Enhancing the explainability and fairness of tree ensembles

Mayuri Mhetre
Advanced machine learning
July 10th, 2025

BUSINESS SCHOOL | RWTH AACHEN UNIVERSITY

Outline

Introduction

- Simple decision trees are easy to interpret with “if-then” rules.
- Modern tree ensembles (e.g., random forests, XGBoost) are accurate but harder to explain.
- They often ignore fairness, leading to biased decisions.
- This paper proposes a new method: the **explainable and fair tree ensemble (EFTE)**, which improves both explainability and fairness.

Limitations of tree ensembles

- **Low explainability:**
  - Predictions depend on hundreds of split rules across many trees.
  - Hard to trace why a specific decision was made.
- **No built-in fairness:**
  - Models may unintentionally discriminate based on sensitive attributes like race, gender, or income.
  - Fairness is not part of the training process.
- **No control over feature usage:**
  - Almost all features are often used.

What is explainable and tree ensemble(EFTE)?

- Start with a tree ensemble $\mathcal{T}$ (e.g., random forest).
- Assign a weight to each tree in $\mathcal{T}$.
- Optimize the weights to:
  - Minimize overall misclassification error
  - Minimize error on the sensitive group
- Use of mixed interger linear optimization(MILO) to:
  - Introduce binary variables for feature selection
  - Introduce variables to assign weights to trees in $\mathcal{T}$
  - Soft-margin modeling with continuous slack variables
  - Enables solving on large datasets with commercial solvers like gurobi

EFTE: variable definitions

$\mathcal{I}$	Set of all training observations, $\|\mathcal{I}\| = I$
$\mathcal{I}_S \subset \mathcal{I}$	Subset of sensitive individuals, $\|\mathcal{I}_S\| = I_S$
$\mathcal{K} \in \{1, \ldots, K\}$	Total K classes
k_i	True class label of observation i and $k_i \in \mathcal{K}$
$\mathrm{x}_i \in \mathbb{R}^p$	The feature vector characterizing observation i
$\omega_t \in [0, 1]$	Weight assigned to tree t in the ensemble $\mathcal{T}$
$\phi_j \in \{0, 1\}$	1 if feature j is selected by EFTE, 0 otherwise
$\xi_{ik} \geq 0$	Slack variable for margin violation for observation i, class $k \neq k_i$
$f_j^t \in \{0, 1\}$	1 if tree t uses feature j, 0 otherwise
$y_k^t(\mathbf{x}_i) \in \{0, 1\}$	1 if tree t assigns class k to input $\mathbf{x}_i$
$C_{kk'} \geq 0$	Cost of misclassifying a class k observation as class k'
$\varepsilon > 0$	Margin parameter to enforce conservative classification
$\eta \in (0, 1]$	Upper bound on maximum weight per tree
$\nu \in \{1, \ldots, p\}$	Maximum number of features used (sparsity control)
α	Hyperparameter for sensitive group

Fairness objective and conservation prediction rule

Fair misclassification cost

$$\text{fairmisclas}(\boldsymbol{\omega}, \mathcal{T}, \mathcal{I}; \alpha) := \text{misclas}(\boldsymbol{\omega}, \mathcal{T}, \mathcal{I}) + \alpha \cdot \text{misclas}(\boldsymbol{\omega}, \mathcal{T}, \mathcal{I}_S) \qquad (1)$$

- Combines overall and sensitive-group misclassification costs.
- Parameter $\alpha \geq 0$: higher values increase fairness emphasis.

Conservative prediction rule

$$\sum_{t=1}^{T} \omega_t y_{k_i}^t(\mathbf{x}_i) \geq \sum_{t=1}^{T} \omega_t y_k^t(\mathbf{x}_i) + \varepsilon \quad \forall k \neq k_i \qquad (2)$$

- Ensures the predicted class has a score gap of at least ε.
- Misclassifications are penalized when the margin is too small.

Illustrative example: conservative prediction rule

Problem setting:

- Two classes $\mathcal{K} = \{1, 2\}$, true class $k_i = 1$
- Three trees $T = 3$, weights: $\omega = [0.4, 0.3, 0.3]$, $\varepsilon = 0.05$

Tree predictions:

$$
\begin{aligned}
y^1(\mathbf{x}_i) &= [1, 0] \quad \text{(predicts class 1)} \\
y^2(\mathbf{x}_i) &= [0, 1] \quad \text{(predicts class 2)} \\
y^3(\mathbf{x}_i) &= [1, 0] \quad \text{(predicts class 1)}
\end{aligned}
$$

Class scores (weighted):

$$\text{Score for true class } k_i = 1\text{:} \quad \sum_{t=1}^{3} \omega_t y^t_{k_i}(\mathbf{x}_i) = 0.4 \cdot 1 + 0.3 \cdot 0 + 0.3 \cdot 1 = 0.7$$

$$\text{Score for other class } k = 2\text{:} \quad \sum_{t=1}^{3} \omega_t y^t_{k}(\mathbf{x}_i) = 0.4 \cdot 0 + 0.3 \cdot 1 + 0.3 \cdot 0 = 0.3$$

Illustrative example: conservative prediction rule

Constraint (2):

$$\sum_{t=1}^{T} \omega_t y_{k_i}^t(\mathbf{x}_i) \geq \sum_{t=1}^{T} \omega_t y_k^t(\mathbf{x}_i) + \varepsilon \quad \forall k \neq k_i$$

Apply to example:

$$0.7 \geq 0.3 + 0.05 \quad \Rightarrow \quad \text{Satisfied}$$

Drawback of using only ε:

- This hard-margin constraint must be enforced for every observation and class pair.
- Modeling this directly requires binary variables and scales poorly with large training samples.

Why introduce $\xi_{ik} \geq 0$:

- ξ_{ik} is a slack variable that measures margin violations.
- It enables a soft formulation that uses only continuous variables.

Mixed integer linear optimization(MILO) formulation

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

$$\text{s.t} \sum_{t=1}^{T}\omega_t y_{k_i}^t(\mathbf{x}_i) \geq \sum_{t=1}^{T}\omega_t y_k^t(\mathbf{x}_i) - \xi_{ik} + \varepsilon \quad \forall i = 1,\ldots,I,\ \forall k \neq k_i \tag{4}$$

$$\sum_{t=1}^{T}\omega_t = 1 \tag{5}$$

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

$$\omega_t \leq \eta\cdot\phi_j \qquad \forall t, j \text{ such that } f_j^t = 1 \tag{7}$$

$$\omega_t \geq 0 \qquad \forall t = 1,\ldots,T \tag{8}$$

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

$$\xi_{ik} \geq 0 \qquad \forall i = 1,\ldots,I,\ k \neq k_i \tag{10}$$

Illustrative Example: margin violation ξ_{ik}

MILO constraint (4):

$$\xi_{ik} \geq \sum_t \omega_t y_k^t(x_i) - \sum_t \omega_t y_{k_i}^t(x_i) + \varepsilon$$

Example 1: Misclassified (constraint violated)

- True class: $k_i = 2$
- $\text{Score}_{k=2} = 0.40$, $\text{Score}_{k=1} = 0.45$ $\varepsilon = 0.01$
- Then: $\xi_{i1} \geq 0.45 - 0.40 + 0.01 = 0.06$
- Margin violated, $\xi_{i1} = 0.06$ (penalty)

Example 2: Correctly classified (constraint satisfied)

- True class: $k_i = 2$
- $\text{Score}_{k=2} = 0.60$, $\text{Score}_{k=1} = 0.50$ $\varepsilon = 0.01$
- Then: $\xi_{i1} \geq 0.50 - 0.60 + 0.01 = -0.09$
- Margin satisfied, $\xi_{i1} = 0$ (no penalty)

Stump tree construction

Stump tree has one root node and two leaf nodes. Each stump tree is defined by one feature and a threshold π:

$$\mathcal{T} = \bigcup_{j=1}^{p} \left(\mathcal{T}_j^L \cup \mathcal{T}_j^R\right) \tag{11}$$

Explanation:

- This defines how the collection of trees $\mathcal{T}$ is constructed.
- For each feature $j = 1, \ldots, p$:
  - $\mathcal{T}_j^L$: stumps where $x_j \leq \pi$ assign class 1 (left node)
  - $\mathcal{T}_j^R$: stumps where $x_j > \pi$ assign class 1 (right node)
- The overall set $\mathcal{T}$ is the union of all such stumps, grouped by feature and split direction.

Datasets used

- Five publicly available binary classification datasets are used:
  - **German credit:** Credit risk prediction (gender-based sensitive group)
  - **Law school:** Bar exam pass prediction (race-based sensitive group)
  - **PIMA:** Diabetes diagnosis (patients with diabetes as sensitive group)
  - **Adult:** Income prediction (females as sensitive group)
  - **COMPAS:** Criminal recidivism prediction (race-based sensitive group)

German credit dataset

- Credit scoring dataset and used to predict defaults on consumer loans.
- Sensitive group: females with bad credit risk.

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 1: Description of the features, classes, and sensitive group in the german credit dataset

Law school dataset

- It is a survey of U.S. law students from 1991 used to predict if a student passed the bar exam on their first attempt.
- Sensitive group: non-white students.

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 2: Description of the features, classes, and sensitive group in the law school dataset

PIMA dataset

- Contains patient records and is used to predict whether a patient has diabetes.
- Sensitive group: the set of individuals with diabetes.

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

Table 3: Description of the features, classes, and sensitive group in the PIMA dataset

Adult dataset

- Used to predict whether income exceeds $50,000 annually.
- Sensitive group: females.

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

Table 4: Description of the features, classes, and sensitive group in the adult dataset

COMPAS dataset

- Contains record of crime defendants and predicts if defendants will be rearrested within 2 years.
- Sensitive group: African-Americans who were not rearrested.

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 5: Description of the features, classes, and sensitive group in the COMPAS dataset

Experimental design for random forest(RF)

- Each dataset is split into:
    - 67% training, 16.5% validation, 16.5% test
- Random forest(RF) with:
    - 500 trees
    - Unlimited depth
    - Trained using all features
- Evaluated on:
    - Overall misclassification error
    - Misclassification error on sensitive group (fairness)

Performance of random forest

- Number of samples and features
- Size of the sensitive group
- Binary class distribution
- RF overall error and sensitive group error

Dataset	I	$\frac{I_s}{I} \times 100$	p	class split	error_{RF}	$\text{error}_{RF,\mathcal{I}_s}$
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

Table 6: Comparison of RF performance on overall vs. sensitive group

Feature importance using mean decrease in impurity

- Mean decrease in impurity(MDI) measures how much each feature contributes to reducing impurity (e.g., gini impurity) across all trees in the forest.
- MDI for a feature is calculated using the following steps:
    - Train a random forest model.
    - For each node in each tree, compute the impurity before and after the split based on a feature.
    - Calculate the impurity decrease and weight it by the proportion of samples reaching the node.
    - Sum the weighted impurity decreases for each feature across all trees.
    - Average the total over all trees to obtain the MDI.
    - Normalize the MDI scores so they sum to one.
- Features with higher MDI values are considered more important for the model's predictions.

Illustrative example: MDI with gini impurity

Gini impurity formula:

$$\text{Gini} = 1 - \sum_{i=1}^{C} p_i^2$$

where p_i is the proportion of class i in the node, and C is the number of classes.

ID	Gender	Income	Buys product
1	Male	High	Yes
2	Female	Low	No
3	Female	High	Yes
4	Male	Low	No
5	Male	High	Yes
6	Female	Medium	No

Table 7: Sample dataset used for illustration

Illustrative example: MDI with gini impurity

Step 1: Root gini (before split)

3 yes, 3 no $\Rightarrow$ gini $= 1 - (3/6)^2 - (3/6)^2 = 0.5$

Step 2: Split by gender

- Male: [yes, no, yes] $\rightarrow$ gini $= 1 - (2/3)^2 - (1/3)^2 = 0.444$
- Female: [no, yes, no] $\rightarrow$ gini $= 1 - (1/3)^2 - (2/3)^2 = 0.444$

Weighted gini after split $= 0.444$
Impurity decrease (gender) $= 0.5 - 0.444 = 0.056$

Step 3: Split male by income

- High: [yes, yes] $\rightarrow$ gini $= 0$
- Low: [no] $\rightarrow$ gini $= 0$

Weighted gini $= 0$, Impurity decrease (income-male) $= 0.444$

Illustrative example: MDI with gini impurity

Step 4: Split female by income

- Low: [no] $\rightarrow$ gini $= 0$
- Medium: [no] $\rightarrow$ gini $= 0$
- High: [yes] $\rightarrow$ gini $= 0$

Weighted gini $= 0$, Impurity decrease (income-female) $= 0.444$

Total impurity decrease:

- Gender $= 0.056$
- Income $= 0.444 + 0.444 = 0.888$

Normalized MDI:

- Gender $= 0.056/(0.056 + 0.888) = 0.059$
- Income $= 0.888/(0.056 + 0.888) = 0.941$

Feature importance MDI on COMPAS dataset

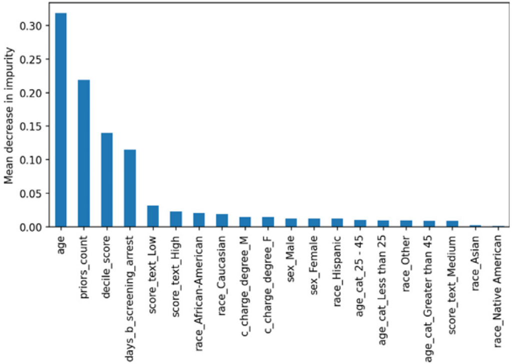


Figure 1: Feature importance using MDI

- `Age` has the highest importance, followed by `priors_count`, `decile_score`, and `days_b_screening_arrest`.
- The remaining 16 features, including categorical ones like `Race`, show significantly lower importance values.

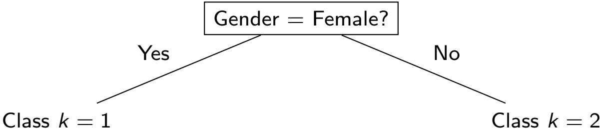
EFTE Experimental setup

- Five monte carlo simulations performed.
- $\epsilon \in \{2^{-3}, 2^{-2}, 2^{-1}\}$
- $\eta \in \{2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 1\}$
- To solve the MILO formulation, Gurobi with Python was used.
- For each continuous feature:
  - We build up to **200 trees**.
  - We choose **percentile values** (e.g., 5%, 10%, ..., 95%) as split points.
  - Example: For age with split $\pi = 30$: tree 1: age $\leq 30 \rightarrow$ left node, age $> 30 \rightarrow$ right node.

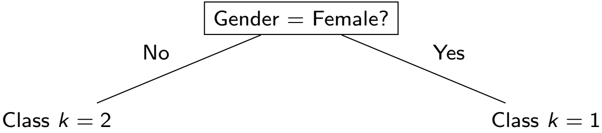
EFTE Experimental setup

For categorical feature: **Gender** {**male, female**}.

Tree 1: Female → Left node



Tree 2: Female → Right node



- We build two trees per category.

COMPAS results: overall misclassification error

- For $\alpha = 0, 0.125$, and 0.25, EFTE achieves comparable or even lower overall misclassification error than RF.
- For $\alpha = 0.5$ and 1, EFTE shows a slight increase in error.

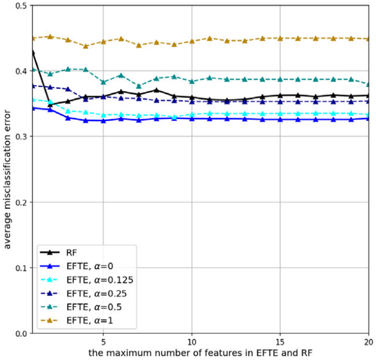


Figure 1: Overall misclassification error

COMPAS results: fairness on sensitive group

- EFTE significantly reduces misclassification error for the sensitive group as α increases, outperforming RF across all α levels.

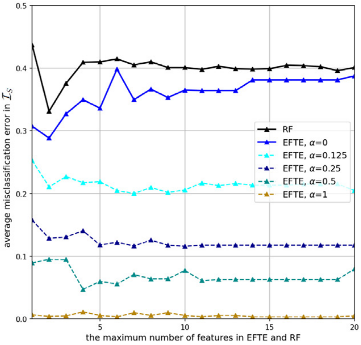


Figure 2: Fairness on sensitive group

COMPAS results: number of features used

- EFTE never uses more than 12 out of the 20 available features, showing effective sparsity.

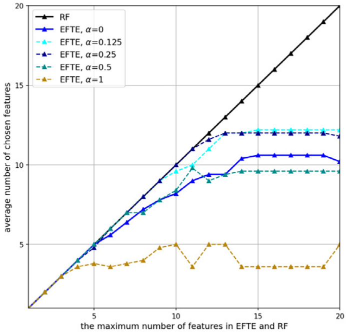


Figure 3: Average number of features used by EFTE

COMPAS results: number of active trees

- Initially, we start with $T = 340$ for the COMPAS dataset.
- EFTE uses no more than 70 out of the 340 available stump trees, highlighting its simplicity and explainability.

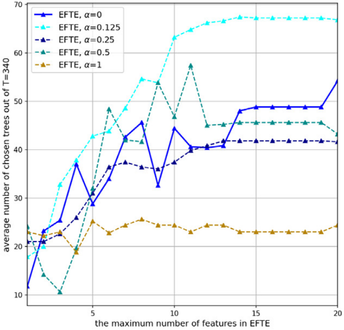


Figure 4: Average number of trees used by EFTE

Conclusion

- EFTE is highly efficient and effective, especially with limited features.
- Numerical experiments on 5 datasets show that EFTE consistently outperforms random forest in terms of fairness, while maintaining competitive accuracy.
- MILO formulation ensures scalability, regardless of the size of the dataset.

Future work

- Deciding the initial tree $\mathcal{T}$, instead of fixing it in advance.
- Exploring more fairness definitions (e.g., disparate mistreatment).
  - For example, if group A has a higher false positive or false negative rate than group B, the classifier suffers from disparate mistreatment.
- Extending EFTE to multiclass and regression problems.
- Improving MILO solver speed for large scale datasets.

References

- Carrizosa, E., Kurishchenko, K., Romero Morales, D. (2025). On enhancing the explainability and fairness of tree ensembles. https://www.sciencedirect.com/science/article/pii/S0377221725000335
- Supplementary material: `https://ars.els-cdn.com/content/image/1-s2.0-S0377221725000335-mmc1.pdf`
- Scikit-learn documentation for mean decrease in impurity: `https://scikit-learn.org/stable/modules/permutation_importance.html`

Slides contributions

- All slides and content prepared and presented by Mayuri Mhetre
- Matriculation number: 468178