Tera

Aula #23

Execução de Projeto & Class Imbalance

Gabriel Cypriano

23/jun/2018

Como será?

Como será?

Como será?

Como será?

Reviewed Dataset

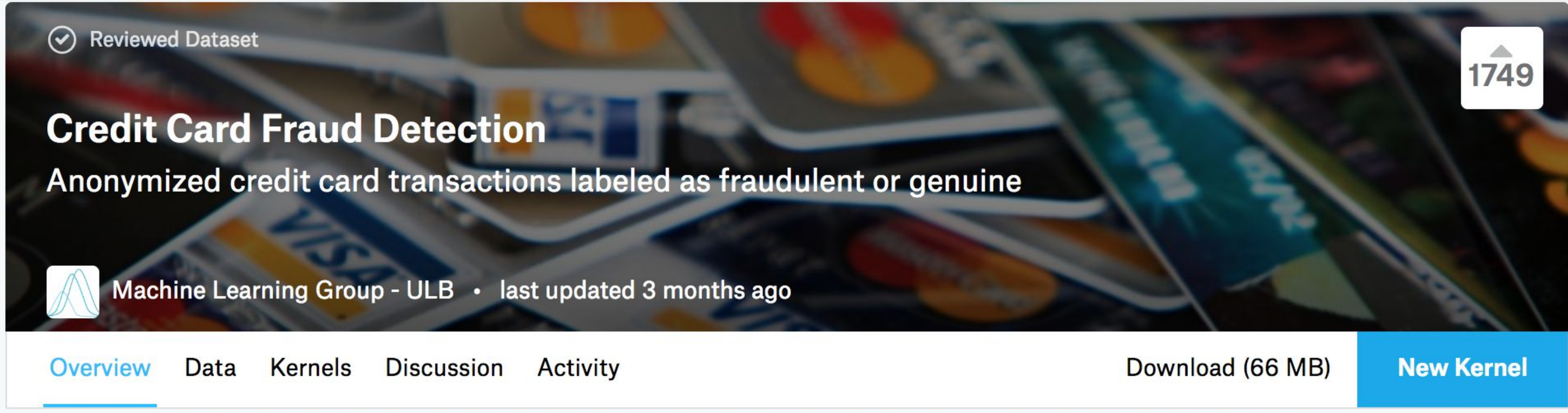
1749

Credit Card Fraud Detection

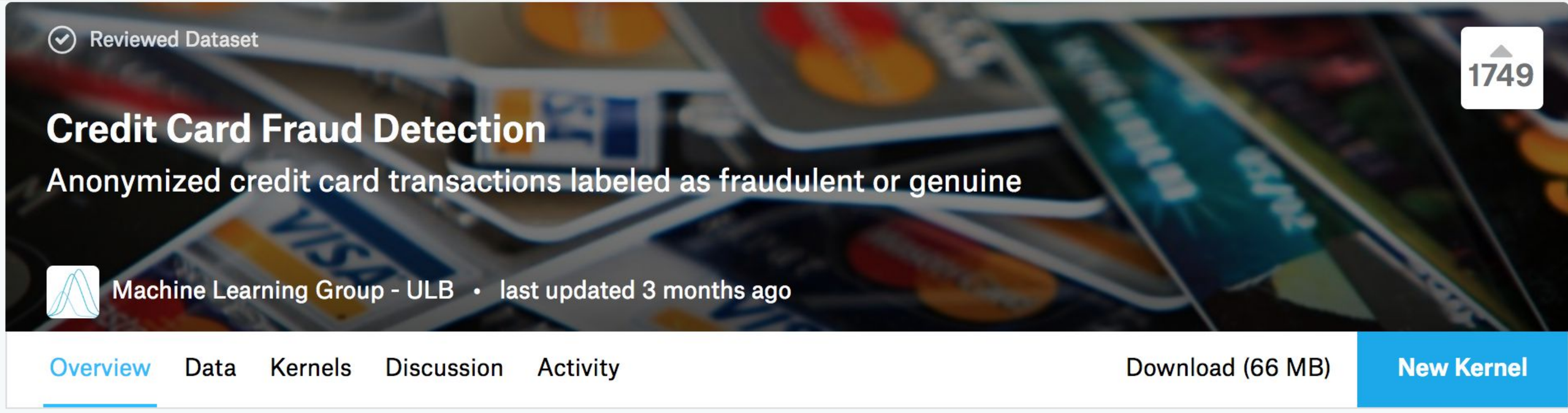
Anonymized credit card transactions labeled as fraudulent or genuine

Machine Learning Group - ULB • last updated 3 months ago

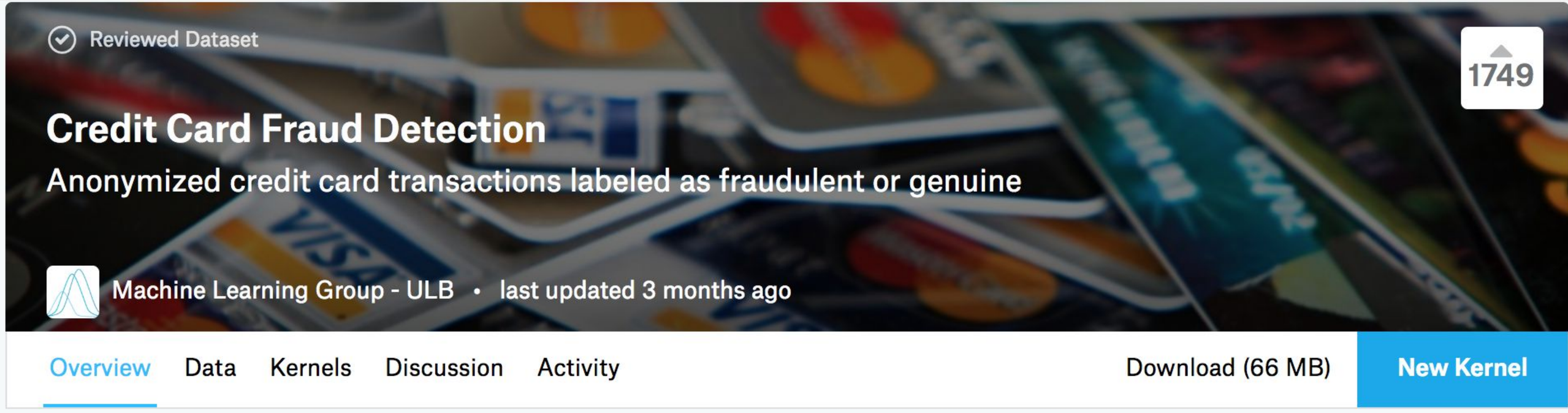
Overview Data Kernels Discussion Activity Download (66 MB) New Kernel



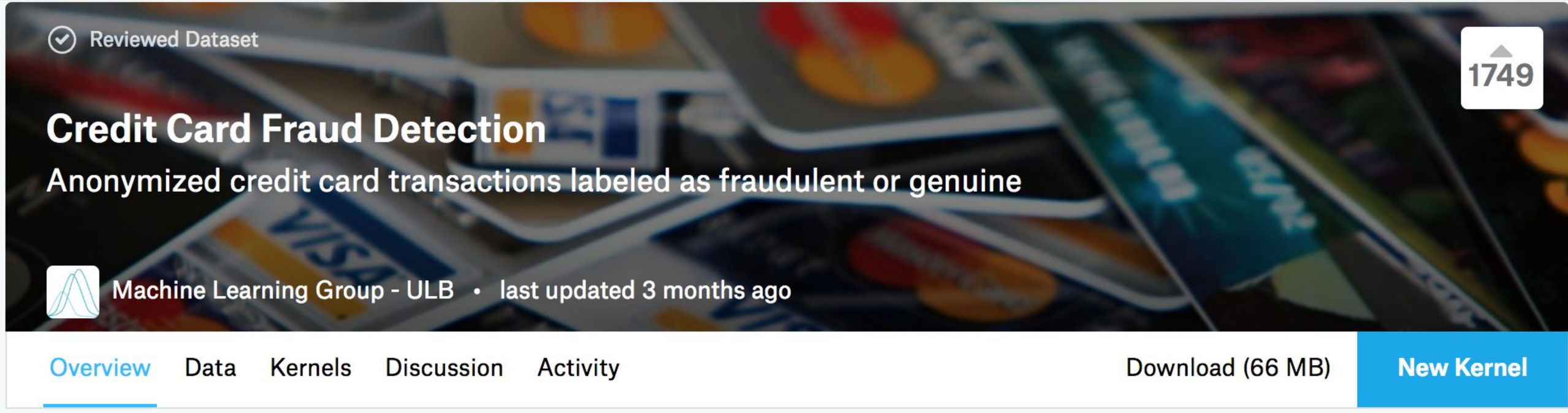
- Genuínas: 0
- Fraudulentas: 1



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



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



- As outras features são anonimizadas
- Análise exploratória básica e foco em modelagem

Familiarização com o dataset

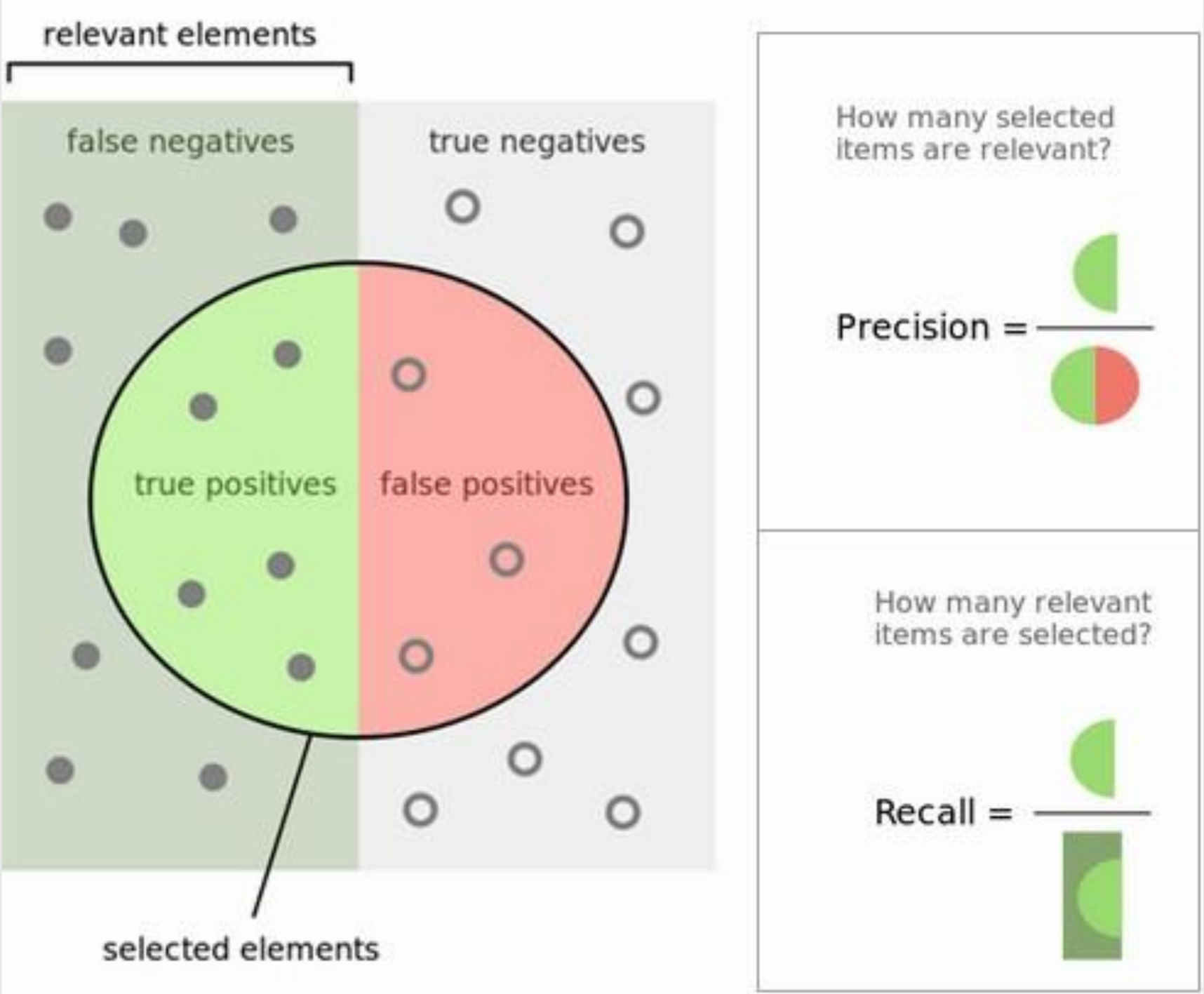
5 minutos

Discussão

Qual métrica utilizar?

Accuracy Paradox

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


relevant elements
false negatives
true negatives
true positives
false positives
selected elements
How many selected items are relevant?
Precision =
How many relevant items are selected?
Recall =

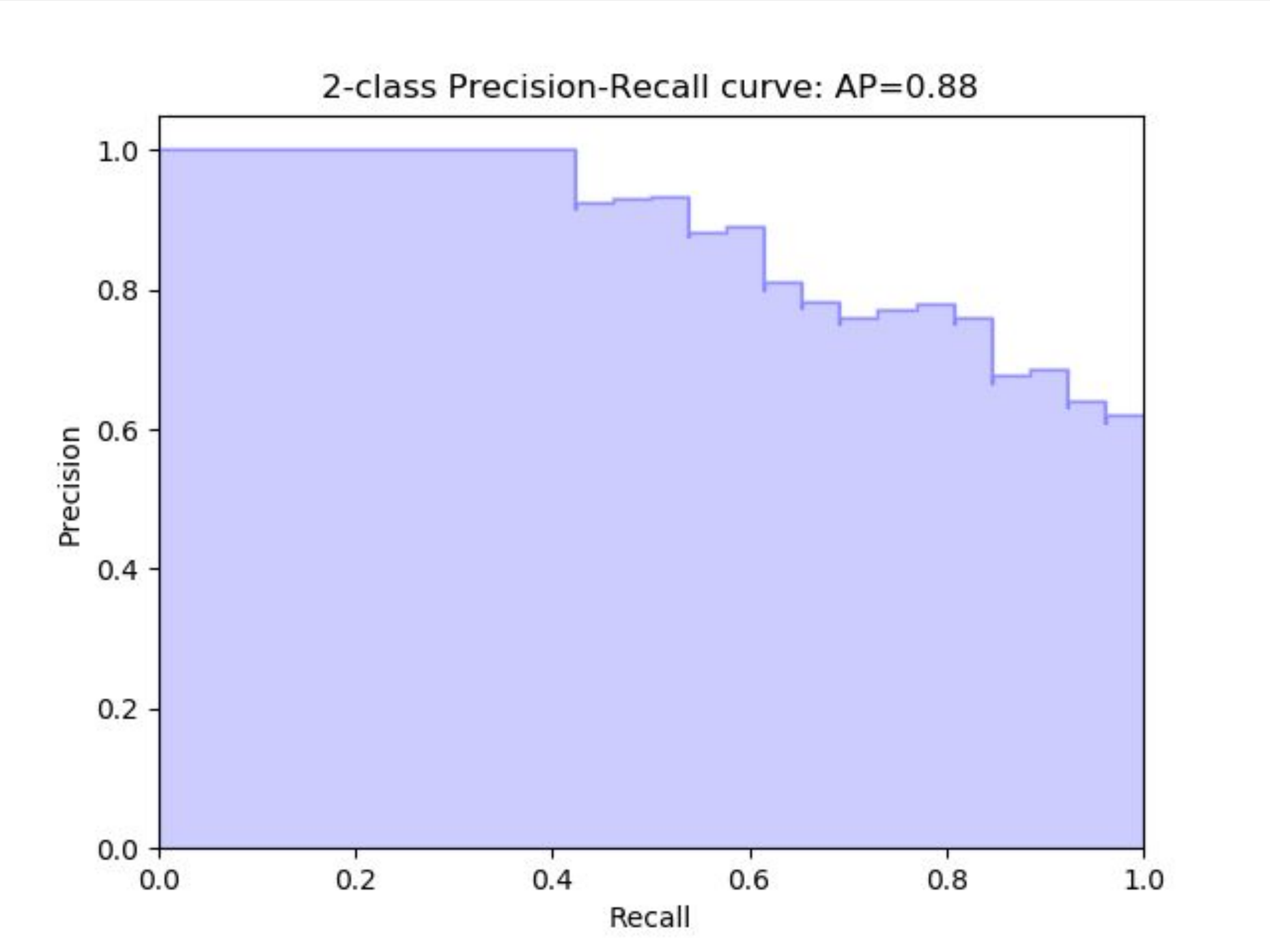
$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Manipulação dos dados

10 minutos

Treinar modelo e obter scores

20 minutos


2-class Precision-Recall curve: AP=0.88
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

Definir ponto de corte ideal

15 minutos

Criar predições finais utilizando o ponte de corte

8 minutos

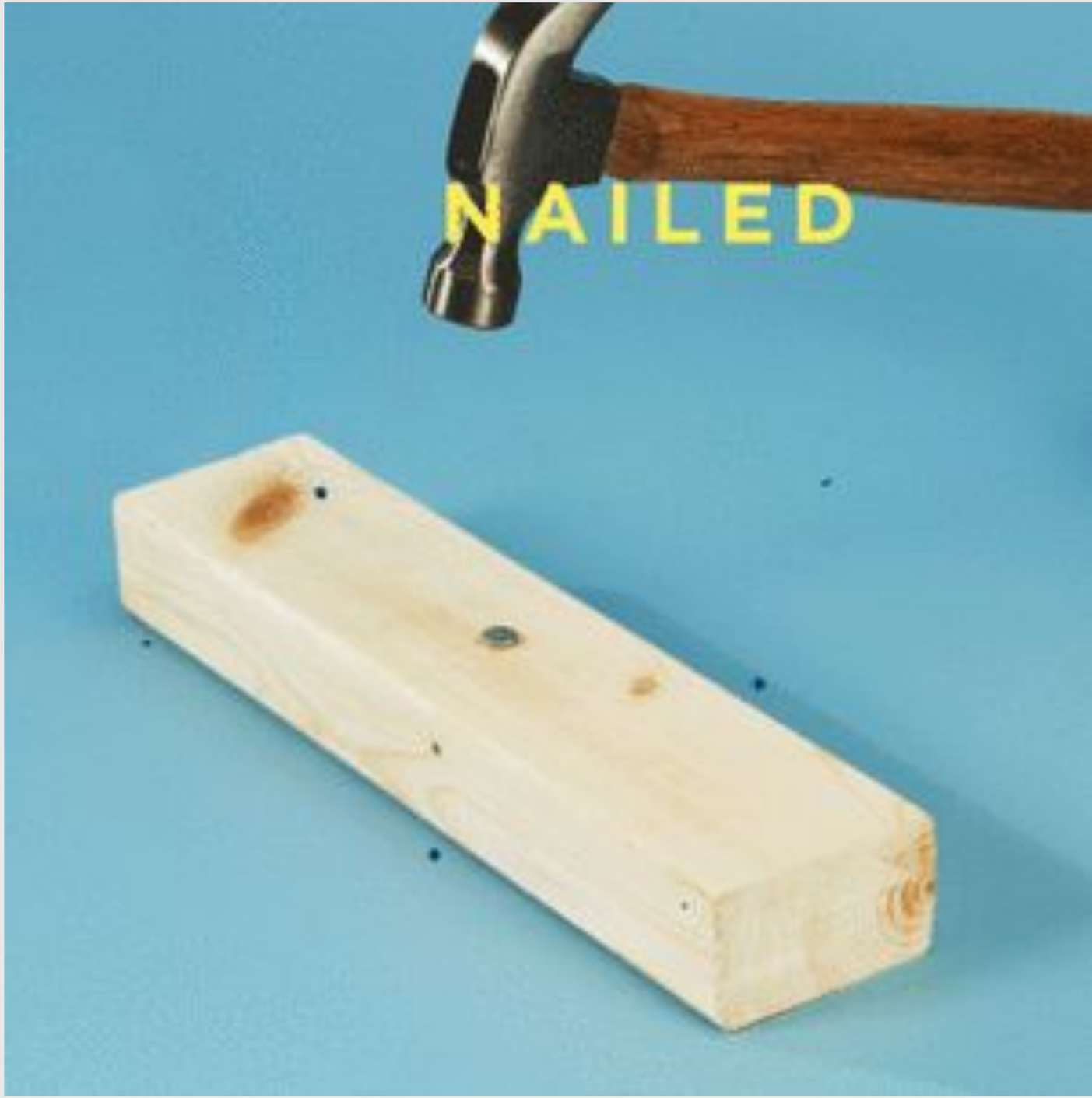
Complementar avaliação com classification report e matriz de confusão

8 minutos

Apresentações

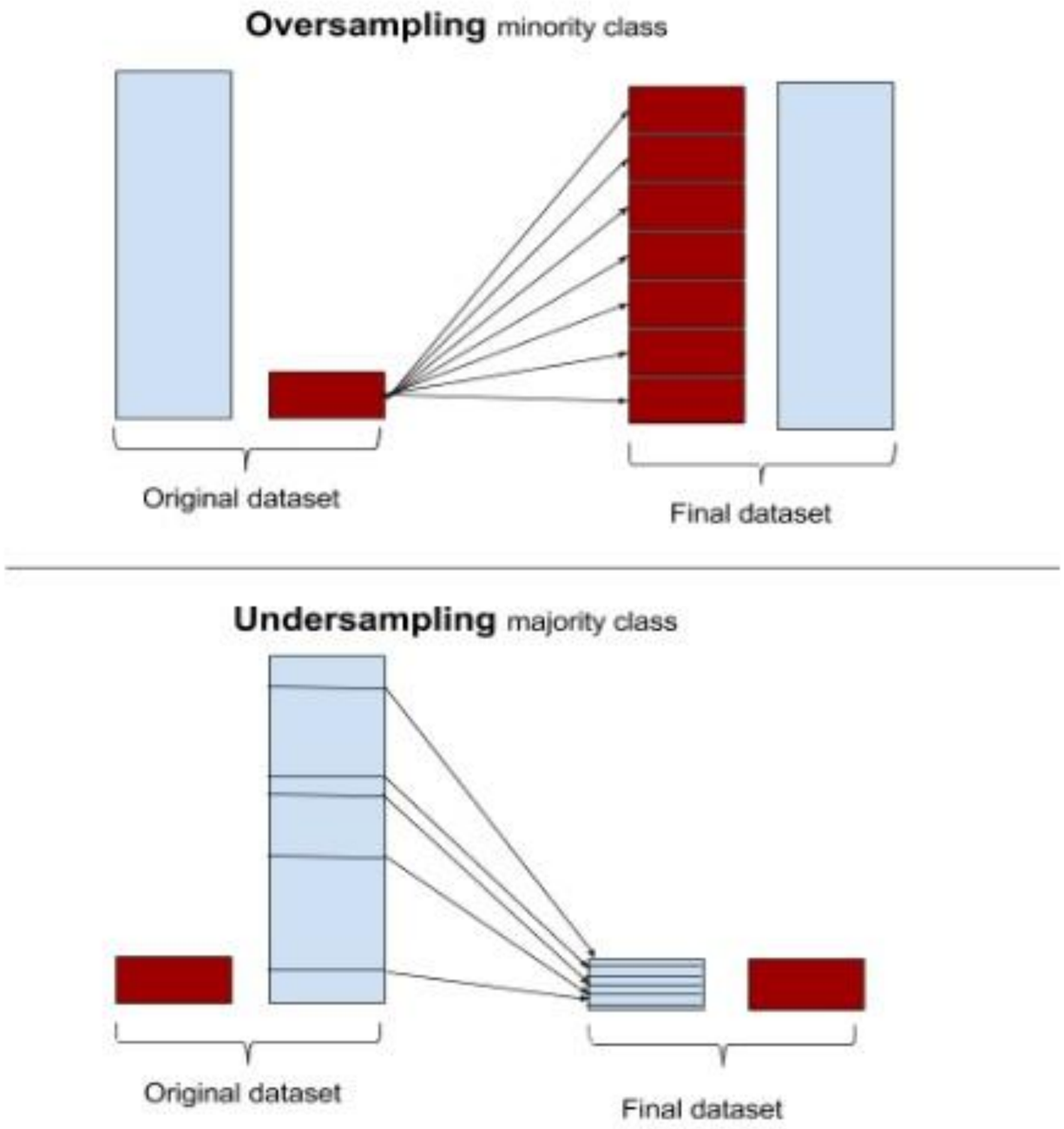
Intervalo

Congrats!

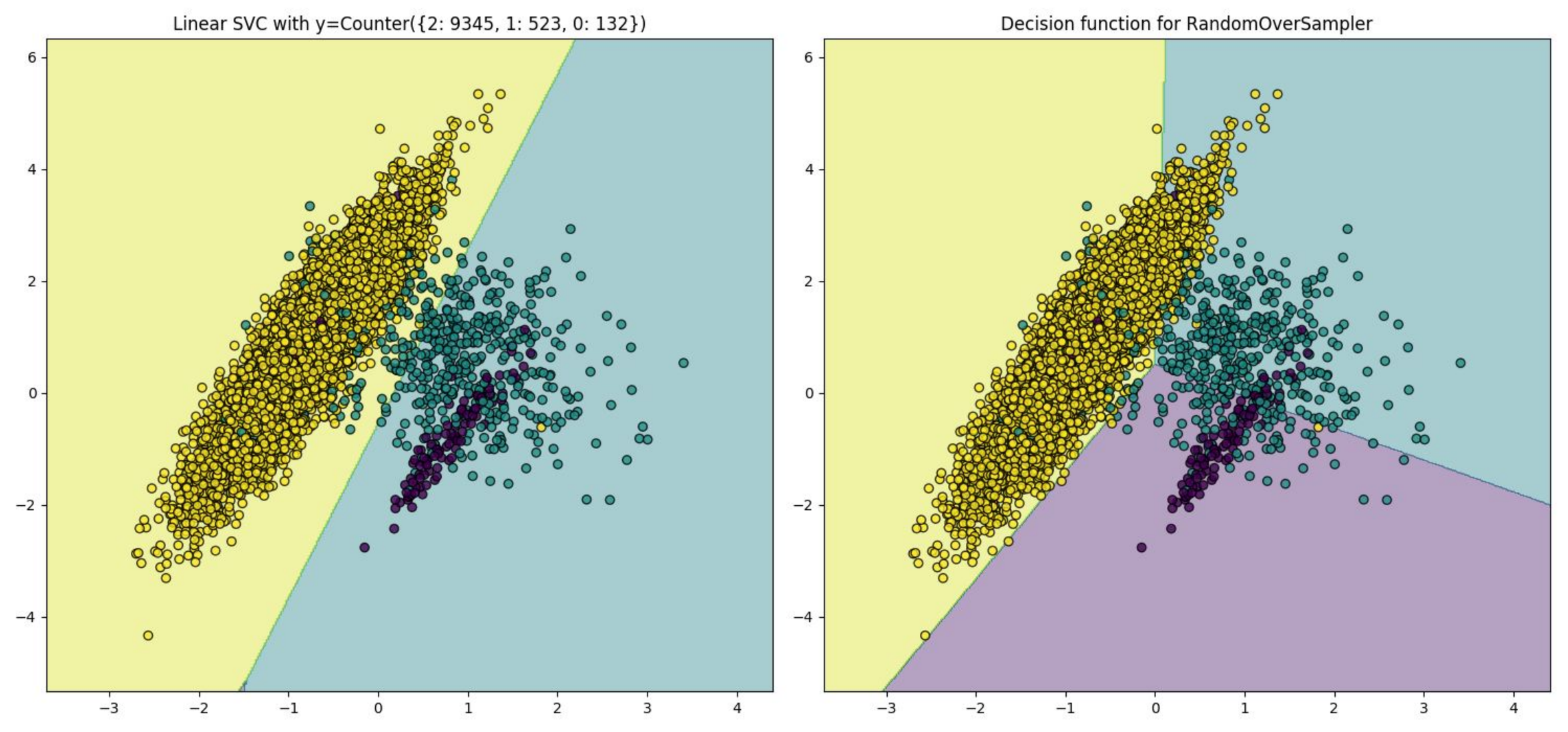


Material extra: Resampling

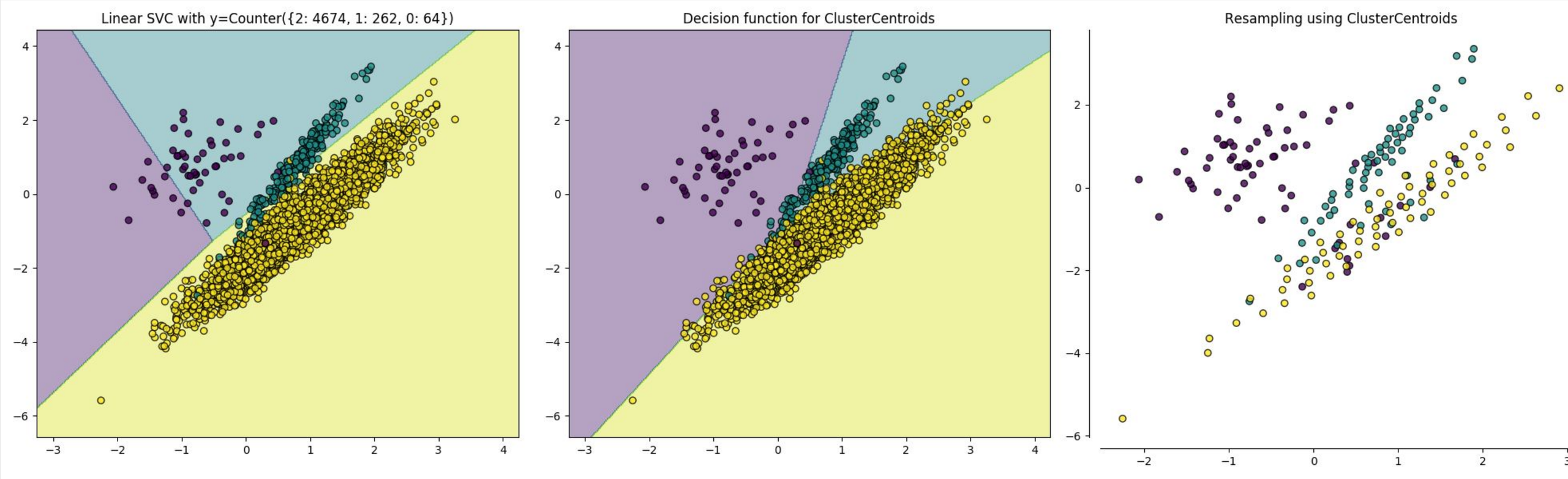
Resampling


Oversampling minority class
Original dataset
Final dataset
Undersampling majority class
Original dataset
Final dataset

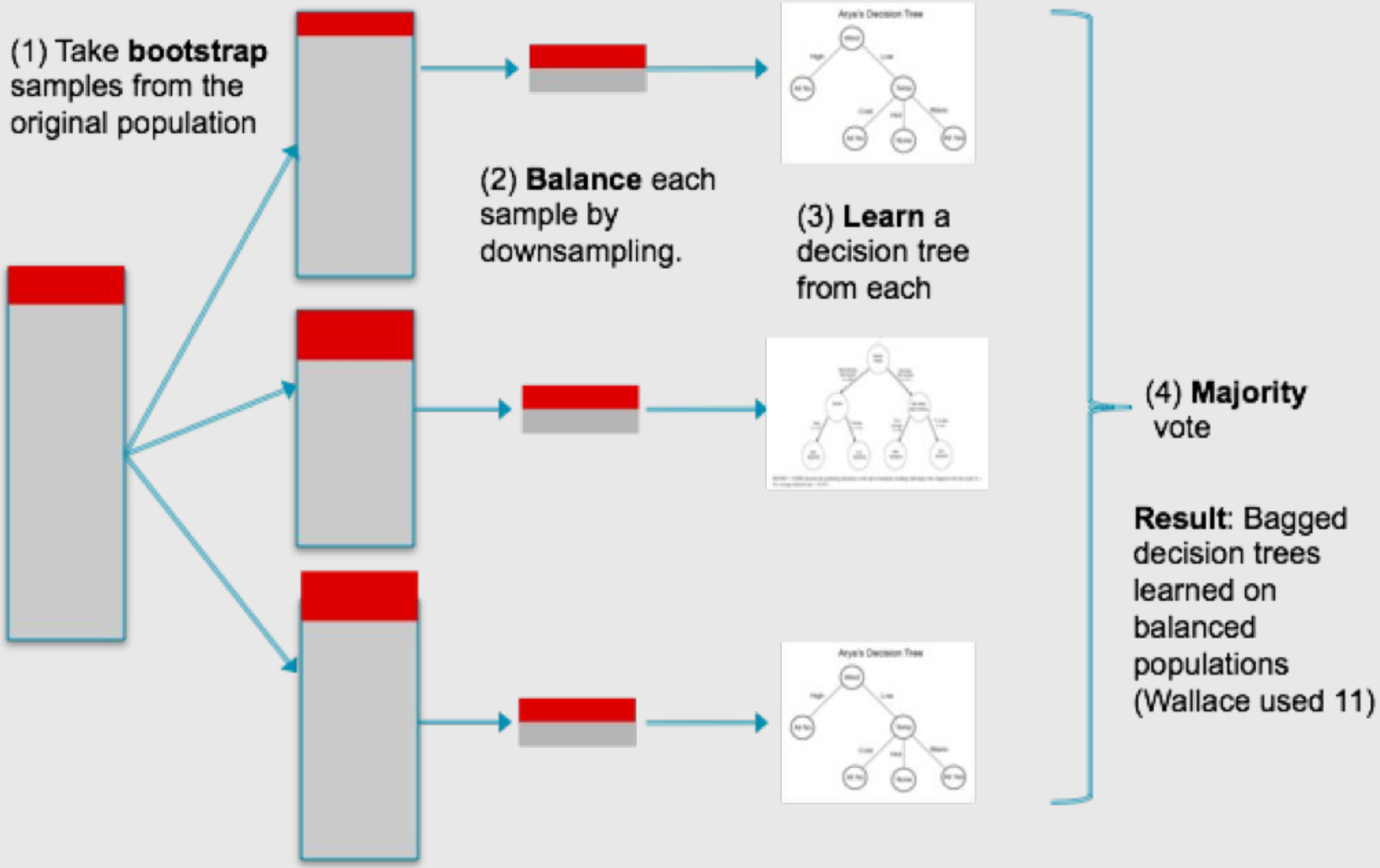
Random Oversampling



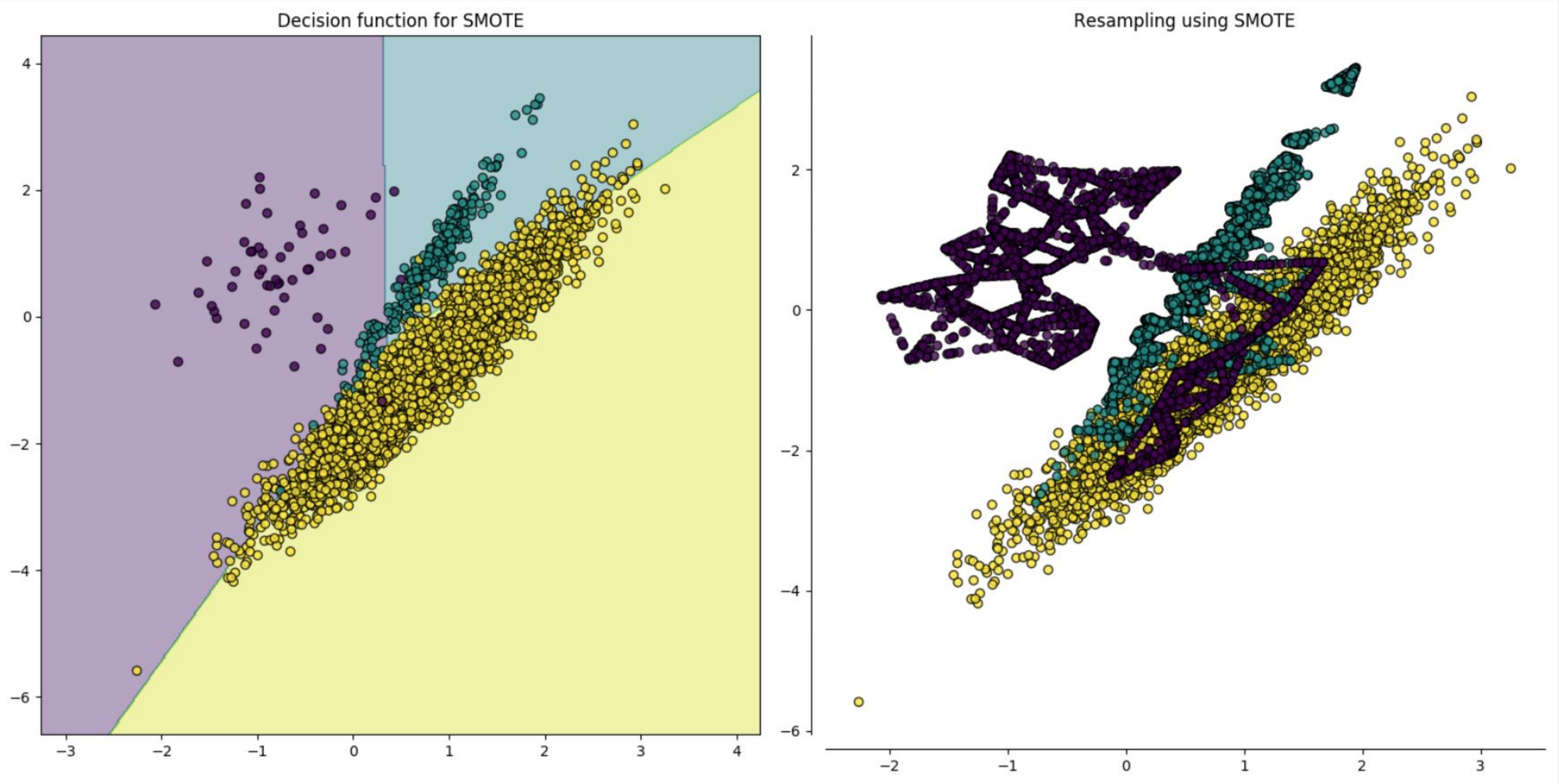
Random Undersampling



Balanced Bagging



SMOTE (Synthetic Oversampling)



Pacote imblearn

- RandomOverSampler
- RandomUnderSampler
- SMOTE
- EasyEnsemle & BalancedBaggingClassifier

Algoritmos com suporte a balanceamento

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

DÚVIDAS?