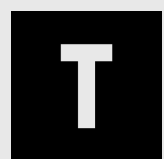


Tera

Aula #23

Execução de Projeto & Class Imbalance

Gabriel Cypriano
23/jun/2018



Como será?

T

Como será?



T

Como será?



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
- Análise exploratória básica e foco em modelagem

Familiarização com o dataset

5 minutos

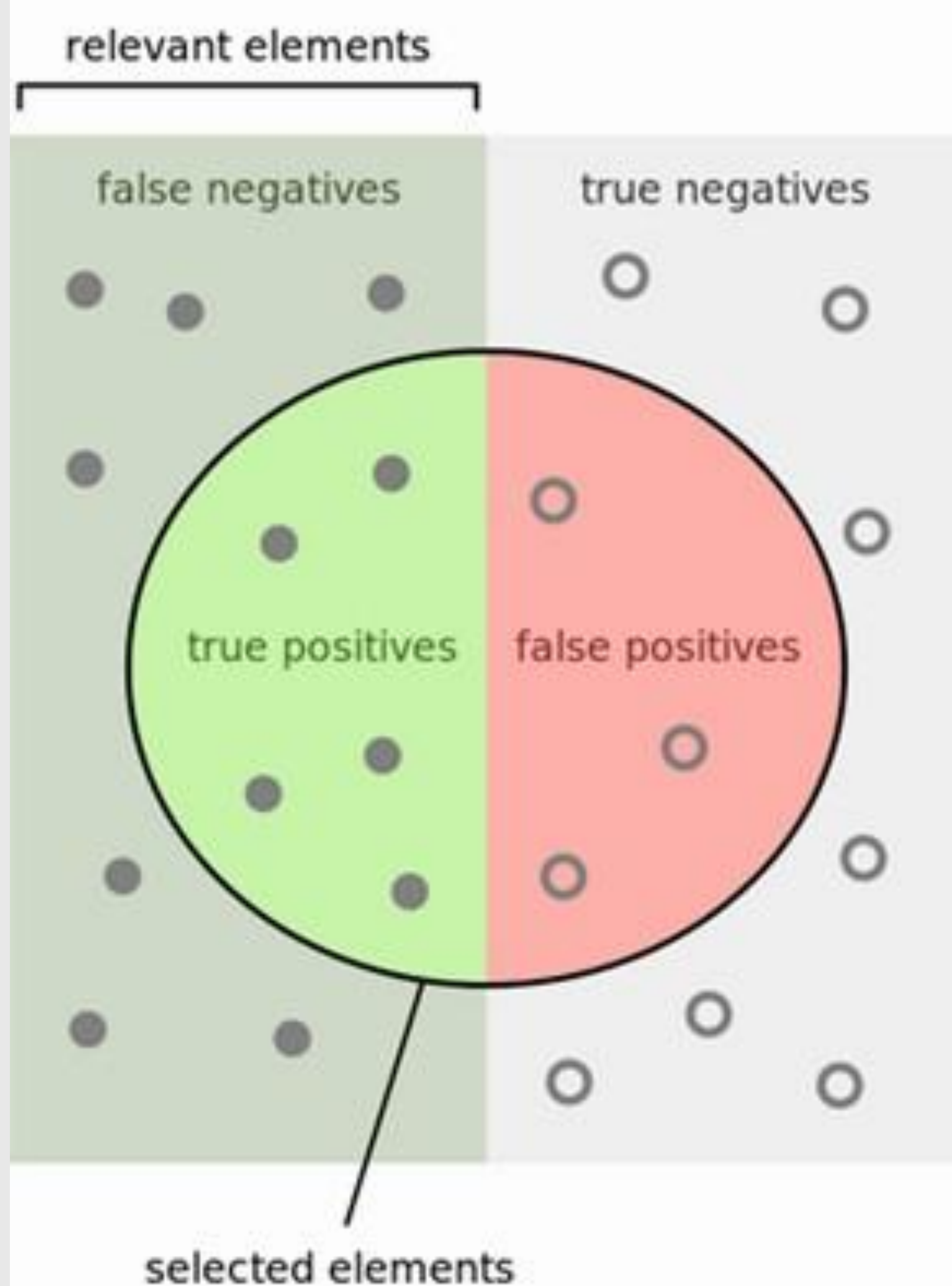
Discussão

Qual métrica utilizar?

Accuracy Paradox

Predictive models with a given level of accuracy may have greater predictive power than models with higher accuracy.

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?

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

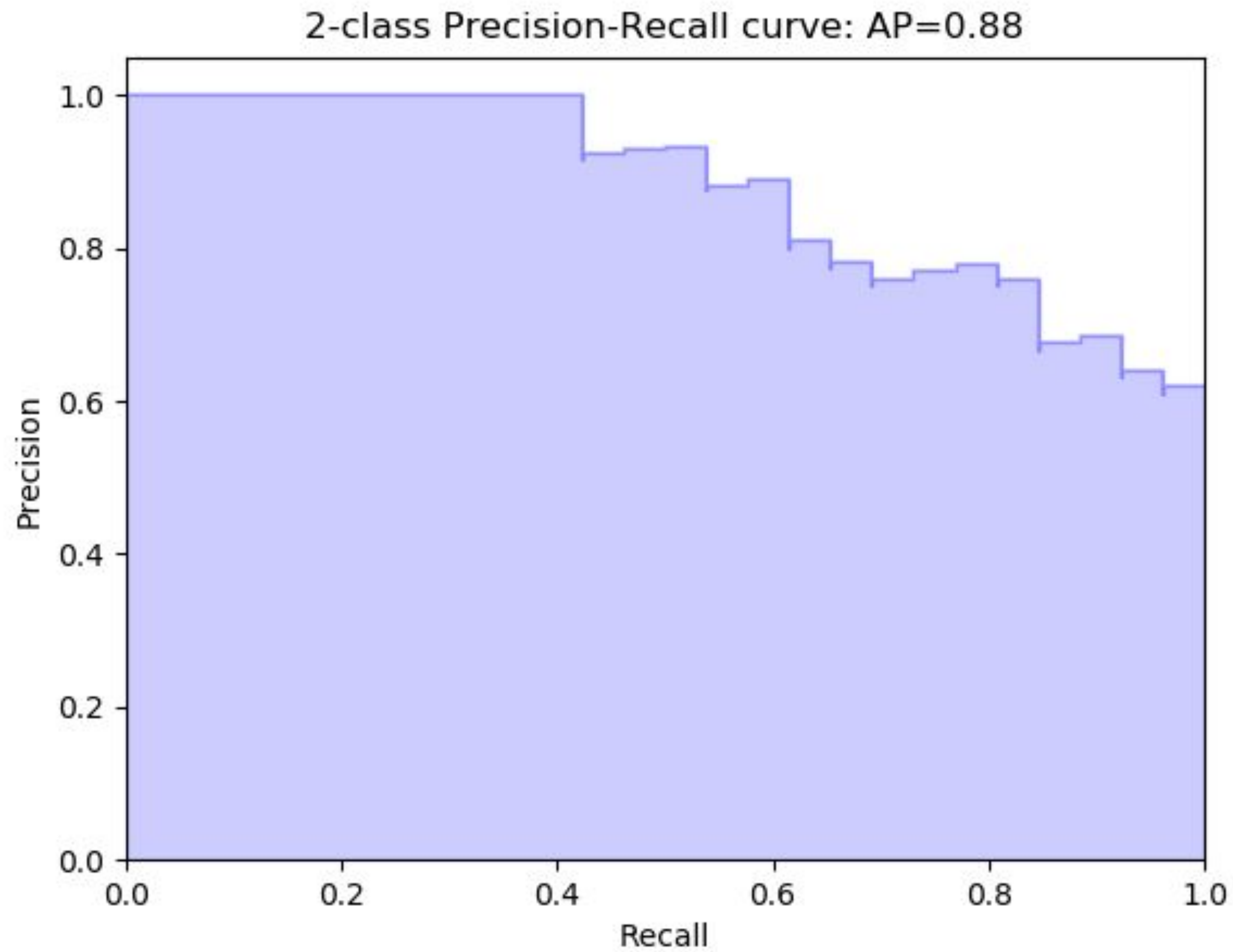
$$F_1 = 2 * \frac{\textit{precision} * \textit{recall}}{\textit{precision} + \textit{recall}}$$

Manipulação dos dados

10 minutos

Treinar modelo e
obter scores

20 minutos



Definir ponto de corte ideal

15 minutos

Criar previsões finais
utilizando o ponto de corte

8 minutos

Complementar avaliação com classification report e matriz de confusão

8 minutos

Apresentações



T

Intervalo



T

Congrats!



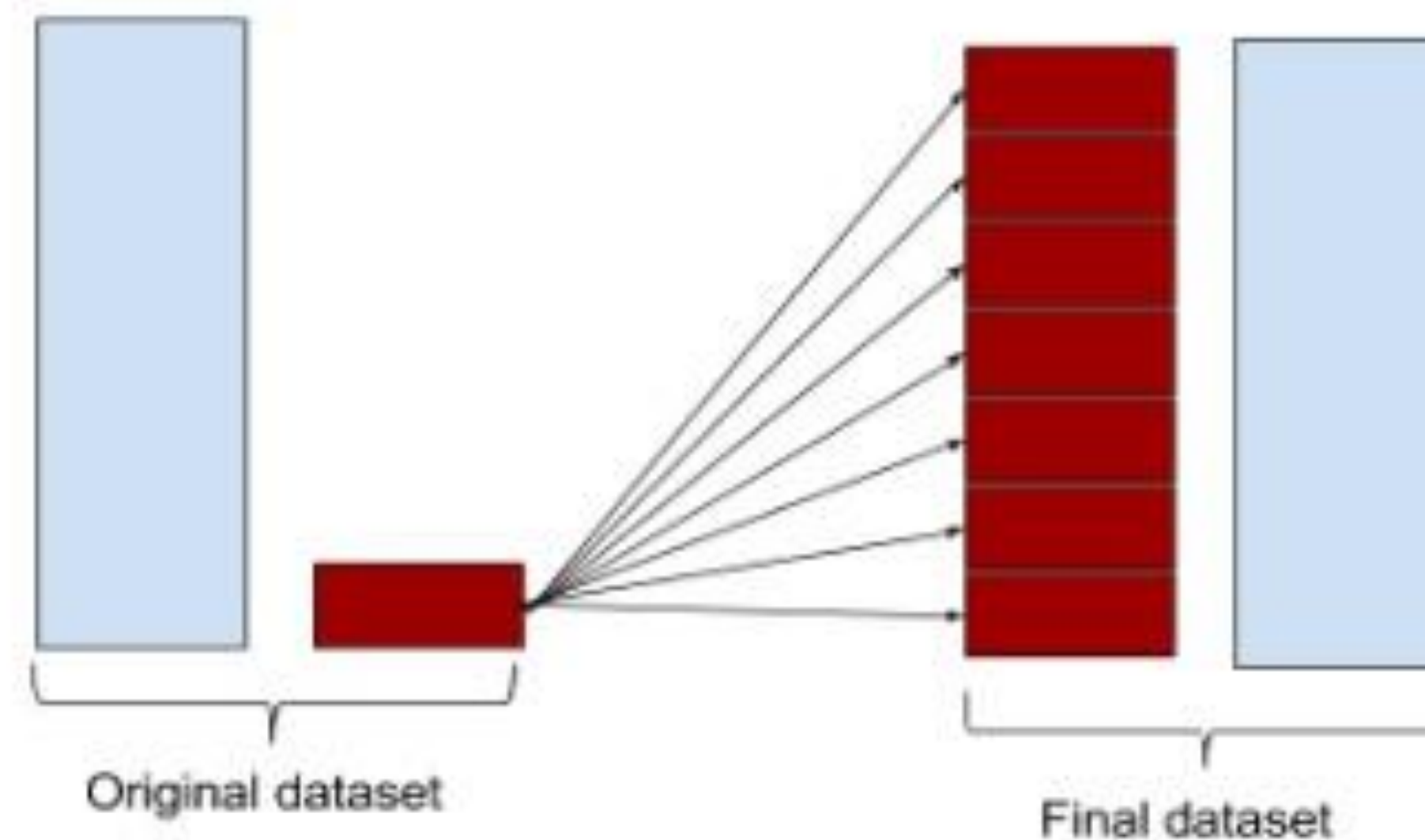
T

Material extra: Resampling

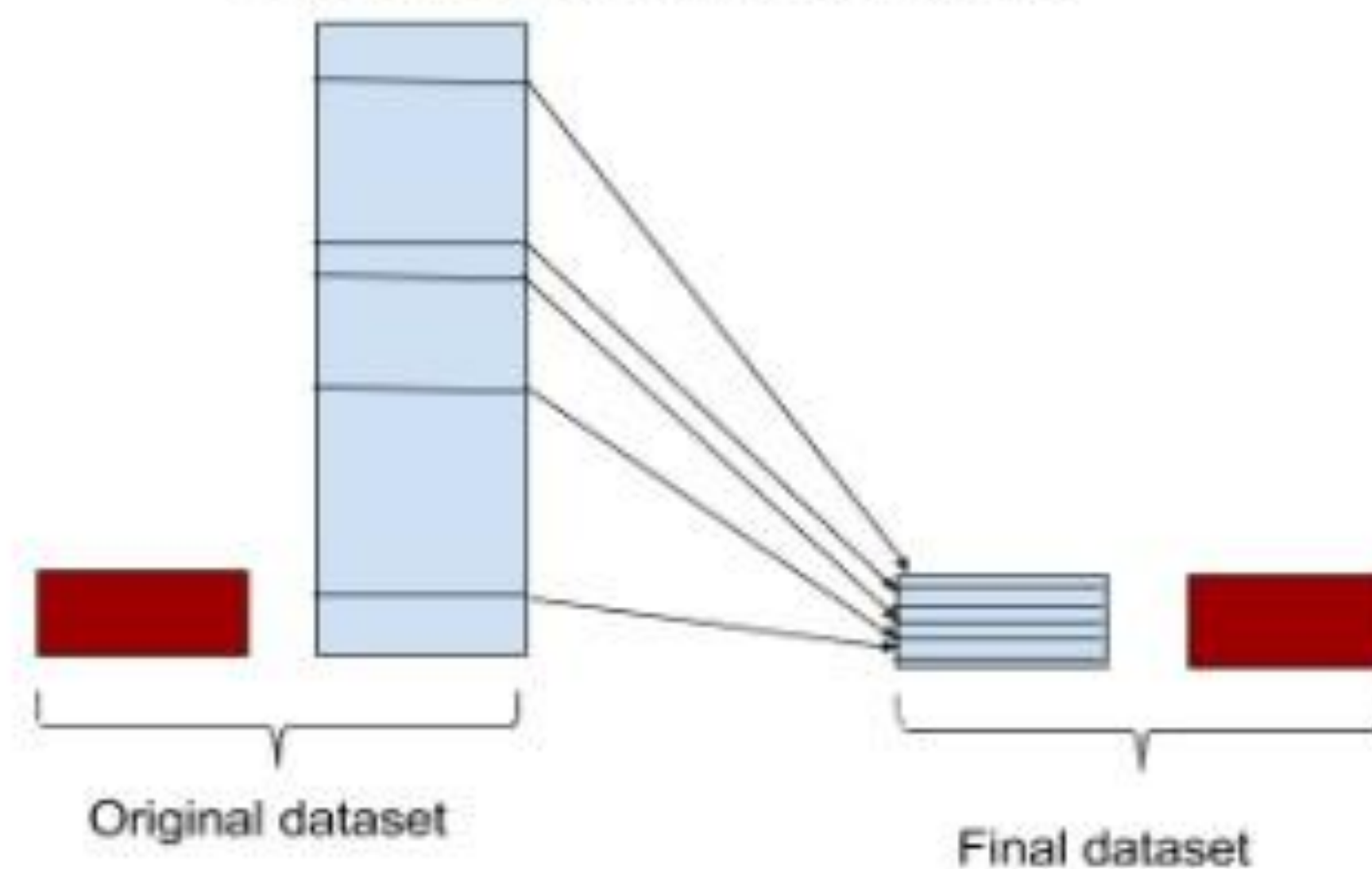


Resampling

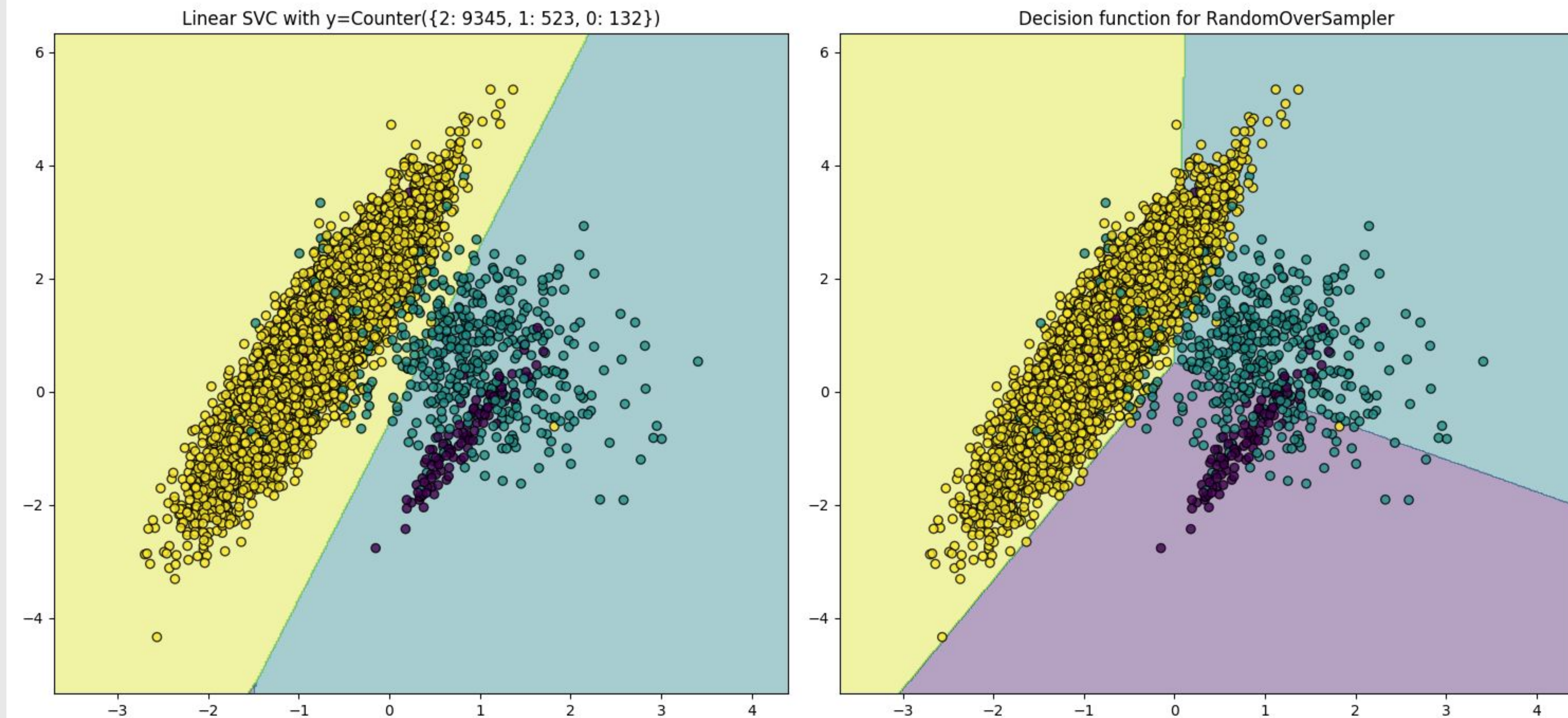
Oversampling minority class



Undersampling majority class



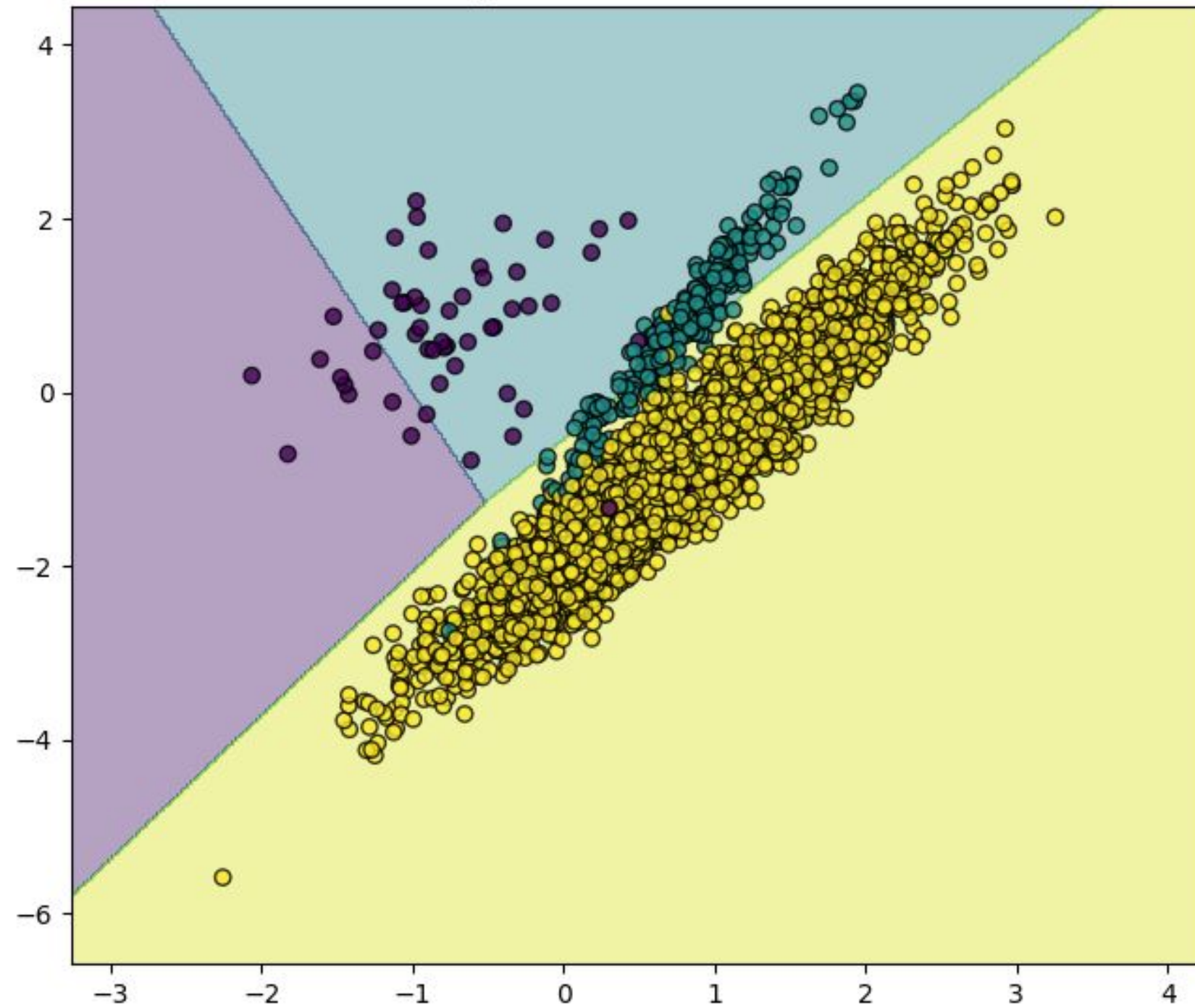
Random Oversampling



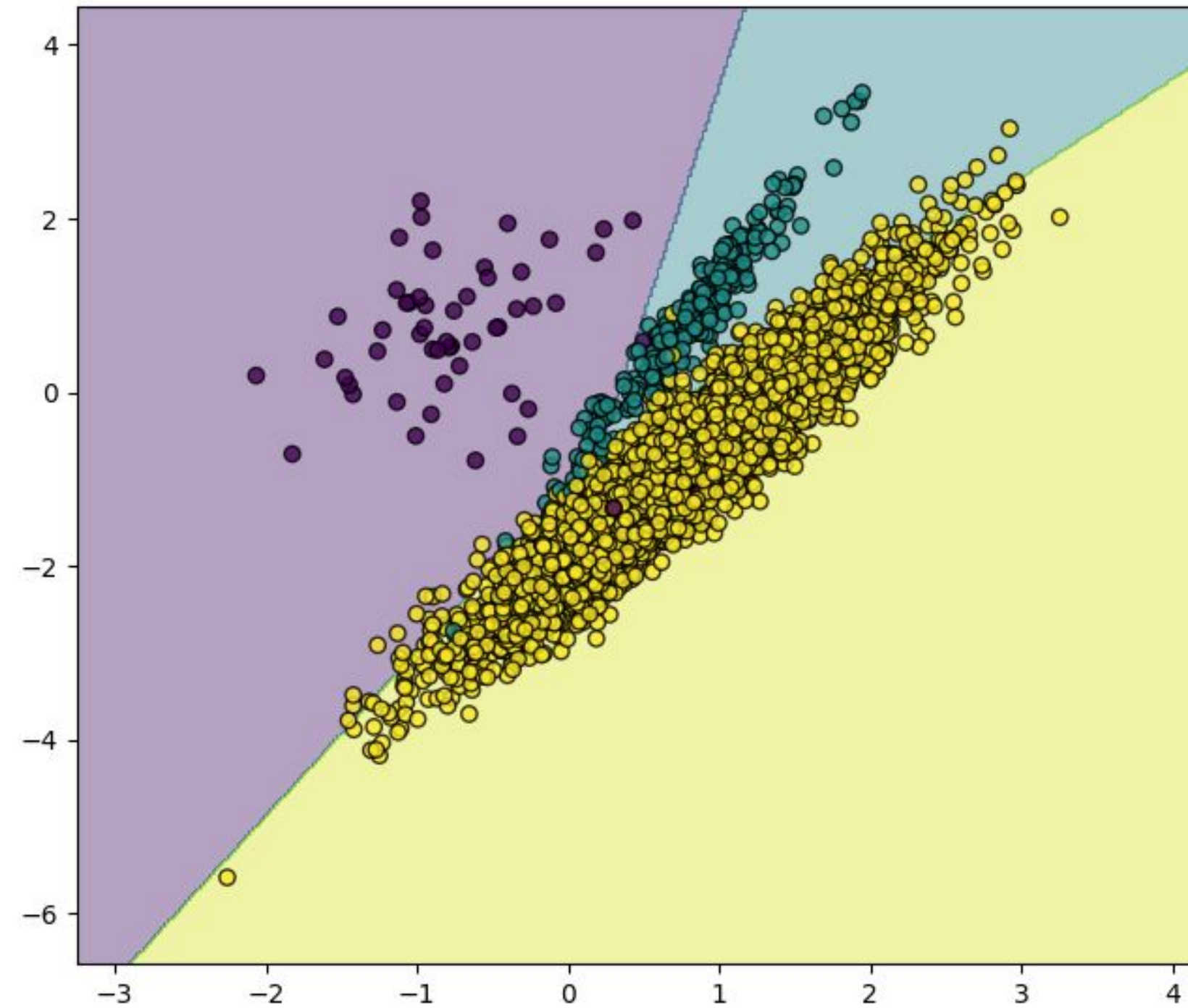
T

Random Undersampling

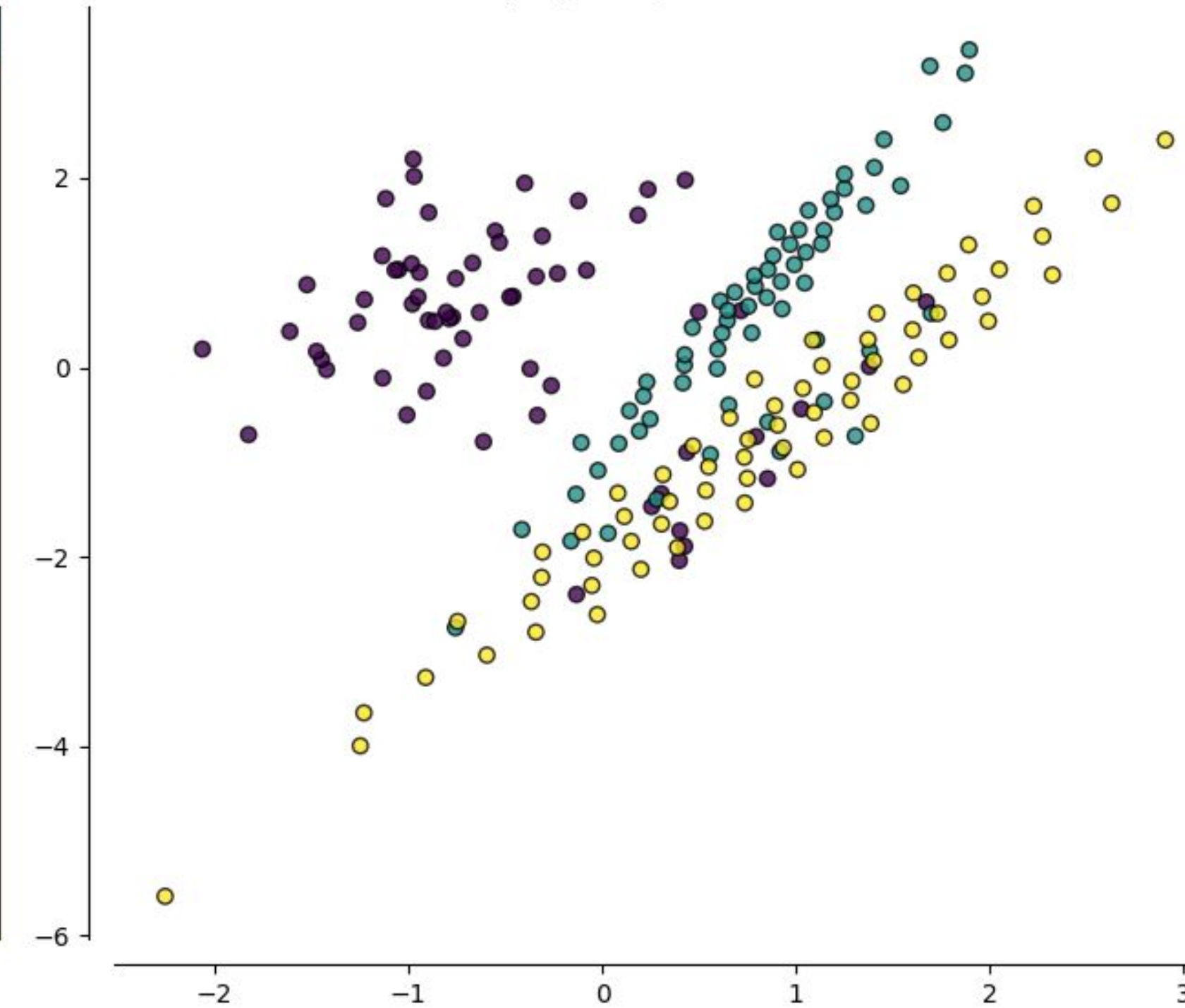
Linear SVC with y=Counter({2: 4674, 1: 262, 0: 64})



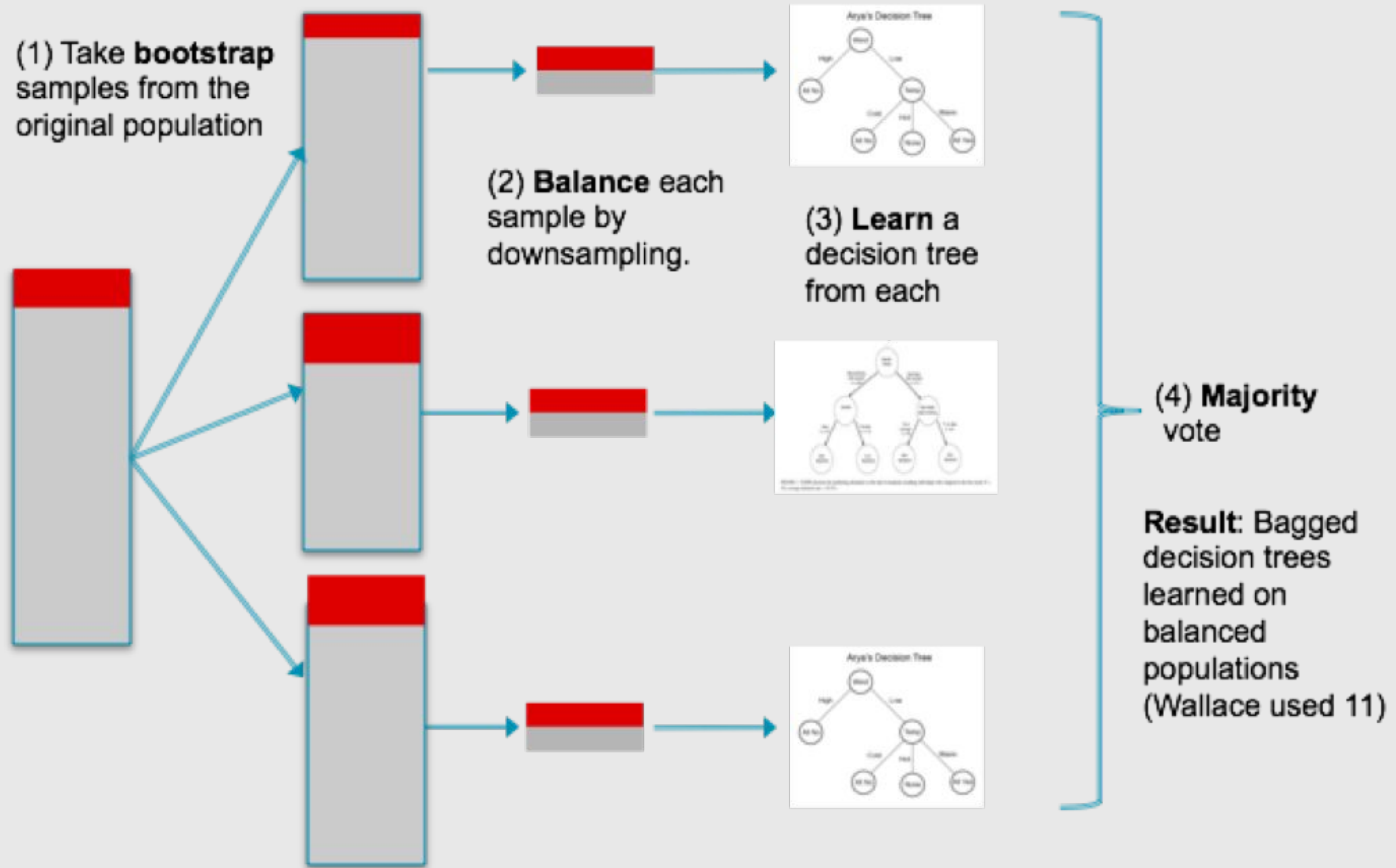
Decision function for ClusterCentroids



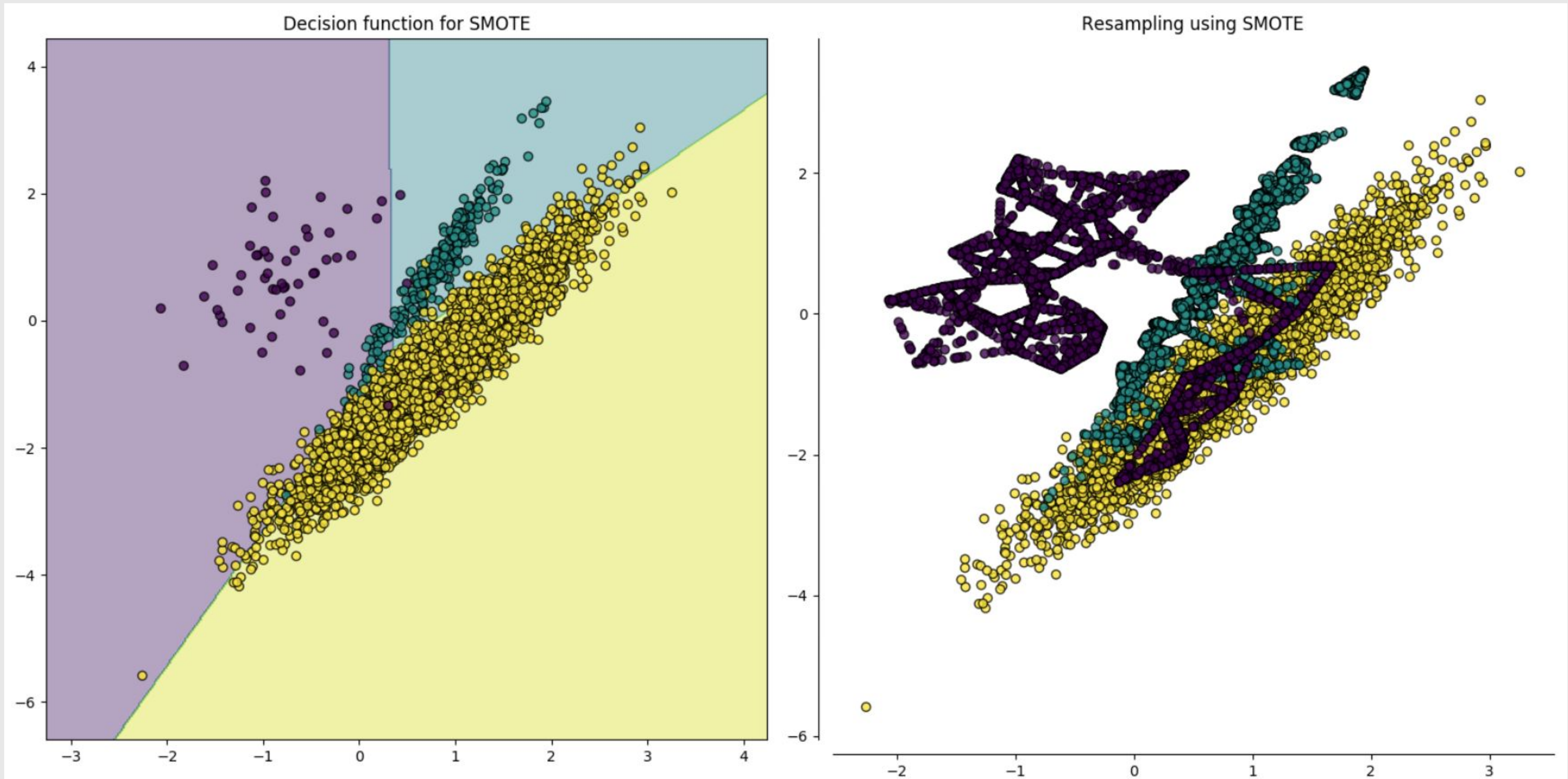
Resampling using ClusterCentroids



T Balanced Bagging

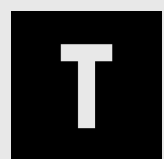


T SMOTE (Synthetic Oversampling)



Pacote imblearn

- RandomOverSampler
- RandomUnderSampler
- SMOTE
- EasyEnsemble & BalancedBaggingClassifier



DÚVIDAS?

