European Journal of Operational Research 323 (2025) 599–608

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Analytics, Computational Intelligence and Information Management

On enhancing the explainability and fairness of tree ensembles

Emilio Carrizosa [a], Kseniia Kurishchenko [b,*], Dolores Romero Morales [b]

[a] Instituto de Matemáticas de la Universidad de Sevilla, Seville, Spain
[b] Department of Economics, Copenhagen Business School, Frederiksberg, Denmark

ARTICLE INFO

Keywords:
(R) Machine learning
Tree ensembles
Explainability
Fairness
Mixed integer linear optimization

ABSTRACT

Tree ensembles are one of the most powerful methodologies in Machine Learning. In this paper, we investigate how to make tree ensembles more flexible to incorporate explainability and fairness in the training process, possibly at the expense of a decrease in accuracy. While explainability helps the user understand the key features that play a role in the classification task, with fairness we ensure that the ensemble does not discriminate against a group of observations that share a sensitive attribute. We propose a Mixed Integer Linear Optimization formulation to train an ensemble of trees that, apart from minimizing the misclassification cost, controls for sparsity as well as the accuracy in the sensitive group. Our formulation is scalable in the number of observations since its number of binary decision variables is independent of the number of observations. In our numerical results, we show that for standard datasets used in the fairness literature, we can dramatically enhance the fairness of the benchmark, namely the popular Random Forest, while using only a few features, all without damaging the misclassification cost.

1. Introduction

The use of Artificial Intelligence (AI) and Machine Learning (ML) to aid Data Driven Decision Making is increasing dramatically. The wide availability of AI/ML algorithms brings important advantages, such as the improved accuracy of decisions and the reduction in the resources required to make them (Athey, 2017; Bertsimas et al., 2022; Jordan & Mitchell, 2015). However, the literature has also reported some controversial/unfair decisions made with AI/ML algorithms when, e.g., assessing the risk of potential recidivism or making social benefit allocations (Rudin, 2019). This, together with the need of users (e.g., physicians, judges, civil servants, citizens) to understand why the model made a decision, calls for enhancing the transparency of Machine Learning algorithms (Blanquero et al., 2020; European Commission, 2020; Goodman & Flaxman, 2017; Panigutti et al., 2023; Rudin et al., 2022). In this paper, we contribute to this stream of literature enhancing the transparency of tree ensembles by means of Optimization (Carrizosa, Molero-Río et al., 2021; Carrizosa & Romero Morales, 2013; Gambella et al., 2021).

Classification trees are seen as the benchmark methodology when pursuing a transparent model (Carrizosa, Molero-Río et al., 2021). As is customary in Supervised Classification, we have observations split into K classes and characterized by p features, either numerical or categorical. A classification tree is defined by a series of if-then queries, which are easy to explain/interpret, in which features are compared against cutoff values. However, classification trees may not be that accurate and they may also suffer from instability, i.e., negligible changes in one feature may yield a rather different accuracy. To overcome these shortcomings, tree ensembles, in which a collection of classification trees are combined (Gambella et al., 2021), have been proposed.

The most common strategies to train tree ensembles are bagging or boosting (Friedman, 2001). In the former one, bootstrapping defines the training sample for each tree, while random sampling on the set of features is used to reduce the number of if-then rules checked in each of the branch nodes. A classic example of this is the Random Forest (Biau & Scornet, 2016; Breiman, 2001). In boosting, a sequential approach is used in which a new classification tree is added in each iteration with the aim to improve the error made by the tree ensemble at hand. A classic example of this is the XGBoost (Chen & Guestrin, 2016), but there are other approaches in the literature based, e.g., on linear programming (Demiriz et al., 2002). By construction, tree ensembles are far less explainable than classification trees, since, in general, almost all features are used in prediction with many cutoffs (Vidal & Schiffer, 2020). Tree ensembles are not flexible enough to incorporate other desirable properties. Of interest to this paper is the case where some of the observations share a sensitive attribute, such as race or low income. One needs to ensure that the classifier does not discriminate against them and/or amplify the biases that may be present in the dataset (Romei & Ruggieri, 2014; Zafar et al., 2017). However, the

* Corresponding author.
E-mail addresses: ecarrizosa@us.es (E. Carrizosa), kk.eco@cbs.dk (K. Kurishchenko), drm.eco@cbs.dk (D. Romero Morales).

https://doi.org/10.1016/j.ejor.2025.01.008
Received 21 September 2023; Accepted 10 January 2025
Available online 16 January 2025

strategies to train tree ensembles above do not allow one to address fairness directly and be able to avoid discrimination.

In this paper, the Explainable and Fair Tree Ensemble (EFTE) methodology is introduced: we assume that we already have a tree ensemble $\mathcal{T}$ at hand, which is modified to, possibly at the expense of a decrease in accuracy, improve its explainability and fairness. The source of these trees can vary. They may have been obtained by training a Random Forest or an XGBoost, or they can simply be a collection of weak learners each of them with a tree structure. In any case, for each classification tree in $\mathcal{T}$, we have the if-then rule associated with each branch node and the class assigned to each leaf node. The EFTE associates a weight to each tree in $\mathcal{T}$. The weights are optimized to ensure a good tradeoff between classification accuracy, explainability, and fairness. Our measure of explainability is sparsity, the typical surrogate (Carrizosa, Molero-Río et al., 2021), and thus we impose an upper bound on the number of features used by EFTE. Our fairness measure is the classification accuracy for the sensitive observations, which we aim to have as high as possible. See Besse et al. (2022), Carrizosa et al. (2022b), Gupta and Kamble (2021), Mehrabi et al. (2022), Miron et al. (2020) for other fairness metrics. Inspired by the models to train Support Vector Machines (Carrizosa & Romero Morales, 2013; Vapnik, 1995, 1998), we model the EFTE using a Mixed Integer Linear Optimization (MILO) formulation, where there are only binary decision variables to model the sparsity, thus enabling scalability in the number of records in the training set. In our numerical results, we show that for standard datasets used in the fairness literature, we can dramatically enhance the fairness of the benchmark, namely the popular Random Forest, while using only a few features, all without damaging the misclassification cost.

The paper is organized as follows. In Section 2 we introduce the EFTE and the MILO formulation. In Section 3 we illustrate the performance of the EFTE in terms of misclassification cost, fairness, and explainability on real-world datasets. In Section 4 we conclude the paper and propose a number of future lines of research. Due to space constraints, we have created a supplementary material where additional numerical results can be found.

2. The explainable and fair tree ensemble

In this section, we introduce the Explainable and Fair Tree Ensemble (EFTE) classifier and a Mixed Integer Linear Optimization (MILO) formulation that is scalable in the number of observations. We start by presenting the information available from the collection $\mathcal{T}$ of classification trees at hand that will be combined to yield an EFTE.

We have a classification problem with K classes indexed by the set $\mathcal{K} = \{1, \dots, \mathrm{K}\}$, defined in a feature space $\mathcal{X} \subset \mathbb{R}^p$. Note that we can handle both numerical and categorical features, where for the latter ones we transform them into 0–1 features using the one-hot encoding. Let $C_{k\,k'} \geq 0$ be the cost incurred when misclassifying an individual from class k in class k', $k \neq k'$. Hereafter, we will refer to them as the unit misclassification costs.

Our methodology requires as a starting point a set $\mathcal{T}$ of T classification trees. Since the trees are in place, we know what features are used in the branch nodes. We use for this the notation f_j^t, $j = 1, \dots, p$ and $t = 1, \dots, \mathrm{T}$, such that f_j^t is equal to 1 if feature j is used at least once in tree t and 0 otherwise. We also know the class assignment rule used by each of the trees. Note that one of the most popular ways to make the class assignment is using the majority rule, where any individual in a given leaf node of the tree is associated with the most frequent class in such a leaf node. We use for this the notation $\psi_t : \mathcal{X} \to \mathcal{K}$, $t = 1, \dots, \mathrm{T}$, where $\psi_t(\boldsymbol{x})$ denotes the class assigned by tree t to datapoint $\boldsymbol{x} \in \mathcal{X}$. We would like to stress that both types of data f_j^t and $\psi_t(\boldsymbol{x})$ are obtained prior to the training of the EFTE.

To build the EFTE we have a training sample $\mathcal{I}$ of $\mathrm{I} = |\mathcal{I}|$ observations, namely, $\{(\boldsymbol{x}_i, k_i)\}_{i \in \mathcal{I}}$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ is the feature vector characterizing observation i and $k_i \in \mathcal{K}$ is its class membership. Recall that we have the so-called sensitive individuals, sharing a sensitive attribute such as race or low income, that we want to protect against an unfair treatment in terms of misclassification cost in the training process. An unfair treatment would normally mean a higher value of the misclassification cost for the sensitive group than for the total. In any case, our approach aims at minimizing the misclassification cost for the sensitive group independently of how this compares to the total cost. Therefore, we introduce notation $\mathcal{I}_S \subset \mathcal{I}$ for the individuals in $\mathcal{I}$ that belong to the sensitive group, with $\mathrm{I}_S = |\mathcal{I}_S|$. As abovementioned, we know the class assigned by each tree to each observation in the training sample, namely, $\psi_t(\boldsymbol{x}_i)$ for $i = 1, \dots, \mathrm{I}$ and $t = 1, \dots, \mathrm{T}$. For each tree t, we define the parameter $y_k^t(\boldsymbol{x}_i)$ that is equal to 1 if observation $i = 1, \dots, \mathrm{I}$ is predicted class k by tree t and 0 otherwise, $i = 1, \dots, \mathrm{I}$, $k = 1, \dots, \mathrm{K}$ and $t = 1, \dots, \mathrm{T}$. This notation will be convenient when defining the class score that EFTE uses to make predictions.

The EFTE has two goals, namely, to achieve a good and fair classification accuracy as well as a good sparsity, a surrogate of explainability. To this aim, we propose a mathematical optimization model to select only a few features in the classifier, to eliminate the trees using features outside the set of selected ones, and to weigh the remaining trees to achieve a good and fair classification accuracy. Therefore, we define the following decision variables and parameters. Let $\omega^t \in [0, 1]$ be a continuous decision variable that models the weight that the EFTE allocates to tree t, $t = 1, \dots, \mathrm{T}$. Let ϕ_j be a binary decision variable equal to 1 if feature j is used in the model and 0 otherwise. Let $\eta \in (0, 1]$ be an upper bound on the maximum weight allocated to each of the trees and $\nu \in \{1, 2, \dots, p\}$ an upper bound on the number of features used by the EFTE. Parameter η can be seen as a regularization parameter on the tree ensemble (Hastie et al., 2009). If $\eta = 1$ only a few trees may be chosen, with the risk of overfitting to the training sample and of malicious manipulation. However, as we decrease the value of this parameter, we force more trees to be part of EFTE decreasing those risks.

Usually, one requires a good classification accuracy by minimizing the misclassification cost in the training sample, hereafter $\mathrm{misclas}(\boldsymbol{\omega}, \mathcal{T}, \mathcal{I})$. In this paper, we also aim to have a fair misclassification cost. Therefore, in addition, we propose to minimize the misclassification cost in the subsample of sensitive individuals of the training sample, i.e., $\mathcal{I}_S$, hereafter $\mathrm{misclas}(\boldsymbol{\omega}, \mathcal{T}, \mathcal{I}_S)$. We follow a weighted approach and combine these two terms using the parameter $\alpha \geq 0$, defining the fair misclassification cost as:

$$\mathrm{fairmisclas}(\boldsymbol{\omega}, \mathcal{T}, \mathcal{I}; \alpha) := \mathrm{misclas}(\boldsymbol{\omega}, \mathcal{T}, \mathcal{I}) + \alpha\, \mathrm{misclas}(\boldsymbol{\omega}, \mathcal{T}, \mathcal{I}_S). \tag{1}$$

With this, our performance measure *fairmisclas* gives weight $1 + \alpha$ to the misclassification cost incurred on an individual of the sensitive group and 1 on an individual outside the protected group. The higher α the more stress we put in ensuring the correct classification in the individuals from the sensitive group.

Once we know the weights ω^t for each tree in the ensemble, the EFTE makes class assignment using $\sum_{t=1}^{\mathrm{T}} \omega^t y_k^t(\boldsymbol{x}_i)$, hereafter referred to as the score associated to class k for individual i. Note that, for a given individual, the scores are non-negative and sum up to one, a property that is inherited from the weights. We assign class $\tilde{k}_i$ to individual i if

$$\tilde{k}_i \in \arg\max_k \sum_{t=1}^{\mathrm{T}} \omega^t y_k^t(\boldsymbol{x}_i).$$

With this, we consider the prediction is correct if

$$\sum_{t=1}^{\mathrm{T}} \omega^t y_{k_i}^t(\boldsymbol{x}_i) \geq \sum_{t=1}^{\mathrm{T}} \omega^t y_k^t(\boldsymbol{x}_i) + \varepsilon \quad \forall k \neq k_i, \tag{2}$$

with $\varepsilon > 0$. Note that this parameter ε is used to model a conservative approach, such that for a record where we have a tie, and thus the difference between the best score and the second best score is below ε, we count this as a misclassification error.

The most straightforward way to measure the misclassification cost would be to introduce binary decision variables for each individual and

each class (Carrizosa, Molero-Río et al., 2021), to check whether the inequalities in (2) are satisfied. However, this hard way of modeling errors is not scalable for large training samples since it includes as many binary decision variables as observations in the data set, and it can overfit the training data. Instead, we propose a soft approach using continuous decision variables $\xi_{ik} \geq 0$, with $k \neq k_i$, to measure the violation of the inequalities in (2). This strategy is similar to the one used to train Support Vector Machines (Carrizosa & Romero Morales, 2013; Vapnik, 1995, 1998).

The second goal of the EFTE is to ensure that only a few features are used, i.e., the classifier is sparse. This is achieved by imposing an upper bound on the number of features used by the EFTE.

The MILO formulation of the EFTE that we will use in the numerical section reads as follows:

$$\min_{\omega,\phi,\xi} \quad \frac{1}{\mathrm{I}} \sum_{i=1}^{\mathrm{I}} \sum_{k \neq k_i} C_{k_i k}\, \xi_{ik} + \alpha \frac{1}{\mathrm{I}_S} \sum_{i=1}^{\mathrm{I}_S} \sum_{k \neq k_i} C_{k_i k}\, \xi_{ik} \tag{3}$$

$$\text{s.t.} \quad \sum_{t=1}^{\mathrm{T}} \omega^t y^t_{k_i}(\boldsymbol{x}_i) \geq \sum_{t=1}^{\mathrm{T}} \omega^t y^t_k(\boldsymbol{x}_i) - \xi_{ik} + \varepsilon, \quad i = 1, \ldots, \mathrm{I},\ k = 1, \ldots, \mathrm{K} : k \neq k_i, \tag{4}$$

$$\sum_{t=1}^{\mathrm{T}} \omega^t = 1, \tag{5}$$

$$\sum_{j=1}^{p} \phi_j \leq \nu, \tag{6}$$

$$\omega^t \leq \eta \phi_j, \quad j = 1, \ldots, p,\ t = 1, \ldots, \mathrm{T} : f^t_j = 1, \tag{7}$$

$$\omega^t \geq 0, \quad t = 1, \ldots, \mathrm{T} \tag{8}$$

$$\phi_j \in \{0, 1\}, \quad j = 1, \ldots, p, \tag{9}$$

$$\xi_{ik} \geq 0, \quad i = 1, \ldots, \mathrm{I},\ k = 1, \ldots, \mathrm{K} : k \neq k_i. \tag{10}$$

Let us discuss the objective function (3) and constraints (4) together. Constraints (4) ensure that ξ_{ik} is well-defined. If $\sum_{t=1}^{\mathrm{T}} \omega^t y^t_{k_i}(\boldsymbol{x}_i) \geq \sum_{t=1}^{\mathrm{T}} \omega^t y^t_k(\boldsymbol{x}_i) + \varepsilon$ for all $k \neq k_i$, then without loss of optimality we can choose $\xi_{ik} = 0$, i.e., there is no misclassification error. However, if this is not the case, then $\xi_{ik} > 0$, for some $k \neq k_i$. Now, it is clear that the objective function (3) minimizes a proxy for the fair misclassification cost with the help of deviation variables ξ_{ik}. Note that the deviations of the protected observations in $\mathcal{I}_S$ are weighed with $1+\alpha$ and the rest with 1, with the goal of being more fair towards those observations. Constraint (5) ensures that the weights ω^t sum up to 1 across the T trees. Therefore ω^t is the fraction of the total weight, and thus the importance, allocated to tree t. Constraint (6) ensures that the EFTE can use at most ν features. Constraints (7) are twofold. First, they ensure that ϕ_j is well-defined, i.e., if feature j cannot be used in the EFTE, namely $\phi_j = 0$, then $\omega^t = 0$ for each tree using that feature. Second, they impose the upper bound η on the weight ω^t, for each t. Constraints (8)–(10) specify the nature of the decision variables $\boldsymbol{\omega}, \boldsymbol{\phi}$ and $\boldsymbol{\xi}$. In sum, EFTE has been formulated as a MILO problem with at most $\mathrm{T}\,p + 2 + \mathrm{I}(\mathrm{K}-1)$ linear constraints, $\mathrm{T} + \mathrm{I}(\mathrm{K}-1)$ non-negative decision variables, and p binary decision variables.

A few remarks can be made about the EFTE formulation (3)–(10). The first one is on the feasibility of the formulation. For small values of ν, the problem may be infeasible. This is probably the case for trees coming from training a random forest in which no pruning has been applied, and therefore there may not be trees using only a few features. This is less of an issue in XGBoost where many trees of very small depth are combined. The same holds for η, namely, for small values of this parameter the problem is infeasible. This will certainly be the case if $\eta < \frac{1}{\mathrm{T}}$, as even by taking the maximum possible value of the weights would violate constraint (5). Note that since we have used a weighted approach to model the misclassification error and the unfairness criteria, the parameter α does not affect the feasibility of the EFTE formulation. The second remark is on the level of fairness we can achieve. This clearly depends not only on the parameter α, the higher the value of this parameter the higher the fairness, but also on the fairness of the individual classification trees in the ensemble $\mathcal{T}$ we start from. If the trees were obtained by training a Random Forest or an XGBoost to have a good classification accuracy, and thus ignoring the fairness criterion as is the case for off-the-shelf statistical packages, we cannot expect that many of these trees will be fair towards the sensitive observations. In our numerical section, the EFTE starts from a collection of weak learners in the form of stump trees (i.e., trees of depth one) such that, for a set of cutoffs, we construct all possible stumps in terms of splitting rules and class predictions. In the extreme case, when for each feature we take as cutoffs the middlepoints between each pair of consecutive observations, this strategy should ensure the existence of fair stumps that the EFTE can use to enhance the fairness. Of course, this may be computationally costly, and in our numerical section we see that we can get good results when we reduce this to using the percentiles.

Once the EFTE has been trained, we make class predictions in new individuals in the following manner. For an individual with feature vector $\boldsymbol{x}$, recall that $y^t_k(\boldsymbol{x})$ is equal to 1 if tree t assigns class k to it, and otherwise 0, $t = 1, \ldots, \mathrm{T}$ and $k = 1, \ldots, \mathrm{K}$. Then, the EFTE predicts class

$$\tilde{k} \in \arg\max_k \sum_{t=1}^{\mathrm{T}} \omega^t y^t_k(\boldsymbol{x}),$$

where in case of ties we can break them randomly. Note that in binary classification, using that the two class scores sum up to one, we can see that the EFTE predicts class $k = 1$ if the corresponding score is above a threshold β, usually $\beta = \frac{1}{2}$, and class $k = 2$ otherwise.

The EFTE has a visual appeal for binary classification when $\mathcal{T}$ is a collection of weak learners in the form of stump trees. Before we can show this, we arrange the stumps in the following way. In all stump trees, and without loss of generality, the left node predicts class $k = 1$ and the right one class $k = 2$, otherwise we can swap the roles of both leaves. Therefore, the stumps only differ on the splitting rule, yielding:

$$\mathcal{T} = \bigcup_{j=1}^{p} (\mathcal{T}^L_j \cup \mathcal{T}^R_j). \tag{11}$$

For a given j, the set $\mathcal{T}^L_j$ contains the stump trees where the splitting rule is of the form $x_j \leq \pi$, while in $\mathcal{T}^R_j$ we have stump trees where the splitting rule is $x_j > \pi$. Note that for j binary we only need to consider $\pi = 0.5$. The class assigned to $\boldsymbol{x}$ is determined by the score of class $k = 1$, namely, $\sum_{t=1}^{\mathrm{T}} \omega^t y^t_1(\boldsymbol{x})$, which, since each tree involves only one feature, can be expressed as

$$\sum_{j=1}^{p} \left(\sum_{t \in \mathcal{T}^L_j} \omega^t \mathbf{1}_{\{x_j \leq \pi\}} + \sum_{t \in \mathcal{T}^R_j} \omega^t \mathbf{1}_{\{x_j > \pi\}} \right), \tag{12}$$

where $\mathbf{1}_{\cdot}$ is the indicator function. We can see that for each j the inner summation in (12) is the contribution of this feature to the score of class $k = 1$, and thus to the classification in class $k = 1$. Fig. 1 illustrates (12) for one of the datasets used in the numerical section, namely, the COMPAS that assesses potential recidivism risk (Angwin et al., 2016). As can be seen from the top left plot, the higher the level of Age the lower the contribution to the score of class $k = 1$ (the defendant is rearrested within 2 years). This is the opposite for the other three features. This means that a young individual with high values in the other three features is predicted to be in class $k = 1$.

We now discuss important extensions of the EFTE. First, our methodology can easily incorporate more sophisticated forms of sparsity. In the current formulation, we control the total number of features used in the EFTE. We can also control the number of features used from a given group (Benítez-Peña et al., 2021; Friedman et al., 2010). This is meaningful, for instance, for categorical features, where for each j categorical we have a group of 0–1 features (one per category) associated with j coming from its one-hot encoding. We may want to

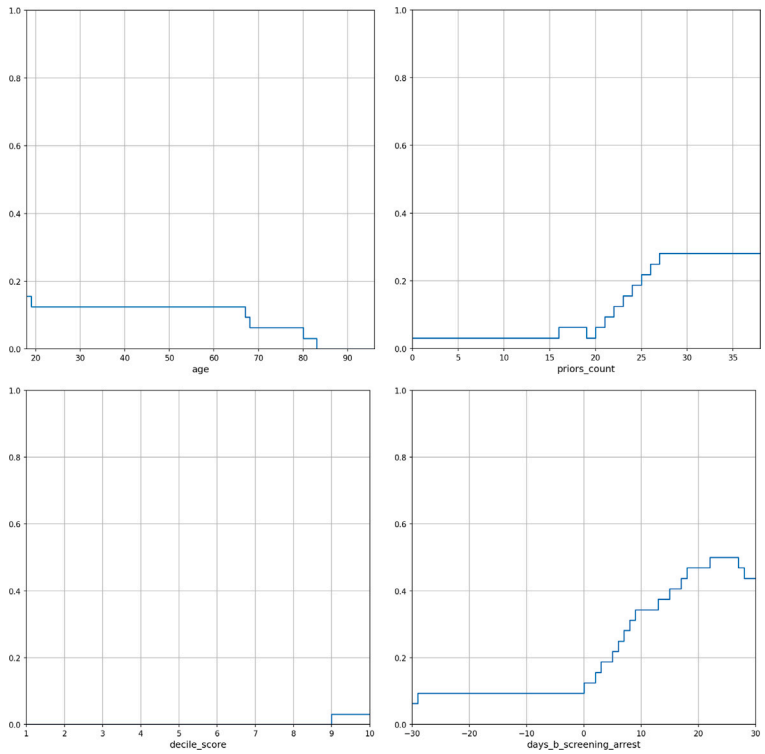


Fig. 1. COMPAS dataset and EFTE with $\alpha = 0.5$, $\nu = 4$, $\epsilon = 0.5$ and $\eta = 0.03125$. Plot of the contribution of feature j to the score of class $k = 1$ in (12). Plots correspond to one of the five Monte Carlo simulations.

impose sparsity within the group, and thus using as few categories as possible from j. Second, although our methodology has been presented for classification problems, building a sparse and fair ensemble tree from an ensemble of classification trees, the EFTE can also deal with regression tasks, where the response variable is a continuous amount. In this case, for each tree in $\mathcal{T}$, we would know the predicted response for each individual, as well as the features used at least once. The EFTE would combine these predictions with the weights of the trees. The goal of the EFTE in regression would be to make these predictions as accurate as possible, as fair as possible, while using as few as possible features. To train this model we need to solve a Mixed Integer Convex Quadratic Problem with linear constraints, where again we only have binary decision variables associated with the selection of features. Indeed, the decision variables are still the weights associated to the trees ω^t and the 0–1 variables ϕ_j to decide which features can be used by the EFTE. The objective function minimizes the mean squared error in $\mathcal{I}$, as opposed to the misclassification cost, and similarly for the sensitive individuals in $\mathcal{I}_S$. As for the feasible region, we only need constraints (5)–(9), as the remaining ones we saw above relate to the definition of the misclassification cost and are not needed in the regression task.

Notice that there is a strand of literature where tree ensembles are being increasingly used outside Machine Learning. This is the case in customer choice modeling (Chen & Mišić, 2022; Mišić, 2020), where a tree represents the product choice of a customer type and the ensemble represents the mix of customers, or in Prescriptive Analytics (Biggs et al., 2023; Biggs & Perakis, 2020), where, for instance, the tree ensemble approximates the (complex) objective function used to make decisions. Needless to say that the optimization formulations in those papers involving tree ensembles, as well as other constraints stemming from the application, do not apply to our setting in which the focus is Supervised Classification and Regression.

3. Numerical results

In this section we illustrate the performance of the EFTE on five publicly available datasets in terms of misclassification error (by taking all unit misclassification costs equal to 1), fairness, and explainability, benchmarking our approach against a very well-known class of tree ensembles, namely, Random Forests (Breiman, 2001). The outline of the remainder of this section is as follows. In Section 3.1, we describe the characteristics of the datasets, in Section 3.2 we present the design of experiments, and we end with Section 3.3 where the results for EFTE and its benchmark are discussed.

3.1. Datasets

We illustrate our methodology on binary classification datasets, i.e., with K = 2 classes, often used in the fairness literature (Le Quy et al., 2022), namely the `COMPAS` (Angwin et al., 2016), the `German credit` (Dua & Graff, 2017), the `Law School` (Komiyama et al., 2018), the `PIMA` diabetes (Dua & Graff, 2017) and the `Adult` (Dua & Graff, 2017) datasets.

The dimension of the datasets is provided in Table 1, including the number of observations (I), the percentage of observations in the sensitive group ($\frac{I_S}{I} \times 100$), the number of features (p), and the class split. Note that we have both numerical and categorical features and therefore p refers to the number of features after the categorical ones have been coded through binary features, having one for each category. The last two columns of Table 1 refer to the out-of-sample misclassification error of a standard Random Forest (RF) with 500 trees of unlimited depth, trained using the *scikit-learn* library (Pedregosa et al., 2011), reported for all individuals as well as for the sensitive individuals only. The feature identified as sensitive is identified below. It is worth noting that $\text{error}_{\text{RF},\mathcal{I}_S} \geq \text{error}_{\text{RF}}$ for the first four datasets,

Table 1
The dimension of the benchmark datasets and the out-of-sample misclassification error of the standard RF.

Dataset	I	$\frac{I_S}{I} \times 100$	p	class split	error_{RF}	error_{RF,I_S}
COMPAS	6172	25%	20	46%/54%	0.36	0.40
German credit	1000	11%	61	70%/30%	0.25	0.52
Law School	20 800	15%	15	89%/11%	0.11	0.24
PIMA	768	35%	8	35%/65%	0.24	0.43
Adult	30162	32%	104	75%/25%	0.15	0.07

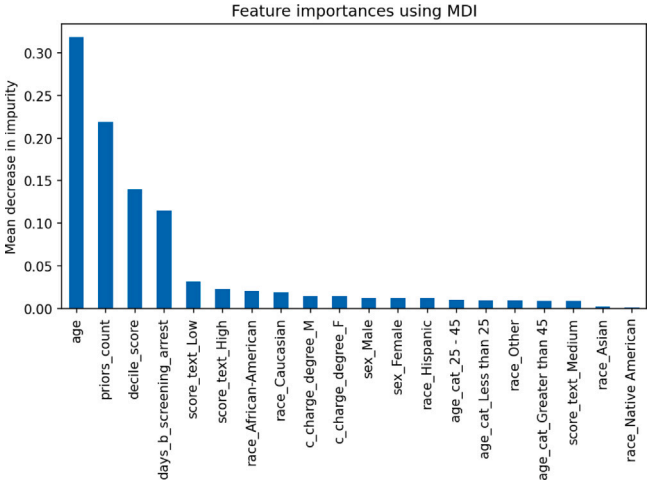


Fig. 2. Variable importance plot for a standard RF trained on the COMPAS dataset.

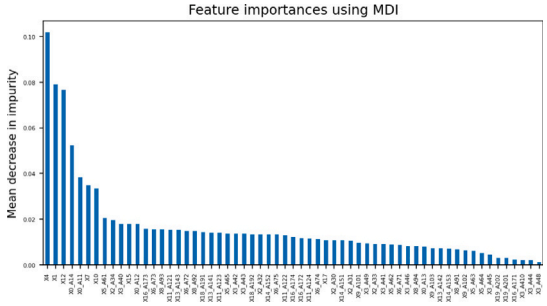


Fig. 3. Variable importance plot for a standard RF trained on the German credit dataset.

while the inequality is the other way around for the Adult dataset. As mentioned in Section 2, our approach works for both cases, as our aim is to minimize the misclassification error for the sensitive group independently of how this compares to the total error.

The COMPAS dataset contains records of crime defendants and is used to assess potential recidivism risk, and the sensitive group is the set of African-Americans not being rearrested within 2 years, i.e., African-Americans in class $k = 2$. The German credit dataset is a credit scoring dataset that is used to predict defaults on consumer loans, and the sensitive group is females with bad credit risk, i.e., females in class $k = 2$. The Law School dataset is a survey among students attending law school in the U.S. in 1991 and is used to predict whether the student passed the bar exam on the first try, while the sensitive group is non-white students. The PIMA dataset contains patient records and is used to predict whether a patient has diabetes, and the sensitive group is the set of individuals with diabetes, i.e., those in class $k = 1$. The Adult dataset is used to predict whether income exceeds $50,000 annually ($k = 1$) or not ($k = 2$), based on census data, while the sensitive group is females. For a detailed description of the features of the datasets, we refer the reader to Tables 2–6.

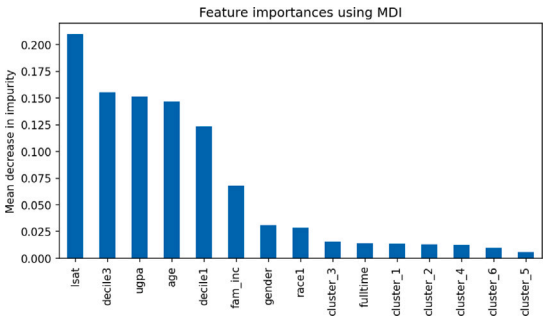


Fig. 4. Variable importance plot for a standard RF trained on the Law School dataset.

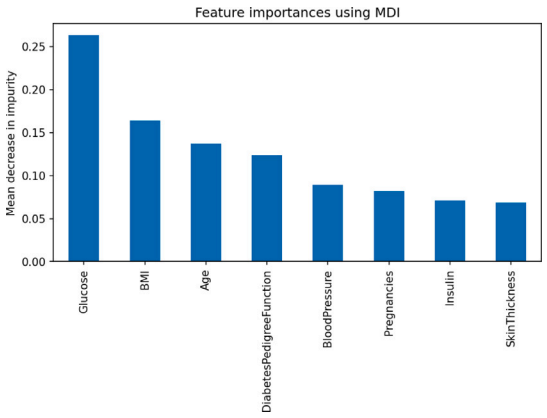


Fig. 5. Variable importance plot for a standard RF trained on the PIMA dataset.

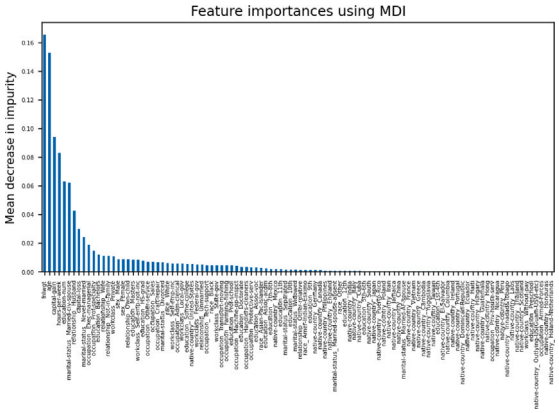


Fig. 6. Variable importance plot for a standard RF trained on the Adult dataset.

Note that in the PIMA dataset we are performing a classical cost-sensitivity analysis (Carrizosa, Molero-Río et al., 2021; Turney, 1995), where EFTE focuses on ensuring that the misclassification error is small for individuals at risk of diabetes, class $k = 1$, while ensuring that the overall misclassification error is also small. This is not the case in the other datasets. Indeed, for the COMPAS dataset, EFTE focuses on those individuals of class $k = 2$ at risk of racial discrimination when predicting recidivism, namely, African-Americans, while in the German credit dataset, it is those in class $k = 2$ at risk of gender discrimination, namely females. For the Law School dataset, we protect non-white students in both classes as in Le Quy et al. (2022), while in the Adult dataset we protect females also for both classes.

Variable importance metrics have been developed to enhance the transparency of Random Forests and other tree ensemble models

Table 2
Description of the features, classes, and sensitive group in the COMPAS dataset.

Features	Description
Sex	Sex
Age	Age in years
Age_cat	Age category
Race	Race
Days_b_screening_arrest	The number of days between COMPAS screening and arrest. If the value is negative, that indicates the screening date happened before the arrest date
Decile_score	A continuous variable, the decile of the COMPAS score
Priors_count	The prior offenses count
C_charge_degree	Charge degree of original crime
Score_text	ProPublica-defined category of decile score
Classes	Defendant is rearrested within 2 years ($k = 1$) or not ($k = 2$)
Sensitive group	African-Americans not being rearrested within 2 years

Table 3
Description of the features, classes, and sensitive group in the German credit dataset.

Features	Description
X0	Status of existing checking account
X1	Duration
X2	Credit history
X3	Purpose
X4	Credit amount
X5	Savings account/bonds
X6	Present employment since
X7	Installment rate in percentage of disposable income
X8	Personal status and sex
X9	Other debtors/guarantor
X10	Present residence since
X11	Property
X12	Age
X13	Installment plans
X14	Housing
X15	Number of existing credits at this bank
X16	Occupation Job
X17	Number of people being liable to provide maintenance for
X18	Telephone
X19	Foreign worker
Classes	Good ($k = 1$) and bad ($k = 2$) credit risk
Sensitive group	Females with bad credit risk

Table 4
Description of the features, classes, and sensitive group in the Law School dataset.

Features	Description
age	The student's age in years
decile1	The student's decile in the school given his grades in Year 1
decile3	The student's decile in the school given his grades in Year 3
fam_inc	Student's family income bracket (from 1 to 5)
lsat	The student's LSAT score
ugpa	The student's undergraduate GPA
gender	Gender
race1	Race
cluster	Encoding the tiers of law school prestige
fulltime	Whether the student will work full-time or part-time
Classes	Whether the student passed the bar exam on the first try ($k = 1$) or not ($k = 2$)
Sensitive group	Non-white students

Table 5
Description of the features, classes, and sensitive group in the PIMA dataset.

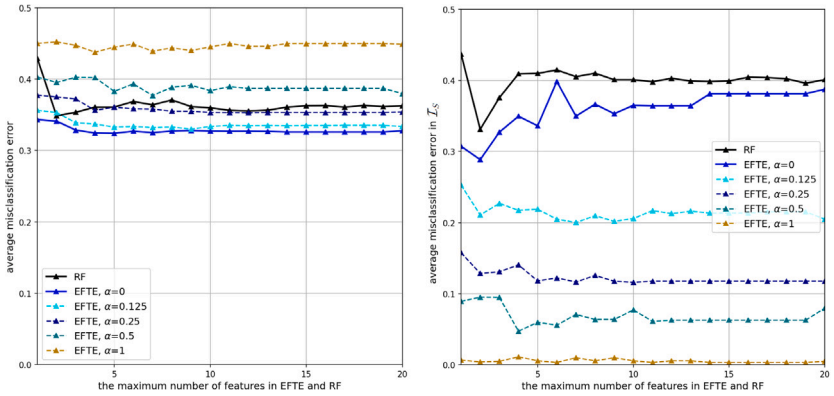
Features	Description
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration a 2 h in an oral glucose tolerance test
BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body mass index (weight in kg/(height in m)^2)
DiabetesPedigreeFunction	Diabetes pedigree function
Age	Age (years)
Classes	Diabetes ($k = 1$) or not ($k = 2$)
Sensitive group	Individuals with diabetes

(Altmann et al., 2010). To give a first impression on the importance of the features for the classification task, we report the variable importance metric that is given by the *scikit-learn* library when training the random forest in Table 1, namely, the Mean Decrease in Impurity (MDI). This can be found in Figs. 2–6. For the COMPAS dataset, from Fig. 2 we can see that Age is the feature with the highest importance, closely followed by Priors_count, Decile_score, and Days_b_screening_arrest. The remaining sixteen features, many of them associated with categorical features such as Race, have a much lower value of the importance metric. For the German credit dataset, Fig. 3 shows that X4 is the feature with the highest importance, and together with X1 and X12 they are the top three of the plot. For the Law School dataset, from Fig. 4 we can see that lsat is the feature with the highest importance, closely followed by decile3, ugpa, age, and decile1. For the PIMA dataset, Fig. 5 shows that Glucose is by far the feature with the highest importance, BMI, Age, and DiabetesPedigreeFunction follow, while the remaining four features have a lower value of importance. For the Adult dataset, Fig. 6 shows that Fnlwgt is the feature with the highest importance, and together with Age, Capital-gain and Hours-per-week they are the top four of the plot.

3.2. The design of experiments

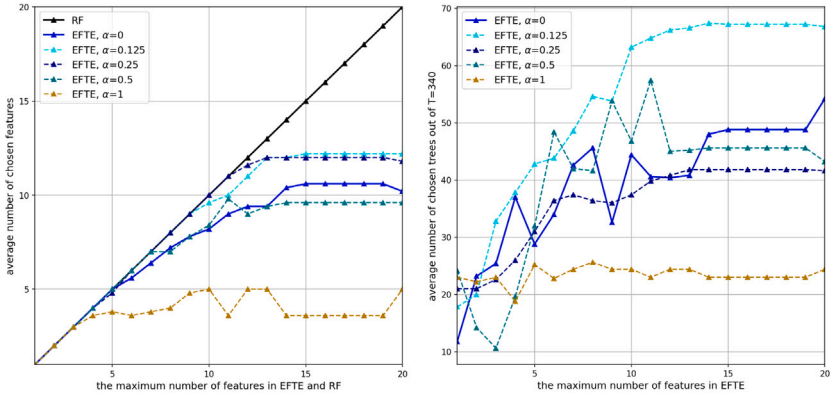
The design of the experiments is as follows. We split the dataset into 3 samples: training (67%), validation (16.5%), and testing (16.5%).

The training sample is used to build and reweight the initial trees. For the EFTE, we consider stump trees as in (11), where recall that we predict class $k = 1$ in the left node and class $k = 2$ in the right node. For each continuous feature, we construct (at most) 200 trees based on the percentiles of the corresponding feature where we split the observations below the percentile from the rest. Given a percentile π, we construct the stump tree with splitting rule $x_j \leq \pi$ and add this tree to $\mathcal{T}^L$, and the one with splitting rule $x_j > \pi$ and add it to $\mathcal{T}^R$. Note that we may have fewer than 200 trees, as there may be repeated trees if the percentiles coincide. For each categorical feature, we build two trees per category. The first one is added to $\mathcal{T}^L$, where the observations showing that category can be found in the left node. In the second stump, which is added to $\mathcal{T}^R$, these observations can be found in the right node.



(a) Misclassification error (b) Fairness

Fig. 7. Out-of-sample misclassification error (left) and out-of-sample misclassification error in sensitive observations (right) in the COMPAS dataset.



(a) Number of used features (b) Number of used trees

Fig. 8. Average number of features (left) and average number of trees (right) used by EFTE in the COMPAS dataset.

The validation sample is used to choose the best values of the EFTE parameters ε and η. We consider $\varepsilon \in \{2^{-3}, 2^{-2}, 2^{-1}\}$ and $\eta \in \{2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^{0}\}$.

We use the testing sample to report the performance of the EFTE in terms of misclassification error, fairness (misclassification error in the sensitive group), and explainability (number of used features and number of used trees).

In our results, $C_{kk'} = 1$ for all $k \neq k'$ and 0 otherwise, and therefore our three criteria are misclassification error, fairness, and explainability. To illustrate the tradeoff between those, we show results for a set of values of the parameters α and ν. Recall that $1 + \alpha$ is the weight we give to the misclassification error incurred in individuals from the sensitive group while this weight is 1 for the remaining individuals, and ν is the maximum number of features that the EFTE can use. We consider $\alpha \in \{0\} \cup \{2^{-3}, 2^{-2}, 2^{-1}, 2^{0}\}$, where the higher the value of α the fairer we are towards individuals in the sensitive group. For ν, we report results for all possible values, i.e., $\nu \in \{1, 2, \ldots, p\}$.

As mentioned earlier, we benchmark our methodology against an RF with 500 trees of unlimited depth. As for the EFTE, we consider the RF with a limited number of features, namely the ν features with the highest value of the variable importance of the standard RF with all features. For instance, when $\nu = 1$, the RF is trained only with Glucose for the PIMA dataset, as seen in Fig. 5. Note that when $\nu = p$ the RF is trained using all the features and thus it coincides with the standard RF, for which the out-of-sample accuracies were reported in Table 1.

Table 6
Description of the features, classes, and sensitive group in the Adult dataset.

Features	Description
Age	Age
Workclass	The employment status
Fnlwgt	Final weight, the number of people the entry represents
Education	Level of education
Education-num	Level of education in numerical form
Marital-status	Marital status
Occupation	The general type of occupation
Relationship	Whether the individual is in a relationship
Race	Race
Capital-gain	Capital gains
Capital-loss	Capital loss
Hours-per-week	The working hours per week
Native-country	The country of origin
Classes	Whether an individual makes more than \$50,000 annually ($k = 1$) or not ($k = 2$)
Sensitive group	Females

To solve the MILO formulation (3)–(10) that trains the EFTE, we use *Gurobi* (Gurobi Optimization, 2020) with *Python* (Python Core Team, 2015) on a PC Intel®Core TM i7-8665U, 16 GB of RAM. Each of the MILO instances, for different training samples and different values of the parameters, was solved to optimality in less than 12 s for the PIMA

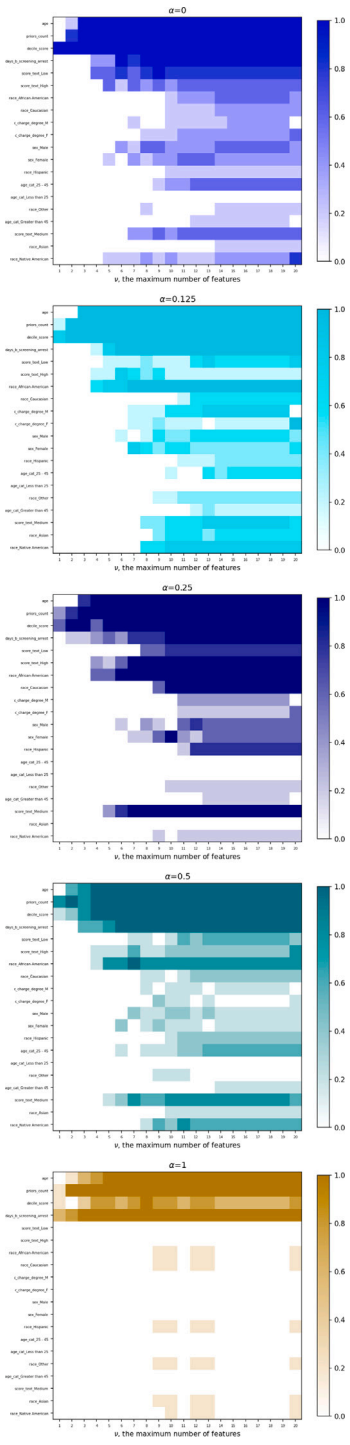


Fig. 9. Heatmap of the average number of folds in which a feature is used by the EFTE in the COMPAS dataset.

dataset, less than 120 s for the COMPAS one; while for the German credit, Law School and Adult datasets some instances reached the time limit of 300 s. Due to space constraints, the discussion of the MIPGAP obtained within the 300 s can be found in the supplementary material, as well as those obtained when imposing a higher time limit.

To end, we run five Monte Carlo simulations and report the average performance for each criterion across all runs. The same training/validation/testing splits are used for both the EFTE and the benchmark. To give an idea about the magnitude of T, the number of classification trees we start from, we report its value for the first run. We have that T is equal to 340 for the COMPAS dataset, 887 for German credit, 284 for Law School, 990 for the PIMA and 928 for the Adult dataset. The values of T for other runs are similar to these. Note that since we are using stump trees and the number of stump trees associated with a given feature is large enough, the trivial solution in which we choose the feature with the largest amount of stump trees and give weight 2^{-5} to the first 2^5 of them and zero to the rest is feasible to the MILO formulation (3)–(10) for any value of ν and η we have tested. This means that all the instances of this formulation are feasible.

3.3. *The results for the* COMPAS *dataset*

This section is devoted to the discussion of the results on misclassification error, fairness, and explainability for the EFTE and its benchmark. Due to space constraints, the explanations below refer to the COMPAS dataset, while those for the German credit, Law School, PIMA and Adult datasets can be found in the supplementary material.

We start with the misclassification error and the fairness of the COMPAS dataset. Fig. 7(a) plots the average out-of-sample misclassification error, while Fig. 7(b) refers to the average out-of-sample misclassification error for the sensitive observations in the different testing samples, our measure of fairness, for the EFTE and its benchmark. Note that the misclassification error of the standard RF, i.e., when all features can be used and given in Table 1, corresponds to the point at the far right of the RF line. It is natural to see in Fig. 7(a) that the larger the value of α the higher the value of the misclassification error of EFTE. The EFTE gives similar results to the RF in terms of out-of-sample misclassification error, or even better, for small values of the parameter α tested, namely, $\alpha \in \{0, 2^{-3}, 2^{-2}\}$, while for $\alpha = 2^{-1}$ the EFTE is only slightly worse. In terms of fairness, we see in Fig. 7(a) that the larger the value of α the lower the out-of-sample misclassification error in the sensitive observations. The EFTE shows better results in terms of out-of-sample misclassification error in the sensitive observations compared to the RF for all values of α. In sum, this means that there are values of α for which the EFTE gives a similar overall misclassification error to RF and is much more fair towards the sensitive group than its benchmark.

In addition, we note that our methodology outperforms RF in terms of fairness for $\alpha = 0$ and in terms of misclassification error. This outperformance may be explained by the fact that we only use a few thresholds of the continuous features as well as a few categories of the categorical features, and thus we are less prone to overfit (Carrizosa et al., 2021a, 2022).

We now discuss the explainability of the EFTE in the COMPAS dataset. We start with the number of features used by the EFTE and by our benchmark. Recall that in the MILO formulation (3)–(10), we have the parameter ν which is an upper bound on the number of features used in the EFTE. Our benchmark is allowed to use the top ν features in terms of feature importance. In Fig. 8(a), we count the actual number of features used and display the average number across the five folds, for each value of ν and α tested. We can see that the benchmark uses exactly ν features, while our methodology uses, in general, fewer features and never more than 12 of the 20 available. Moreover, combining Figs. 7 and 8(a), we can see that the misclassification error and the fairness are very similar for all values of $\nu \geq 4$, meaning that with 4 features we can find a good tradeoff between these two criteria.

To complement Figs. 8(a), 9 displays the average number of folds in which a feature is used. We have a plot for each value of α, where the color coding in each plot is the same as in Fig. 7. To ease the visualization of these averages, we use a heatmap with a cell for each value of ν (horizontal) and each feature (vertical), where the features are in decreasing order of the variable importance metric in Fig. 2. The darker the cell, the more folds are using that feature for the corresponding value of ν. Note that when $\nu = 1$, Age, the feature with the largest value of the variable importance metric in Fig. 2, is never

chosen by the EFTE. This means that for this value of ν the ranking of the variable importance is not necessarily followed by EFTE, but this is the case for other values too.

We continue discussing another explainability metric of the EFTE, namely, the number of stump trees actually used in the tree ensemble, i.e., those for which the continuous decision variable $\omega^t \neq 0$. Fig. 8(b) displays the results for the COMPAS dataset, where we count the number of trees active in the ensemble and plot the average results across the five folds, for each value of α and ν tested. We can see that from the 340 trees we start with, no more than 70 trees are used by the EFTE. Since we are using stumps, this means that only a few thresholds of the (continuous) features play a role in the classification task. This information was visualized in Fig. 1 in Section 2 for a specific choice of the set of parameters.

4. Conclusions

Tree ensembles, such as Random Forest and XGBoost, are popular in Machine Learning. In this paper, we trade off some of the accuracy of the tree ensemble to enhance its sparsity, ensuring that we use fewer features, and its fairness towards a group sharing a sensitive attribute, ensuring that the accuracy in this group is high. This means that the feature selection is not only guided by the overall misclassification cost but also the misclassification cost in the sensitive group. We propose a MILO formulation to train the Explainable and Fair Tree Ensemble (EFTE), where the misclassification cost is modeled through continuous decision variables as opposed to binary ones. Therefore, our formulation has the advantage of being scalable in the number of observations. Our numerical results illustrate the EFTE built from a pool of stump trees. For datasets often used in the fairness literature, we can show that the EFTE dramatically improves the fairness of the ensemble without harming the overall misclassification cost and that this is true even if we use less than half of the features.

As for future research, there are two interesting directions. The first one is about the collection of classification trees $\mathcal{T}$ at hand to train the EFTE. In this paper, we have assumed that this collection is known in advance. For example, in our numerical section we have considered an exhaustive list of weak learners in the form of stump trees obtained from all possible percentiles of the features. A more general problem of interest is the one where we construct the EFTE from scratch, i.e., where we need to decide $\mathcal{T}$ too. The second line of future research is about the fairness measure optimized by the EFTE. We have modeled the misclassification cost in the sensitive group, but there are other criteria we could have considered such as the disparate mistreatment (Miron et al., 2020), as well as fairness metrics beyond classification (Kallus et al., 2022; Kallus & Zhou, 2019). The study of efficient mathematical optimization formulations for these problems is left as an open question.

CRediT authorship contribution statement

Emilio Carrizosa: Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Conceptualization. **Kseniia Kurishchenko:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Conceptualization. **Dolores Romero Morales:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Conceptualization.

Acknowledgments

The authors would like to thank the anonymous reviewers for the comments provided in their reports, which have been very valuable to enhance earlier versions of this paper. This research has been financed in part by research projects EC H2020 MSCA RISE NeEDS (Grant agreement ID: 822214), FQM-329 (Junta de Andalucía), and PID2022-137818OB-I00 (Ministerio de Ciencia, Innovación *y* Universidades, Spain). This support is gratefully acknowledged.

Appendix A. Supplementary data

Supplementary material with additional numerical experiments can be found online at https://doi.org/10.1016/j.ejor.2025.01.008.

References

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. In *Ethics of data and analytics* (pp. 254–264). Auerbach Publications.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, *355*(6324), 483–485.

Benítez-Peña, S., Carrizosa, E., Guerrero, V., Jiménez-Gamero, M. D., Martín-Barragán, B., Molero-Río, C., Ramírez-Cobo, P., Romero Morales, D., & Sillero-Denamiel, M. R. (2021). On sparse ensemble methods: An application to short-term predictions of the evolution of COVID-19. *European Journal of Operational Research*, *295*(2), 648–663.

Bertsimas, D., Pauphilet, J., Stevens, J., & Tandon, M. (2022). Predicting inpatient flow at a major hospital using interpretable analytics. *Manufacturing & Service Operations Management*, *24*(6), 2809–2824.

Besse, P., del Barrio, E., Gordaliza, P., Loubes, J.-M., & Risser, L. (2022). A survey of bias in machine learning through the prism of statistical parity. *The American Statistician*, *76*(2), 188–198.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227.

Biggs, M., Hariss, R., & Perakis, G. (2023). Constrained optimization of objective functions determined from random forests. *Production and Operations Management*, *32*(2), 397–415.

Biggs, M., & Perakis, G. (2020). Dynamic routing with tree based value function approximations. Available At SSRN 3680162.

Blanquero, R., Carrizosa, E., Molero-Río, C., & Romero Morales, D. (2020). Sparsity in optimal randomized classification trees. *European Journal of Operational Research*, *284*(1), 255–272.

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

Carrizosa, E., Galvis Restrepo, M., & Romero Morales, D. (2021a). On clustering categories of categorical predictors in generalized linear models. *Expert Systems with Applications*, *182*, Article 115245.

Carrizosa, E., Galvis Restrepo, M., & Romero Morales, D. (2022b). *Improving fairness of Generalized Linear Models by feature shrinkage*: *Technical Report*, https://www.researchgate.net/publication/358614960_Improving_fairness_of_Generalized_Linear_Models_by_feature_shrinkage.

Carrizosa, E., Molero-Río, C., & Romero Morales, D. (2021). Mathematical optimization in classification and regression trees. *TOP*, *29*(1), 5–33.

Carrizosa, E., Mortensen, L., Romero Morales, D., & Sillero-Denamiel, M. (2022). The tree based linear regression model for hierarchical categorical variables. *Expert Systems with Applications*, *203*(7), Article 117423.

Carrizosa, E., & Romero Morales, D. (2013). Supervised classification and mathematical optimization. *Computers & Operations Research*, *40*(1), 150–165.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785–794).

Chen, Y.-C., & Mišić, V. (2022). Decision forest: A nonparametric approach to modeling irrational choice. *Management Science*, *68*(10), 7090–7111.

Demiriz, A., Bennett, K. P., & Shawe-Taylor, J. (2002). Linear programming boosting via column generation. *Machine Learning*, *46*, 225–254.

Dua, D., & Graff, C. (2017). *UCI machine learning repository*. University of California, Irvine, School of Information and Computer Sciences, URL: http://archive.ics.uci.edu/ml.

European Commission (2020). White Paper on Artificial Intelligence : a European approach to excellence and trust. URL: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. arXiv preprint arXiv:1001.0736.

Gambella, C., Ghaddar, B., & Naoum-Sawaya, J. (2021). Optimization models for machine learning: A survey. *European Journal of Operational Research*, *290*(3), 807–828.

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, *38*(3), 50–57.

Gupta, S., & Kamble, V. (2021). Individual fairness in hindsight. *Journal of Machine Learning Research*, *22*(144), 1–35.

Gurobi Optimization (2020). Gurobi optimizer reference manual. URL: http://www.gurobi.com.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). New York: Springer.

Jordan, M., & Mitchell, T. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Kallus, N., Mao, X., & Zhou, A. (2022). Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, *68*(3), 1959–1981.

Kallus, N., & Zhou, A. (2019). The fairness of risk scores beyond classification: Bipartite ranking and the xAUC metric. *Advances in Neural Information Processing Systems*, *32*.

Komiyama, J., Takeda, A., Honda, J., & Shimao, H. (2018). Nonconvex optimization for regression with fairness constraints. In *International conference on machine learning* (pp. 2737–2746). PMLR.

Le Quy, T., Roy, A., Iosifidis, V., Zhang, W., & Ntoutsi, E. (2022). A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *12*(3), Article e1452.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*, 1–35.

Miron, M., Tolan, S., Gómez, E., & Castillo, C. (2020). Addressing multiple metrics of group fairness in data-driven decision making. arXiv preprint arXiv:2003.04794.

Mišić, V. (2020). Optimization of tree ensembles. *Operations Research*, *68*(5), 1605–1624.

Panigutti, C., Hamon, R., Hupont, I., Fernandez Llorca, D., Fano Yela, D., Junklewitz, H., Scalzo, S., Mazzini, G., Sanchez, I., Soler Garrido, J., & Gomez, E. (2023). The role of explainable AI in the context of the AI act. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency* (pp. 1139–1150). New York, NY, USA.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Python Core Team (2015). *Python: A dynamic, open source programming language*. Python Software Foundation, URL: https://www.python.org.

Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, *29*(5), 582–638.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, *16*, 1–85.

Turney, P. (1995). Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of Artificial Intelligence Research*, *2*, 369–409.

Vapnik, V. (1995). The nature of statistical learning theory. Springer Verlag.

Vapnik, V. (1998). Statistical learning theory. Wiley.

Vidal, T., & Schiffer, M. (2020). Born-again tree ensembles. In *International conference on machine learning* (pp. 9743–9753). PMLR.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web* (pp. 1171–1180).