CTU
CZECH TECHNICAL UNIVERSITY IN PRAGUE

Improving the validity of Decision Trees as Explanations

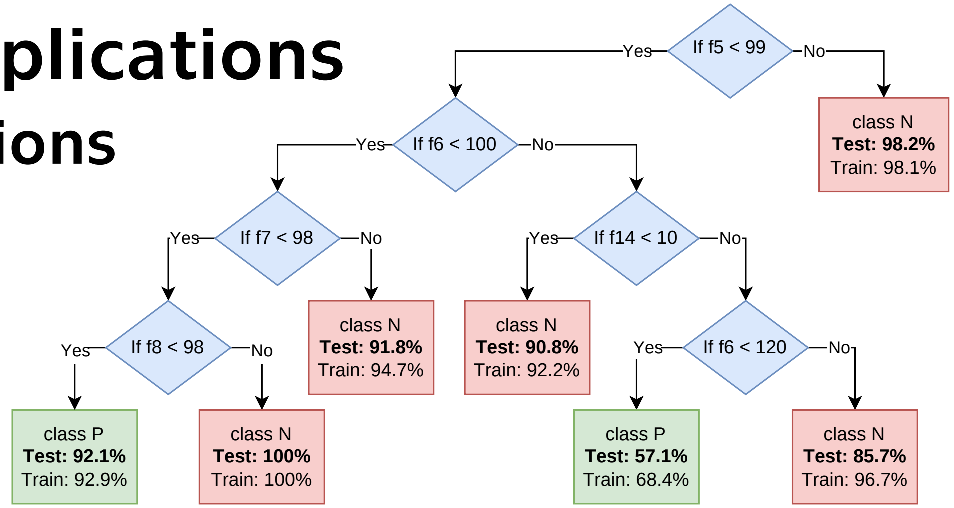
Jiří Němeček, Tomáš Pevný, Jakub Mareček

contact@nemecekjiri.cz

arXiv:2306.06777

What are we talking about?

- Decision trees
  - Classification
- Explainability applications
  - Univariate decisions
  - Shallow



Funded by the European Union

Decision Trees

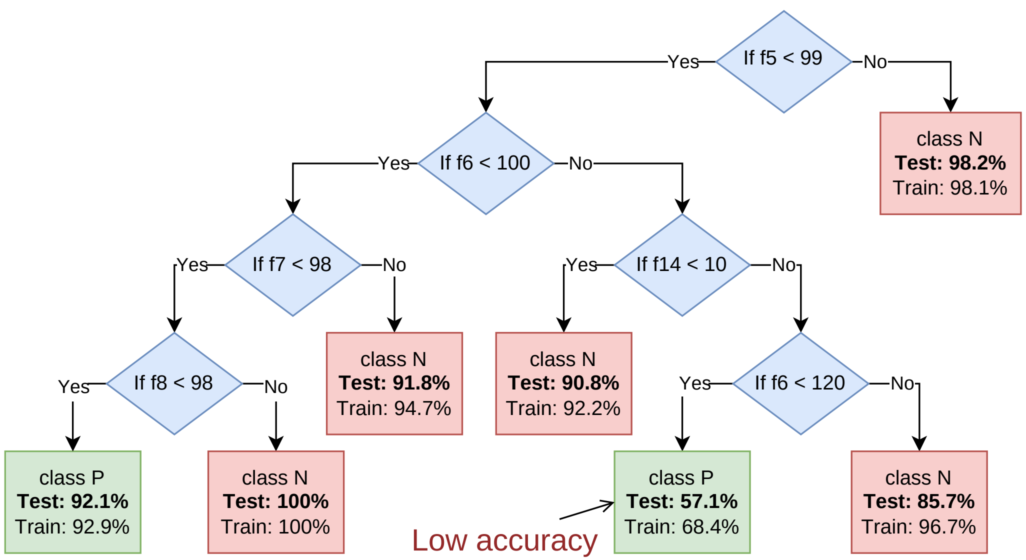
- Global optimality NP-hard
- Heuristic algorithms
  - Good empirically
  - Greedy top-down
    - Information gain
    - Gini impurity

  \+ Pruning

Funded by
the European Union

The problem

- CART creates a leaf with low accuracy
  → Misleading (~unfair) explanation



Funded by
the European Union

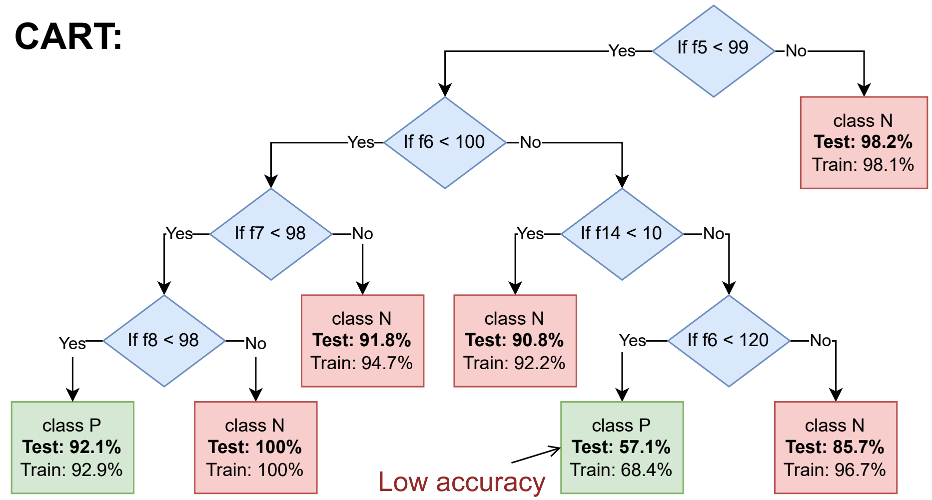
CTU
CZECH TECHNICAL UNIVERSITY IN PRAGUE

Proposed solution

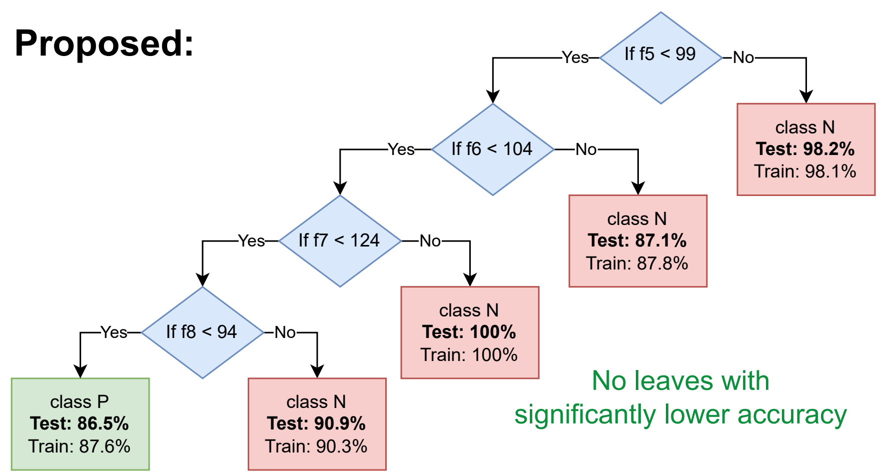
- Maximize leaf accuracy
  - Minimal accuracy across leaves

$$A_L(T) := \min_{l \in \mathcal{L}(T)} \frac{1}{|X_l|} \sum_{(\boldsymbol{x},y) \in X_l} [\![y = C_l]\!]$$

CART:



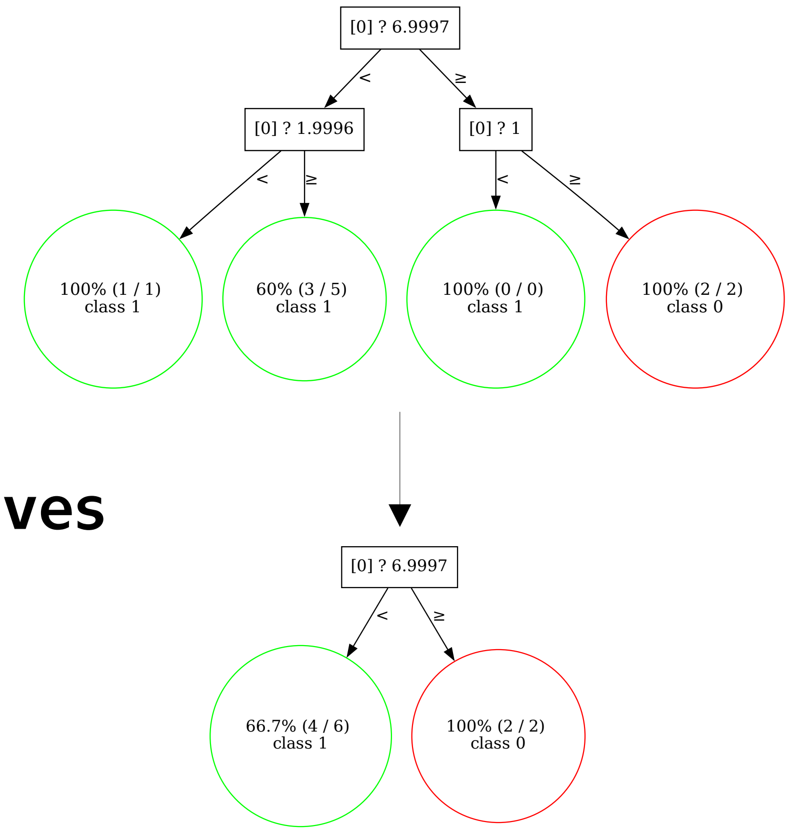
Proposed:



Funded by the European Union

Training Process

- Create a Tree
    - MIP formulation
- Reduce
    - Remove redundant leaves
- Extend leaves
    - Any ML model
    - Improve the total model accuracy



Funded by
the European Union

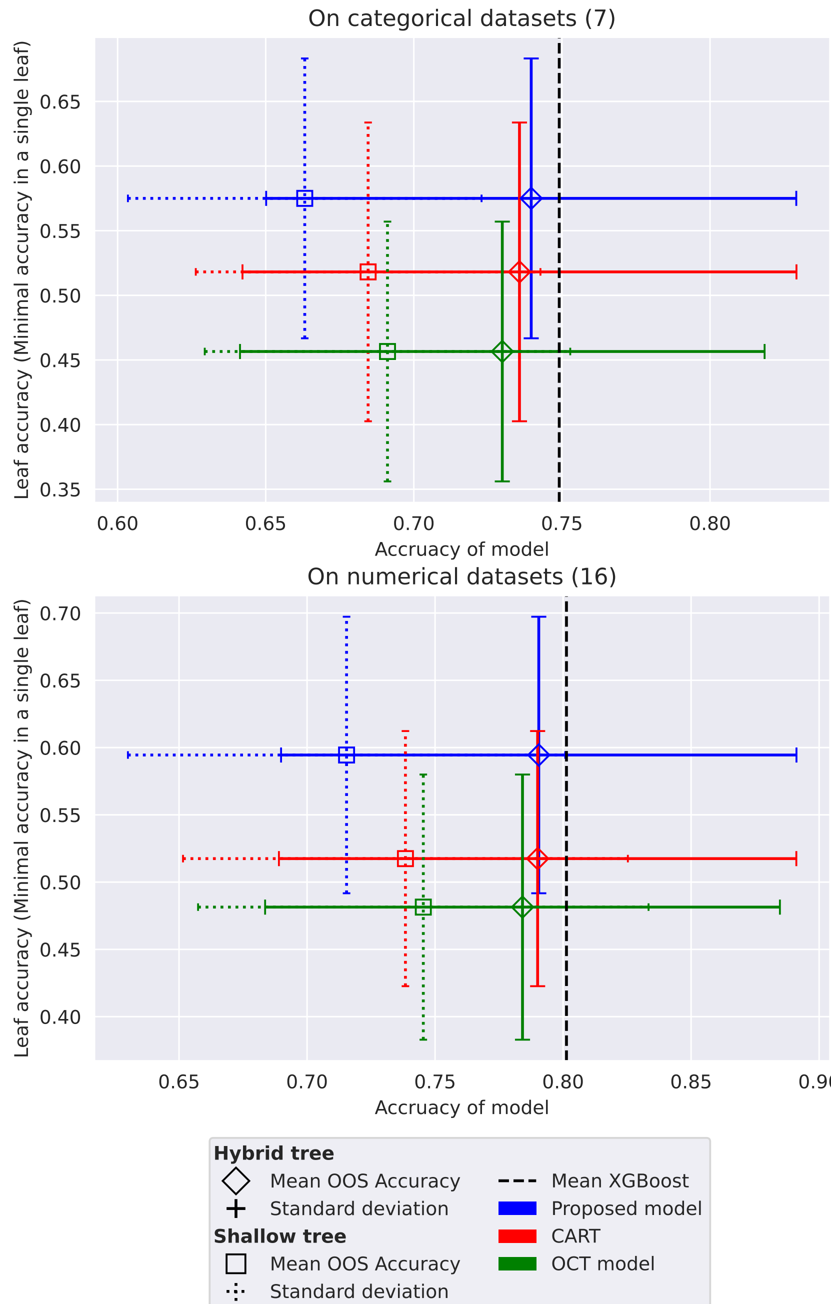
MIP formulation

- Based on Optimal Classification Tree (OCT) formulation [Bertsimas and Dunn, 2017]
- Change the objective to Leaf Accuracy
  - Details in the paper

Funded by the European Union

Results

- Categorical and Numerical tabular data [Grinsztajn et al., 2022]
- At least 50 samples in each leaf
- Tree depth = 4
- 10 random runs
- MIP time limited to 8 hours



CTU
CZECH TECHNICAL UNIVERSITY IN PRAGUE

Summary

- Top-down algorithms for trees can make unbalanced leaves (in terms of accuracy)
- Maximizing leaf accuracy improves this
- Trade-off between leaf and model accuracy
- When extended, the model has comparable performance + added explainability

- MIP limitations (dataset, tree depth)

Funded by
the European Union

CTU
CZECH TECHNICAL UNIVERSITY IN PRAGUE

Questions?

Improving the Validity of Decision Trees as Explanations

Jiří Němeček, Tomáš Pevný, Jakub Mareček

contact@nemcekjiri.cz

https://arxiv.org/abs/2306.06777

Work was funded by the AutoFair project
https://doi.org/10.3030/101070568

Funded by
the European Union

The access to the computational infrastructure
of the OP VVV funded project
CZ.02.1.01/0.0/0.0/16_019/0000765
"Research Center for Informatics"
is also gratefully acknowledged.

References

- Dimitris Bertsimas and Jack Dunn. **Optimal classification trees.** *Machine Learning*, 106(7):1039–1082, July 2017. ISSN 1573-0565. doi: 10.1007/s10994-017-5633-9.
- Léo Grinsztajn, Edouard Oyallon, and Gael Varoquaux. **Why do tree-based models still outperform deep learning on typical tabular data?** *In Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=Fp7 phQszn

Funded by the European Union

CTU
CZECH TECHNICAL UNIVERSITY IN PRAGUE

Tabulated results

	Data Type	Min	Mean ($\pm$ std)	Max
Compared to CART				
Leaf Accuracy	categorical	-0.0142	0.0569 ± 0.0533	0.1206
	numerical	-0.0061	0.0770 ± 0.0556	0.1841
Hybrid-tree Acc.	categorical	-0.0078	0.0040 ± 0.0071	0.0147
	numerical	-0.0244	0.0004 ± 0.0082	0.0087
Compared to XGBoost				
Hybrid-tree Acc.	categorical	-0.0228	-0.0095 ± 0.0064	-0.0036
	numerical	-0.0276	-0.0108 ± 0.0076	0.0005

Funded by
the European Union