Annals of Operations Research
https://doi.org/10.1007/s10479-025-06758-7

ORIGINAL RESEARCH

Predicting depression in Italy using random forest through the E2Tree methodology

Massimo Aria[1] · **Agostino Gnasso**[1] · **Rebecca Rivieccio**[2] · **Roberta Siciliano**[3]

Received: 23 December 2024 / Accepted: 8 July 2025

Abstract
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.

Keywords Explainability · Interpretability · Machine learning · EHIS · Mental health

✉ Agostino Gnasso
agostino.gnasso@unina.it

Massimo Aria
massimo.aria@unina.it

Rebecca Rivieccio
rebecca.rivieccio@unina.it

Roberta Siciliano
roberta.siciliano@unina.it

[1] Department of Economics and Statistics, University of Naples Federico II, Naples, Italy

[2] Department of Physics, University of Naples Federico II, Naples, Italy

[3] Department of Electrical Engineering and Information Technology, University of Naples Federico II, Naples, Italy

Published online: 25 July 2025

1 Introduction

Machine Learning (ML), a branch of Artificial Intelligence, utilizes computational statistics to identify patterns within data. It is an automated analytical approach characterized by iterative algorithms that enable computer systems to learn from experience, minimizing human intervention (Mitchell, 1997). ML models are inherently data-driven, exhibiting a high degree of adaptability to training datasets. This flexibility makes them particularly advantageous for predictive tasks where human perception struggles to discern patterns in large, complex datasets. However, the trade-off between accuracy and interpretability poses a significant challenge, as the complexity of these models often comes at the expense of transparency. This study focuses on supervised ML techniques, particularly ensemble methods, which enhance predictive performance by combining multiple weak learners into a single composite predictor (James et al., 2021). Classification and Regression Trees (CART), introduced by Breiman et al. (1984), are commonly employed as weak learners in these methods. While CART is valued for its interpretability, it often suffers from overfitting, which limits its predictive accuracy. To address this limitation, Bagging (Bootstrap Aggregating) (Breiman, 1996) aggregates predictions from multiple bootstrap samples to reduce variance. Building on Bagging, Random Forest (RF) was introduced by Breiman (2001). RF enhances predictive performance by introducing randomness in the selection of split variables, ensuring that individual decision trees are decorrelated. This randomization mitigates overfitting and leads to robust, reliable models. However, RF and similar ensemble methods are often criticized as "black boxes" due to their lack of interpretability, raising concerns about their trustworthiness in high-stakes decision-making. In such contexts, the concepts of interpretability and explainability become crucial to ensure the transparency and reliability of model predictions. The growing demand for interpretable ML has driven the development of Explainable Machine Learning (a subset of Explainable Artificial Intelligence, or XAI). XAI seeks to illuminate the decision-making processes of complex models, making their predictions comprehensible and trustworthy. In health-related predictive modeling, a systematic integration of interpretable and explainable tools capable of reconstructing the internal logic of such models remains underexplored. This study want to addresses this gap by employing the E2Tree methodology to enhance the transparency of RF applied to mental health prediction. In this work, we apply the RF algorithm to analyze a dataset from the Italian National Institute of Statistics (Istituto Nazionale di Statistica, Istat), specifically the European Health Interview Survey (EHIS), to predict depression in Italy. Subsequently, we employ the Explainable Ensemble Trees (E2Tree) method to interpret the inner workings of the RF model. By providing insights into the decision process, E2Tree aims to uncover the underlying patterns associated with depression and support evidence-based This paper is structured as follows: Section 2 provides an overview of interpretability and explainability, with a specific focus on explainability in the context of RF models. Section 3 details the E2Tree methodology. Section 4 describes the data and presents the analysis results.

2 Interpretability versus explainability

The concepts of interpretability and explainability are fundamental in Machine Learning (ML), especially as models grow in complexity and are applied to critical domains such as finance, healthcare, and public policy (Aria et al., 2024a). Although these terms are often used interchangeably, they represent distinct yet complementary aspects of model transparency. Understanding their differences is essential for addressing the challenges posed by the "black box" nature of many modern ML models. Interpretability refers to the degree to which a human can understand the causal relationships within an ML model (Linardatos et al., 2020). It addresses the question "Why?" a specific output is produced, focusing on the direct associations between inputs and outputs. For example, in a multivariate regression model predicting inflation rates, interpretability allows an economist to analyze the regression coefficients, which represent the causal contributions of independent variables, such as unemployment or monetary policy, to the target variable. This understanding helps evaluate the importance of each factor and ensures that the model aligns with domain knowledge (Gilpin et al., 2018). Interpretability is crucial for building trust, validating predictions, and implementing policy-relevant interventions. It is particularly valued in applications where simplicity and clarity are prioritized over raw predictive power, such as in decision-making contexts that require transparency and accountability. Explainability, on the other hand, goes deeper into the inner workings of an ML model, addressing the question "How?" a specific output is derived. It reveals the underlying processes and logic by which a model transforms inputs into predictions (Linardatos et al., 2020). In the same regression example, explainability involves understanding how the coefficients are mathematically combined in the model's functional form-typically linear and additive relationships-to produce the final prediction. In more complex models, such as RF or Neural Networks, explainability tools like (Shapley Additive Explanations) (Lundberg et al., 2018) or LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro et al., 2016) are often employed to illuminate the intricate decision-making pathways. Explainability is essential for fostering trust in complex systems, enabling error analysis, and addressing ethical concerns, particularly in high-stakes domains like healthcare or criminal justice, where understanding how a decision was reached is as important as the decision itself. The distinction between interpretability and explainability becomes increasingly important in real-world applications where transparency and performance must be balanced (Marcinkevičs & Vogt, 2023). Simpler models, such as linear regression and decision trees, excel in interpretability but often sacrifice predictive accuracy compared to complex approaches like Neural Networks or Ensemble Methods. Conversely, explainability techniques enable these complex models to be more transparent, bridging the gap between high performance and low interpretability. However, achieving explainability can introduce computational and conceptual complexities, making it essential to evaluate the specific needs of each application domain. For example, a policymaker using an ML model to assess poverty risk factors might prioritize interpretability to identify and act on key drivers like income and education. Meanwhile, a medical professional employing a Neural Network for disease diagnosis would rely on explainability tools to understand how the system integrates patient symptoms into a final prediction. Table 1 highlights the fundamental distinctions between these two key concepts in ML transparency.

SHAP, LIME, and E2Tree offer complementary perspectives on model explainability. SHAP provides both global and local feature attributions grounded in game theory, captur-

Table 1 Key differences between interpretability and explainability

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

ing complex interactions but suffering from high computational cost and assumptions of feature independence. LIME focuses on local approximations through interpretable surrogate models, offering intuitive instance-level explanations, though often unstable and lacking global insights. E2Tree differs by reconstructing a simplified global representation of a RF, highlighting its structural decision logic while maintaining predictive performance. It captures how features interact hierarchically across the ensemble, making the overall model behavior more transparent. However, E2Tree's scalability can be limited by the computational burden of computing pairwise co-occurrences, and its framework is currently specific to RF. Together, these methods could form a robust toolkit: SHAP and LIME validate and enrich E2Tree's structural insights, while E2Tree contextualizes and organizes feature contributions identified by SHAP and LIME. Their integration enhances both the explanatory power and interpretability of complex models, especially in high-stakes domains.

2.1 Explaining random forests

The explanation of RF predictions and classifications is predominantly achieved through Variable Importance (VI) methods, which assess the influence of input features on the model's outcomes. Breiman (2001) introduced two widely-used VI metrics: Mean Decrease Impurity (MDI) and Mean Decrease Accuracy (MDA). In a recent simulation study, Vannucci et al. (2024) demonstrated that alternative approaches, such as Conditional VI (Strobl et al., 2008) and Sobol-MDA (Bénard et al., 2022), surpass MDA in capturing causal effects in RF models. Another frequently employed VI method is TreeSHAP (Lundberg et al., 2018), which uses Shapley values from game theory to provide both global and local explanations of tree-based models. In addition to VI methods, the literature highlights rule extraction approaches designed to mimic the predictive performance of opaque models. Aria et al. (2021) conducted a comparative analysis of two rule extraction techniques: Node Harvest (Meinshausen, 2010) and inTrees (Deng, 2019). Their findings indicate that inTrees outperforms Node Harvest in balanced prediction problems across varying dataset sizes. However, for imbalanced datasets, no significant performance difference was observed. A more recent advancement in rule-based explanation is the Transparent Rule Generator-Random Forest (TRG-RF) (Boruah et al., 2023), which ranks the ensemble's trees by accuracy and extracts rules solely from the top-performing trees. Another dimension of RF explanation focuses on reducing the size of the ensemble. Zhang and Wang (2009) proposed identifying an optimal subset of weak learners capable of delivering the same predictive performance as the full RF model. This process involves eliminating redundant trees in favor of most representative ones, thereby enhancing interpretability without compromising accuracy. Similarly, Sagi and Rokach (2020) introduced the Forest-Based Tree (FBT) method. This post-hoc

approach involves extracting conjunctive rules from an ensemble and organizing them into a structured decision tree. Recent work by Aria et al. (2024a) aims to summarize the internal workings of RF into a tree-like structure, providing a clear and interpretable representation of the model's decision-making process. This methodology has also been extended to regression contexts (Aria et al., 2024b). The specifics of this approach will be elaborated in the following section, as it forms the basis for analyzing the case study.

3 Explainable Ensemble Trees recall

The Explainable Ensemble Trees (E2Tree) (Aria et al., 2024a) is a recent proposed methodology designed to explore the decision-making process of tree-based ensemble models. This work will implement this approach to explain RF for predicting a binary response variable. Let *t* be a generic node in a decision tree, and let $u_i = (X_i, Y_i)$ and $u_j = (X_j, Y_j)$ denote two observations from the training set $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of size *n*, where *i* and *j* index individual data points. Each base learner H_b in the RF ensemble $H = \{H_1, H_2, \ldots, H_B\}$ partitions the input space into a collection of terminal nodes. The E2Tree methodology begins by constructing a similarity matrix $O \in \mathbb{R}^{n \times n}$, where each element O_{ij} quantifies how often observations u_i and u_j fall together in a terminal node across the ensemble, weighted by node-level classification accuracy. Formally, the weighted co-occurrence between u_i and u_j is defined as:

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

where $I(u_i \wedge u_j) = 1$ if both observations belong together in the terminal node $t_{ij|b}$ in tree H_b, and $W_{t_{ij}|b}$ denotes the classification accuracy of that node. The terms $W_{t_i|b}$ and $W_{t_j|b}$ refer to the classification accuracy of the terminal nodes containing u_i and u_j, respectively, in tree H_b. The authors set $W_{t_{ij}|b}$ equal to the correct classification rate R_t at a generic node *t*, defined as:

$$R_t = \sum_{i=1}^{n_t} \frac{I(y_i = \hat{y}_t)}{n_t} \tag{2}$$

where n_t is the size of the node t and $\hat{y}_t$ is the classification assigned to the same node. The dissimilarity matrix *D* is subsequently defined as the complement of the co-occurrence matrix *O*, given by: $D = 1 - O$. The explanatory tree is constructed on the basis of data in matrix *D*. The values d_{ij} represent the entries of the dissimilarity matrix *D*, which summarizes the pairwise dissimilarities between observations. Each entry $d_{ij} \in D$ quantifies how dissimilar observations u_i and u_j are, based on their co-occurrence in terminal nodes across the ensemble of trees. This matrix *D* is the core structure used to guide the construction of the E2Tree. The E2Tree algorithm's binary segmentations are also the most similar according to the ensemble method, effectively reconstructing its decision rule. The split variable

is selected based on the decrease in impurity for each candidate, favoring the segmentation that divides matrix D into two submatrices with the minimum dissimilarity.

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 $X > s$. The impurity measure $\bar{d}_t = \sum_i \sum_j \frac{d_{ij}}{n_t(n_t-1)}$ quantifies the average pairwise dissimilarity between all observations within node *t*. More formally, the index sets for *i* and *j* must be restricted to the set of observations belonging to node *t*. The quantity $\bar{d}_{t|s}$ represents the pooled impurity measure after applying the candidate split *s* at node *t*. It is computed as:

$$\bar{d}_{t|s} = \left\{ \left[\frac{1}{n_t} \cdot \left(\frac{\sum_i^{n_{t_L}} \sum_j^{n_{t_L}} d_{ij}}{(n_{t_L} - 1)} + \frac{\sum_i^{n_{t_R}} \sum_j^{n_{t_R}} d_{ij}}{(n_{t_R} - 1)} \right) \right] \middle| \, s \right\} \tag{3}$$

This formula aggregates the within-node dissimilarities in both child nodes t_L and t_R, normalized by their respective sizes, and scaled by the size of the parent node n_t. The goal is to find the split *s* that minimizes $\bar{d}_{t|s}$, thereby producing child nodes that are internally more homogeneous in terms of the pairwise dissimilarity values d_{ij}. This approach generalizes classical impurity measures by incorporating relational information captured by the ensemble model. The decrease in impurity at node *t*, derived from *s*, is calculated as follows:

$$\Delta\bar{d}_{t|s} = (\bar{d}_t - \bar{d}_{t|s}) = \left\{ \bar{d}_t - \left[(\bar{d}_{t_L} \cdot p_{t_L} + \bar{d}_{t_R} \cdot p_{t_R}) \mid s \right] \right\} \tag{4}$$

where $p_{t_L} = \frac{n_{t_L}}{n_t}$ and $p_{t_R} = \frac{n_{t_R}}{n_t}$ are the proportions of observations in the left and right child nodes, respectively. The best split at node t, denoted by s^*, is the one that maximizes the decrease in impurity $\Delta\bar{d}_{t|s^*}$. In addition to the conventional stopping criteria of node size and minimum impurity decrease, an additional criterion is used to stop the segmentation process when child nodes attain a high level of accuracy, such as 95% or 99%. This approach mitigates the risk of uninformative splits that may arise from the random selection of variables in the RF algorithm, leading to an increase in the dissimilarity measure.

Figure 1 illustrates the main steps of the E2Tree method. We start by training a RF model (panel a). For each pair of observations, using the proximity matrix O_i (panel b) derived from RF, we track how often they appear together in a terminal node across all trees. This produces a proximity matrix $\hat{O}_{ij}$ estimated, that reflects how similar observations are, based on the model's internal structure. The idea behind the proximity matrix, is that proximities are calculated for each pair of sample observations. If two observations are in a terminal node in a tree, their proximity is increased by one point. If a pair of observations is often at the same terminal node of several trees, this means that both explain an underlying relationship. Using the matrix $\hat{O}_{ij}$ estimated, E2Tree builds a single representative tree (panel c) that captures the most common patterns across the forest and makes them easier to interpret. The final aim will be to explain the relationship between predictors and response, reconstructing the same structure in $\hat{O}_{ij}$ (panel d) as the proximity matrix O_i output of the decision tree ensemble. This process helps to uncover the main relationships between predictors and the response in a clear, interpretable way. We would like to highlight that the heatmap in panel b shows the proximity matrix obtained from the RF model, which appears dense and affected by background noise. This is a natural result of how the ensemble is built, aggregating co-occurrences from hundreds of trees (e.g., 500), each introducing small variations and overlaps. In contrast, panel d shows the heatmap generated by E2Tree, which is much

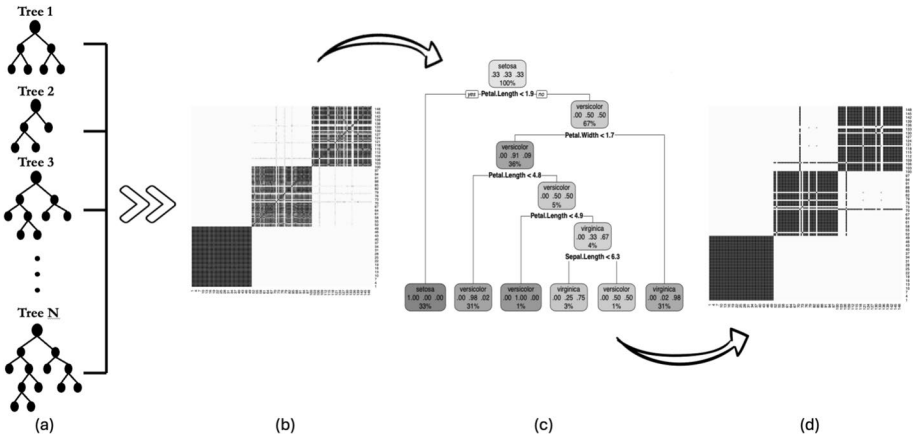


Fig. 1 E2Tree steps

cleaner and structured. This is because E2Tree does not average over multiple trees, but instead reconstructs a single representative tree based on the co-occurrence patterns.

4 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) (Amarasinghe et al., 2023). 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 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 services. 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 indi-

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

viduals 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 2, 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

Table 2 Description and encoding of variables

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

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 (Istat, 1989) 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.

4.2 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 (1997), 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 3 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 (Akosa, 2017).

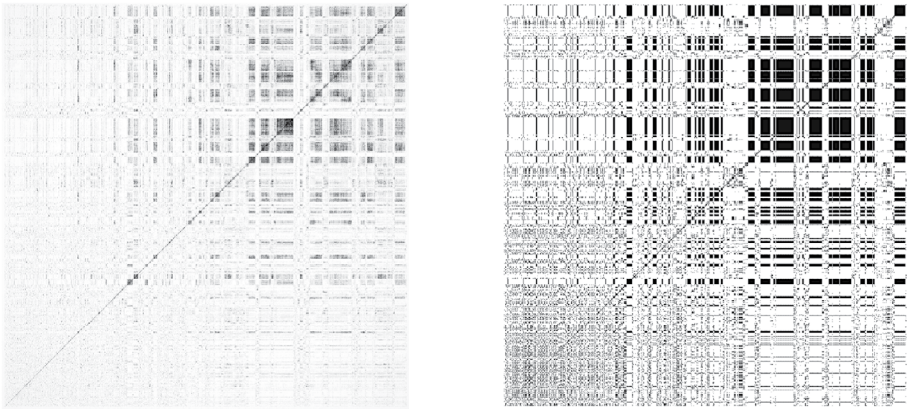
Table 4 presents the evaluation metrics derived from the test set confusion matrix. Balanced Accuracy, which is the average of *sensitivity* (true depressed rate) and *specificity* (true

Table 3 Test set confusion matrix

Observed	Predicted	
	Yes	No
Yes	472	265
No	187	1915

Table 4 Evaluation metrics (values in bold are those discussed in the text)

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)



(a) (b)

Fig. 2 Heatmap of the matrix O_{ij} (sub-plot **a**) and heatmap of the matrix $\hat{O}_{ij}$ estimated by E2Tree (sub-plot **b**). The higher the values in the matrix cells, the darker the colour of the heatmap cells

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 predictions 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.

4.3 Results

The generation of the E2Tree begins with the construction of the matrix O_{ij}, as defined in Eq. (1), derived from the outputs of the RF algorithm. The heatmap in Fig. 2a visually repre-

sents 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, Fig. 2b 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 Fig. 2a, depicting the RF-derived O_{ij}, and

Fig. 2b, 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.

Having 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 Fig. 4, 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 (Mantel, 1967) was conducted to assess the correlation between the pairwise similarity matrices generated by both approaches. The results, depicted in the density

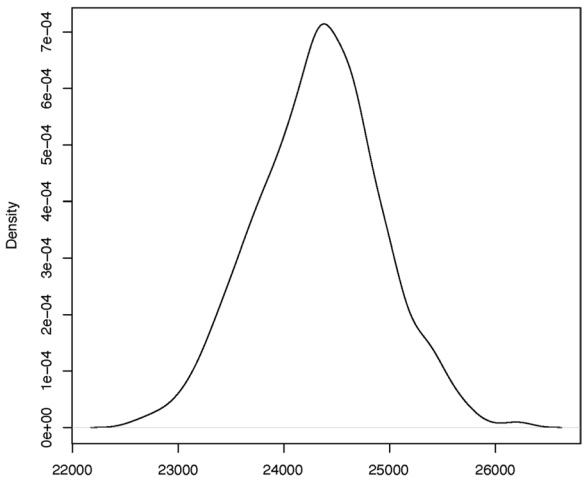


Fig. 3 Density plot of Mantel test permutations ($N = 999$, $Bandwidth = 140$)

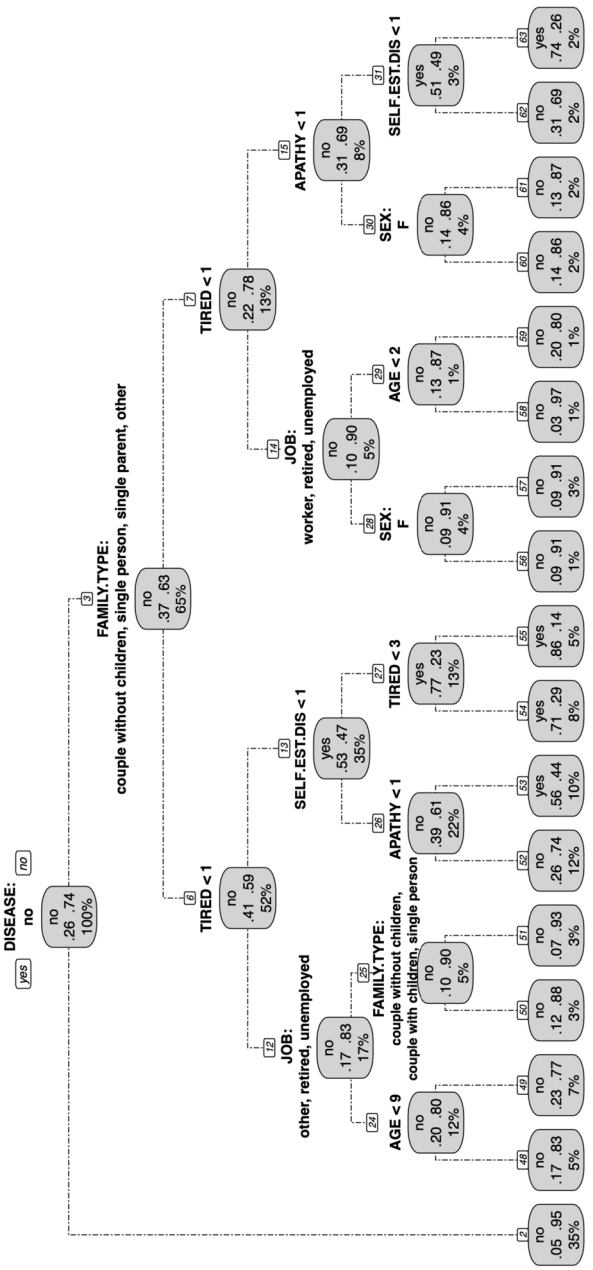


Fig. 4 E2Tree output. Each node displays the proportion of observations belonging to each class of the target variable, as well as the percentage of the total dataset represented by the node

Table 5 Results of E2Tree (a)

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
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

Table 5 (continued)

t	n_t	*Class*	R_t	s^*	*NodeType*	*Path*
53	659	Yes	0.56	–	Terminal	DISEASE = yes & FAMILY.TYPE = (couple without children, other, single parent, single person) & TIRED ≥ 1 & SELF. EST.DIS ≤ 1 & APATHY ≥ 1
27	845	Yes	0.77	TIRED ≤ 3	Non-Terminal	DISEASE = yes & FAMILY.TYPE = (couple without children, other, single parent, single person) & TIRED ≥ 1 & SELF. EST.DIS ≥ 1

plot in Fig. 3, 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 visually represents the explanation tree. Instead, Tables 5 and 6 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 family. 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.

Table 6 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 $\geq$ 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
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 $\geq$ 2
15	498	No	0.69	APATHY $\leq$ 1	Non-Terminal	DISEASE = yes & FAMILY.TYPE = couple with children & TIRED $\geq$ 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 $\leq$ 1 & SEX = M
31	226	Yes	0.51	SELF.EST.DIS $\leq$ 1	Non-Terminal	DISEASE = yes & FAMILY.TYPE = couple with children & TIRED $\geq$ 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

5 Conclusion

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 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.

Acknowledgements This research has been financed by the following research projects: PRIN-2022 "SCIKHEALTH" (Project Code: 2022825Y5E; CUP: E53D2300611006); PRIN-2022 PNRR "The value of scientific production for patient care in Academic Health Science Centres" (Project Code: P2022RF38Y; CUP: E53D23016650001).

Author contributions Massimo Aria: Conceptualization, Methodology, Review & Editing, Supervision. Agostino Gnasso: Conceptualization, Methodology, Writing—Original Draft, Writing—Review & Editing, Software, Formal analysis, Visualization, Supervision. Rebecca Rivieccio: Conceptualization, Methodology, Writing—Original Draft, Writing—Review & Editing, Software, Data Curation, Formal analysis, Visualization. Roberta Siciliano: Writing—Review & Editing.

Funding Open access funding provided by Università degli Studi di Napoli Federico II within the CRUI-CARE Agreement.

Data availability The data are freely available on the platform from where it was extracted, and the dataset in specific can be requested from the corresponding author with a justification for the need to access the dataset.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Akosa, J. (2017). Predictive accuracy: A misleading performance measure for highly imbalanced data. *Proceedings of the SAS Global Forum, 12*, 1–4.

Amarasinghe, K., Rodolfa, K. T., Lamba, H., & Ghani, R. (2023). Explainable machine learning for public policy: Use cases, gaps, and research directions. *Data & Policy, 5*, e5. https://doi.org/10.1017/dap.2023.2

Aria, M., Gnasso, A., Iorio, C., & Fokkema, M. (2024). Extending explainable Ensemble Trees (E2Tree) to regression contexts. arXiv preprint arXiv:2409.06439. https://doi.org/10.48550/arXiv.2409.06439.

Aria, M., Cuccurullo, C., & Gnasso, A. (2021). A comparison among interpretative proposals for random forests. *Machine Learning with Applications, 6*, 100094. https://doi.org/10.1016/j.mlwa.2021.100094

Aria, M., Gnasso, A., Iorio, C., & Pandolfo, G. (2024). Explainable ensemble trees. *Computational Statistics, 39*, 3–19. https://doi.org/10.1007/s00180-022-01312-6

Bénard, C., Da Veiga, S., & Scornet, E. (2022). Mean decrease accuracy for random forests: Inconsistency, and a practical solution via the Sobol-MDA. *Biometrika, 109*, 881–900. https://doi.org/10.1093/biomet/asac017

Boruah, A. N., Biswas, S. K., & Bandyopadhyay, S. (2023). Transparent rule generator random forest (TRG-RF): An interpretable random forest. *Evolving Systems, 14*, 69–83. https://doi.org/10.1007/s12530-022-09434-4

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC Press.

Deng, H. (2019). Interpreting tree ensembles with intrees. *International Journal of Data Science and Analytics, 7*, 277–287. https://doi.org/10.1007/s41060-018-0144-8

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th international conference on data science and advanced analytics (DSAA)*, pp. 80–89. https://doi.org/10.1109/DSAA.2018.00018.

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

Istat. (1989). *Manuale di Tecniche di Indagine 6 - Il Sistema di Controllo della Qualità dei Dati*. Roma: Istituto Nazionale di Statistica.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning*. Springer.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A review of machine learning interpretability methods. *Entropy, 23*(1), 18. https://doi.org/10.3390/e23010018

Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888. https://doi.org/10.48550/arXiv.1802.03888.

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

Marcinkevičs, R., & Vogt, J. E. (2023). Interpretable and explainable machine learning: A methods-centric overview with concrete examples. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 13*(3), e1493. https://doi.org/10.1002/widm.1493

Meinshausen, N. (2010). Node harvest. *The Annals of Applied Statistics, 4*, 2049–2072. https://doi.org/10.1214/10-AOAS367

Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill.
Ribeiro, M.T., Singh, S., & Guestrin, C. (2016) "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778.
Sagi, O., & Rokach, L. (2020). Explainable decision forest: Transforming a decision forest into an interpretable tree. *Information Fusion, 61*, 124–138. https://doi.org/10.1016/j.inffus.2020.03.013
Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics, 9*, 307. https://doi.org/10.1186/1471-2105-9-307
Vannucci, G., Siciliano, R., & Saltelli, A. (2024). Enhancing variable importance in random forests: A novel application of global sensitivity analysis. arXiv preprint arXiv:2407.14194. https://doi.org/10.48550/arXiv.2407.14194.
Zhang, H., & Wang, M. (2009). Search for the smallest random forest. *Statistics and its Interface, 2*, 381. https://doi.org/10.4310/sii.2009.v2.n3.a11