10th UCAAT User Conference on Advanced Automated Testing

ETSI The Standards People

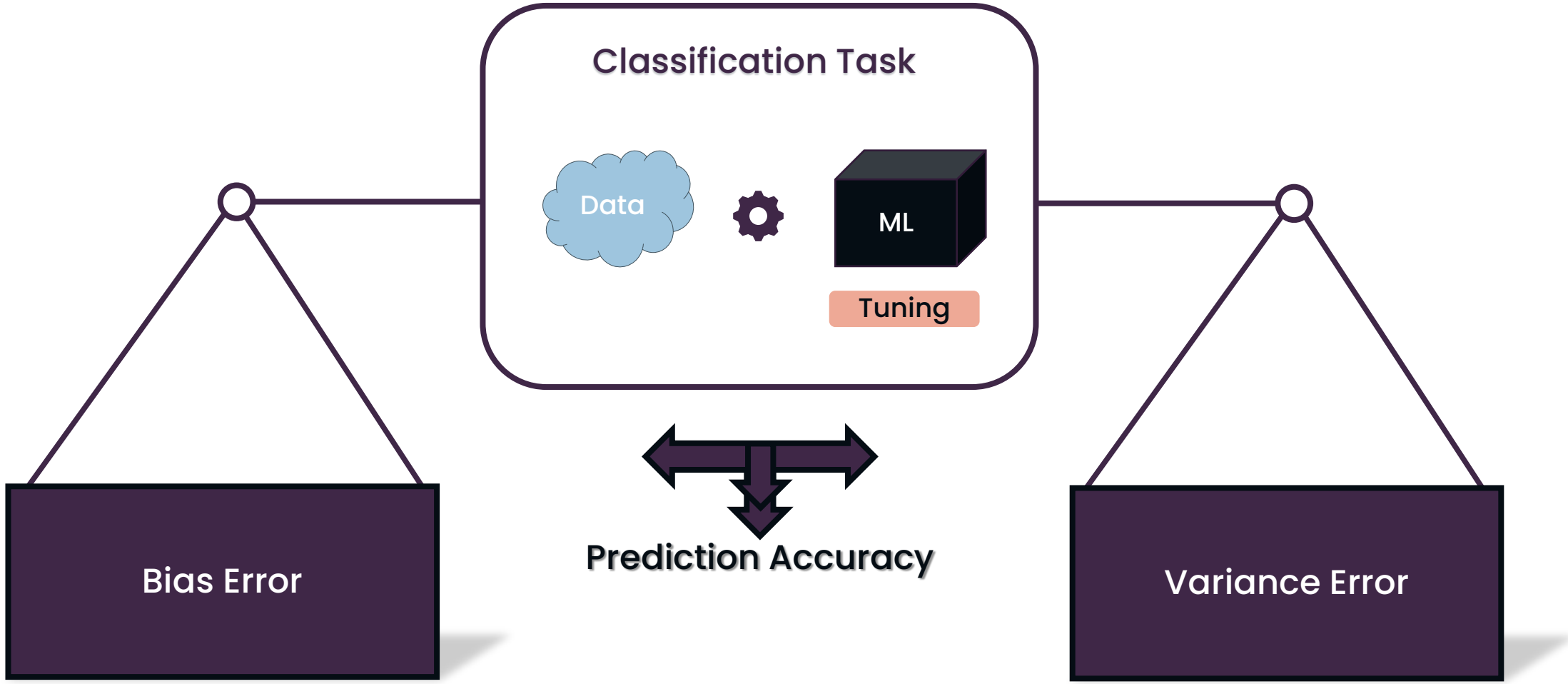
Hyperparameter Importance: Rethinking Bias-Variance with Meta-features

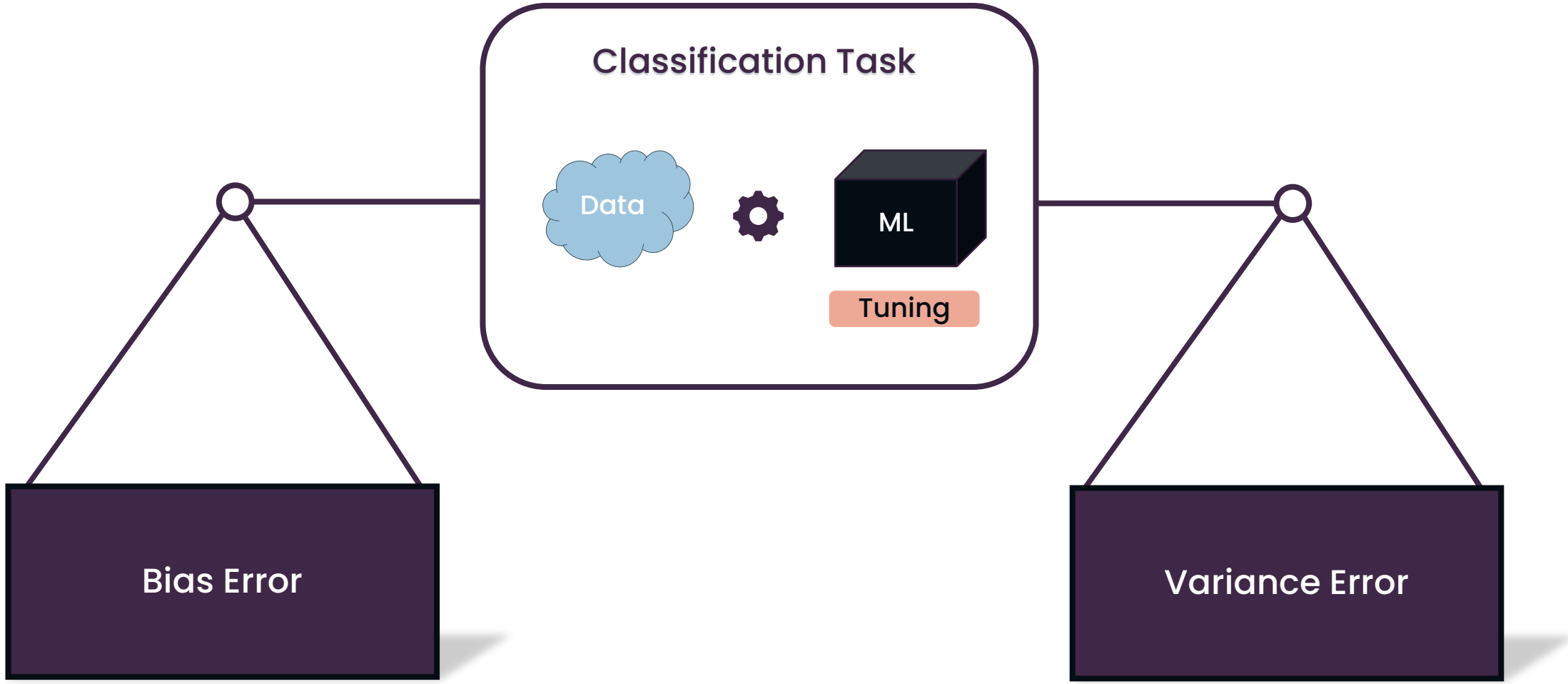
Emmanuel Dapaah, Marianne Njifon &

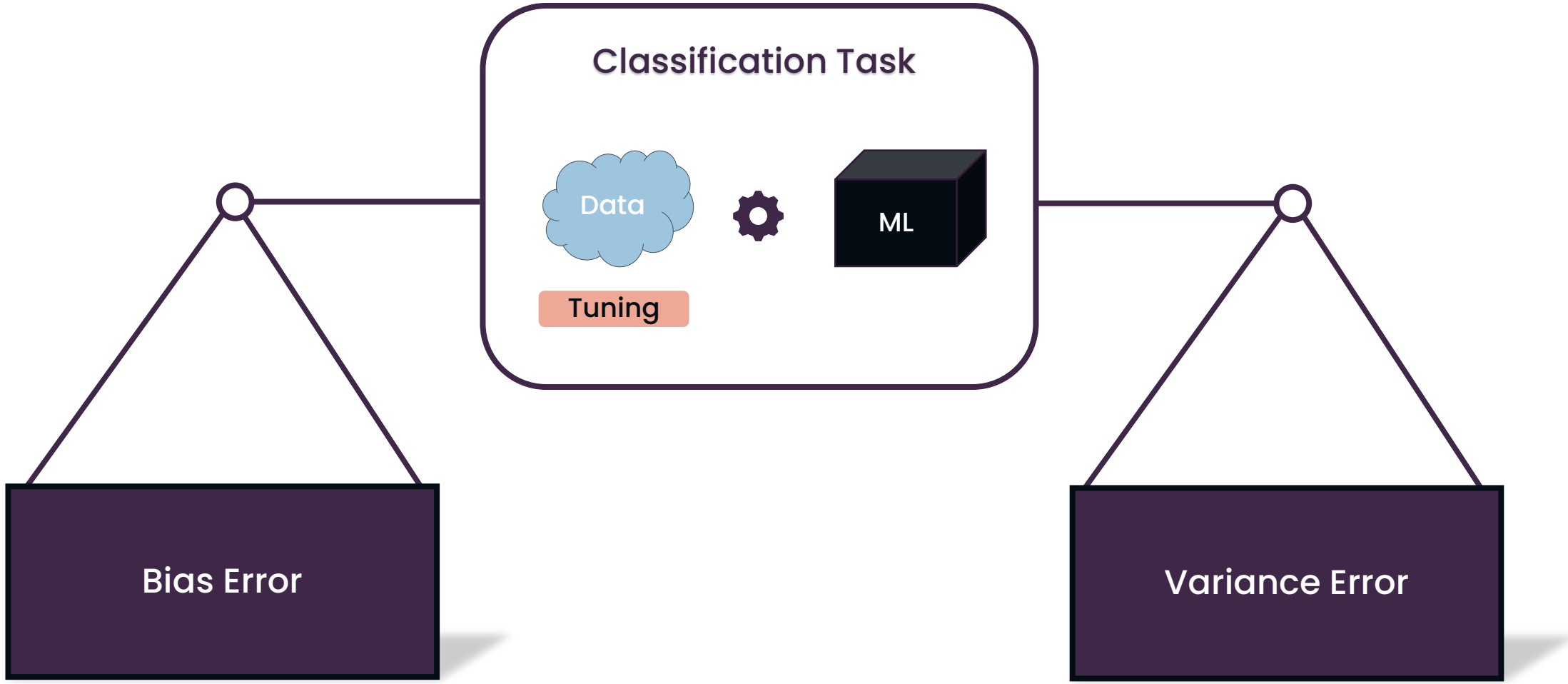
Prof. Jens Grabowski

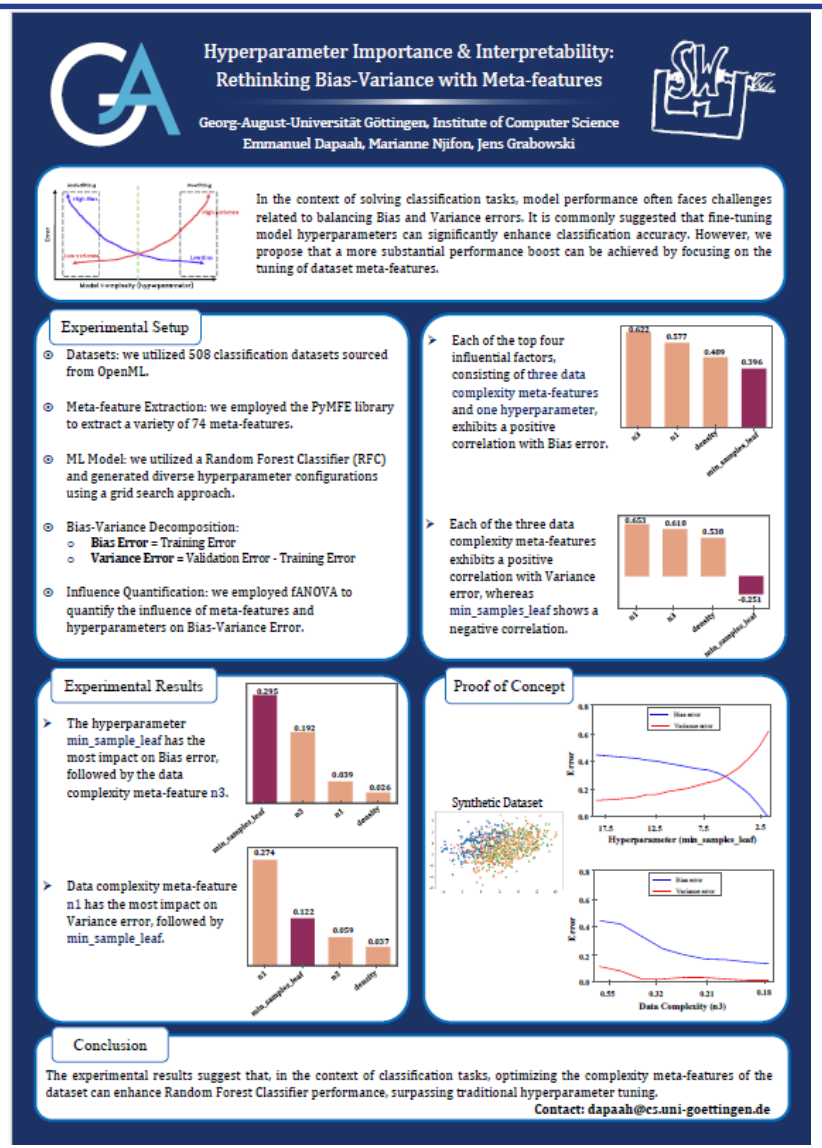
UNIVERSITÄT GÖTTINGEN

15/11/2023


Classification Task
Data
ML
Tuning
Bias Error
Prediction Accuracy
Variance Error


Classification Task
Data
ML
Tuning
Bias Error
Variance Error


Classification Task
Data
Tuning
ML
Bias Error
Variance Error



Hyperparameter Importance & Interpretability: Rethinking Bias-Variance with Meta-features

Georg-August-Universität Göttingen, Institute of Computer Science
Emmanuel Dapaah, Marianne Njifon, Jens Grabowski

In the context of solving classification tasks, model performance often faces challenges related to balancing Bias and Variance errors. It is commonly suggested that fine-tuning model hyperparameters can significantly enhance classification accuracy. However, we propose that a more substantial performance boost can be achieved by focusing on the tuning of dataset meta-features.

Experimental Setup

- Datasets: we utilized 508 classification datasets sourced from OpenML.
- Meta-feature Extraction: we employed the PyMFE library to extract a variety of 74 meta-features.
- ML Model: we utilized a Random Forest Classifier (RFC) and generated diverse hyperparameter configurations using a grid search approach.
- Bias-Variance Decomposition:
  - **Bias Error** = Training Error
  - **Variance Error** = Validation Error - Training Error
- Influence Quantification: we employed fANOVA to quantify the influence of meta-features and hyperparameters on Bias-Variance Error.

- Each of the top four influential factors, consisting of three data complexity meta-features and one hyperparameter, exhibits a positive correlation with Bias error.

- Each of the three data complexity meta-features exhibits a positive correlation with Variance error, whereas min_samples_leaf shows a negative correlation.

Experimental Results

- The hyperparameter min_sample_leaf has the most impact on Bias error, followed by the data complexity meta-feature n3.

- Data complexity meta-feature n1 has the most impact on Variance error, followed by min_sample_leaf.

Proof of Concept

Synthetic Dataset

Conclusion

The experimental results suggest that, in the context of classification tasks, optimizing the complexity meta-features of the dataset can enhance Random Forest Classifier performance, surpassing traditional hyperparameter tuning.

Contact: dapaah@cs.uni-goettingen.de

Visit Our Poster …