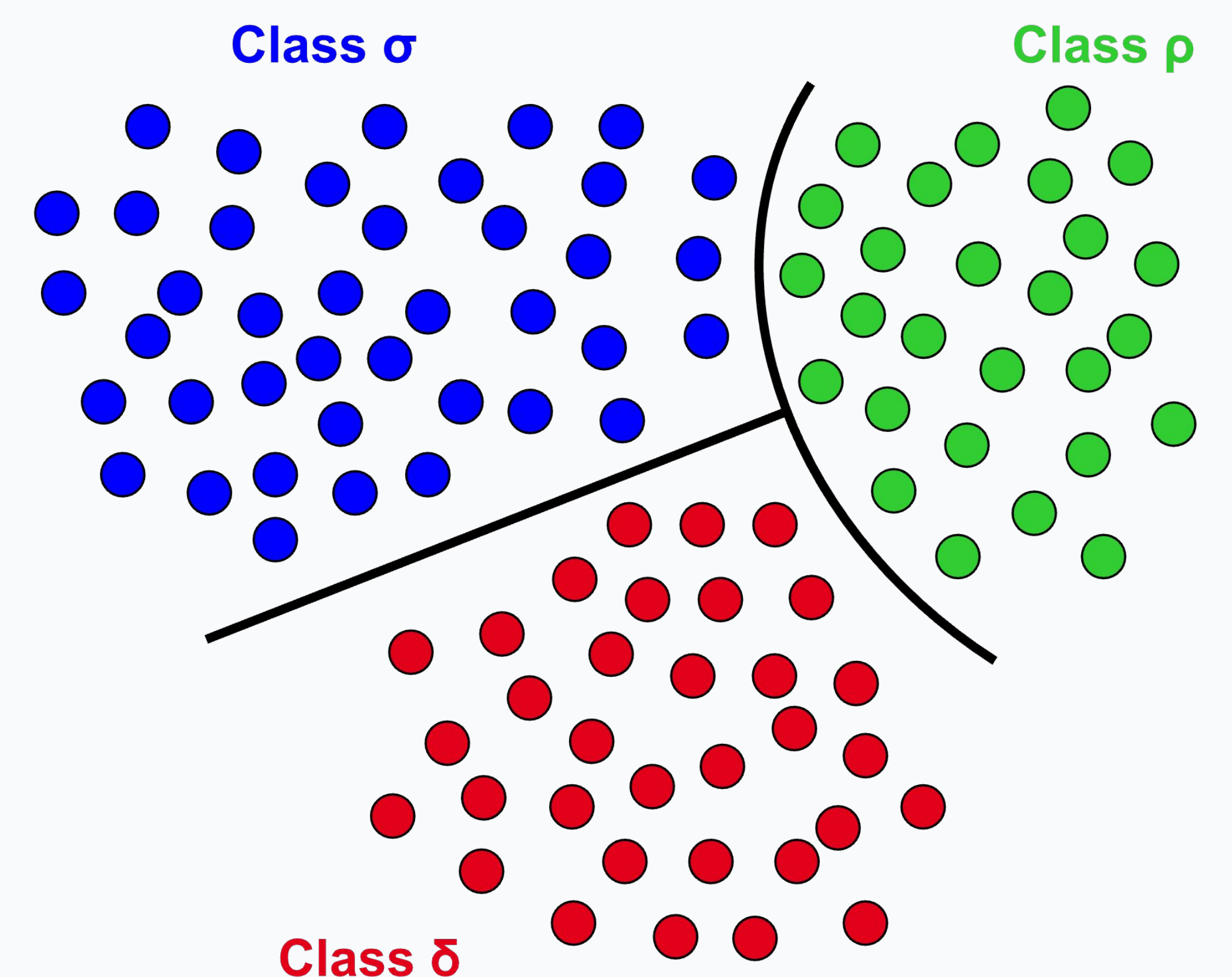


ICFI: a Feature Importance Measure For Multi-Class Classification

Tommaso Amico* , Pernille Matthews* , Ira Assent*

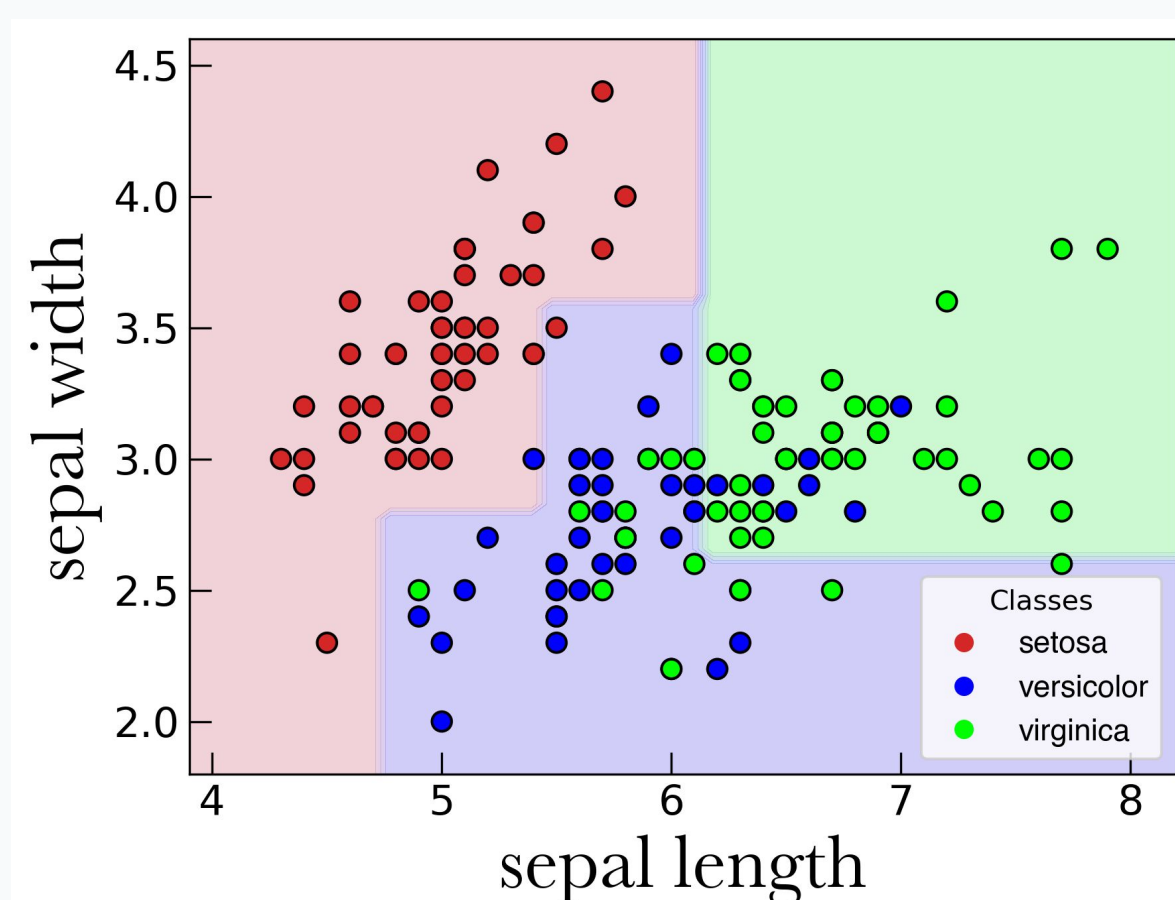
Inter-Class Feature importance

- Feature importance methods don't deal with inter-class relationships.
- Global methods (e.g. PFI [1]) summarize the average model behavior, making them subject to aggregation bias.
- Local methods (e.g. SHAP [2]) explain one instance. A unique ranking is outputted in binary classification only.
- A feature might carry discriminative power to separate just two features while being otherwise not relevant.

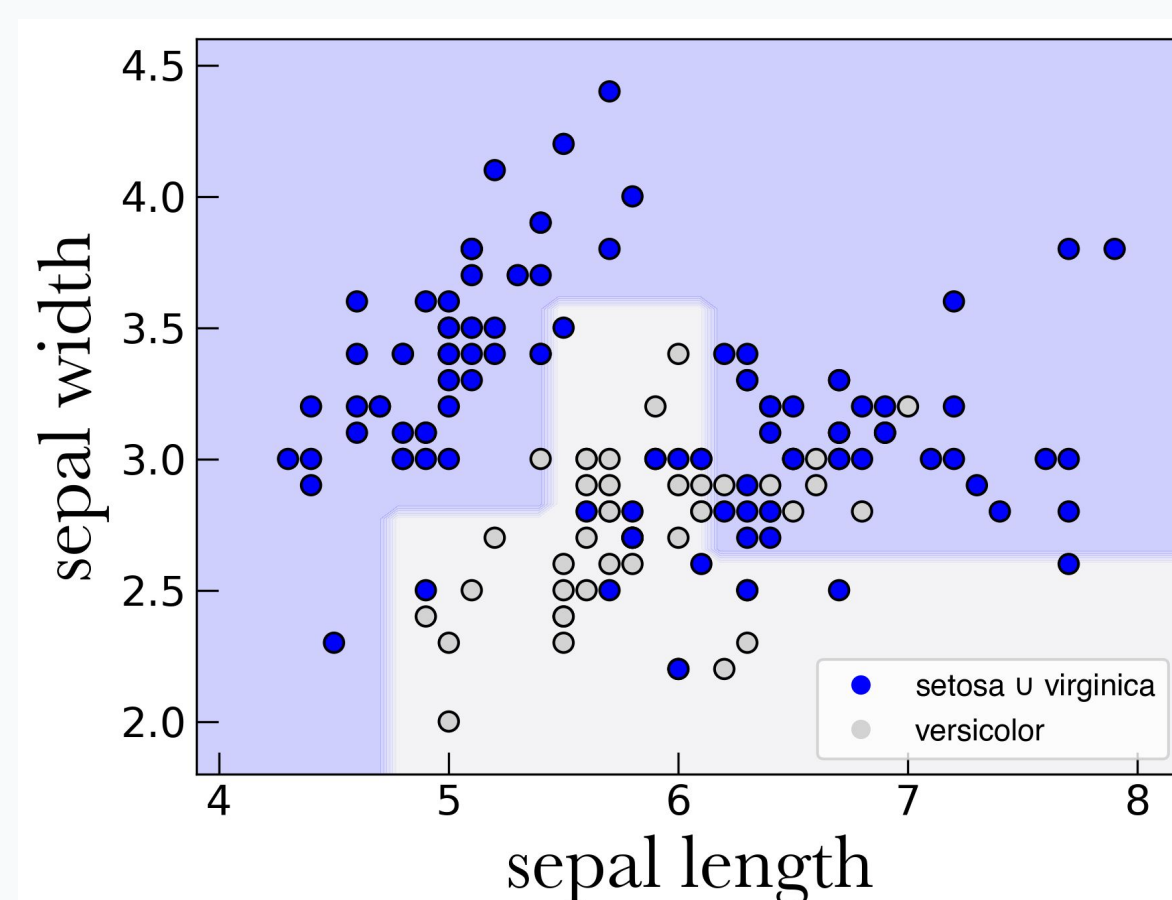


Method

- To quantify the model's performance in separating classes σ and ρ , we merge σ and ρ to then look at the improvement in empirical risk $\Delta R^{\sigma\rho}$
- We mimick the absence of feature j by permuting it. We assess the new model performance $\Delta \tilde{R}_j^{\sigma\rho}$

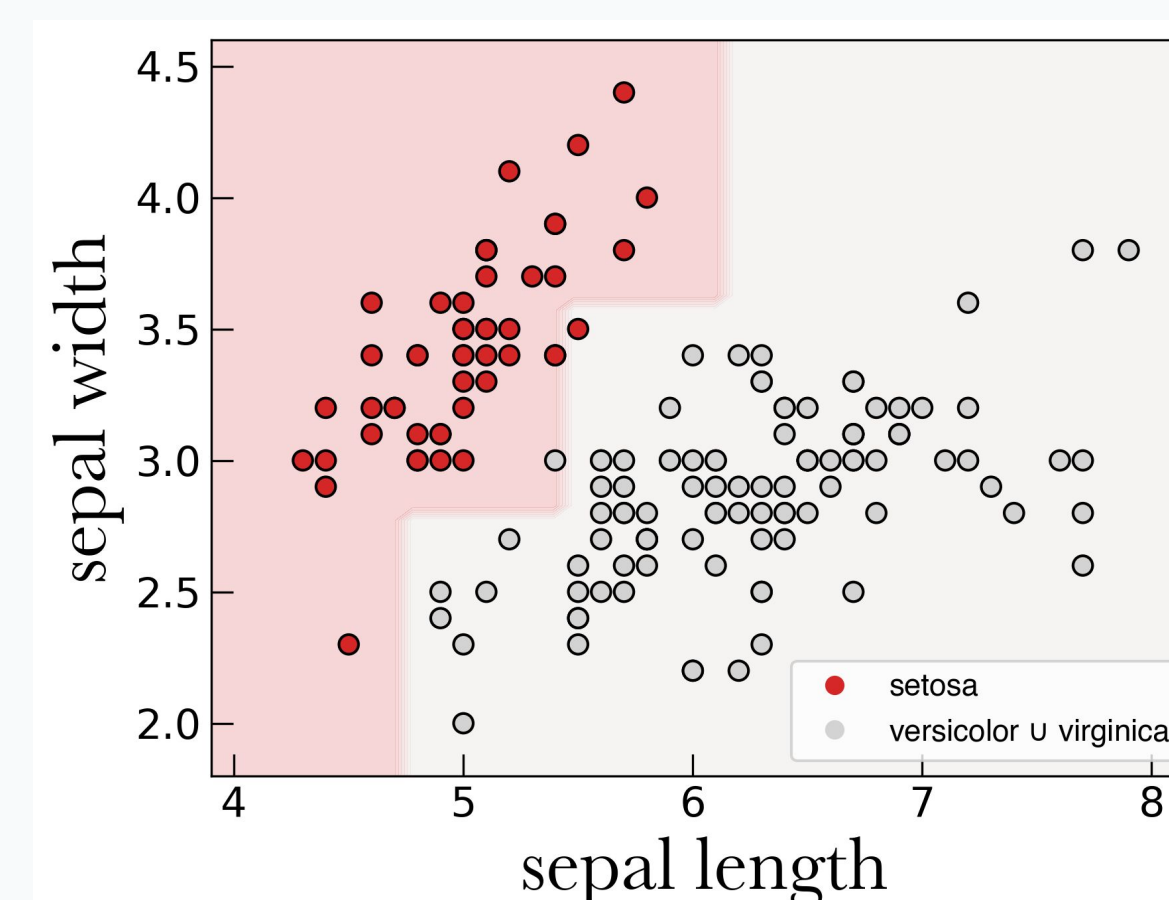


Caption. Decision tree fitted on the sepal length and sepal width features of the Iris dataset



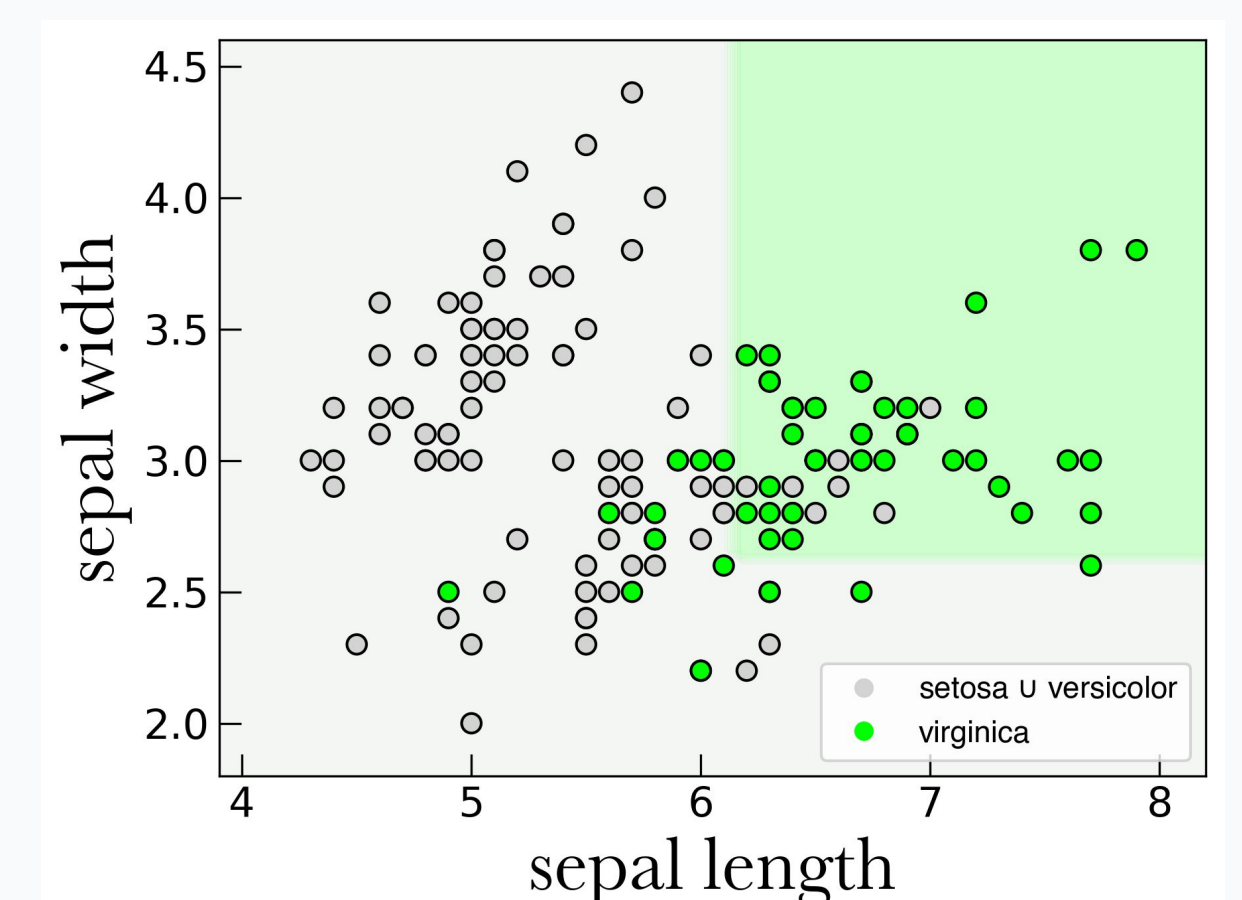
Caption. Merging setosa and virginica

$$\Delta R^{\sigma\rho} = 0$$



Caption. Merging versicolor and virginica

$$\Delta R^{\sigma\rho} = 0.20$$



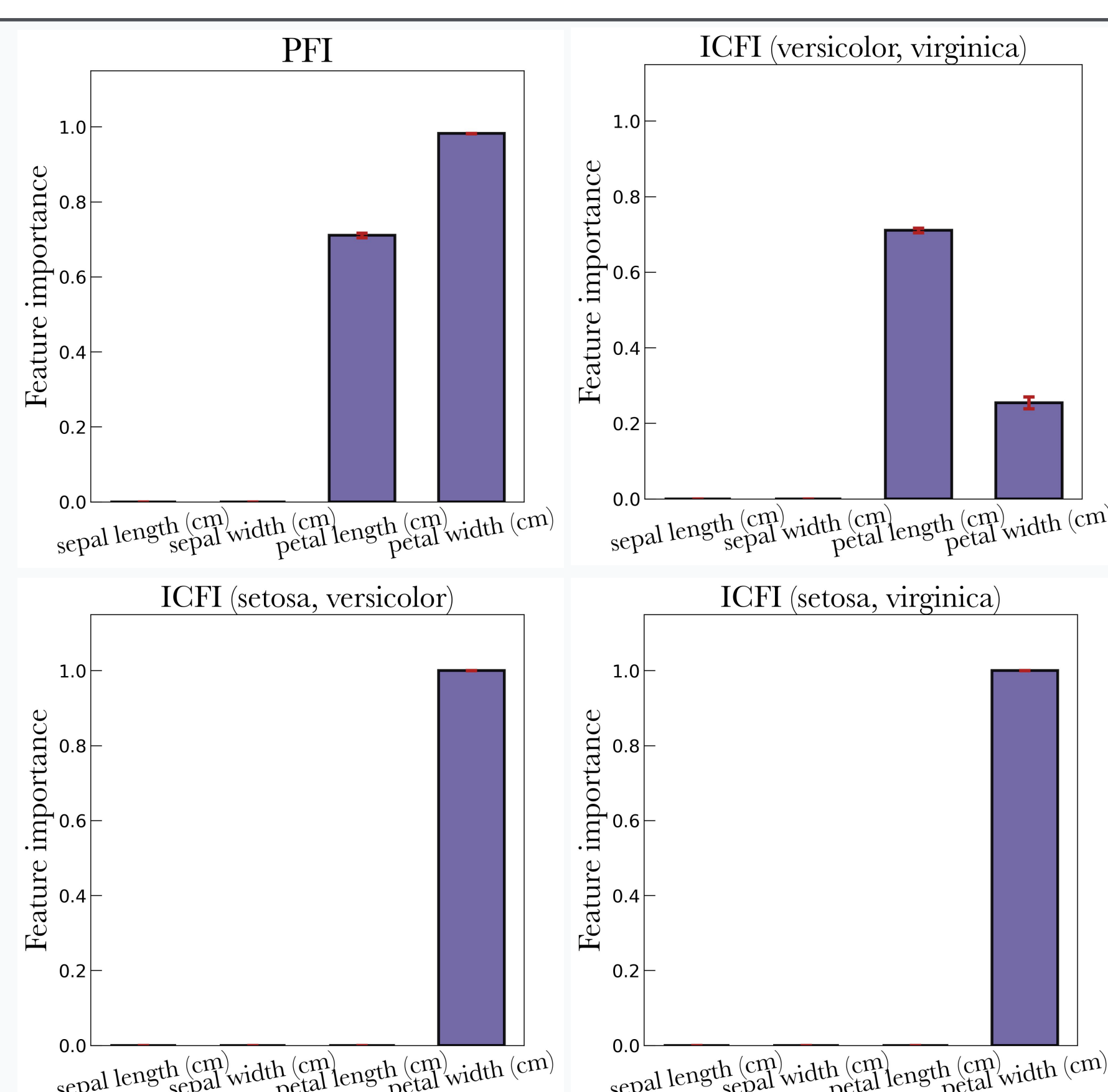
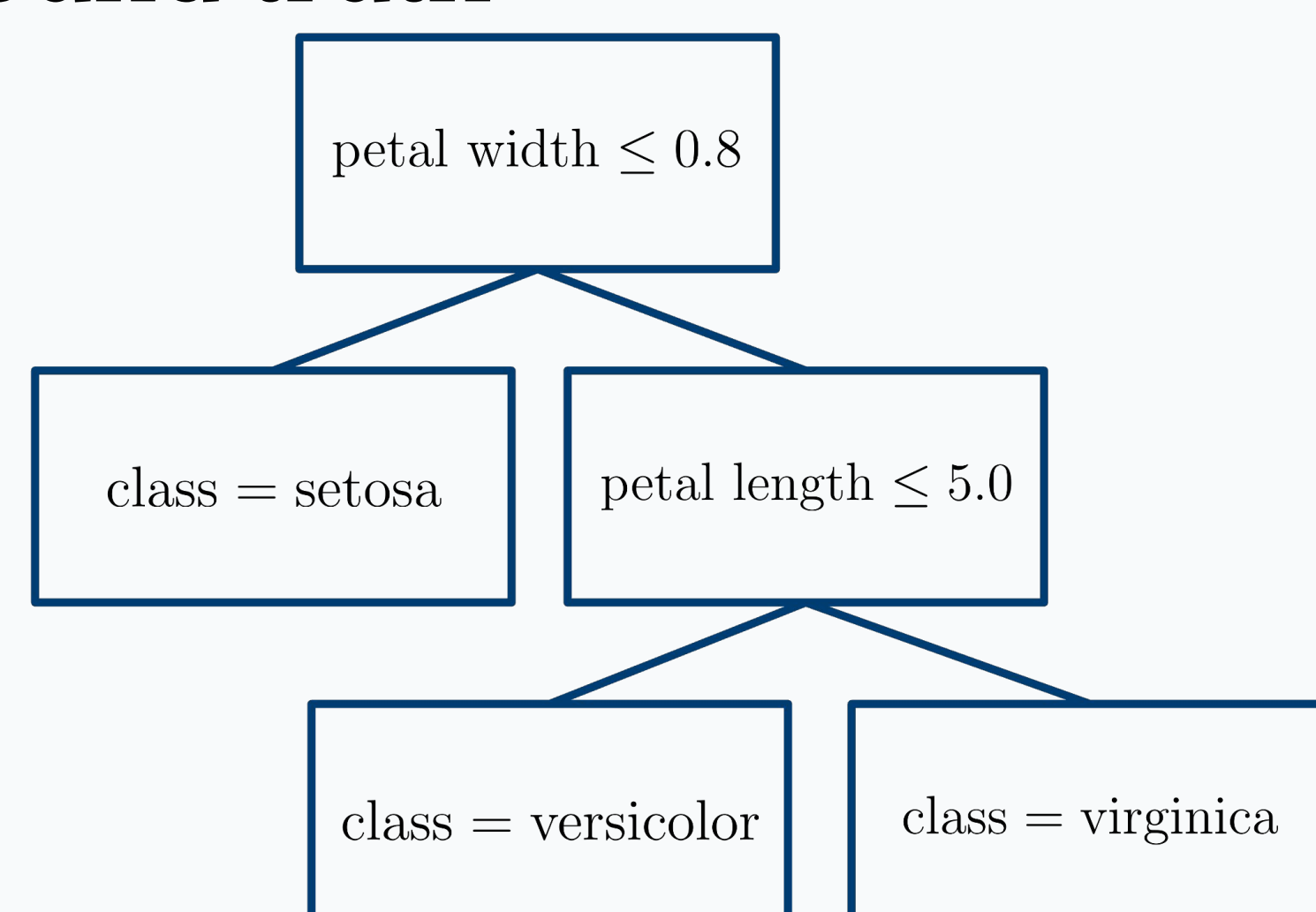
Caption. Merging setosa and versicolor

$$\Delta R^{\sigma\rho} = 0.01$$

$$ICFI_j^{\sigma\rho} = 1 - \frac{1}{1 + |\Delta \tilde{R}_j^{\sigma\rho} - \Delta R^{\sigma\rho}| / \Delta R^{\sigma\rho}}$$

Results

- An interpretable decision tree is fitted on the Iris dataset to have ground truth



References

- [1] Breiman, Leo. "Random forests." Machine learning 45 (2001): 5-32.
- [2] Lundberg, Scott. "A unified approach to interpreting model predictions." arXiv preprint arXiv:1705.07874 (2017).

*Department of Computer Science, Aarhus University