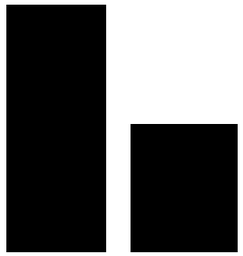
Impact of data balancing on model behaviour with Explainable Artificial Intelligence tools in imbalanced classification problems

Author: Adrian Stańdo

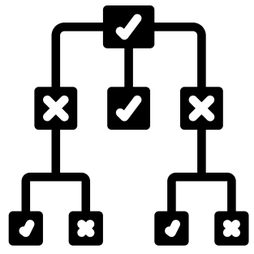
Supervisor: dr Mustafa Çavuş
Supervising professor: dr hab. inż. Przemysław Biecek

21 unbalanced datasets
(OpenML-100, OpenML-CC18, imblearn)

6 balancing methods
(oversampling, undersampling, hybrid)

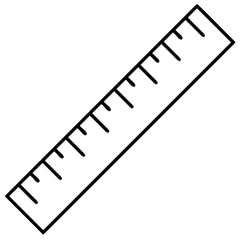


5 unbalancedness values
(based on the percentage of the original Imbalance Ratio)



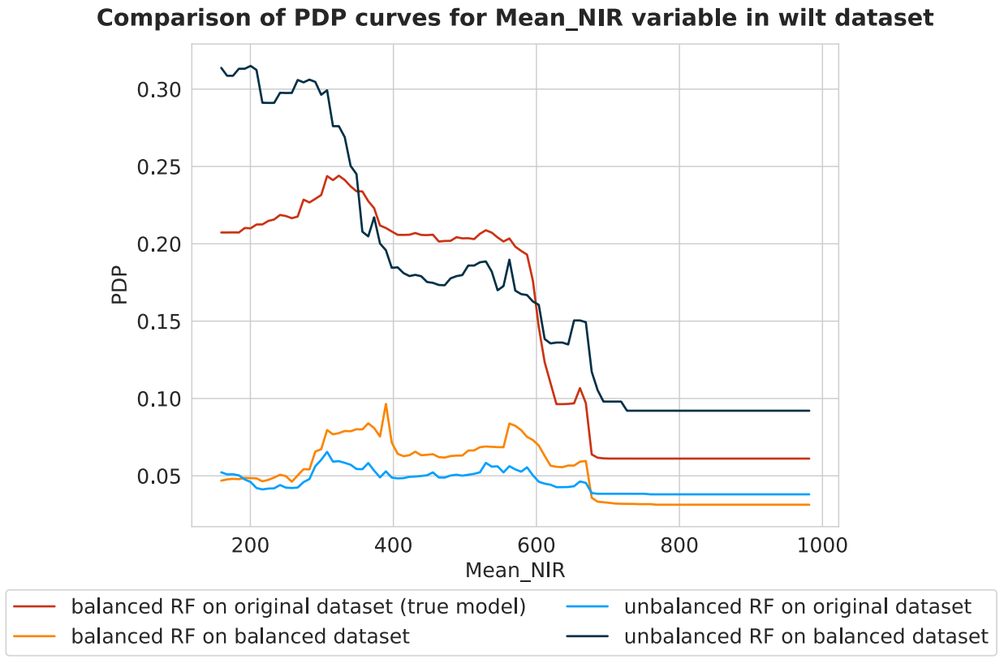
2 Random Forest models
(original and weighted)

3 XAI model explanations
(PDP, ALE, VI)



2 explanation comparison metrics
(SDD, Wilcoxon test)

The main aim of the paper was to analyse the differences in model behaviours after balancing datasets in imbalanced binary classification tasks. In order to automate the procedures, a Python package called *edgaro* was created. Moreover, two novel explanation comparison metrics were proposed. The results show that the balancing methods mainly reduce model bias towards the majority class, however, they may also change the existing relationships in data.



github.com/adrianstando/edgaro