted for publication as: Aria, M., Gnasso, A., Iorio, C., & Fokkema, M. (2025). Extending Explainable Ensemble Trees to regression contexts. Applied Stochastic Models in Business and Industry.

As discussed in Chapter 3, the E2Tree methodology provides a transparent reconstruction of RF models in the form of a single transparent and interpretable tree structure. While the previous chapters introduced the method and illustrated its application to a classification problem of public health relevance, many real-world applications involve continuous response variables, where the objective is to explain variations in magnitude rather than discrete classes. This chapter extends the E2Tree framework to regression contexts. The key idea remains the reconstruction of a single, explainable and interpretable tree structure that summarizes the logic of a RF ensemble. Nonetheless, specific adaptations are required to account for the continuous nature of the target variable. In particular, a new definition of local goodness-of-fit is introduced to evaluate node quality, together with tailored stopping rules that ensure meaningful splits and avoid overfitting. The proposed extension demonstrates how E2Tree can maintain the predictive accuracy of RF while enhancing transparency in regression problems. By providing a graphical and interpretable representation of complex models, it broadens the scope of E2Tree and reinforces its role as a versatile tool for explainable ML.

5.1 E2Tree in the regression context

The proposed extension of E2Tree to regression contexts introduces significant advantages over existing explainable ML techniques. By utilizing a dissimilarity matrix, E2Tree quantifies and visualizes interactions among predictors, a feature absent in many standard methods. Furthermore, the methodology generates a single tree-like structure that encapsulates the ensemble model's logic, ensuring interpretability without compromising predictive accuracy. This approach bridges the gap between user comprehension and model complexity, providing an intuitive framework for understanding regression models. In both classification and regression, the dissimilarity matrix is computed from the frequency with which two observations fall into the same terminal nodes across the RF ensemble. However, how this co-occurrence is weighted differs between the classification and the regression setting.
The matrix D (3.1) can be interpreted as a fuzzy partition of the data obtained by the RF procedure. In this context, d_{ij} is a weighted mem-

bership value of observations i and j belonging to the same group. This results in a novel representation of the RF model, where O_{ij} is a weighted co-occurrence measure between observations u_i and u_j along the sequence $H_1, \ldots, H_B$. Hence, the values of matrix O are a weighted measure of how often a pair of observations fall into the same node in an RF. Specifically, we define O_{ij} as:

$$O_{ij} = \frac{\sum_{b=1}^{B} I(u_i \wedge u_j) \cdot W_{(t)_{ij}|b}}{\max(\sum_b u_i W_{(t)_i|b}; \sum_b u_j W_{(t)_j|b})}, \qquad 0 \leq O_{ij} \leq 1 \tag{5.1}$$

where $W_{(t)_{ij}|b}$ is a normalized local goodness-of-fit measure of node t of the b-th learner, such that $0 \leq W_{(t)_{ij}|b} \leq 1$. In the classification context, goodness of fit is given by the correct classification rate [67].
In the regression context, we propose the following local goodness of fit measure:

$$W_{(t)_{ij}|b} = \begin{cases} 0 & \text{if } 1 - NMSE_{(t)_{ij}|b} \leq 0 \\ 1 - NMSE_{(t)_{ij}|b} & \text{otherwise} \end{cases}$$

where $NMSE_{(t)_{ij}|b}$ is the normalized measure of the mean square error at node t:

$$NMSE_{(t)_{ij}|b} = \frac{MSE_{(t)_{ij}|b}}{MSE_{max} \cdot p(t)}, \tag{5.2}$$

where $MSE_{(t)_{ij}|b}$ is the mean square error at node t, $p(t) = n(t)/N$ is the proportion of observations falling into the t-th node, and MSE_{max} is the theoretical maximum of the MSE obtained under the assumption formulated by Jacobson [68]:

$$MSE_{max} = \frac{(\max(Y|b) - \min(Y|b))^2}{9}.$$

In the literature, several upper bounds on variance as a measure of maximum dispersion from the mean have been proposed that also require knowledge about the underlying distributions. Here, variance bounds themselves are not the principal subjects of interest, but we refer the reader to Samuelson [69], Arnold [70] and Seaman et al. [71].

Alternative measures for $W_{(t)_{ij}|b}$ can be based on different approaches to normalize *NMSE* at node t. For example, following the concept of coefficient of variation, an alternative proposal for *NMSE* could be:

$$NMSE_{(t)_{ij}|b} = \frac{MSE_{(t)_{ij}|b}}{\mid \hat{Y}_t \mid}.$$

To construct the tree, E2Tree uses a greedy splitting approach in order to minimize the within-node dissimilarities according to the algorithm in Section 3. In line with the methodology proposed by Aria et al. [67], we also establish a set of stopping criteria that reflect the traditional approaches of tree estimation. In particular, we consider thresholds related to the size and impurity of terminal nodes. It is well-known that RFs tend to include many non-informative splits due to the random selection of split variables. Consequently, it can be expected that with an increasing number of nodes in a tree, the dissimilarity values d_{ij} between two observations (i, j) in a fitted RF will tend to converge to one. To address this issue, our proposed algorithm incorporates two additional stopping rules: first, after identifying the optimal split based on dissimilarity, we verify that the value of $NMSE_{(t)_{ij}|b}$ meets a specified threshold γ. The node t becomes terminal if $NMSE_{(t)_{ij}|b} \leq \gamma$. In our experiments, we set this threshold parameter to $\gamma = 0.05$. This choice reflects a balance between tree complexity and interpretability. The use of error-based thresholds for stopping rules is well established in the literature on regression trees, where similar criteria are commonly employed to avoid overfitting [72]. The value $\gamma = 0.05$ was selected as a conservative cutoff and was empirically validated in preliminary experiments on the Auto MPG and Boston Housing datasets, where it consistently provided stable and interpretable structures without sacrificing predictive accuracy. Second, for each split, the distributions between the child nodes are verified to be identical through the Mann-Whitney test, as proposed by Mann et al. [73]. If the predicted values demonstrate identical distributions in the child nodes after splitting, the branch is removed and the parent node is regarded as a terminal node. Otherwise, splitting proceeds.

While E2Tree is similar as classical single-tree induction methods such as CART or C4.5 because all these methods maximize a measure of within-node purity. However, for CART and C4.5, the measures of within-node

purity are solely based on prediction accuracy. By contrast, E2Tree reconstructs the logic of the RF into a single decision tree by minimizing both RF-based dissimilarities as well as prediction error, as reflected in . In doing so, it retains the predictive fidelity of the underlying ensemble while also providing explainability. In this sense, E2Tree should not be regarded as a competitor of traditional tree inducers, but rather as a complementary tool that bridges the gap between ensemble accuracy and single-tree interpretability.

5.2 Illustrative examples

The example presented in this section are based on two widely used benchmark datasets. First, we employed the Auto MPG dataset [74], available from the UCI Machine Learning Repository, as an illustrative example to demonstrate the functionality of E2Tree. Next, we analyzed the Boston Housing dataset [75], which includes data collected by the U.S. Census Service on housing in the Boston, Massachusetts area. We discuss the results achieved by E2Tree, emphasizing its ability to explain the decision-making process of RF models for regression tasks.
Data analysis was performed using the `R` software and our *e2tree* package, published on CRAN.
The RF models were trained on datasets split into 70% for training and 30% for testing, following the methodology described by Guyon [65]. Each model was configured with 500 trees, evaluating $p/3$ randomly selected features at each node. Additionally, we set the algorithm's threshold parameter to $\gamma = 0.05$. To validate fidelity, we rely on the Mantel test, which statistically demonstrates that the partitions reconstructed by E2Tree are consistent with those produced by the Random Forest ensemble.

5.2.1 A toy example: Auto MPG data set

The Auto MPG dataset contains 398 entries, each representing a unique automobile model. The dataset is designed to predict the fuel efficiency of cars, measured in miles per gallon (MPG), using several predictor variables. These variables include quantitative features (see Table 5.1). The dataset provides a set of attributes, making it suitable for regression analysis and

Feature	Mean	SD	Min	1st Quartile	Median	3rd Quartile	Max
mpg	23.51	7.82	9.00	17.50	23.00	29.00	46.60
cylinders	-	-	3	-	-	-	8
displacement	193.40	104.27	68.0	104.20	148.50	262.00	455.00
horsepower	103.60	38.86	46.0	75.00	92.00	125.00	230.00
weight	2970.00	846.84	1613.0	2224.00	2804.00	3608.00	5140.00
acceleration	15.57	2.76	8.00	13.82	15.50	17.18	24.80
model_year	76.01	3.70	70.00	73.00	76.00	79.00	82.00
origin	-	-	1	-	-	-	3

Table 5.1. Summary statistics of the Auto MPG dataset features.

Type of RF:	regression
Number of trees:	500
No. of variables tried at each split:	2
Mean of squared residuals:	7.95
% Var explained:	86.62

Table 5.2. Auto MPG data: Results of RF regression.

other ML tasks.

Here, we conduct a regression analysis on the MPG variable. We built the RF regression model using the training set, as previously described. The RF results are summarized in Table 5.2. The initial stage in the generation of the E2tree is the construction of O_{ij}, as defined in Equation 5.1, using the outcomes produced by the RF algorithm. The heatmap of the matrix O_{ij} is presented in Figure 5.1. The matrix O_{ij} enables us to quantify the frequency with which pairs of observations co-occur within the same node of an RF, with the result weighted by the normalised generic measure of local goodness of fit, $W_{(t)_{ij}|b}$. The normalised mean square error, $NMSE_{(t)_{ij}|b}$, is computed in accordance with the procedure set out in Equation 5.2. Given that E2Tree is based on the concept of a dissimilarity matrix D, this matrix is calculated next as $1-O_{ij}$ (Section 3.1). Figure 5.2 and Table 5.3 illustrate the E2Tree and provide insight into the path leading to each terminal node, thereby elucidating the predictions made by the RF. E2Tree naturally can handles both numerical or ordinal predictors, as well as ordered-categorical predictors. In addition to accounting for the effects of predictor variables on the response, E2Tree also incorporates

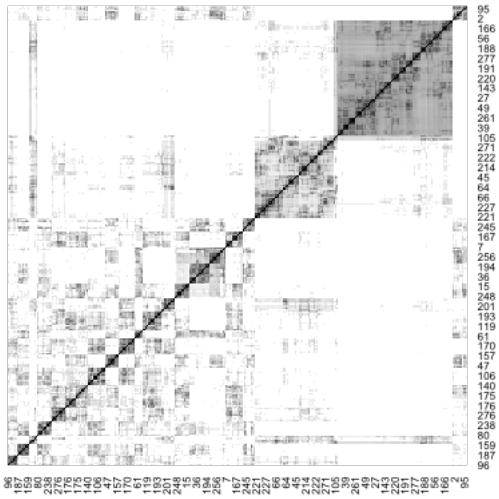


Figure 5.1. Auto MPG data: heatmap of the matrix O_{ij}.

the associations between predictor variables through the computation and utilisation of dissimilarity measures. In order to demonstrate the efficacy of E2Tree in explaining the internal workings of RF, Figure 5.3 presents a comparison between the heatmap of the matrix O_{ij} obtained from the RF output and the heatmap of the matrix $\hat{O}_{ij}$ estimated by E2Tree. We would like to point out that the figure 5.3 clearly shows the heatmap of the O_{ij} matrix (panel a) obtained by RF, which is rather fuzzy. This is due to the heatmap being based on co-occurrences obtained from 500 trees. On the other hand, in figure 5.3, the heatmap obtained by E2Tree (panel b) is more clear-cut, because it is built on the co-occurrences in only a single tree.

To assess whether the matrix $\hat{O}_{ij}$ estimated by E2Tree accurately reflects the structure of the matrix O_{ij} derived from the RF, we conducted a Mantel test [66] to evaluate the correlation between the pairwise similarity matrices produced by the two methods. The findings, illustrated in the density plot in Figure 5.4, highlight the statistical significance of the observed correlation, $z = 190.2009$; $p < .001$ (two-sided). These results demonstrate E2Tree's capability to faithfully reconstruct the structure and relationships encapsulated in the RF model, while simultaneously provid-

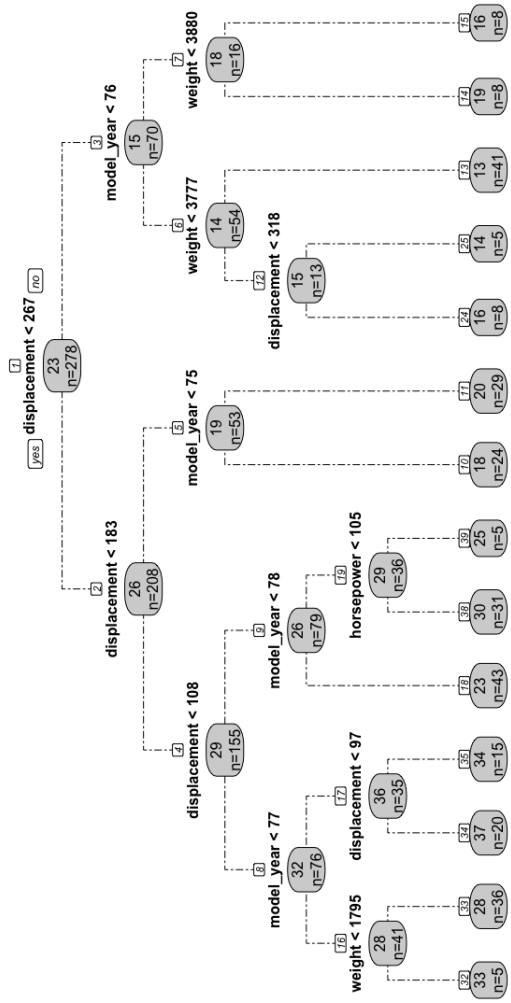
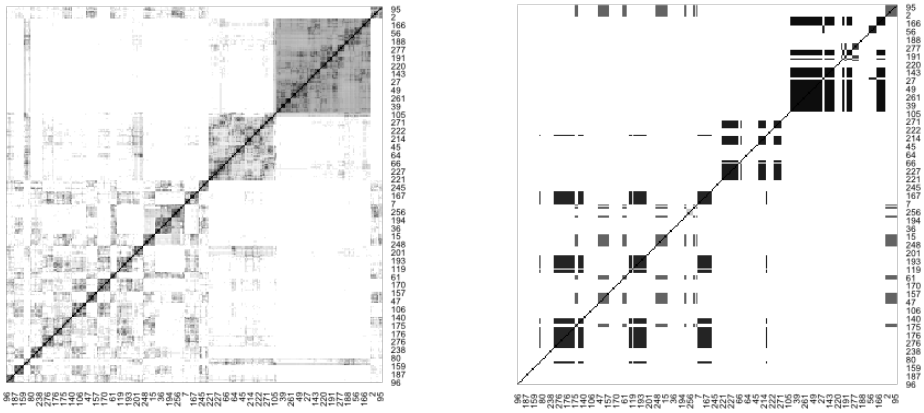


Figure 5.2. Auto MPG data: path visualization of E2Tree. The color intensity indicates the magnitude of the predicted values: darker shades typically correspond to higher predicted values, while lighter shades correspond to lower predicted values.



(a) (b)

Figure 5.3. Auto MPG data: heatmap of the matrix O_{ij} (panel a) and heatmap of the matrix $\hat{O}_{ij}$ estimated by E2Tree (panel b).

ing enhanced interpretability and transparency.

The E2Tree output related to the Auto MPG dataset is presented in Figure 5.2, provides a graphical representation of the E2Tree structure, and Table 5.3 presents the node-specific numerical results for the E2Tree. The results show the comprehensive nature of our proposal, highlighting the entire decision paths and providing crucial information at each node for predictive analysis. Table 5.3 outlines the primary information and specifies the prediction paths made by the RF. One of the significant advantages of the E2Tree model, as depicted in Figure 5.2, is the immediate visualization of the importance of splits. This straightforward visualisation explains the relational structure between the response variable and the predictors, along with their interactions, which cannot be directly obtained from a fitted RF model.

The tree-like structure provides a comprehensive understanding of the determinants of vehicle fuel efficiency, measured in miles per gallon (MPG). The results identify displacement, weight, horsepower and model year as critical predictors of MPG, providing insight into the complex relationships between vehicle attributes and fuel economy. Engine displacement emerges as the most significant variable, with vehicles with larger engines

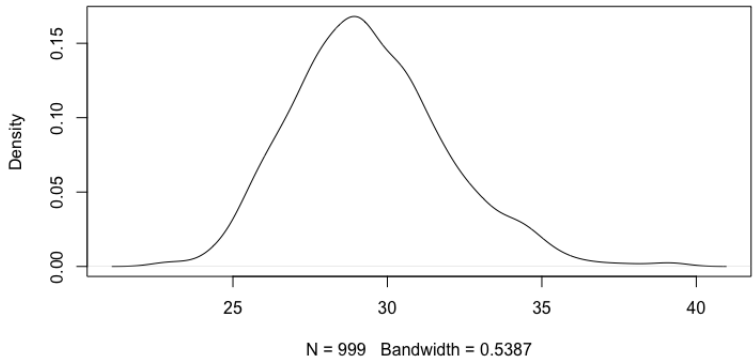


Figure 5.4. Auto MPG data: Density plot of Mantel test permutations, visualizing the frequency of correlations obtained with permuted data, where there is no association.

consistently showing lower fuel economy. This finding reflects the higher fuel consumption associated with larger engines, which are designed to produce more power at the expense of efficiency. Similarly, weight is shown to have a significant negative impact on MPG, as heavier vehicles require more energy to operate, further emphasising the role of mass reduction in improving fuel economy. Model year has a positive effect on MPG, highlighting the importance of technological advances in vehicle design. Newer models benefit from innovations such as more efficient engines, lightweight materials and aerodynamic designs, all of which contribute to better fuel economy. This trend underlines the progress made by the automotive industry in response to growing environmental concerns and regulatory pressures. Horsepower also plays a role in determining MPG, albeit with a more nuanced effect. While higher horsepower generally correlates with reduced efficiency, certain configurations appear to optimise power output without significant losses in fuel economy, suggesting that modern engineering practices can mitigate some of the negative effects traditionally associated with higher horsepower. These results underline the importance of reducing engine displacement and weight, while exploiting technological advances to improve vehicle efficiency. The results highlight the need for continued innovation in lightweight materials and powertrain technologies to further improve fuel economy. Policy-makers can support

these efforts by providing incentives for the adoption of fuel-efficient designs and by maintaining stringent regulatory standards. For consumers, the results suggest that choosing newer, lighter vehicles with optimised engines is a practical approach to reducing fuel consumption and environmental impact. This study could provide critical insights into the interplay between physical and technological factors in determining fuel efficiency, providing valuable guidance for advancing sustainable automotive practices. We want to stress that one of the most important features of the E2Tree model is its ability to provide clear interpretations of the "If-Then" pathways. These pathways show how interactions between predictors culminate in the final prediction at each terminal node. This interpretative capability is provided through both graphical and tabular representations, thus offering a comprehensive understanding of the predictive mechanisms within the model.

5.2.2 Boston Housing data

The Boston Housing dataset is a well-known dataset derived from information collected by the U.S. Census Service to understand housing characteristics in the Boston metropolitan area. It contains 506 records and has been widely used in the ML literature for benchmarking algorithms. Originally published in 1978 by Harrison and Rubinfeld [75] in their study of hedonic pricing and demand for clean air, this dataset remains a cornerstone of predictive modelling and regression tasks. Each entry in the dataset contains 14 attributes describing various socio-economic, geographical and structural factors related to housing. Key variables include crime rate per capita (CRIM), nitrogen oxide concentration (NOX), average number of rooms per dwelling (RM), property tax rates (TAX) and median value of owner-occupied homes (MEDV). The MEDV variable, which represents the median value of owner-occupied dwellings in thousands of dollars, is the target variable for our regression analysis. However, note that MEDV is partially censored at $50,000, with 16 cases reporting this maximum value. Additionally, 15 cases have values between $40,000 and $50,000, rounded to the nearest hundred, suggesting potential censoring in the dataset. Table 5.4 provides a detailed statistical summary of the variables used to construct the RF model for the Boston Housing dataset. Table 5.5 summarizes the results of the RF regression performed. The RF model was

t	n_t	Prediction	$W_{t_{i,j}}$	s^*	Node Type	Path
1	278	23,33	0,54	displacement $\leq$ 267	Non-Terminal	
2	208	26,20	0,54	displacement $\leq$ 183	Non-Terminal	displacement $\leq$ 267
4	155	28,59	0,54	displacement $\leq$ 108	Non-Terminal	displacement $\leq$ 267 & displacement $\leq$ 183
8	76	31,68	0,21	model_year $\leq$ 77	Non-Terminal	displacement $\leq$ 267 & displacement $\leq$ 183 & displacement $\leq$ 108
16	41	28,28	0,30	weight $\leq$ 1795	Non-Terminal	displacement $\leq$ 267 & displacement $\leq$ 183 & displacement $\leq$ 108 & model_year $\leq$ 77
32	5	32,60	0,03	-	Terminal	displacement $\leq$ 267 & displacement $\leq$ 183 & displacement $\leq$ 108 & model_year $\leq$ 77 & weight $\leq$ 1795
33	36	27,68	0,28	-	Terminal	displacement $\leq$ 267 & displacement $\leq$ 183 & displacement $\leq$ 108 & model_year $\leq$ 77 & weight $>$ 1795
17	35	35,67	0,06	displacement $\leq$ 97	Non-Terminal	displacement $\leq$ 267 & displacement $\leq$ 183 & displacement $\leq$ 108 & model_year $>$ 77
34	20	36,91	0,00	-	Terminal	displacement $\leq$ 267 & displacement $\leq$ 183 & displacement $\leq$ 108 & model_year $>$ 77 & displacement $\leq$ 97
35	15	34,02	0,00	-	Terminal	displacement $\leq$ 267 & displacement $\leq$ 183 & displacement $\leq$ 108 & model_year $>$ 77 & displacement $>$ 97
9	79	25,62	0,44	model_year $\leq$ 78	Non-Terminal	displacement $\leq$ 267 & displacement $\leq$ 183 & displacement $>$ 108
18	43	22,82	0,64	-	Terminal	displacement $\leq$ 267 & displacement $\leq$ 183 & displacement $>$ 108 & model_year $\leq$ 78
19	36	28,97	0,05	horsepower $\leq$ 105	Non-Terminal	displacement $\leq$ 267 & displacement $\leq$ 183 & displacement $>$ 108 & model_year $>$ 78
38	31	29,55	0,00	-	Terminal	displacement $\leq$ 267 & displacement $\leq$ 183 & displacement $>$ 108 & model_year $>$ 78 & horsepower $\leq$ 105
39	5	25,38	0,00	-	Terminal	displacement $\leq$ 267 & displacement $\leq$ 183 & displacement $>$ 108 & model_year $>$ 78 & horsepower $>$ 105
5	53	19,21	0,55	model_year $\leq$ 75	Non-Terminal	displacement $\leq$ 267 & displacement $>$ 183
10	24	17,75	0,74	-	Terminal	displacement $\leq$ 267 & displacement $>$ 183 & model_year $\leq$ 75
11	29	20,41	0,00	-	Terminal	displacement $\leq$ 267 & displacement $>$ 183 & model_year $>$ 75
3	70	14,78	0,79	model_year $\leq$ 76	Non-Terminal	displacement $>$ 267
6	54	13,87	0,87	weight $\leq$ 3777	Non-Terminal	displacement $>$ 267 & model_year $\leq$ 76
12	13	15,38	0,45	displacement $\leq$ 318	Non-Terminal	displacement $>$ 267 & model_year $\leq$ 76 & weight $\leq$ 3777
24	8	16,25	0,48	-	Terminal	displacement $>$ 267 & model_year $\leq$ 76 & weight $\leq$ 3777 & displacement $\leq$ 318
25	5	14,00	0,00	-	Terminal	displacement $>$ 267 & model_year $\leq$ 76 & weight $\leq$ 3777 & displacement $>$ 318
13	41	13,39	0,88	-	Terminal	displacement $>$ 267 & model_year $\leq$ 76 & weight $>$ 3777
7	16	17,84	0,11	weight $\leq$ 3880	Non-Terminal	displacement $>$ 267 & model_year $>$ 76
14	8	19,20	0,00	-	Terminal	displacement $>$ 267 & model_year $>$ 76 & weight $\leq$ 3880
15	8	16,49	0,72	-	Terminal	displacement $>$ 267 & model_year $>$ 76 & weight $>$ 3880

Table 5.3. Auto MPG data: results of E2Tree.

Feature	Mean	SD	Min	1st Quartile	3rd Quartile	Max
CRIM	3.61	8.60	0.00632	0.08	3.68	88.98
ZN	11.36	23.32	0.00	0.00	12.50	100.00
INDUS	11.14	6.86	0.46	5.19	18.10	27.74
CHAS	-	-	0	-	-	1
NOX	0.55	0.12	0.385	0.45	0.62	0.87
RM	6.29	0.70	3.56	5.89	6.62	8.78
AGE	68.57	28.15	2.90	45.02	94.08	100.00
DIS	3.80	2.11	1.13	2.10	5.19	12.13
RAD	9.55	8.71	1.00	4.00	24.00	24.00
TAX	408.2	168.54	187.0	279.0	666.0	711.0
PTRATIO	18.46	2.16	12.60	17.40	20.20	22.00
B	356.67	91.29	0.32	375.38	396.23	396.90
LSTAT	12.65	7.14	1.73	6.95	16.95	37.97
MEDV	22.53	9.20	5.00	17.02	25.00	50.00

Table 5.4. Summary statistics of the Boston Housing dataset attributes.

Type of RF:	regression
Number of trees:	500
No. of variables tried at each split:	4
Mean of squared residuals:	10.69
% Var explained:	87.04

Table 5.5. Boston Housing Data: Results of RF regression.

configured as a regression model, utilizing 500 trees for robust ensemble learning. At each split, 4 variables were considered, ensuring a diverse and comprehensive exploration of the feature space. The model achieved a mean squared residual of 10.69, indicating a relatively low level of prediction error. Furthermore, the RF explained 87.04% of the variance in the data, demonstrating its strong predictive power and effectiveness in capturing the complex relationships between the features and the target variable. The tree structure in Figure 5.5 and it's summary Table 5.6 provides an interpretable summary of the complex relationships captured by the ensemble model, highlighting the critical predictors and their thresholds. At the root, the tree splits based on the crime rate (CRIM), with a threshold of 4.6. This indicates that neighbourhoods with lower crime

rates tend to have higher median house values, highlighting the strong link between safety and house prices. Further breakdowns reveal other significant predictors, including the average number of rooms per dwelling (RM), the percentage of lower status population (LSTAT) and the concentration of nitrogen oxide (NOX). The number of rooms (RM) appears repeatedly in the tree, demonstrating its strong influence on house prices. Higher RM values are associated with higher MEDV, reflecting the preference for spacious dwellings in determining property values. Similarly, the percentage of lower status population (LSTAT) is a critical predictor, with lower percentages correlating with higher house values. This relationship highlights socio-economic factors as important determinants of housing market trends. Environmental factors, represented by nitrogen oxide (NOX) levels, also play a significant role. Higher NOX levels, indicating poorer air quality, are associated with lower house values, in line with the general preference for healthier living conditions. The thresholds of 0.77 and 0.61 for NOX highlight its fading effect over different ranges. Other predictors, such as the black population index (B) and additional splits on CRIM, provide further refinement, suggesting that neighbourhood demographics and safety consistently influence property values. End nodes at the bottom of the tree represent the predicted MEDV values for subsets of the data, providing interpretable insights into how combinations of predictors determine house prices. This analysis highlights the importance of socio-economic, structural and environmental variables in shaping housing markets. The results provide actionable insights for policymakers and urban planners, suggesting that improvements in neighbourhood safety, air quality and housing infrastructure could increase property values. Lastly, we evaluate the performance of E2Tree in replicating the structure and relationships captured by the RF model. Specifically, we examine in Figure 5.6 the heatmap of the matrix O_{ij} (panel a) and compare it with the heatmap of the matrix $\hat{O}_{ij}$ estimated by E2Tree (panel b). To quantify and test this relationship, we perform the Mantel Test, whose results are summarized in Figure 5.4. The Mantel Test revealssignificant correlation, $z = 178.8469$; $p < .001$, allowing to reject the null hypothesis of no correlation. These results confirm again that E2Tree effectively reconstructs the intricate structure of the RF model while maintaining high interpretability. The statistically significant findings and the robustness of

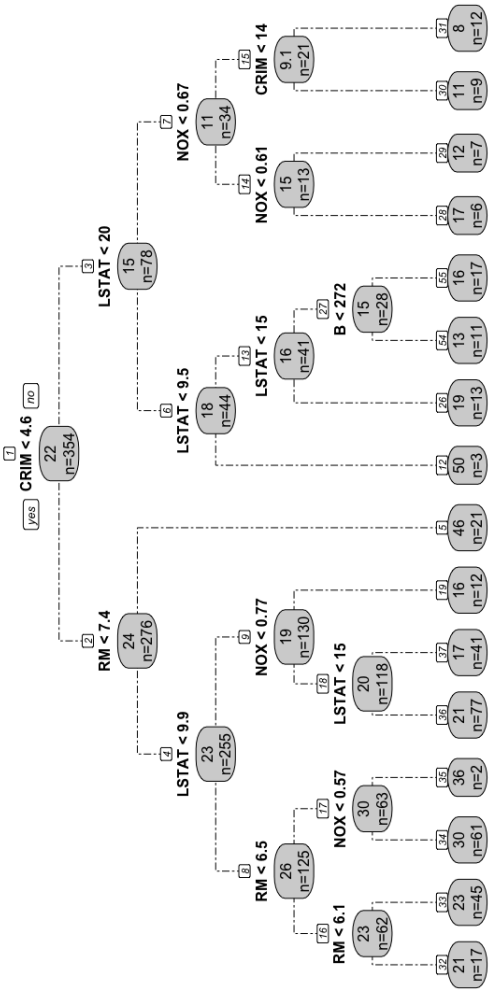
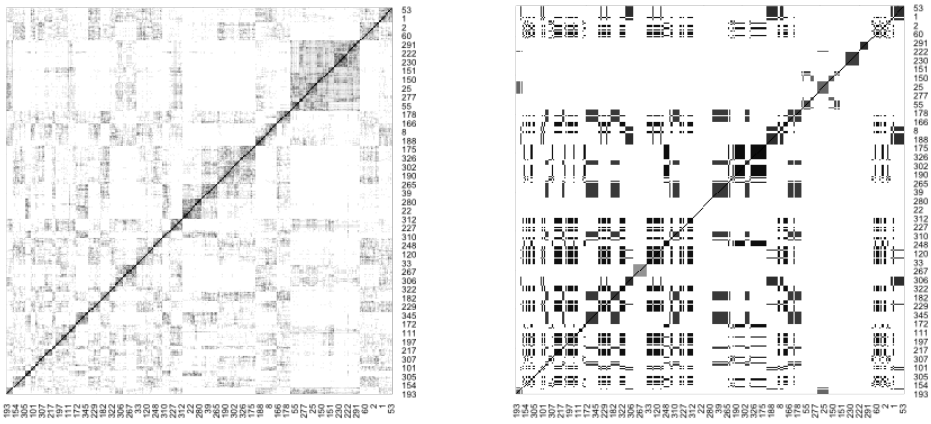


Figure 5.5. Boston Housing Data: path visualization of E2Tree. The color intensity indicates the magnitude of the predicted values: darker shades typically correspond to higher predicted values, while lighter shades correspond to lower predicted values.

the correlation underscore the utility of E2Tree as a reliable method for enhancing transparency in ensemble RF models.



(a) (b)

Figure 5.6. Boston Housing Data: heatmap of the matrix O_{ij} (panel a) and heatmap of the matrix $\hat{O}_{ij}$ estimated by E2Tree (panel b).

Computational performance

To complement the interpretability analysis, we evaluated the computational performance of E2Tree. Table 5.7 reports the average computation times required for constructing the dissimilarity matrix, obtained using the R package `microbenchmark` with 100 replications. We considered both sequential and parallel implementations. For the Auto MPG dataset, the average time decreased from 401 ms (sequential) to 274 ms (parallel). For the Boston Housing dataset, the average time decreased from 573 ms to 381 ms. All computations were executed on a MacBook Pro with Apple M4 Pro chip, 48 GB RAM, and a 14-core CPU. These results highlight the speed of E2Tree as well as the effectiveness of parallel processing in improving computational efficiency.

t	n_t	Prediction	$W_{t_{ij}}$	s^*	Node Type	Path
1	354	22,46	0,63	CRIM $\leq$ 4.64689	Non-Terminal	
2	276	24,49	0,62	RM $\leq$ 7.42	Non-Terminal	CRIM $\leq$ 4.64689
4	255	22,71	0,82	LSTAT $\leq$ 9.93	Non-Terminal	CRIM $\leq$ 4.64689 & RM $\leq$ 7.42
8	125	26,28	0,69	RM $\leq$ 6.54	Non-Terminal	CRIM $\leq$ 4.64689 & RM $\leq$ 7.42 & LSTAT $\leq$ 9.93
16	62	22,78	0,88	RM $\leq$ 6.059	Non-Terminal	CRIM $\leq$ 4.64689 & RM $\leq$ 7.42 & LSTAT $\leq$ 9.93 & RM $\leq$ 6.54
32	17	21,21	0,63	-	Terminal	CRIM $\leq$ 4.64689 & RM $\leq$ 7.42 & LSTAT $\leq$ 9.93 & RM $\leq$ 6.54 & RM $\leq$ 6.059
33	45	23,37	0,88	-	Terminal	CRIM $\leq$ 4.64689 & RM $\leq$ 7.42 & LSTAT $\leq$ 9.93 & RM $\leq$ 6.54 & RM $>$ 6.059
17	63	29,72	0,50	NOX $\leq$ 0.573	Non-Terminal	CRIM $\leq$ 4.64689 & RM $\leq$ 7.42 & LSTAT $\leq$ 9.93 & RM $>$ 6.54
34	61	29,51	0,50	-	Terminal	CRIM $\leq$ 4.64689 & RM $\leq$ 7.42 & LSTAT $\leq$ 9.93 & RM $>$ 6.54 & NOX $\leq$ 0.573
35	2	36,25	0,95	-	Terminal	CRIM $\leq$ 4.64689 & RM $\leq$ 7.42 & LSTAT $\leq$ 9.93 & RM $>$ 6.54 & NOX $>$ 0.573
9	130	19,29	0,86	NOX $\leq$ 0.77	Non-Terminal	CRIM $\leq$ 4.64689 & RM $\leq$ 7.42 & LSTAT $<$ 9.93
18	118	19,58	0,86	LSTAT $\leq$ 14.8	Non-Terminal	CRIM $\leq$ 4.64689 & RM $\leq$ 7.42 & LSTAT $<$ 9.93 & NOX $\leq$ 0.77
36	77	20,71	0,86	-	Terminal	CRIM $\leq$ 4.64689 & RM $\leq$ 7.42 & LSTAT $<$ 9.93 & NOX $\leq$ 0.77 & LSTAT $\leq$ 14.8
37	41	17,46	0,59	-	Terminal	CRIM $\leq$ 4.64689 & RM $\leq$ 7.42 & LSTAT $<$ 9.93 & NOX $\leq$ 0.77 & LSTAT $>$ 14.8
19	12	16,39	0,13	-	Terminal	CRIM $\leq$ 4.64689 & RM $\leq$ 7.42 & LSTAT $<$ 9.93 & NOX $>$ 0.77
5	21	46,01	0,00	-	Terminal	CRIM $\leq$ 4.64689 & RM $>$ 7.42
3	78	15,28	0,00	LSTAT $\leq$ 19.69	Non-Terminal	CRIM $>$ 4.64689
6	44	18,40	0,00	LSTAT $\leq$ 9.53	Non-Terminal	CRIM $>$ 4.64689 & LSTAT $\leq$ 19.69
12	3	50,00	1,00	-	Terminal	CRIM $>$ 4.64689 & LSTAT $\leq$ 19.69 & LSTAT $\leq$ 9.53
13	41	16,09	0,38	LSTAT $\leq$ 14.7	Non-Terminal	CRIM $>$ 4.64689 & LSTAT $\leq$ 19.69 & LSTAT $>$ 9.53
26	13	19,48	0,00	-	Terminal	CRIM $>$ 4.64689 & LSTAT $\leq$ 19.69 & LSTAT $>$ 9.53 & LSTAT $\leq$ 14.7
27	28	14,52	0,56	B $\leq$ 272.21	Non-Terminal	CRIM $>$ 4.64689 & LSTAT $\leq$ 19.69 & LSTAT $>$ 9.53 & LSTAT $>$ 14.7
54	11	13,01	0,45	-	Terminal	CRIM $>$ 4.64689 & LSTAT $\leq$ 19.69 & LSTAT $>$ 9.53 & LSTAT $>$ 14.7 & B $\leq$ 272.21
55	17	15,50	0,28	-	Terminal	CRIM $>$ 4.64689 & LSTAT $\leq$ 19.69 & LSTAT $>$ 9.53 & LSTAT $>$ 14.7 & B $>$ 272.21
7	34	11,24	0,17	NOX $\leq$ 0.671	Non-Terminal	CRIM $>$ 4.64689 & LSTAT $>$ 19.69
14	13	14,76	0,00	NOX $\leq$ 0.614	Non-Terminal	CRIM $>$ 4.64689 & LSTAT $>$ 19.69 & NOX $\leq$ 0.671
28	6	17,48	0,00	-	Terminal	CRIM $>$ 4.64689 & LSTAT $>$ 19.69 & NOX $\leq$ 0.671 & NOX $\leq$ 0.614
29	7	12,43	0,60	-	Terminal	CRIM $>$ 4.64689 & LSTAT $>$ 19.69 & NOX $\leq$ 0.671 & NOX $>$ 0.614
15	21	9,06	0,57	CRIM $\leq$ 13.6781	Non-Terminal	CRIM $>$ 4.64689 & LSTAT $>$ 19.69 & NOX $>$ 0.671
30	9	10,51	0,29	-	Terminal	CRIM $>$ 4.64689 & LSTAT $>$ 19.69 & NOX $>$ 0.671 & CRIM $\leq$ 13.6781
31	12	7,97	0,46	-	Terminal	CRIM $>$ 4.64689 & LSTAT $>$ 19.69 & NOX $>$ 0.671 & CRIM $>$ 13.6781

Table 5.6. Boston Housing Data: Results of E2Tree.

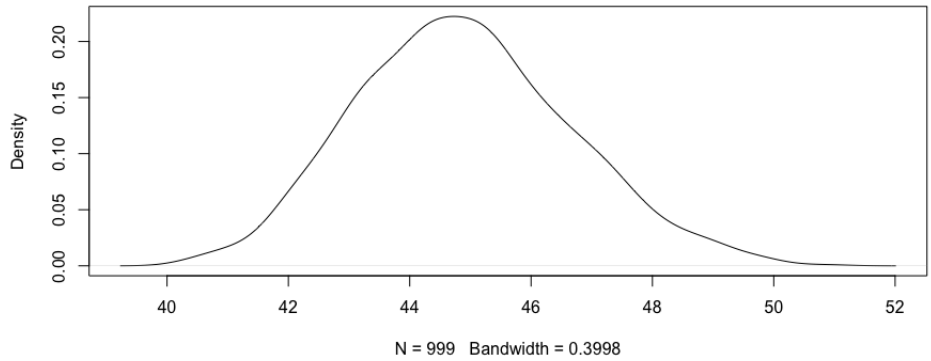


Figure 5.7. Boston Housing Data: Density plot of Mantel test permutations.

Dataset	Mean time (no parallel)	Median (ms)	Mean time (parallel)	Median (ms)
Auto MPG	401.4	395.3	274.0	265.9
Boston Housing	572.9	548.9	381.3	356.7

Table 5.7. Average computation times (100 replications, with and without parallelization)

5.3 Concluding Remarks

In this paper, we introduced Explainable Ensemble Trees (E2Tree) for the regression contexts. We extend the work proposed by Aria et al. (2024) [67] for classification tasks. E2Tree offers a coherent, tree-like structure that facilitates anunderstanding of the relationships between features and predictions while preserving the model's predictive accuracy. The enhanced explainability provided by E2Tree has significant implications for high-stakes applications, such as healthcare, finance, and environmental modeling, where explainability as well as interpretability are paramount. By providing clear, validated insights into predictor relationships and model behavior, our proposal empowers decision-makers to leverage machine learning models with greater confidence and understanding. Its ability to integrate local and global explanations within a single framework addresses a critical need in the field of explainable machine learning, setting a new standard for interpretability in regression modeling. E2Tree provides several advantages. It explains the ensemble process through a

single tree-like structure; it maintains the accuracy performance of the Random Forest ensemble; finally, it provides a powerful tool to explain interactions among predictors determining the final node predictions. It is also worth noting that the proposed approach is not without limitations. The first of these is scalability, since the computation of the dissimilarity matrix may become computationally expensive for large datasets. To address this challenge, several optimization strategies can be considered. First, approximation techniques, such as randomized algorithms or matrix sparsification, could reduce computational complexity while preserving essential structural properties. Second, sampling procedures, including sub-sampling of instances or trees, may provide efficient surrogates of the complete dissimilarity matrix, provided that they retain the core ensemble patterns. Finally, parallelization strategies represent a practical solution to improve computational efficiency. We note that the current implementation of E2Tree already integrates parallel processing and sparse matrix storage, which allow for substantial reductions in computation time and memory requirements. Future research will further explore these directions to enhance the applicability of E2Tree in large-scale contexts. Another important direction concerns the generality of the method. While the present framework has been specifically developed for RF, E2Tree already supports XGBoost models, and ongoing work is focused on improving its computational efficiency in this context. Extending the methodology to additional ensemble-tree algorithms such as Gradient Boosting, LightGBM, or CatBoost could further broaden its scope. However, these algorithms present specific challenges. In the case of Gradient Boosting, the sequential nature of training, where each tree corrects for the errors of its predecessors, differs substantially from the parallel construction of Random Forests and requires adapting the dissimilarity matrix as well as the interpretability strategies. With XGBoost, additional complexities such as regularization and second-order optimization must be explicitly incorporated, as they influence the weighting of splits and the definition of co-occurrence measures. Addressing these algorithm-specific properties represents a promising line of research that will further enhance the versatility of E2Tree and provide comparative insights across alternative ensemble paradigms. E2Tree offers a tool for explaining the predictions of RFs in regression and classification contexts. That is, it is assumed the black-box model (RF) is used

for prediction, while E2Tree is used for explaining these predictions. This explainable ML approach can be contrasted to interpretable ML, in which an interpretable model is fitted to the data in the first place. Although users may mistakenly think of employing E2Tree for predictive tasks, it must be emphasized that this methodology is not intended for prediction or classification of new instances. When the aim is accurate prediction, ensemble methods such as RF should be employed, since they significantly outperform single trees in terms of predictive accuracy. In contrast, the sole purpose of E2Tree is to open the black box of ensemble models, providing a transparent and comprehensive reconstruction of their decision logic. Related, there are no widely agreed-upon quantitative measures of interpretability or explainability [76]. Commonly-used proxies are the total number of decision rules in a node [77], in terms of which E2Tree easily outperforms RF. Ather proxy being used is the number of variables in the model [76]. While RF generally uses all features for prediction, only four out of the seven original features were used by the E2Tree for the MPG data, and five out of thirteen for the Boston Housing data. This comprises a significant reduction compared to the RFs and suggests that the amount of information users need to process to interpret the E2Tree is lower than for RF coupled with other post-hoc explanation measures such as permutation importances, SHAP, SAGE and/or PDPs.

Conclusions and perspective

The framework of this thesis is ensemble methods. The general framework has been presented in Chapter 1. The methodological contributions have been provided in Chapters 2, 3, 4 and 5.

Chapter 1 established the theoretical and methodological foundations necessary for understanding the contributions. It introduced the core concepts of supervised machine learning, with particular attention to tree-based methods and ensemble learning approaches. The chapter formalized the construction and properties of decision trees, including their splitting criteria, impurity measures, and inherent interpretability. It then examined ensemble methodologies (specifically Bagging, Boosting, and Random Forest) highlighting how these techniques enhance predictive performance through variance reduction and model aggregation. The chapter also critically distinguished between interpretability and explainability, two concepts often mixed up in the literature but fundamentally different in scope and purpose. Interpretability was characterized as the direct transparency of a model's structure, while explainability was defined as the post-hoc reconstruction of decision logic in otherwise opaque models. This conceptual clarification provided the epistemological basis for positioning Explainable Ensemble Trees (E2Tree) as a methodology that bridges both dimensions.

Chapter 2 presented a comprehensive and critical review of existing interpretative proposals for Random Forest models. The chapter adopted a structured taxonomy, distinguishing between internal processing approaches (such as variable importance measures, partial dependence plots, and proximity matrices) and post-hoc methodologies, including size reduction techniques, rule extraction methods, and local explanation frameworks such as

LIME and SHAP. An empirical comparative study was conducted across six benchmark datasets with varying characteristics, including class imbalance, sample size, and predictor types. Performance was evaluated using multiple metrics, including balanced accuracy, F1-score, and Cohen's Kappa. This comparative analysis highlighted the fragmentation of existing explainability strategies and underscored the need for a unified, globally interpretable framework capable of faithfully reconstructing ensemble decision logic.

Chapter 3 introduced the E2Tree methodology in the context of classification tasks. The chapter formalized the core algorithmic framework, which is grounded in the construction of a dissimilarity matrix derived from co-occurrence patterns of observations across terminal nodes in the Random Forest ensemble. The dissimilarity measure was weighted by a local goodness-of-fit metric, defined as the correct classification rate at each node. A greedy splitting procedure was then employed to construct a single decision tree that minimizes within-node dissimilarity while maximizing interpretability. Specific stopping criteria were introduced to prevent overfitting and ensure meaningful splits, including thresholds on node purity, sample size, and classification accuracy. The methodology was validated through three illustrative datasets. In each case, E2Tree successfully reconstructed the relational structure of the Random Forest model, producing interpretable decision pathways that aligned with domain knowledge. The chapter demonstrated that E2Tree preserves the predictive fidelity of the ensemble while offering a transparent, graphical representation of the decision-making process, thereby addressing the opacity inherent in traditional ensemble methods.

Chapter 4 demonstrated the practical applicability of E2Tree through a substantive empirical study in the domain of public health. The analysis focused on predicting depression risk factors in the Italian population using microdata from the European Health Interview Survey (EHIS), conducted by the Italian National Institute of Statistics (Istat). A Random Forest model was trained on a stratified sample of 9,463 individuals, with 26% classified as depressed. The E2Tree methodology was then applied to reconstruct the ensemble's decision logic into an interpretable tree structure. The Mantel test confirmed a statistically significant correlation between the original Random Forest similarity matrix and the E2Tree-estimated

matrix, validating the fidelity of the reconstruction. The resulting decision tree revealed that the primary predictor of depression was the presence of chronic illnesses, followed by family structure and psychological symptoms including fatigue, apathy, and low self-esteem. These findings provided actionable insights for public health policy, suggesting the need for targeted interventions addressing both the physical and psychological dimensions of chronic illness, as well as social support systems for vulnerable family structures. The chapter highlighted the capacity of E2Tree to translate opaque ensemble predictions into transparent, policy-relevant insights in high-stakes decision contexts.

Chapter 5 extended the E2Tree methodology to regression contexts, addressing a critical gap in the original framework. This proposal required the development of a novel local goodness-of-fit measure adapted to continuous response variables. Specifically, a normalized mean square error (NMSE) metric was introduced, grounded in Jacobson's theoretical maximum variance bound, to weight co-occurrence patterns in the dissimilarity matrix. Additional stopping criteria were formalized, including a threshold on NMSE and a Mann-Whitney test to ensure that child nodes exhibited statistically distinct distributions following each split. The extended methodology was validated on two benchmark datasets. In both cases, the Mantel test confirmed a statistically significant correlation between the Random Forest similarity structure and the E2Tree reconstruction. Computational performance was evaluated using parallel processing, demonstrating significant reductions in computation time relative to sequential implementations. The chapter established that E2Tree is not only theoretically sound but also computationally feasible and empirically robust in regression settings, thereby broadening its scope and reinforcing its versatility as a general framework for ensemble explainability.

The proposed E2Tree does have some limitations, which in turn define key directions for my future researchs.

First, the computational cost associated with constructing the dissimilarity matrix scales with dataset size and ensemble complexity, raising concerns regarding scalability in high-dimensional or large-sample contexts. While the current implementation incorporates parallel processing, further algorithmic optimization is necessary. Second, its generalization to other ensemble architectures beyond Random Forest and XGBoost, such

as Gradient Boosting Machines or LightGBM, requires methodological refinement to accommodate algorithm-specific characteristics. Thirdly, in view of the absence of a standardised quantitative measure of explainability, there is a necessity to develop metrics that accurately capture and evaluate this crucial attribute.

These limitations suggest several directions for future inquiry. To address scalability, the development of computationally efficient approximation strategies, such as subsampling or low-rank matrix decompositions, could mitigate performance concerns while preserving explanatory fidelity. Furthermore, extending E2Tree to a broader class of ensemble algorithms would enhance its methodological generality and empirical applicability. This would necessitate the formalization of algorithm-specific dissimilarity measures and stopping criteria. A crucial area of research is the establishment of formal evaluation metrics for explainability, grounded in statistical or information-theoretic principles, to provide a rigorous basis for assessing the quality of E2Tree explanations. Finally, integrating E2Tree with complementary post-hoc methods (such as SHAP or LIME) could yield a more comprehensive interpretative ecosystem, enabling multi-faceted analyses of model behavior.

Bibliography

[1] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[2] 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.

[3] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[4] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[5] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[6] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[7] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[8] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168, 2006.

[9] Massimo Aria, Corrado Cuccurullo, and Agostino Gnasso. A comparison among interpretative proposals for random forests. *Machine Learning with Applications*, 6:100094, 2021.

[10] M. Aria, A. Gnasso, C. Iorio, and G. Pandolfo. Explainable ensemble trees. *Computational Statistics*, 39:3–19, 2024.

[11] M. Aria, A. Gnasso, C. Iorio, and M. Fokkema. Extending explainable ensemble trees (e2tree) to regression contexts. *arXiv preprint arXiv:2409.06439*, 2024.

[12] Massimo Aria, Agostino Gnasso, Rebecca Rivieccio, and Roberta Siciliano. Predicting depression in italy using random forest through the e2tree methodology. *Annals of Operations Research*, pages 1–18, 2025.

[13] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees*. CRC Press, 1984.

[14] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[15] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Interpretable machine learning–a brief history, state-of-the-art and challenges. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 417–431. Springer, 2020.

[16] Chidanand Apté and Sholom Weiss. Data mining with decision trees and decision rules. *Future generation computer systems*, 13(2-3):197–210, 1997.

[17] Carl Kingsford and Steven L Salzberg. What are decision trees? *Nature biotechnology*, 26(9):1011–1013, 2008.

[18] Leo Breiman et al. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

[19] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[20] Robi Polikar. Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, 6(3):21–45, 2006.

[21] Marquis de Condorcet. Essay on the application of analysis to the probability of majority decisions. *Paris: Imprimerie Royale*, 1785.

[22] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[23] Leo Breiman. Population theory for boosting ensembles. *The Annals of Statistics*, 32(1):1–11, 2004.

[24] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2025.

[25] Zhicheng Cui, Wenlin Chen, Yujie He, and Yixin Chen. Optimal action extraction for random forests and boosted trees. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 179–188, 2015.

[26] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[27] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[28] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Bari, Italy, 1996.

[29] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[30] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

[31] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[32] Leo Breiman. Out-of-bag estimation. 1996.

[33] Leo Breiman. Manual on setting up, using, and understanding random forests v3. 1. *Statistics Department University of California Berkeley, CA, USA*, 1:58, 2002.

[34] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

[35] Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

[36] Cynthia Rudin and Joanna Radin. Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition. *Harvard Data Science Review*, 1(2), 2019.

[37] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[38] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[39] 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.

[40] Maissae Haddouchi and Abdelaziz Berrado. A survey of methods and tools used for interpreting random forest. In *2019 1st International Conference on Smart Systems and Data Science (ICSSD)*, pages 1–6. IEEE, 2019.

[41] Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. Variable selection using random forests. *Pattern recognition letters*, 31(14):2225–2236, 2010.

[42] Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. Understanding variable importances in forests of randomized trees. In *Advances in neural information processing systems*, pages 431–439, 2013.

[43] Maissae Haddouchi and Abdelaziz Berrado. A survey of methods and tools used for interpreting random forest. In *2019 1st International Conference on Smart Systems and Data Science (ICSSD)*, pages 1–6. IEEE, 2019.

[44] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[45] John Ehrlinger. ggrandomforests: random forests for regression. *arXiv preprint arXiv:1501.07196*, 2016.

[46] 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.

[47] Hui Fen Tan, Giles Hooker, and Martin T Wells. Tree space prototypes: Another look at making tree ensembles interpretable. *arXiv preprint arXiv:1611.07115*, 2016.

[48] HA Chipman, EI George, and RE McCulloh. Making sense of a forest of trees. *Computing Science and Statistics*, pages 84–92, 1998.

[49] Minghui Wang and Heping Zhang. Search for the smallest random forest. *Statistics and its Interface*, 2(3):381–388, 2009.

[50] Robert D Gibbons, Giles Hooker, Matthew D Finkelman, David J Weiss, Paul A Pilkonis, Ellen Frank, Tara Moore, and David J Kupfer. The cad-mdd: a computerized adaptive diagnostic screening tool for depression. *The Journal of clinical psychiatry*, 74(7):669, 2013.

[51] Yichen Zhou, Zhengze Zhou, and Giles Hooker. Approximation trees: Statistical stability in model distillation. *arXiv preprint arXiv:1808.07573*, 2018.

[52] Nicolai Meinshausen. Node harvest. *The Annals of Applied Statistics*, pages 2049–2072, 2010.

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

[54] Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158, 2012.

[55] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[56] Milad Moradi and Matthias Samwald. Post-hoc explanation of black-box classifiers using confident itemsets. *arXiv preprint arXiv:2005.01992*, 2020.

[57] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*, pages 1015–1021. Springer, 2006.

[58] Vicente García, Ramón Alberto Mollineda, and José Salvador Sánchez. Index of balanced accuracy: A performance measure for skewed class distributions. In *Iberian conference on pattern recognition and image analysis*, pages 441–448. Springer, 2009.

[59] Josephine Akosa. Predictive accuracy: A misleading performance measure for highly imbalanced data. In *Proceedings of the SAS Global Forum*, pages 2–5, 2017.

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

[61] Paula Branco, Luís Torgo, and Rita P Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49(2):1–50, 2016.

[62] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[63] K. Amarasinghe, K. T. Rodolfa, H. Lamba, and R. Ghani. Explainable machine learning for public policy: Use cases, gaps, and research directions. *Data & Policy*, 5:e5, 2023.

[64] Istat. Manuale di tecniche di indagine 6 - il sistema di controllo della qualità dei dati, 1989.

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

[66] N. Mantel. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2_Part_1):209–220, 1967.

[67] Massimo Aria, Agostino Gnasso, Carmela Iorio, and Giuseppe Pandolfo. Explainable ensemble trees. *Computational Statistics*, 39(1):3–19, 2024.

[68] Harold I. Jacobson. The maximum variance of restricted unimodal distributions. *The annals of mathematical statistics*, 40(5):1746–1752, 1969.

[69] Paul A Samuelson. How deviant can you be? *Journal of the American Statistical Association*, 63(324):1522–1525, 1968.

[70] Barry C Arnold. Schwarz, regression, and extreme deviance. *The American Statistician*, 28(1):22–23, 1974.

[71] John W Seaman Jr, Patrick S Odell, and Dean M Young. Maximum variance unimodal distributions. *Statistics & probability letters*, 3(5):255–260, 1985.

[72] Wei-Yin Loh. Fifty years of classification and regression trees. *International Statistical Review*, 82(3):329–348, 2014.

[73] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[74] R. Quinlan. Auto MPG. UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C5859H.

[75] 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.

[76] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52, 2021.

[77] Josue Obregon and Jae-Yoon Jung. RuleCOSI+: Rule extraction for interpreting classification tree ensembles. *Information Fusion*, 89:355–381, 2023.

Author's Publications

Articles in Scientific Journals - *Fascia A, Area 13/STAT-01/A1*

1. Aria, M., **Gnasso, A.**, Iorio, C. et al. Explainable Ensemble Trees. *Computational Statistics* 39, 3–19 (2024). DOI: https://doi.org/10.1007/s00180-022-01312-6.

2. Aria, M., Gnasso, A., Rivieccio, R. et al. Predicting depression in Italy using random forest through the E2Tree methodology. *Annals of Operations Research* (2025). https://doi.org/10.1007/s10479-025-06758-7

Articles in Scientific Journals - *Indexed in WoS/SCOPUS*

1. Aria M., Cuccurullo C., **Gnasso A.** (2021). A comparison among interpretative proposals for Random Forests, *Machine Learning with Applications*, DOI: https://doi.org/10.1016/j.mlwa.2021.100094.

2. Adamo D., **Gnasso A.**, et al. (2021). Assessment of Sleep Disturbance in Oral Lichen Planus and Validation of PSQI: a case-control multicenter study from the SIPMO, *Journal of Oral Pathology & Medicine*, DOI: https://doi.org/10.1111/jop.13255.

Working papers

1. Aria, M., **Gnasso, A.**, Iorio, C., & Fokkema, M. (forthcoming, in press). Extending Explainable Ensemble Trees to regression contexts. *Applied Stochastic Models in Business and Industry*. Preprint available on arXiv: https://arxiv.org/abs/2409.06439. (Fascia A, Area 13/STAT-01/A)

Conference Proceedings

1. **Gnasso, A.**, Sacco, D., Celardo, L., Smecca, M.A., Alabiso, C., & Spano, M. (2025). Research excellence and patient perception: investigating the

impact of AHSCs' scientific output. In: Boccuzzo, G., Bovo, E., Manisera, M., Salmaso, L. *IES 2025 – Innovation & Society: Statistics and Data Science for Evaluation and Quality* pp. 1014-1020. CLEUP, Padova.

2. **Gnasso, A.**, Aria, M., Iorio, C., & Fokkema, M. (2025). Explainable Decision Tree Ensembles. In: Pollice, A., Mariani, P. (eds), *Methodological and Applied Statistics and Demography IV. SIS 2024*. Italian Statistical Society Series on Advances in Statistics. Springer, Cham. DOI: 10.1007/978-3-031-64447-4_21

3. **Gnasso, A.**, & Aria, M. (2024). The evolution of Explainable Artificial Intelligence (XAI): a preliminary systematic literature review. In: *Book of Short Papers of the ASA Conference 2024*.

4. Aria M., **Gnasso A.**, D'Aniello L. (2022). Twenty Years of Random Forest: preliminary results of a systematic literature review. In: *IES 2022*, pp. 225–230.

5. Belfiore, A., **Gnasso A.**, Cuccurullo, C., Massimo, A. (2022). AI and ML in accounting and finance: a bibliometric review. In: *JADT 2022 – Proceedings of the 16th International Conference on Statistical Analysis of Textual Data* (Vol. 1, pp. 95–101).

6. Aria M., Cuccurullo C., **Gnasso A.** (2021). Supporting decision-makers in healthcare domain. A comparative study of two interpretative proposals for Random Forests. In: *ASA 2021 – Book of Short Papers* (Vol. 132), pp. 179–184. Firenze University Press.

7. **Gnasso, A.**, Aria, M. (2025). "Can You Explain That?" E2Tree, SHAP, and LIME for Interpretable Random Forests. In: D'Ambrosio, A., de Rooij, M., De Roover, K., Iorio, C., La Rocca, M. (eds) *Supervised and Unsupervised Statistical Data Analysis. CLADAG-VOC 2025. Studies in Classification, Data Analysis, and Knowledge Organization*. Springer, Cham. https://doi.org/10.1007/978-3-032-03042-9_20.

8. **Gnasso, A.**, Aria, M., Siciliano, R. (2025). From Prediction to Explanation: Interpreting Risk Factors in Health Survey Analytics. In: D'Ambrosio, A., de Rooij, M., De Roover, K., Iorio, C., La Rocca, M. (eds) *Supervised and Unsupervised Statistical Data Analysis. CLADAG-VOC 2025. Studies in Classification, Data Analysis, and Knowledge Organization*. Springer, Cham. https://doi.org/10.1007/978-3-032-03042-9_21

Conference Abstracts

1. Iorio, C., **Gnasso, A.**, & Aria, M. (2024). Inside the black-box models through explainable decision tree ensembles. In: *Programme and Abstracts*

– *26th International Conference on Computational Statistics (COMPSTAT 2024)*, pp. 40–40.

2. Aria, M., **Gnasso, A.**, Iorio, C., & Pandolfo, G. (2023). Unlocking explainable in ensemble trees. In: *Programme and Abstracts – CMStatistics 2023 and CFE 2023*, ECOSTA.