Multi-Class Classification With Reject Option and Performance Guarantees Using Conformal Prediction

Alberto García-Galindo[1,2], Marcos López-De-Castro[1,2] and Rubén Armañanzas[1,2]

[1] Institute of Data Science and Artificial Intelligence (DATAI), Universidad de Navarra, Spain
[2] TECNUN School of Engineering, Universidad de Navarra, Spain

13th Symposium on Conformal and Probabilistic Prediction with Applications

September 9-11, 2024

SUMMARY.

INTRODUCTION.

INTRODUCTION.

- Beyond the common classification scenario, an interesting alternative in safety-critical and high-risk applications is classification with reject option.

INTRODUCTION.

- Beyond the common classification scenario, an interesting alternative in safety-critical and high-risk applications is classification with reject option.

- Key idea: only a prediction is made when the model is confident enough.

INTRODUCTION.

- Beyond the common classification scenario, an interesting alternative in safety-critical and high-risk applications is classification with reject option.

- Key idea: only a prediction is made when the model is confident enough.

- The central task lies in developing suitable rejection mechanisms.

INTRODUCTION.

- Beyond the common classification scenario, an interesting alternative in safety-critical and high-risk applications is classification with reject option.

- Key idea: only a prediction is made when the model is confident enough.

- The central task lies in developing suitable rejection mechanisms.

- In this setting, achieving performance guarantees for different rejection rates becomes valuable in decision-making.

INTRODUCTION.

- Beyond the common classification scenario, an interesting alternative in safety-critical and high-risk applications is classification with reject option.
- Key idea: only a prediction is made when the model is confident enough.
- The central task lies in developing suitable rejection mechanisms.
- In this setting, achieving performance guarantees for different rejection rates becomes valuable in decision-making.
- This paper extends previous works on classifiers with reject option and statistical guarantees grounded on the conformal prediction framework.

Previously...

- Binary classification
- Accuracy and precision estimation

This paper...

- Multi-class classification
- Accuracy and recall estimation

Also at COPA24...

- Regression with reject option 😊

BACKGROUND.

BACKGROUND.

- **Multi-class classifier with reject option**

$$h : \mathcal{X} \rightarrow \Delta^J$$

Input space — *Probability simplex*

$$g : \mathcal{X} \rightarrow \{0, 1\}$$

Input space — *Binary rejection*

$$(h, g)(x) = \begin{cases} \arg\max_{\bar{c}_j} \pi_j & \text{if } g(x) = 1 \\ \text{reject} & \text{if } g(x) = 0 \end{cases}$$

BACKGROUND.

- **Multi-class classifier with reject option**

$$h : \mathcal{X} \to \Delta^J$$

Input space — *Probability simplex*

$$g : \mathcal{X} \to \{0, 1\}$$

Input space — *Binary rejection*

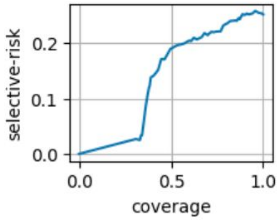
$$(h,g)(x) = \begin{cases} \arg\max_{\bar{c}_j} \pi_j & \text{if } g(x) = 1 \\ \text{reject} & \text{if } g(x) = 0 \end{cases}$$

- **Selective risk**

Risk on the non-rejected samples

$$R(h,g) = \frac{\mathbb{E}_{\mathcal{X}\mathcal{Y}}[\ell(h(x), y) g(x)]}{\mathbb{E}_{\mathcal{X}}[g(x)]}$$

Predictive coverage: Rate of non-rejected samples



BACKGROUND.

BACKGROUND.

- **Calibrated classifier**

$$\mathbb{P}(y = c_j \mid \pi_j) = \pi_j$$

BACKGROUND.

- **Calibrated classifier**

$$\mathbb{P}(y = c_j \mid \pi_j) = \pi_j$$

- **Platt scaling**

$$\hat{\mathbb{P}}(y = c_j \mid \pi_j) = \frac{1}{1 + e^{w_0 + \pi_j w_1}}$$

BACKGROUND.

- **Calibrated classifier**

$$\mathbb{P}(y = c_j \,|\, \pi_j) = \pi_j$$

- **Platt scaling**

$$\hat{\mathbb{P}}(y = c_j \,|\, \pi_j) = \frac{1}{1 + e^{w_0 + \pi_j w_1}}$$

- **Conformal classification** (at a glance)

A non-conformity function $\mathcal{S} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is used to quantify the degree of strangeness of a new sample z_i compared to a set of (labelled) samples $\{z_1, \ldots, z_n\}$.

In practice, the strangeness of z_i is quantified through the ability of a model learned on $\{z_1, \ldots, z_n\}$ to precisely predict its true label.

In the inductive scheme, the non-conformity scores are computed on a hold-out calibration dataset.

For a new sample x_{new}, we tentatively label it with each possible label $\bar{c}_j$ and calculate a valid *p*-value to evaluate the hypothesis that $\bar{c}_j$ is the actual label y_{new}.

These *p*-values are employed to create a set predictor such that $\Gamma_{\text{set}}(x_{\text{new}}) = \{c_j \in \mathcal{Y} \,|\, p_j > \alpha\}$.

This set predictor has coverage guarantees: $\mathbb{P}\big(y_{\text{new}} \in \Gamma_{\text{set}}(x_{\text{new}})\big) = 1 - \alpha, \;\; \alpha \in [0, 1]$

Since prediction sets are uncertainty estimates, can we use them as a basis for a reject option?

Since prediction sets are uncertainty estimates, can we use them as a basis for a reject option?

How about reject everything, but singleton sets?

Since prediction sets are uncertainty estimates, can we use them as a basis for a reject option?

How about reject everything, but singleton sets?

Conformal guarantees only hold marginally → We cannot make any claims about the coverage of singleton predictions!

However...

APPROACH.

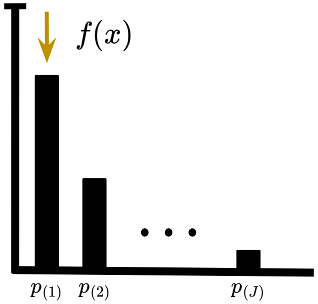
APPROACH.

- We can still use the *p*-values to create, instead of a prediction set, a less typical output for a new test sample, the confidence-credibility prediction: $\Gamma_{\text{cc}}(x_{\text{new}}) = \big(f(x_{\text{new}}), \lambda(x_{\text{new}}), \gamma(x_{\text{new}})\big)$

APPROACH.

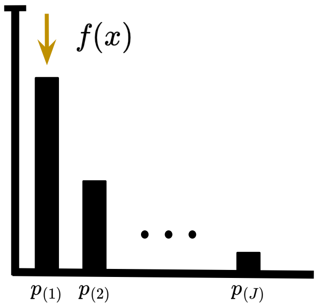
- We can still use the p-values to create, instead of a prediction set, a less typical output for a new test sample, the confidence-credibility prediction: $\Gamma_{\text{cc}}(x_{\text{new}}) = \big(f(x_{\text{new}}), \lambda(x_{\text{new}}), \gamma(x_{\text{new}})\big)$

  - **Forced prediction**

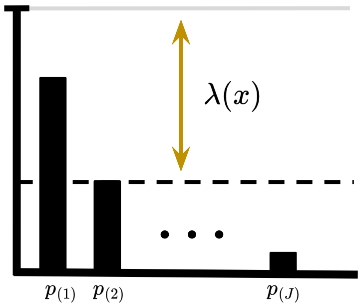


APPROACH.

- We can still use the p-values to create, instead of a prediction set, a less typical output for a new test sample, the confidence-credibility prediction: $\Gamma_{\text{cc}}(x_{\text{new}}) = \big(f(x_{\text{new}}), \lambda(x_{\text{new}}), \gamma(x_{\text{new}})\big)$
  - **Forced prediction**



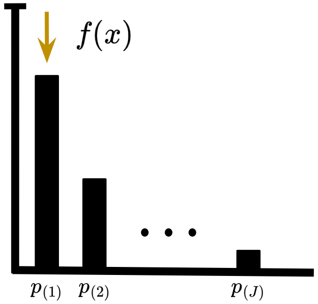
  - **Confidence**



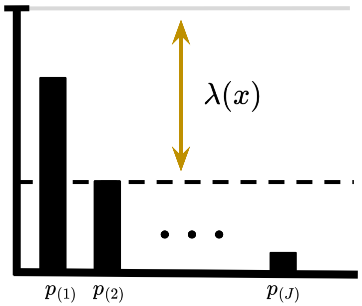
APPROACH.

- We can still use the p-values to create, instead of a prediction set, a less typical output for a new test sample, the confidence-credibility prediction: $\Gamma_{\text{cc}}(x_{\text{new}}) = \big(f(x_{\text{new}}), \lambda(x_{\text{new}}), \gamma(x_{\text{new}})\big)$

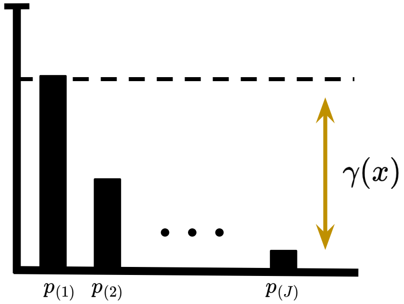
  - **Forced prediction**



  - **Confidence**

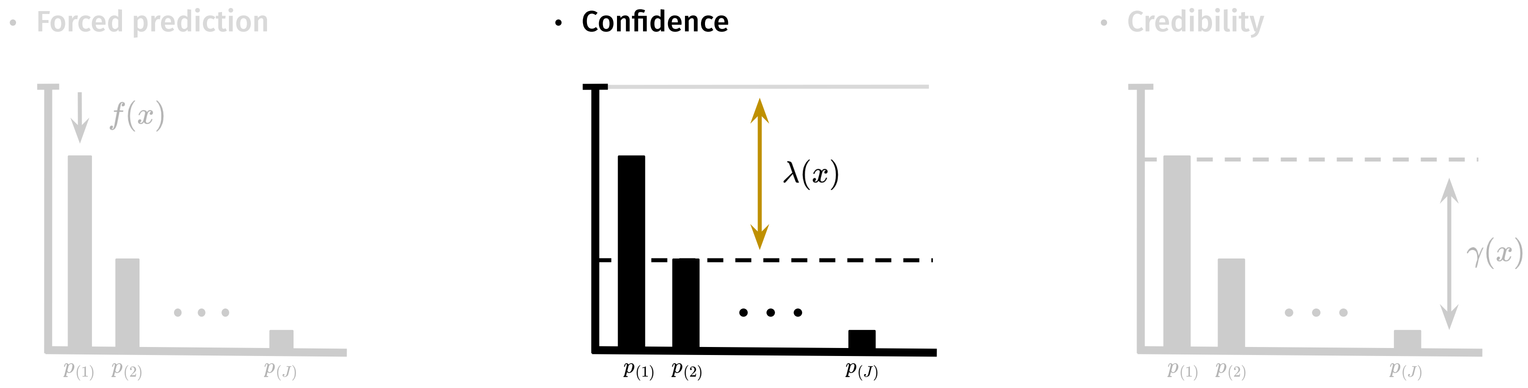


  - **Credibility**



APPROACH.

- We can still use the p-values to create, instead of a prediction set, a less typical output for a new test sample, the confidence-credibility prediction: $\Gamma_{\text{cc}}(x_{\text{new}}) = \big(f(x_{\text{new}}), \lambda(x_{\text{new}}), \gamma(x_{\text{new}})\big)$



- The confidence measure presents a suitable tool for the reject option setting: it is the highest significance level where we get a singleton prediction.

APPROACH.

- Imagine a synthetic case that we have 10 test samples with the following sorted confidence measures:

idx	0	1	2	3	4	5	6	7	8	9
$\hat{y}$	0	0	1	1	0	1	0	1	0	1
confidence	0.70	0.75	0.80	0.83	0.87	0.90	0.93	0.95	0.97	0.99

APPROACH.

- Imagine a synthetic case that we have 10 test samples with the following sorted confidence measures:

idx	0	1	2	3	4	5	6	7	8	9
$\hat{y}$	0	0	1	1	0	1	0	1	0	1
confidence	0.70	0.75	0.80	0.83	0.87	0.90	0.93	0.95	0.97	0.99

- We should approximately expect 70 % accuracy in the whole test set.

APPROACH.

- Imagine a synthetic case that we have 10 test samples with the following sorted confidence measures:

idx	0	1	2	3	4	5	6	7	8	9
$\hat{y}$	0	0	1	1	0	1	0	1	0	1
confidence	0.70	0.75	0.80	0.83	0.87	0.90	0.93	0.95	0.97	0.99

- We should approximately expect 70 % accuracy in the whole test set.
- If we reject the 2 most unconfident predictions, we should expect 80 % accuracy.

APPROACH.

- Imagine a synthetic case that we have 10 test samples with the following sorted confidence measures:

idx	0	1	2	3	4	5	6	7	8	9
$\hat{y}$	0	0	1	1	0	1	0	1	0	1
confidence	0.70	0.75	0.80	0.83	0.87	0.90	0.93	0.95	0.97	0.99

- We should approximately expect 70 % accuracy in the whole test set.
- If we reject the 2 most unconfident predictions, we should expect 80 % accuracy.
- And if we reject the 5 most unconfident predictions, we should expect 90 % accuracy.

APPROACH.

- Imagine a synthetic case that we have 10 test samples with the following sorted confidence measures:

idx	0	1	2	3	4	5	6	7	8	9
$\hat{y}$	0	0	1	1	0	1	0	1	0	1
confidence	0.70	0.75	0.80	0.83	0.87	0.90	0.93	0.95	0.97	0.99

- We should approximately expect 70 % accuracy in the whole test set.
- If we reject the 2 most unconfident predictions, we should expect 80 % accuracy.
- And if we reject the 5 most unconfident predictions, we should expect 90 % accuracy.
- We can estimate the expected accuracy as we start rejecting samples without knowing the true classes.
- If Mondrian calibration is performed, we can estimate the expected per-class recall similarly.

APPROACH.

APPROACH.

$$g : \mathcal{X} \to \{0, 1\}$$

Input space — *Binary rejection*

APPROACH.

$$g : \mathcal{X} \to \{0, 1\}$$

Input space — *Binary rejection*

- **Conformal selection function**

$$g_\lambda(x) = \begin{cases} 1, & \text{if } \lambda(x) \geq \theta; \\ 0, & \text{otherwise.} \end{cases}$$

APPROACH.

$$g : \mathcal{X} \rightarrow \{0, 1\}$$

Input space — *Binary rejection*

- **Conformal selection function**

$$g_\lambda(x) = \begin{cases} 1, & \text{if } \lambda(x) \geq \theta; \\ 0, & \text{otherwise.} \end{cases}$$

- In standard conformal prediction, θ is reported as the expected accuracy in the non-rejected samples.

APPROACH.

$$g : \mathcal{X} \rightarrow \{0, 1\}$$

Input space — *Binary rejection*

- **Conformal selection function**

$$g_\lambda(x) = \begin{cases} 1, & \text{if } \lambda(x) \geq \theta; \\ 0, & \text{otherwise.} \end{cases}$$

- In standard conformal prediction, θ is reported as the expected accuracy in the non-rejected samples.
- In Mondrian conformal prediction, θ is reported as the expected recall in the non-rejected samples.

APPROACH.

$$g : \mathcal{X} \rightarrow \{0, 1\}$$

Input space *Binary rejection*

- **Conformal selection function**

$$g_\lambda(x) = \begin{cases} 1, & \text{if } \lambda(x) \geq \theta; \\ 0, & \text{otherwise.} \end{cases}$$

- In standard conformal prediction, θ is reported as the expected accuracy in the non-rejected samples.
- In Mondrian conformal prediction, θ is reported as the expected recall in the non-rejected samples.

- **Maximum score function**

$$g_\pi(x) = \begin{cases} 1, & \text{if } \max_j \pi_j \geq \theta; \\ 0, & \text{otherwise.} \end{cases}$$

The average maximum posterior score of the non-rejected samples is reported as the expected accuracy.

APPROACH.

$$g : \mathcal{X} \rightarrow \{0, 1\}$$

Input space — *Binary rejection*

- **Conformal selection function**

$$g_\lambda(x) = \begin{cases} 1, & \text{if } \lambda(x) \geq \theta; \\ 0, & \text{otherwise.} \end{cases}$$

- In standard conformal prediction, θ is reported as the expected accuracy in the non-rejected samples.
- In Mondrian conformal prediction, θ is reported as the expected recall in the non-rejected samples.

- **Maximum score function**

$$g_\pi(x) = \begin{cases} 1, & \text{if } \max_j \pi_j \geq \theta; \\ 0, & \text{otherwise.} \end{cases}$$

The average maximum posterior score of the non-rejected samples is reported as the expected accuracy.

- **Class score function**

$$g_\pi(x) = \begin{cases} 1, & \pi_j \geq \theta; \\ 0, & \text{otherwise.} \end{cases}$$

The average class-specific posterior score of the non-rejected samples is reported as the expected recall.

EXPERIMENTAL SETUP.

- **Classification algorithms:** decision tree and random forest (with fixed hyperparameters).
- **Rejection mechanisms:**

  Uncalibrated classifiers (Uncal).
  Platt-scaled classifiers (Platt). | **Maximum score function** (accuracy) and **class score function** (recall)

  Conformal-based rejection (Conf) | **Conformal function with standard** (accuracy) and **Mondrian** (recall)

- **Rejection rates:** $\tau \in \{0.1, 0.2, \ldots, 0.9\}$
- **Non-conformity measure** $\Xi\left(y_i, h(x_i)\right) = 1 - h(x_i)_{y_i},$
- **Testing protocol**: 100x repeated hold-out, 75/25 % train-test split, 66/33 % proper train-calibration split

EXPERIMENTAL SETUP.

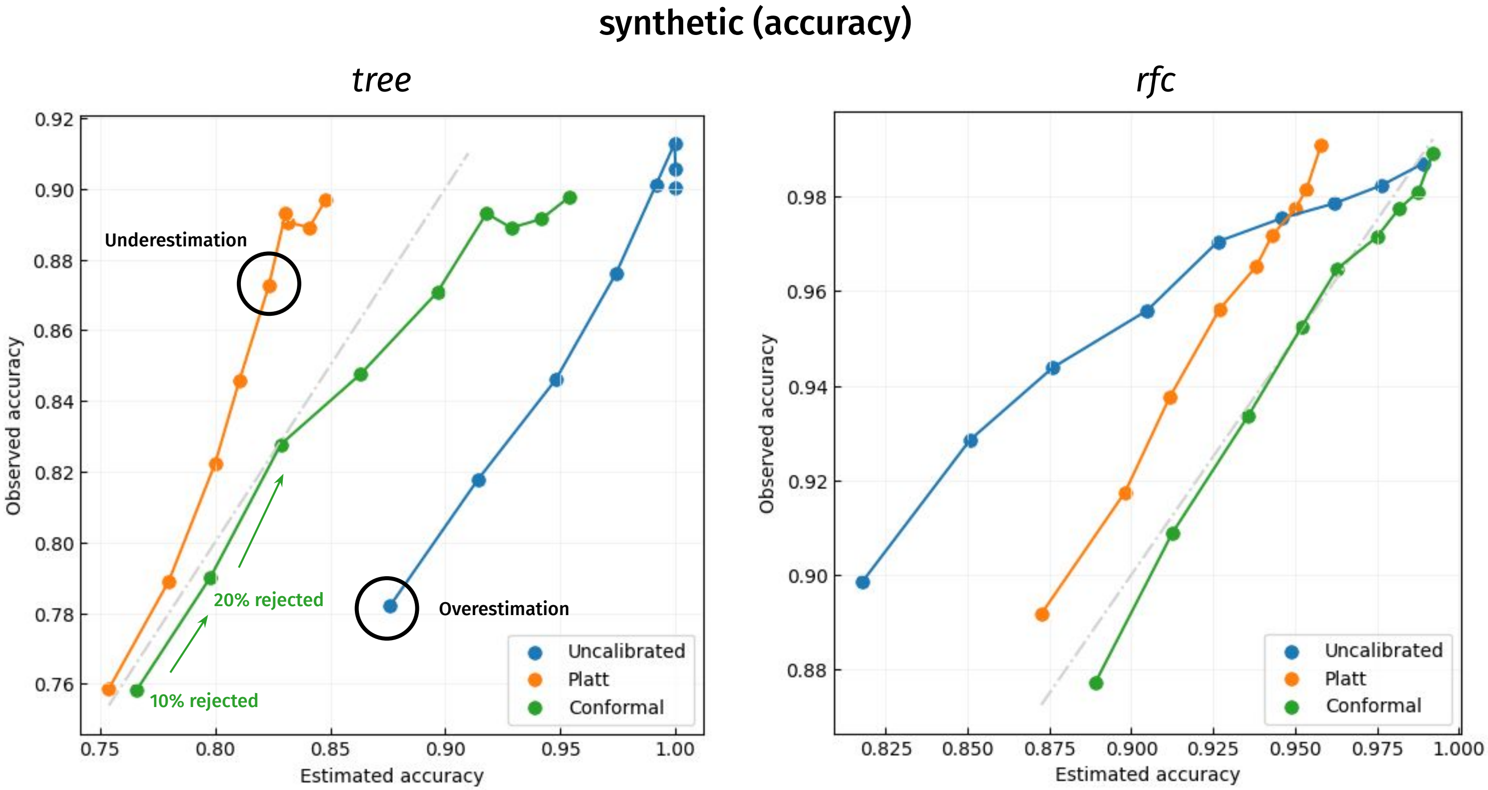
Datasets

Dataset	n	p	$\|\mathcal{Y}\|$
adult	49,531	10	3
beans	13,611	16	6
ocr	5,620	64	9
cars	1,728	6	4
synthetic	1,000	8	3
glass	214	9	6

Target class for each dataset

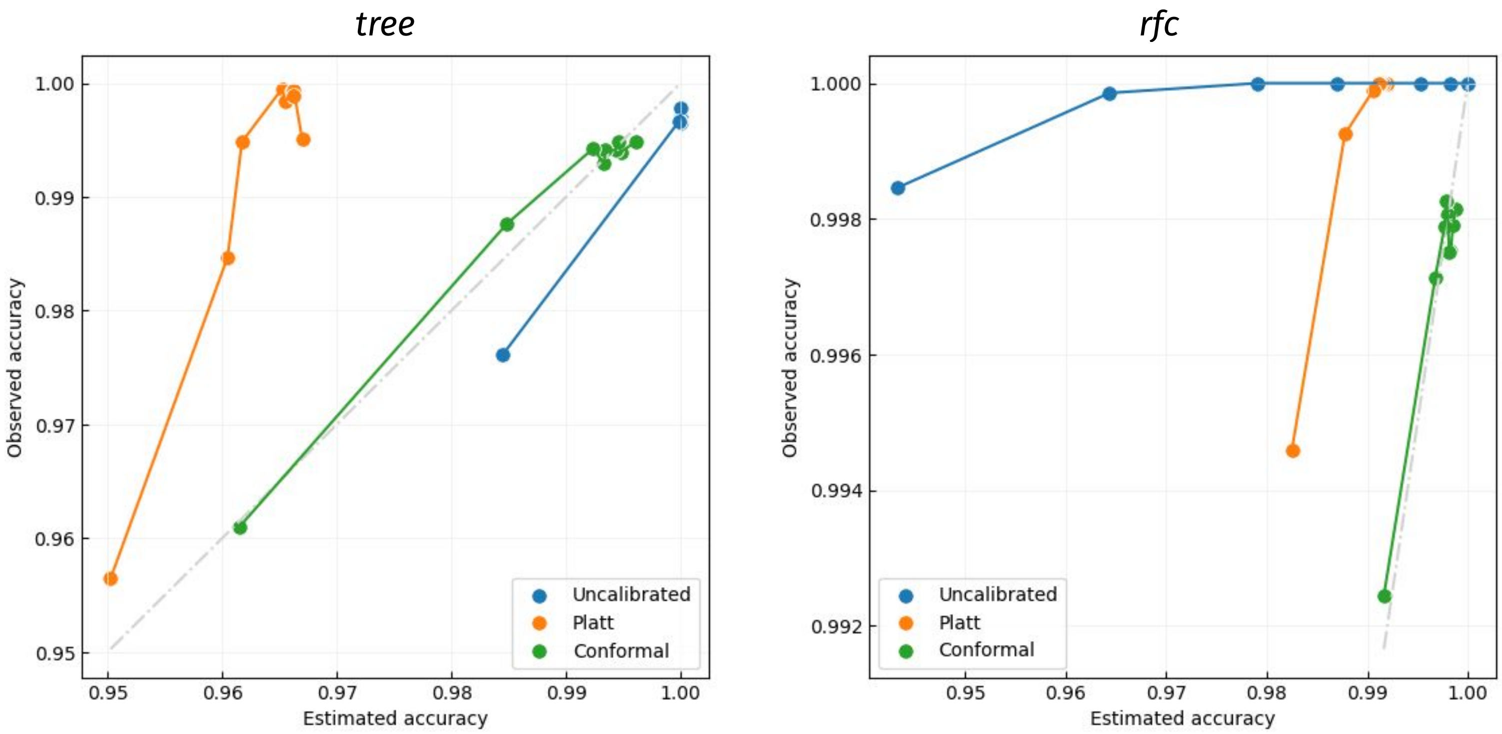
Class description	Proportion
Income $\leq$ \$20K	0.399
Sira dry bean type	0.194
Digit nine	0.100
Car with an acceptable evaluation	0.222
Synthetic class ($y = 1$)	0.333
Headlamp	0.136

RESULTS (some of them).

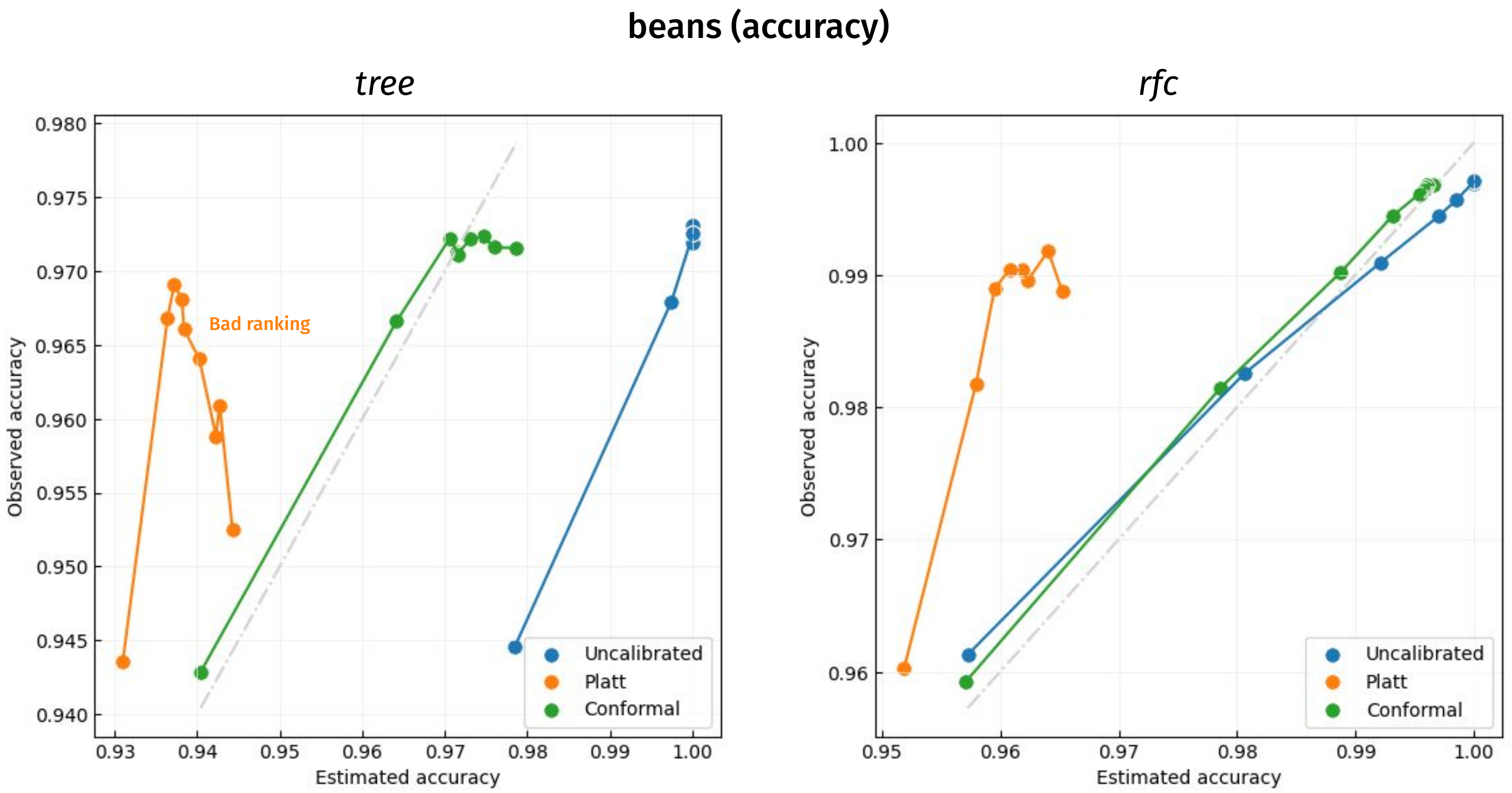


RESULTS (some of them).

cars (accuracy)

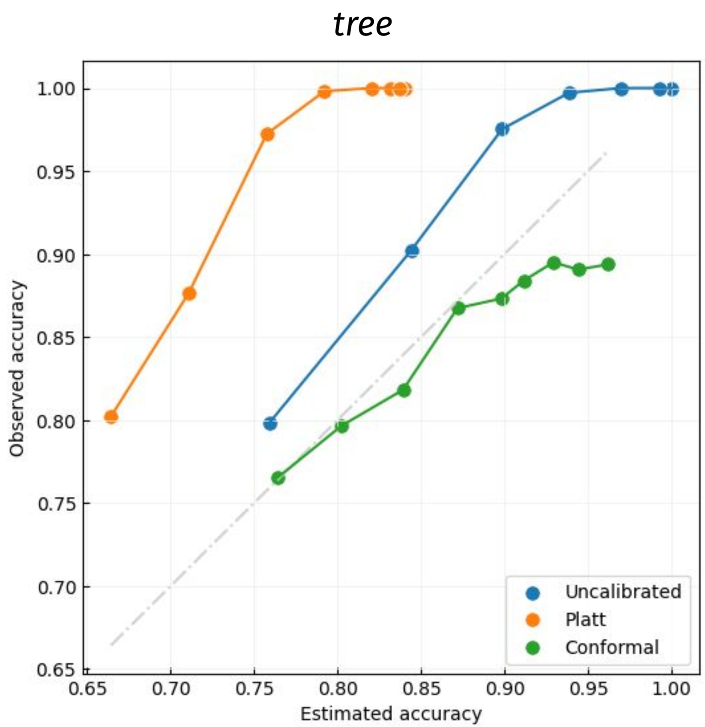


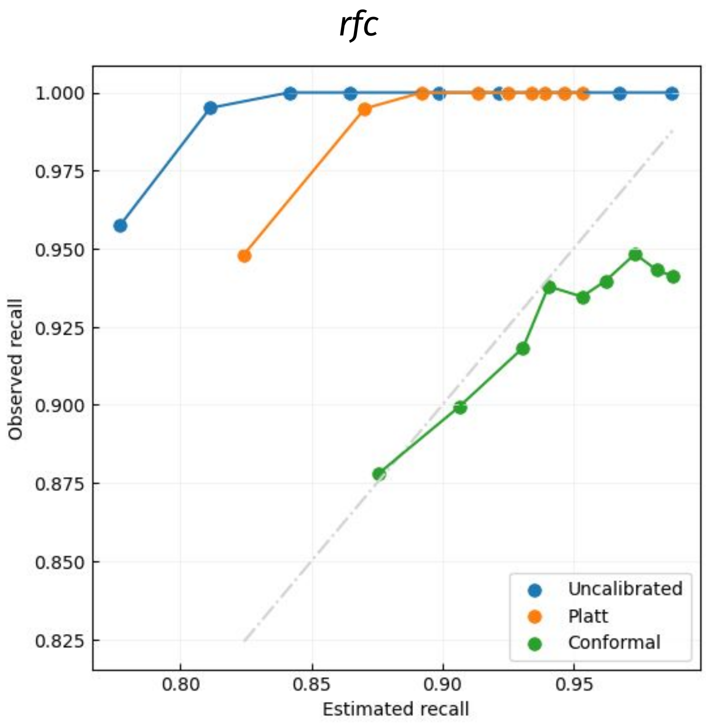
RESULTS (some of them).



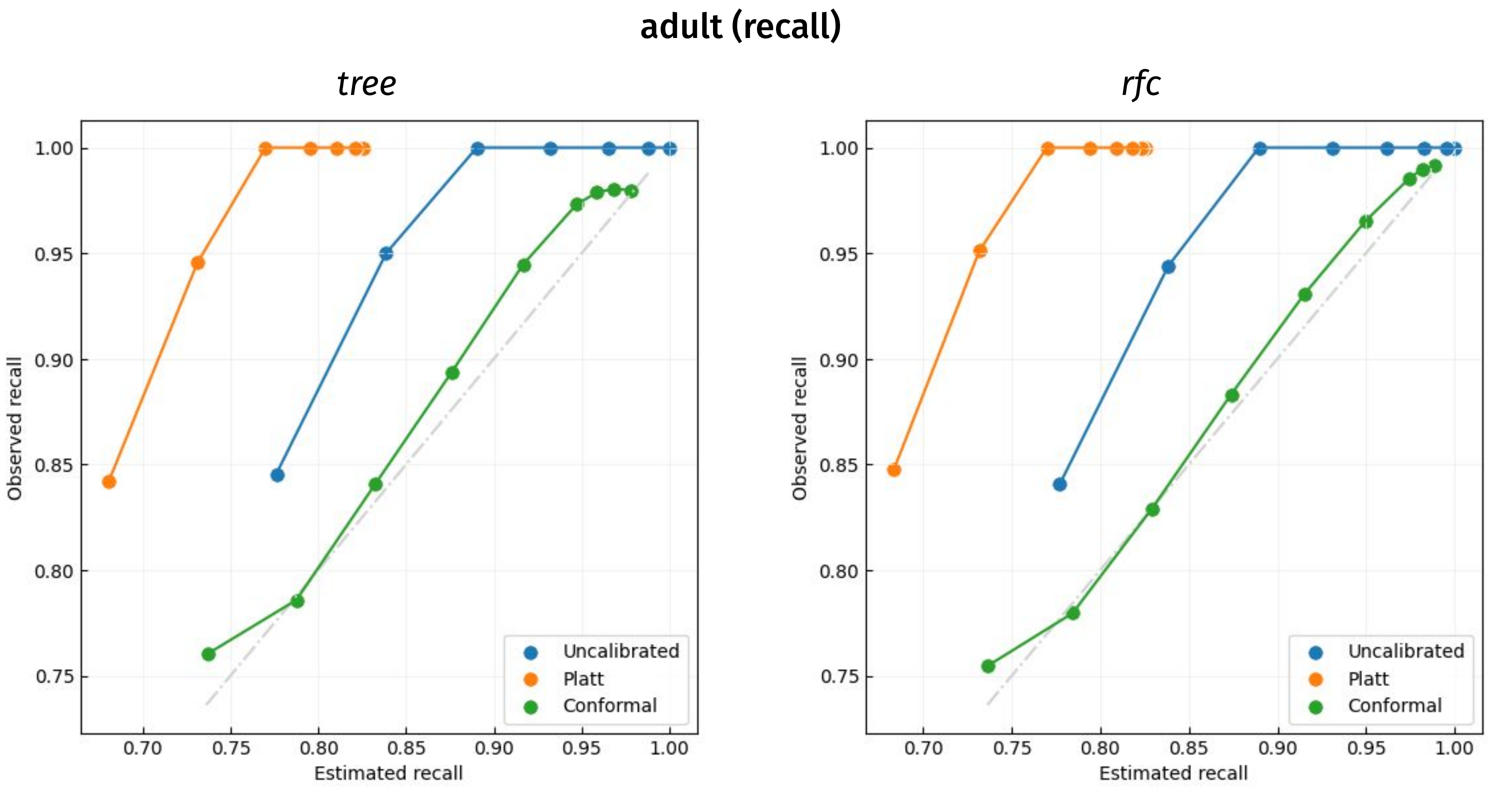
RESULTS (some of them).

synthetic (recall)

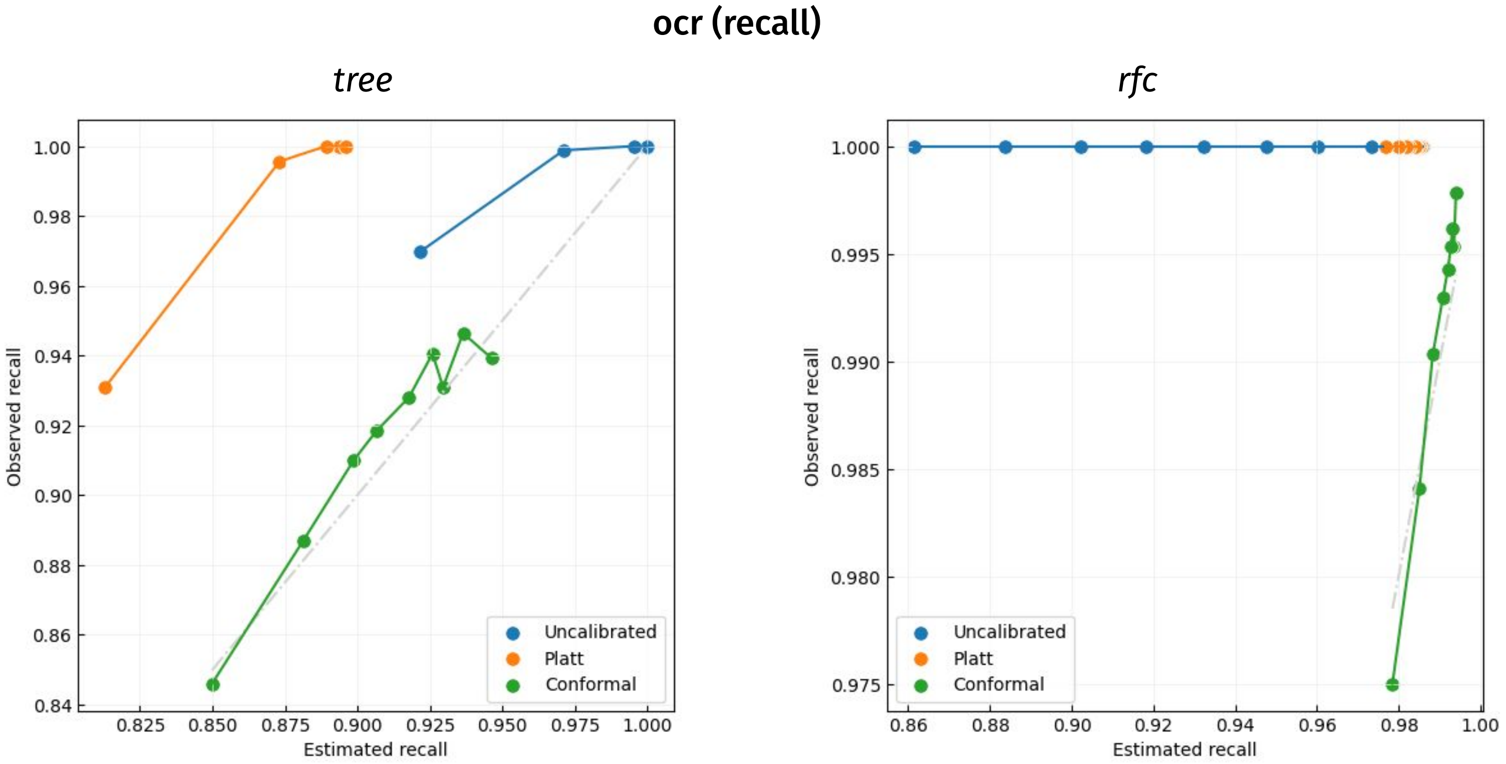




RESULTS (some of them).



RESULTS (some of them).



CONCLUSIONS.

- Conformal prediction can be used in a suitable way in the context of classification with reject option to produce reliable performance estimates.
- In this study, we have extended previous research in the development of confidence classifiers with reject option and covered multi-class classification.
- Our approach, tested in six datasets, consistently delivers reliable accuracy and recall estimates with better results than off-the-shelf uncalibrated classifiers and Platt-scaled models.
- Some limitations on low-data regime scenarios, when only a few calibration samples are available and statistical efficiency is sacrificed.

Acknowledgements

Gobierno de Navarra | Nafarroako Gobernua

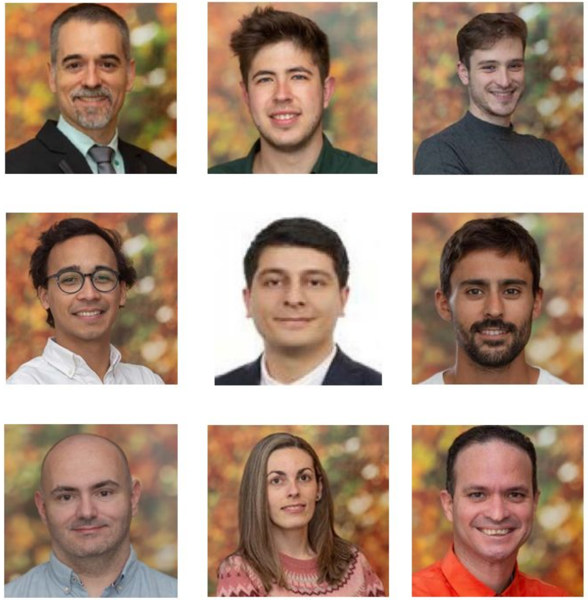
AGENDA 2030

European Commission

Horizon 2020
European Union funding
for Research & Innovation

ERA PerMed

Universidad de Navarra | DATAI
INSTITUTO DE CIENCIA DE LOS
DATOS E INTELIGENCIA ARTIFICIAL



Multi-Class Classification With Reject Option and Performance Guarantees Using Conformal Prediction

THANK YOU FOR YOUR ATTENTION!

Alberto García-Galindo

You can reach me via LinkedIn or email: **agarciagali@unav.es**