UNIVERSITÀ
DEGLI STUDI
DI TRIESTE

SMOTE VARIANT FOR UNBALANCED DATA IN CLASSIFICATION PROBLEMS

Valeria De Stasio, Christian Faccio, Andrea Suklan, Agnese Valentini

January 28, 2025

Introduction

Problem of **unbalanced data** in classification problems:

- Solutions at the **algorithm level**: cost-sensitive learning, ensemble methods
- Solutions at the **data level**: oversampling, undersampling, syntethic data (SMOTE, ROSE)

Aim of the project:

- Generate synthetic data using the variant SMOTE-DIRICHLET
- Compare models trained on unbalanced dataset, balanced with SMOTE and balanced with SMOTE-DIRICHLET

Data Generation Process

Parameter	Values	Description
n_{train}	600, 1000, 5000	Train set size
n_{test}	600	Test set size
π	0.10, 0.05, 0.025	Proportion of rare examples
IR	9, 19, 39	Imbalance ratio

Table: Parameters of the 100 simulations

$$(\mathbf{X}, y) \text{ s.t. } \begin{cases} \mathbf{X} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) & \text{if } y = 0, \\ \mathbf{X} \sim N_2\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}\right) & \text{if } y = 1. \end{cases}$$

Contour Plot of the Data Distribution

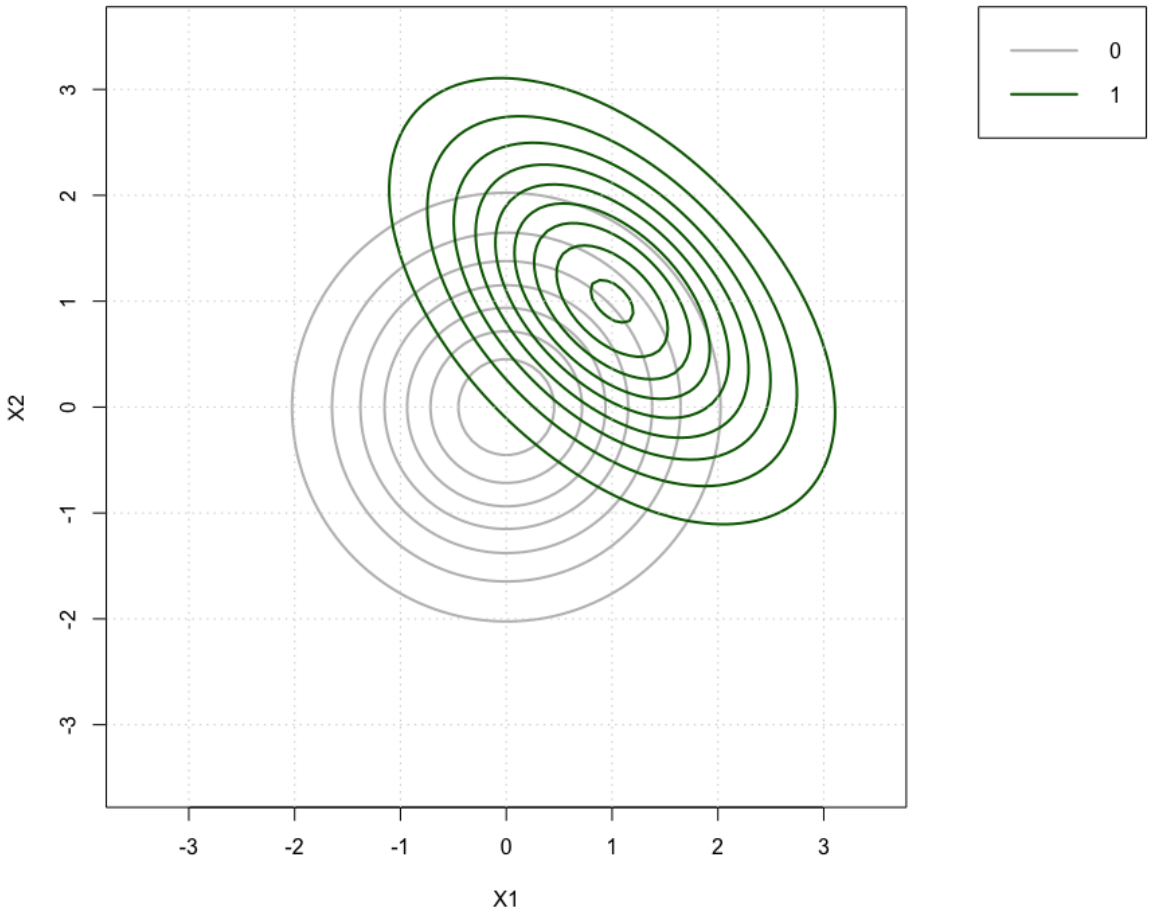


Figure: Contour plot of the data distribution

Example of Generated Data

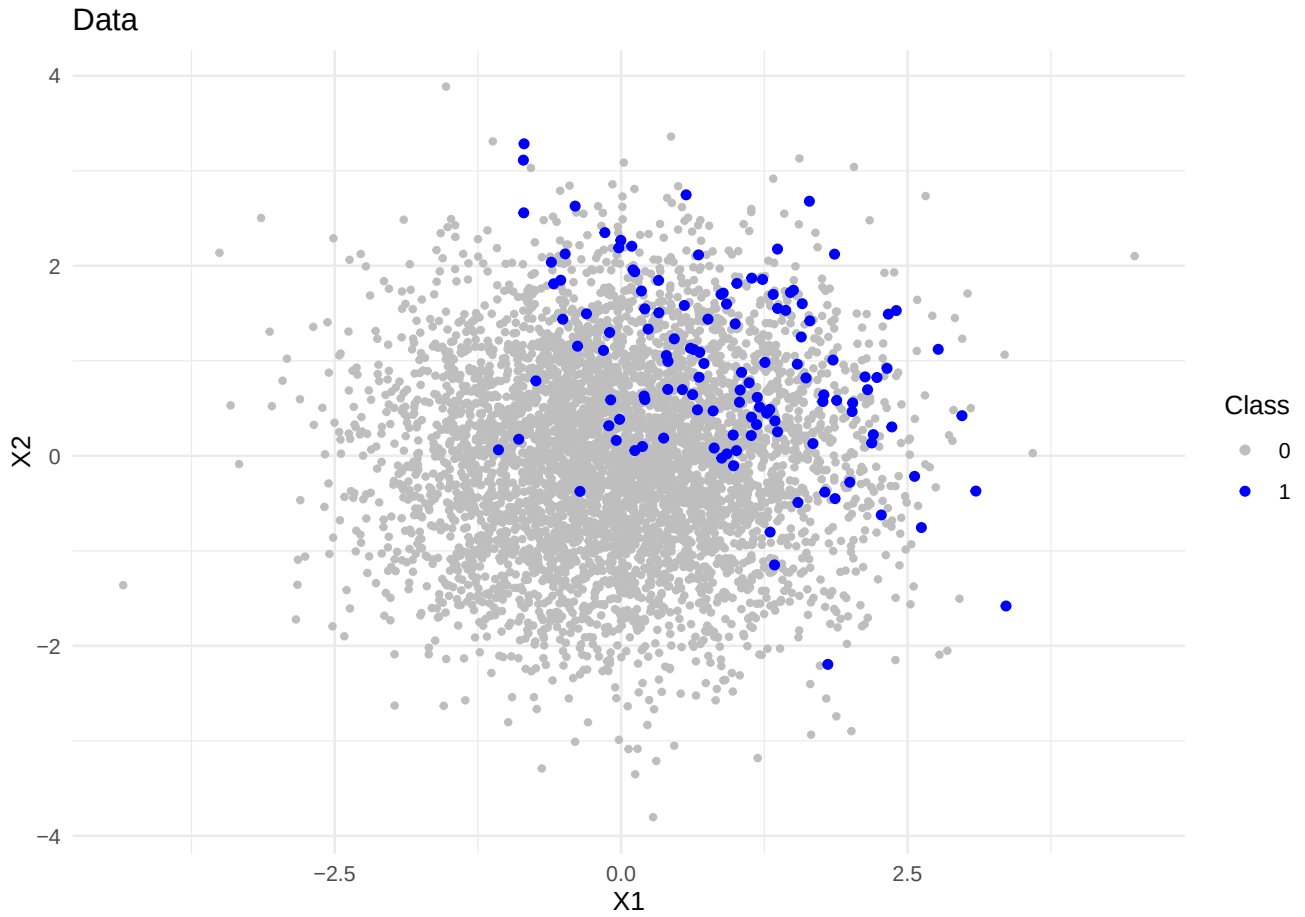


Figure: Data generated with *size* = 5000 and $\pi = 0.025$

Assessment Method

- 100 Static Train/Test Division
- Metrics: Balanced Accuracy, F1 Score
- Models: Decision Tree, Logistic Regression
- Comparisons with Boxplots

SMOTE

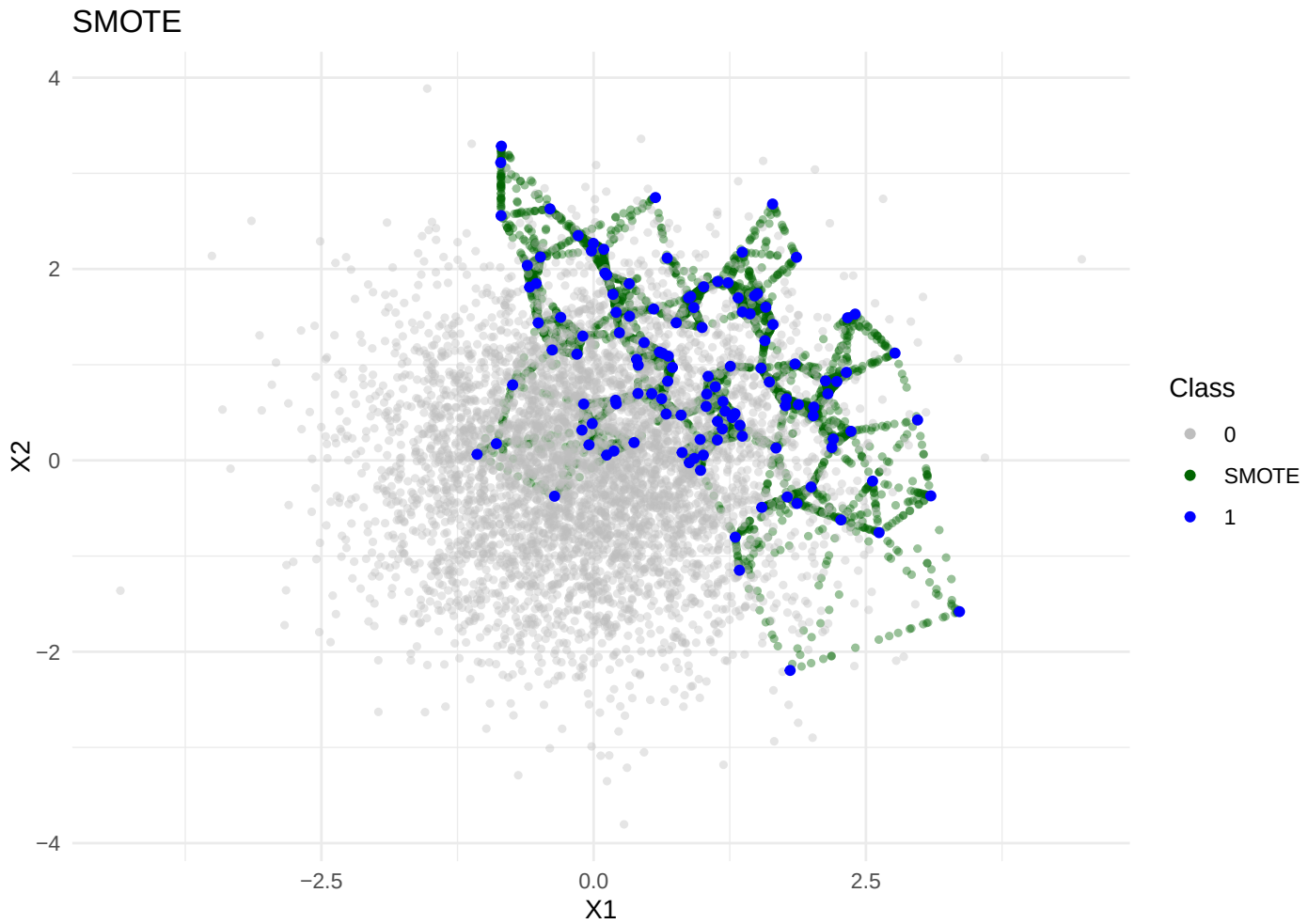


Figure: Synthetic data generated with SMOTE

SMOTE-DIRICHLET

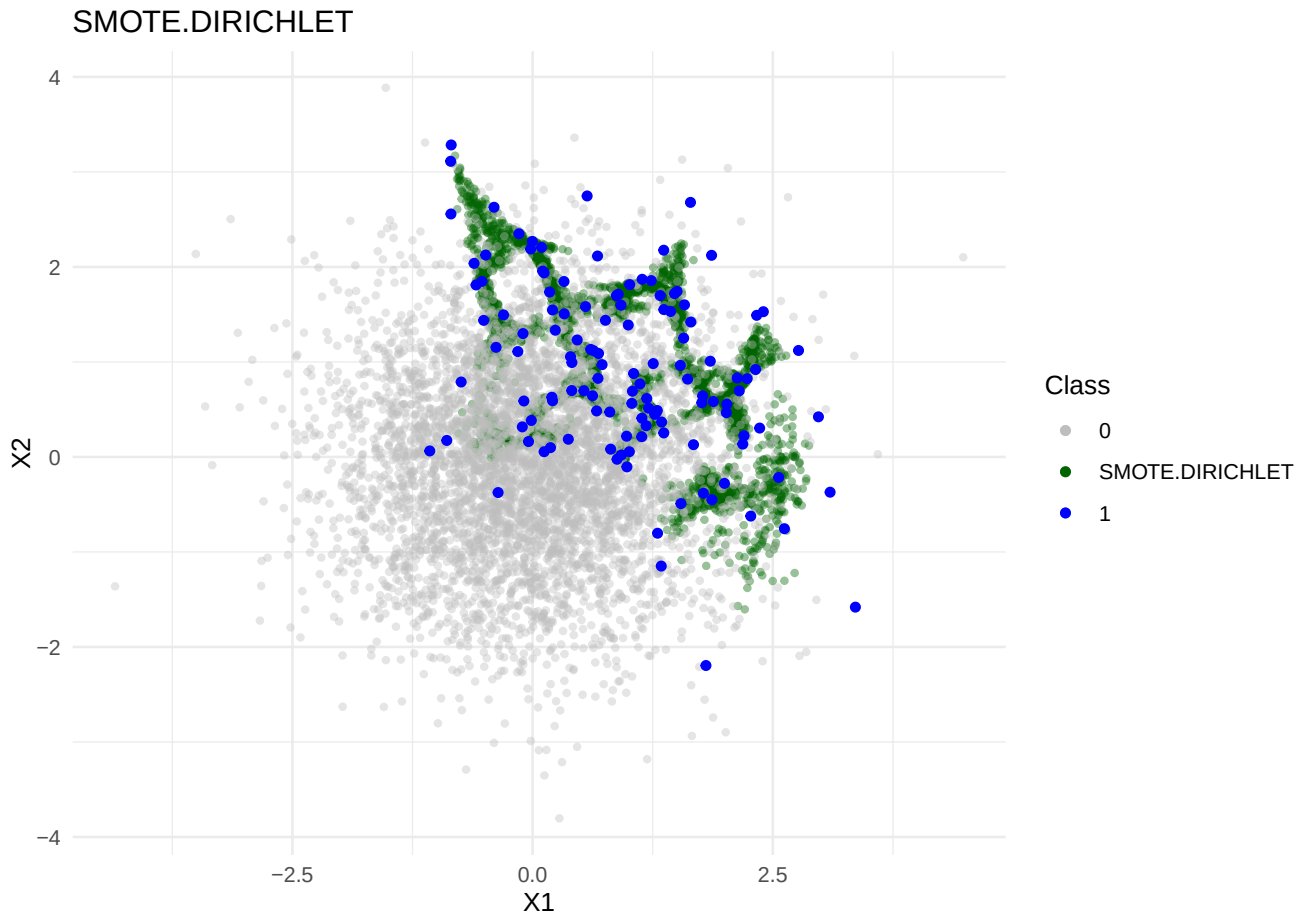
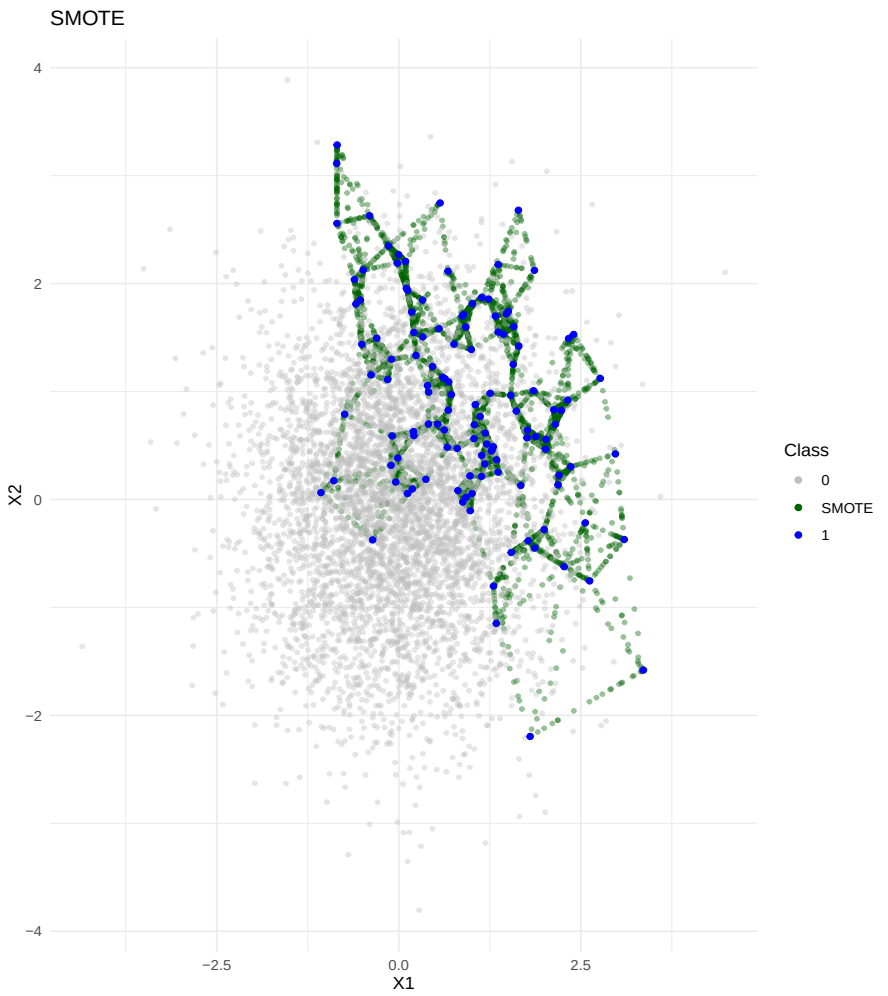
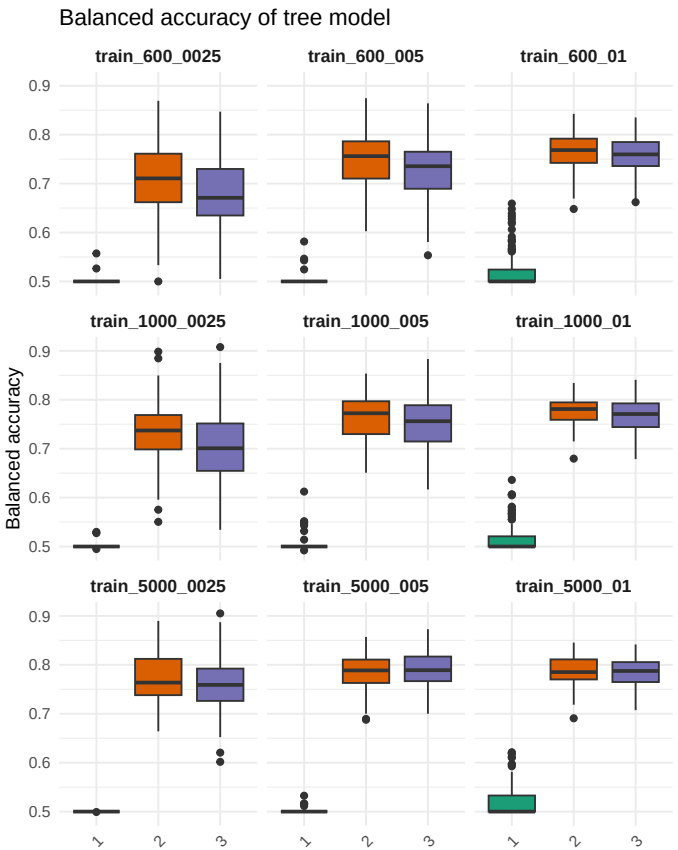


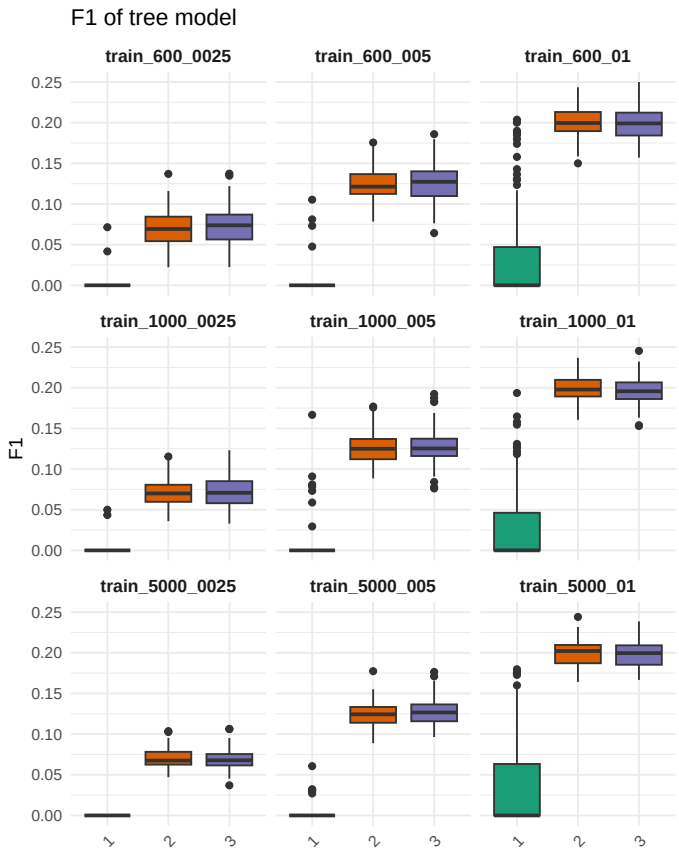
Figure: Synthetic data generated with Dirichlet SMOTE using Dirichlet distribution: $P(p|\alpha) \sim Dir(\alpha_1, ..., \alpha_k) = \frac{\Gamma(\sum_j \alpha_j)}{\Pi_j \Gamma(\alpha_j)} \Pi_{j=1}^{k} p_j^{\alpha_j - 1}$. $\alpha = 1$, $k = 3, 5$. [3]

Comparison between SMOTE and Dirichlet SMOTE



Metrics of Tree Model





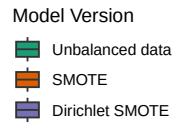
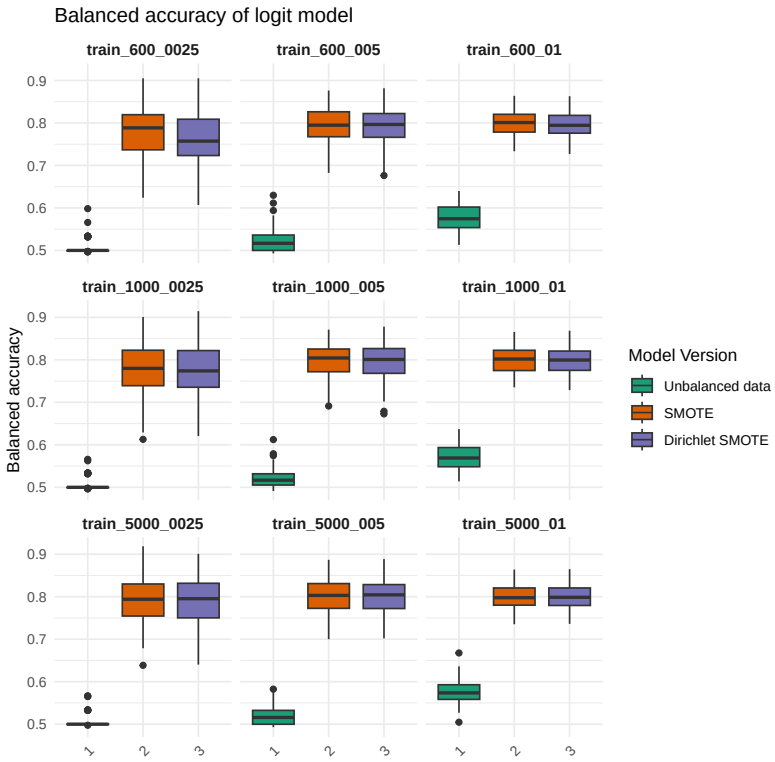


Figure: Distribution of the balanced accuracy and F1 for the classification tree. Each boxplot refers to a different couple of trainset and testset. The threshold for all models is left as default (0.5).

Metrics of Logistic Regression Model



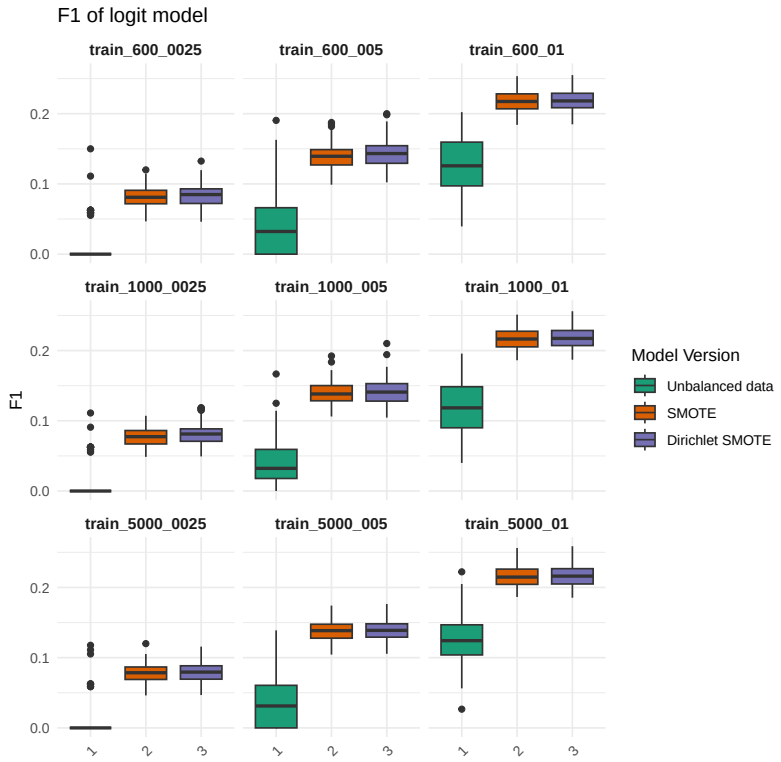
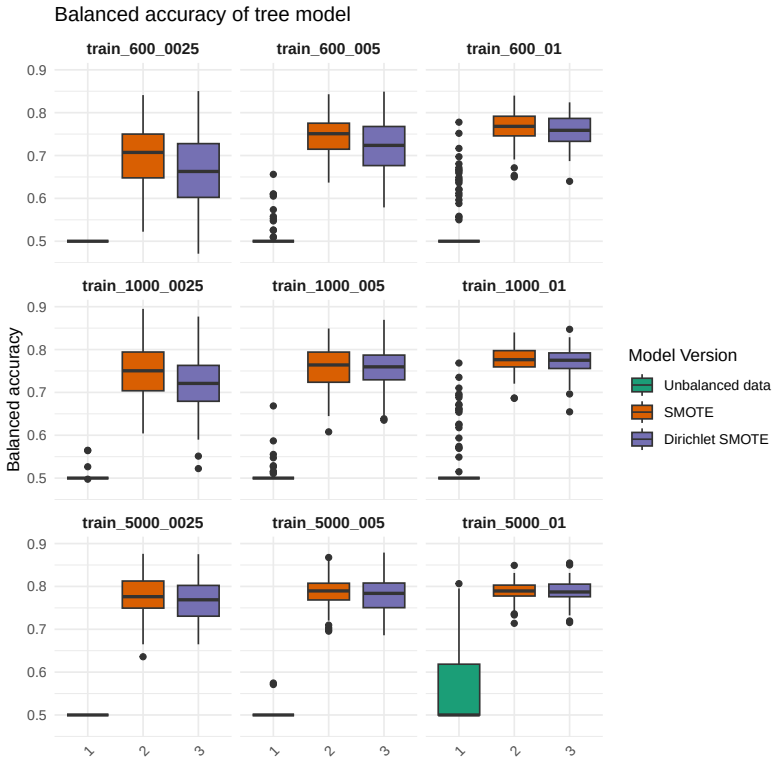


Figure: Distribution of the balanced accuracy and F1 for the logit model. Each boxplot refers to a different couple of trainset and testset. The threshold for all models is left as default (0.5).

Metrics of Tree Model with Optimized Threshold



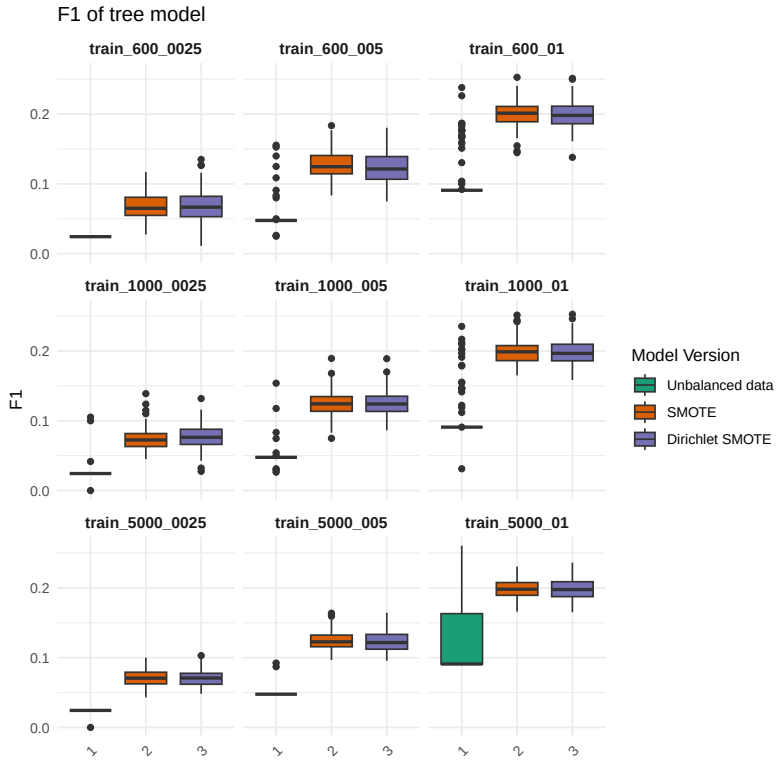
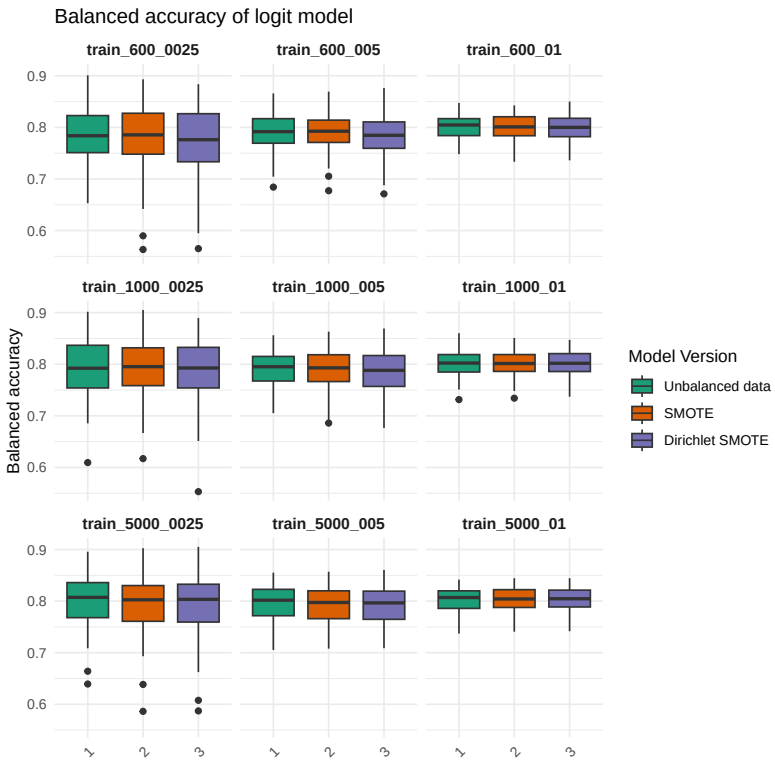


Figure: Distribution of the balanced accuracy and F1 for the classification tree. Each boxplot refers to a different couple of trainset and testset. The threshold for the model learned on unbalanced data is equal to π.

Metrics of Logistic Regression Model with Optimized Threshold



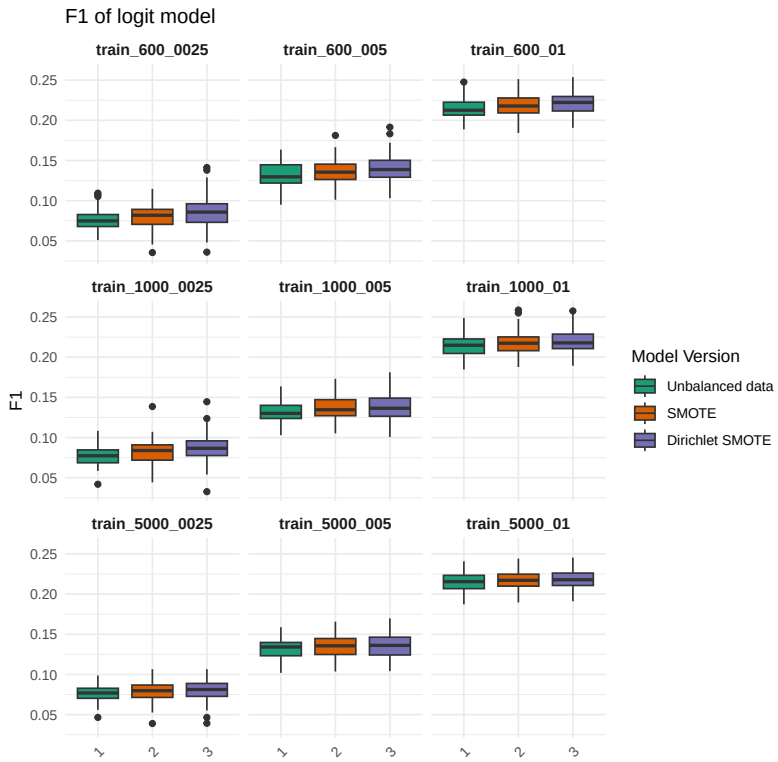
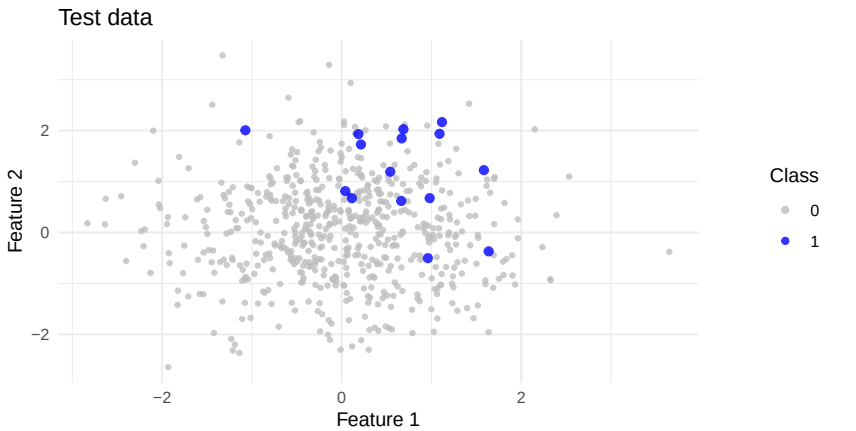
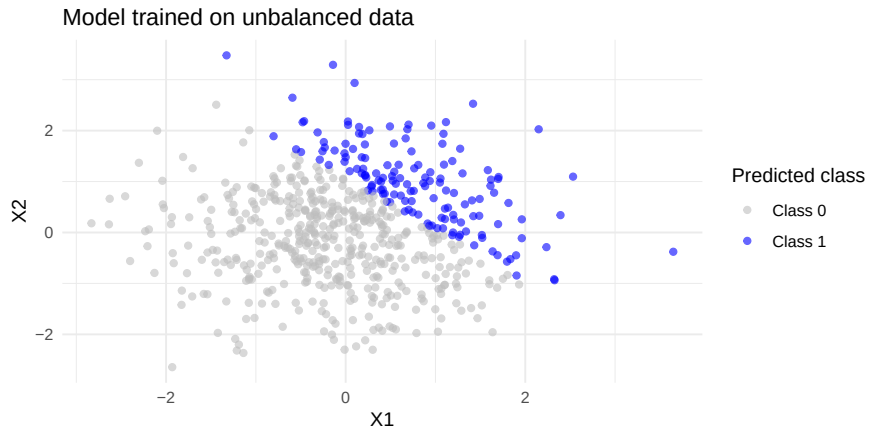
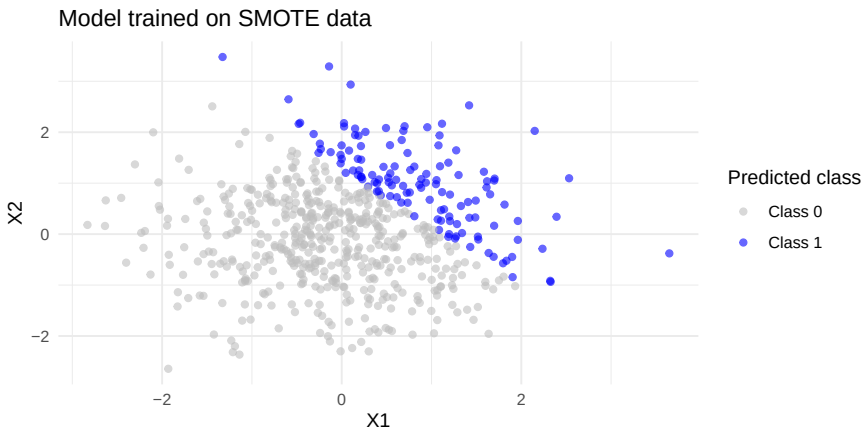


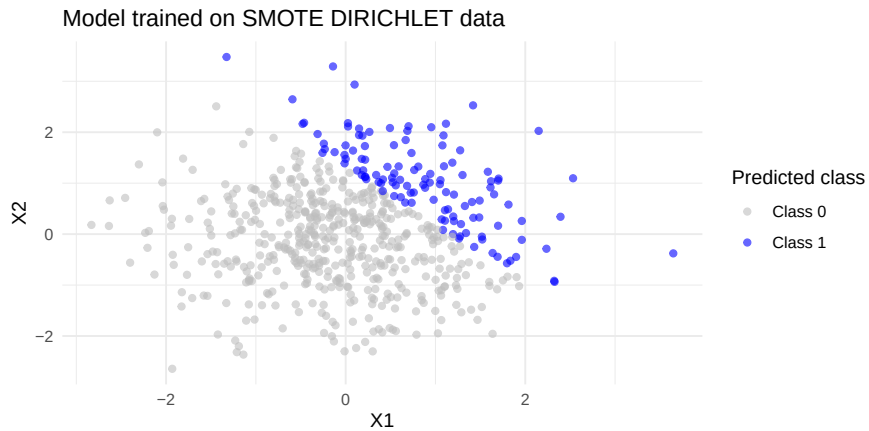
Figure: Distribution of the balanced accuracy and F1 for the logit model. Each boxplot refers to a different couple of trainset and testset. The threshold for the model learned on unbalanced data is equal to π.

Decision boundaries of logistic models









References

- Chawla, Nitesh V et al. (2002). "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16, pp. 321–357.
- Goorbergh, Ruben van den et al. (2022). "The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression". In: *Journal of the American Medical Informatics Association* 29.9, pp. 1525–1534.
- Matharaarachchi, Surani, Mike Domaratzki, and Saman Muthukumarana (2024). "Enhancing SMOTE for imbalanced data with abnormal minority instances". In: *Machine Learning with Applications* 18, p. 100597.
- Menardi, Giovanna and Nicola Torelli (2014). "Training and assessing classification rules with imbalanced data". In: *Data mining and knowledge discovery* 28, pp. 92–122.