

Tera

Aula #23


Class Imbalance & Execução de Projeto

Gabriel Cypriano

T

Como será?



 Reviewed Dataset
1749

Credit Card Fraud Detection

Anonymized credit card transactions labeled as fraudulent or genuine



Machine Learning Group - ULB • last updated 3 months ago

[Overview](#)[Data](#)[Kernels](#)[Discussion](#)[Activity](#)[Download \(66 MB\)](#)[New Kernel](#)

T

✓ Reviewed Dataset

▲
1749

Credit Card Fraud Detection

Anonymized credit card transactions labeled as fraudulent or genuine



Machine Learning Group - ULB • last updated 3 months ago

[Overview](#)

[Data](#)

[Kernels](#)


[Discussion](#)

[Activity](#)

[Download \(66 MB\)](#)

[New Kernel](#)

- Genuínas: 0
- Fraudulentas: 1

 Reviewed Dataset 1749

Credit Card Fraud Detection

Anonymized credit card transactions labeled as fraudulent or genuine



Machine Learning Group - ULB • last updated 3 months ago

[Overview](#)[Data](#)[Kernels](#)[Discussion](#)[Activity](#)[Download \(66 MB\)](#)[New Kernel](#)

- 285 mil transações
- 2 dias
- Na Europa em setembro/2013

T

✓ Reviewed Dataset

▲
1749

Credit Card Fraud Detection

Anonymized credit card transactions labeled as fraudulent or genuine



Machine Learning Group - ULB • last updated 3 months ago

[Overview](#)

[Data](#)

[Kernels](#)


[Discussion](#)

[Activity](#)

[Download \(66 MB\)](#)

[New Kernel](#)

- **Amount:** valor da transação
- **Time:** tempo da transação (em segundos) relativo à primeira transação do dataset

 Reviewed Dataset 1749

Credit Card Fraud Detection

Anonymized credit card transactions labeled as fraudulent or genuine



Machine Learning Group - ULB • last updated 3 months ago

[Overview](#)[Data](#)[Kernels](#)[Discussion](#)[Activity](#)[Download \(66 MB\)](#)[New Kernel](#)

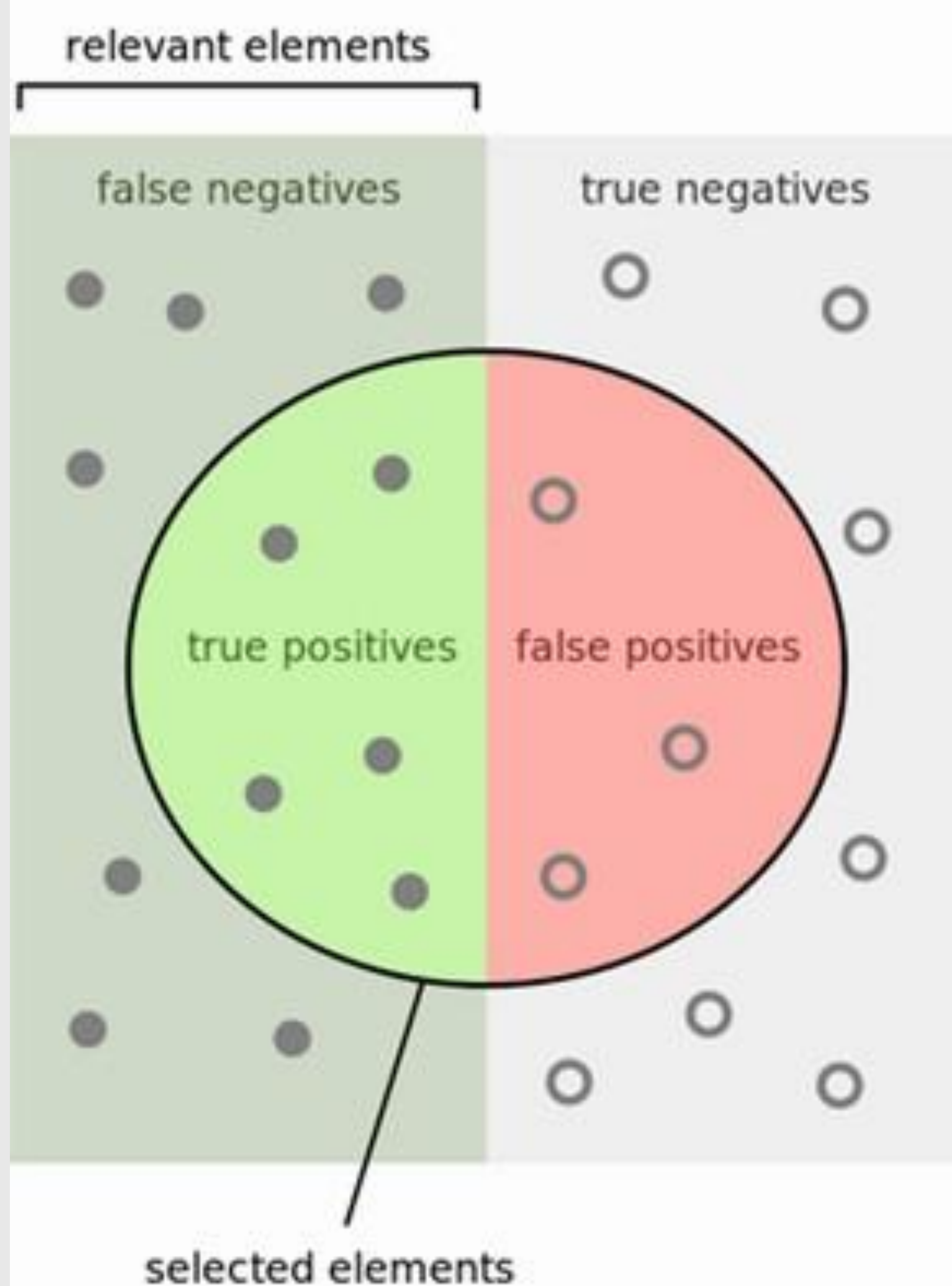
- As outras features são anonimizadas
- Revisão da análise exploratória básica e feature engineering
- Foco em diferentes técnicas de treinamento

Familiarização com o dataset

Discussão

Qual métrica utilizar?

T

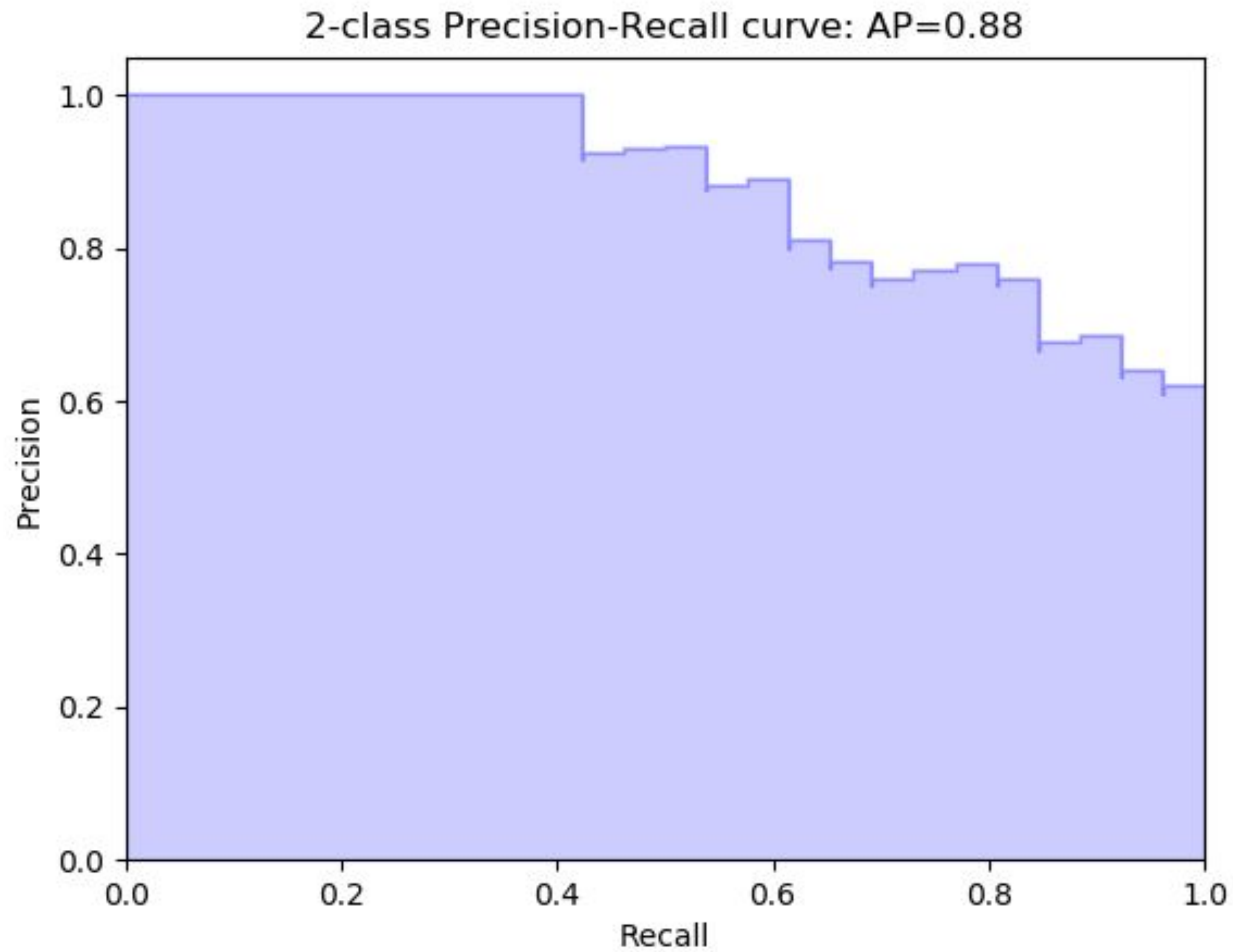


How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$



$$F_1 = 2 * \frac{\textit{precision} * \textit{recall}}{\textit{precision} + \textit{recall}}$$

Modelo naïve
(sem se preocupar com
desbalanceamento)

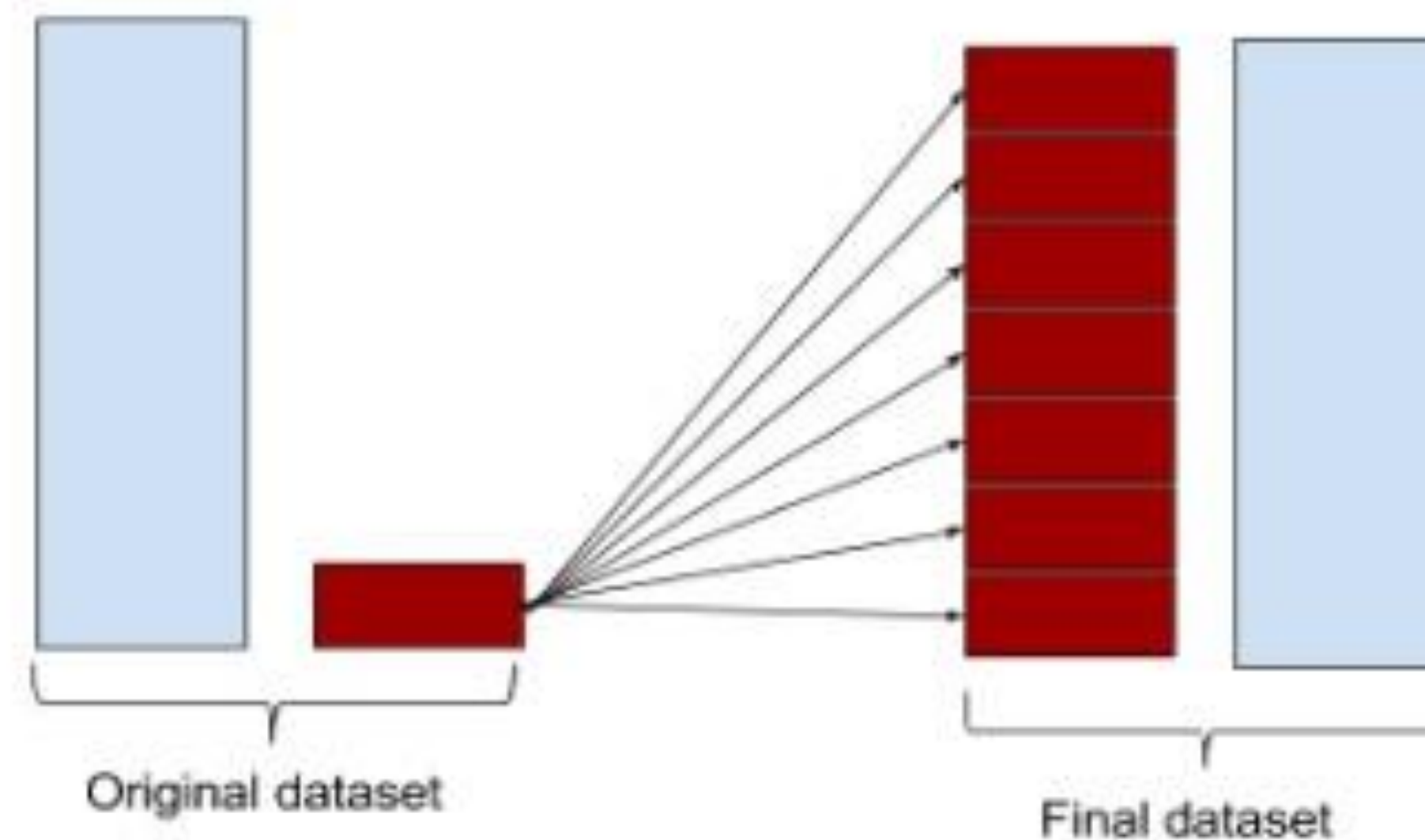
15 minutos

Resampling

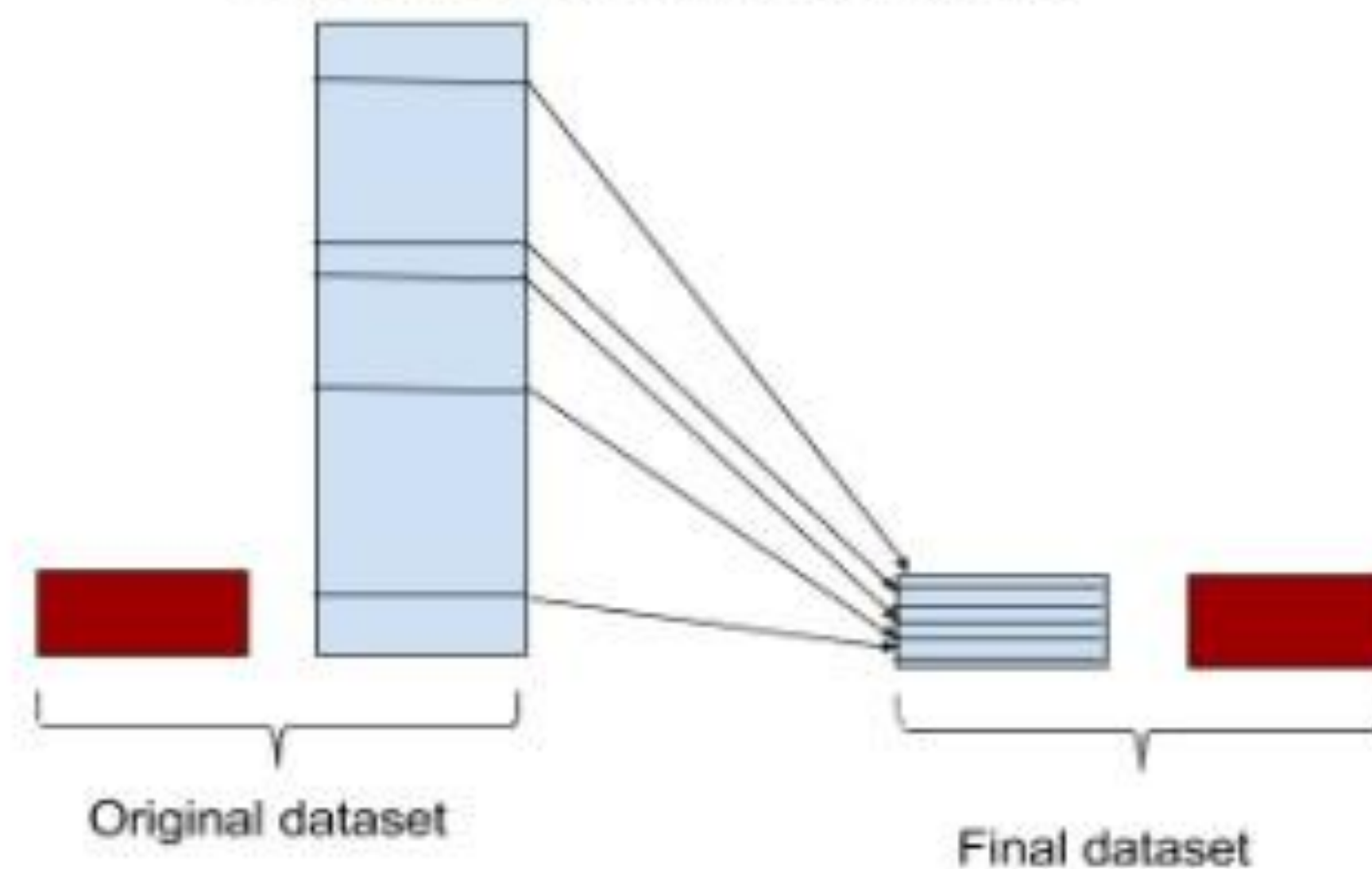


Resampling

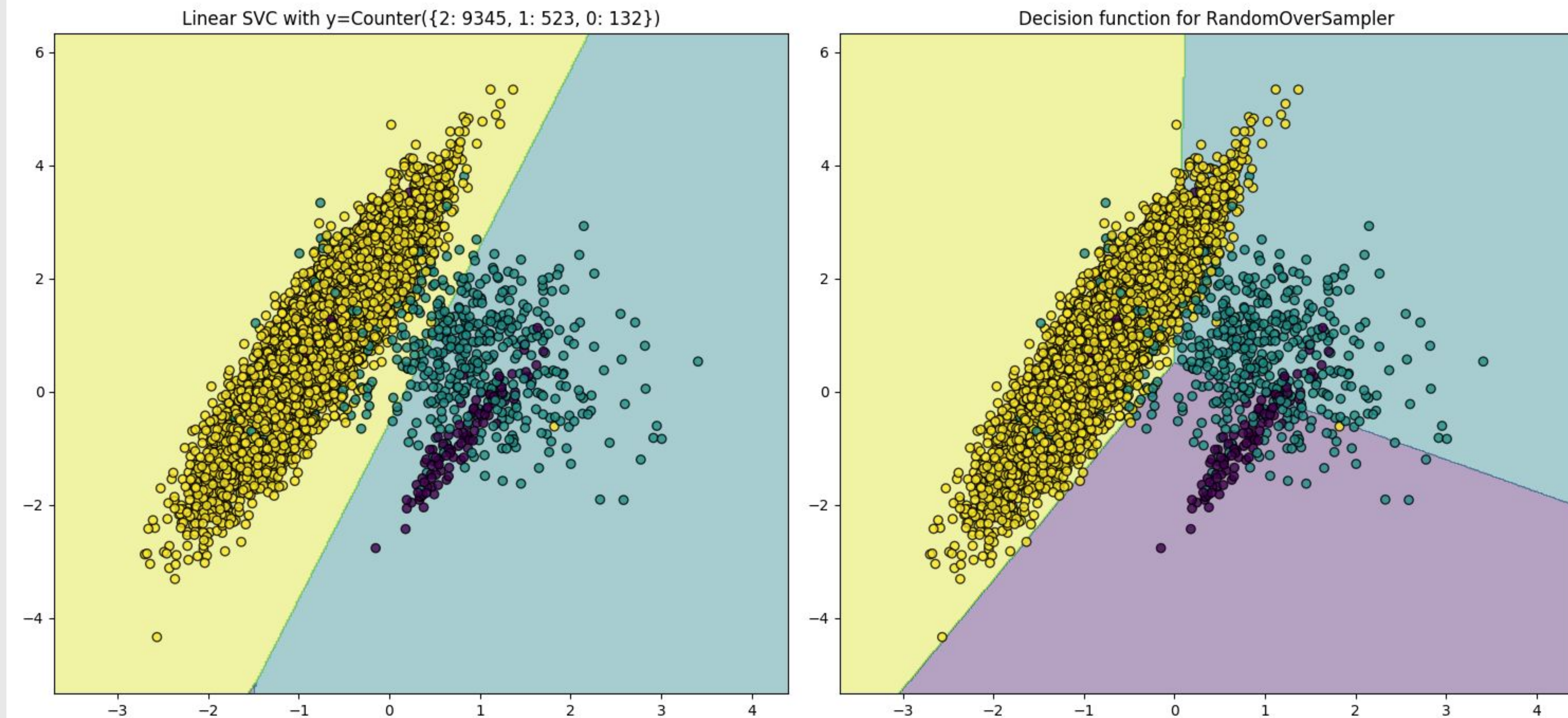
Oversampling minority class



Undersampling majority class



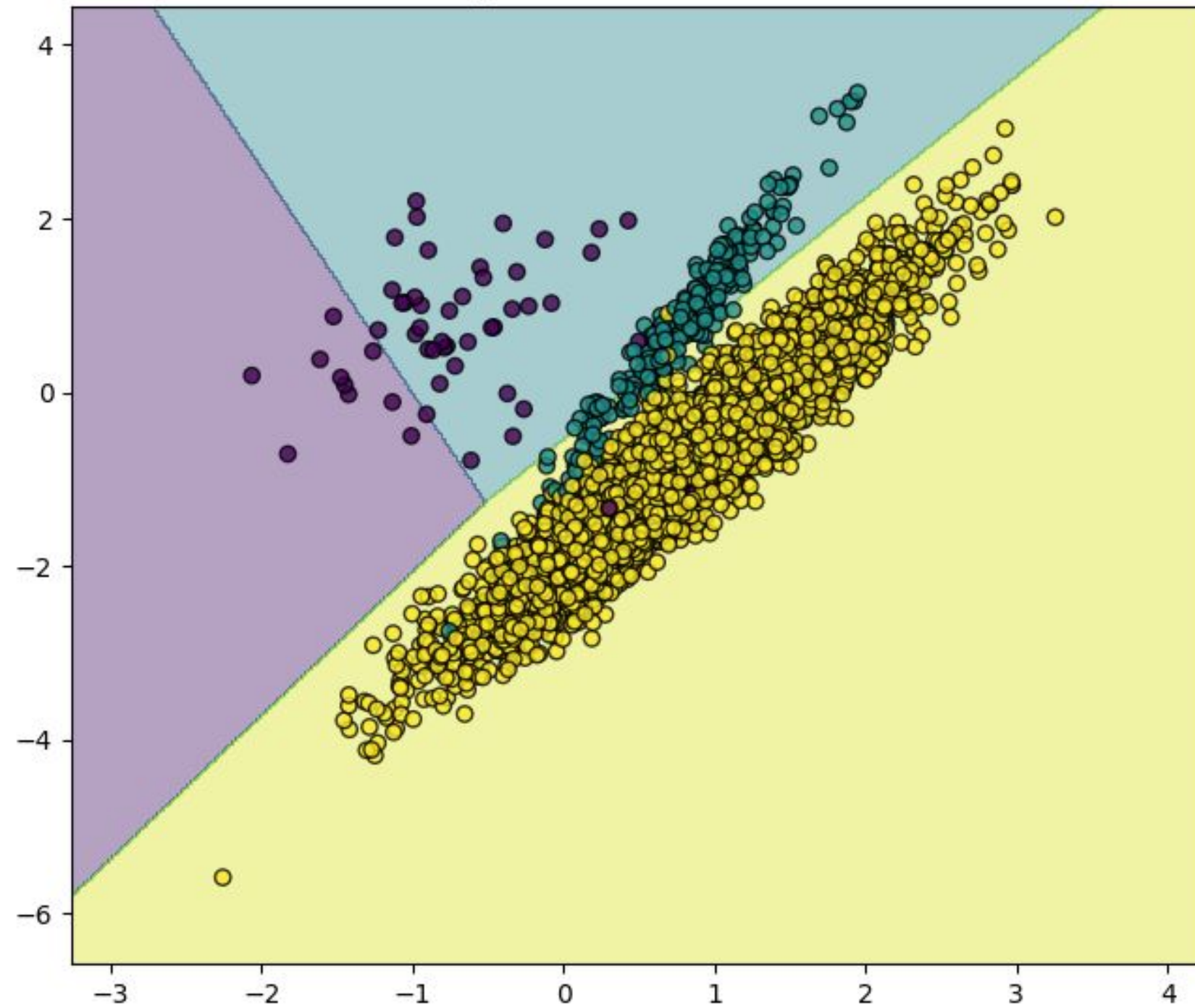
Random Oversampling



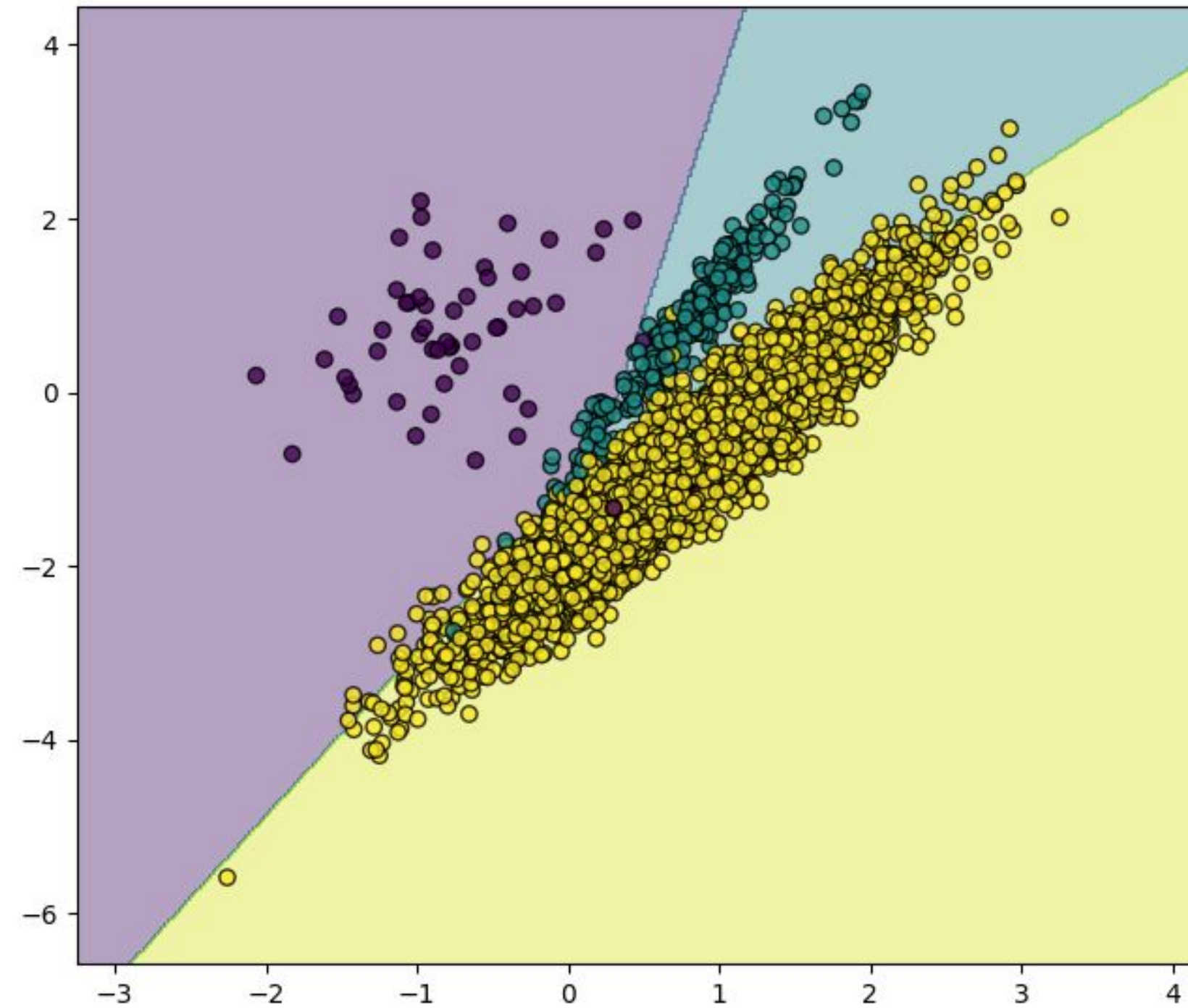
T

Random Undersampling

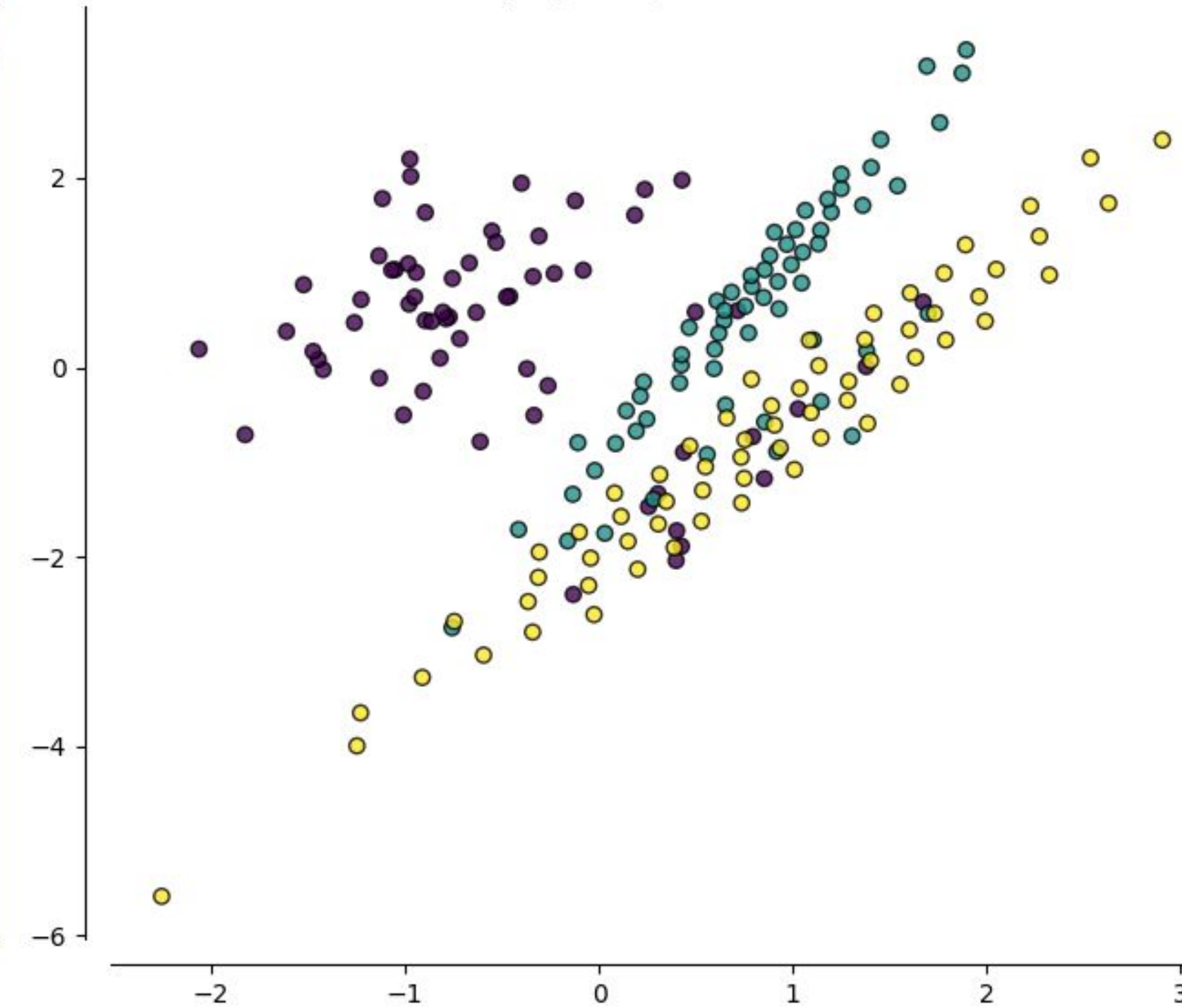
Linear SVC with y=Counter({2: 4674, 1: 262, 0: 64})



Decision function for ClusterCentroids



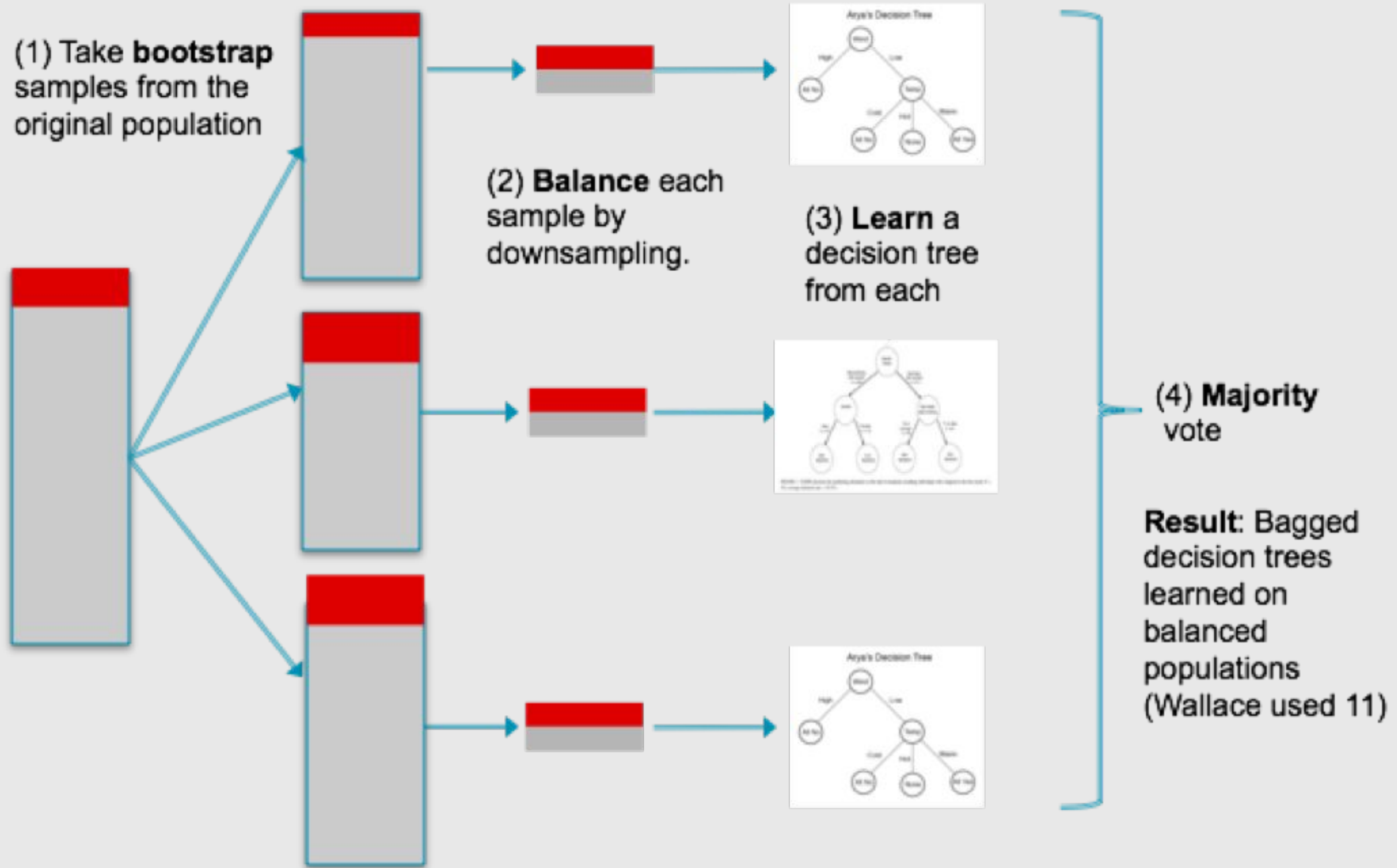
Resampling using ClusterCentroids



Modelo com undersampling

5 minutos

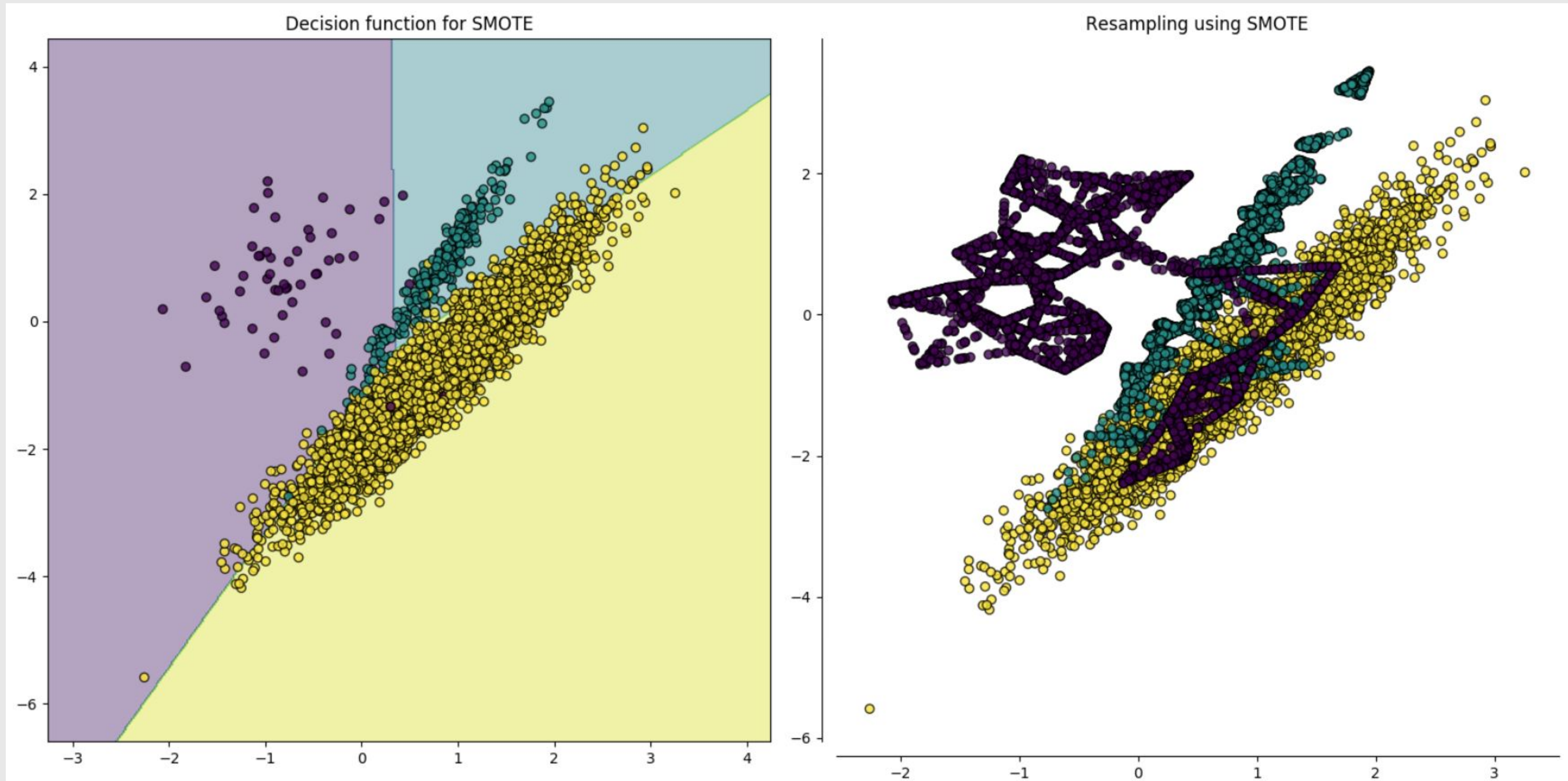
T Balanced Bagging



Modelo com balanced
bagging

5 minutos

T SMOTE (Synthetic Oversampling)



Algoritmos com suporte a balanceamento

- LogisticRegressionClassifier &
RandomForestClassifier
 - **setar** `class_weight='balanced'`
- XGBoostClassifier
 - **setar** `scale_pos_weight=sum(negative cases) / sum(positive cases)`

Modelo com “class weight”

5 minutos

T

Congrats!





Featured Prediction Competition

Zillow Prize: Zillow's Home Value Prediction (Zestimate)

\$1,200,000

Prize Money

Can you improve the algorithm that changed the world of real estate?



Zillow · 3,779 teams · 8 months ago

[Overview](#)

[Data](#)

[Kernels](#)

[Discussion](#)

[Leaderboard](#)

[Rules](#)

[Team](#)

[My Submissions](#)

[Late Submission](#)

T

Intervalo





Featured Prediction Competition

Zillow Prize: Zillow's Home Value Prediction (Zestimate)

\$1,200,000

Prize Money

Can you improve the algorithm that changed the world of real estate?



Zillow · 3,779 teams · 8 months ago

[Overview](#)

[Data](#)

[Kernels](#)

[Discussion](#)

[Leaderboard](#)

[Rules](#)

[Team](#)

[My Submissions](#)

[Late Submission](#)

- Usem o dataset `properties_2016`
- A nossa target será: `taxamount`
- Objetivo: simular a hackathon e treinar um modelo!

Apresentações

