U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Random Forests on Small Imbalanced Datasets

Introduction to Machine Learning and Knowledge Extraction - MDSE

Group 07 - Bruno Fernandes e Hugo Abelheira

Agenda

1. Introduction
2. Random Forests
    - 2.1. Base Learners
    - 2.2. Motivation
    - 2.3. Explanation
        - 2.3.1. Bootstrap
        - 2.3.2. Random Feature Selection
    - 2.4. Issues
        - 2.4.1. Imbalanced Data
        - 2.4.2. Dataset Size
3. Performance Metrics
    - 3.1. Accuracy
    - 3.2. Recall
4. SMOTE
5. Result Analysis
    - 5.1. Comparing Model Performance
    - 5.2. Performance Metrics
6. Conclusions

1. Introduction

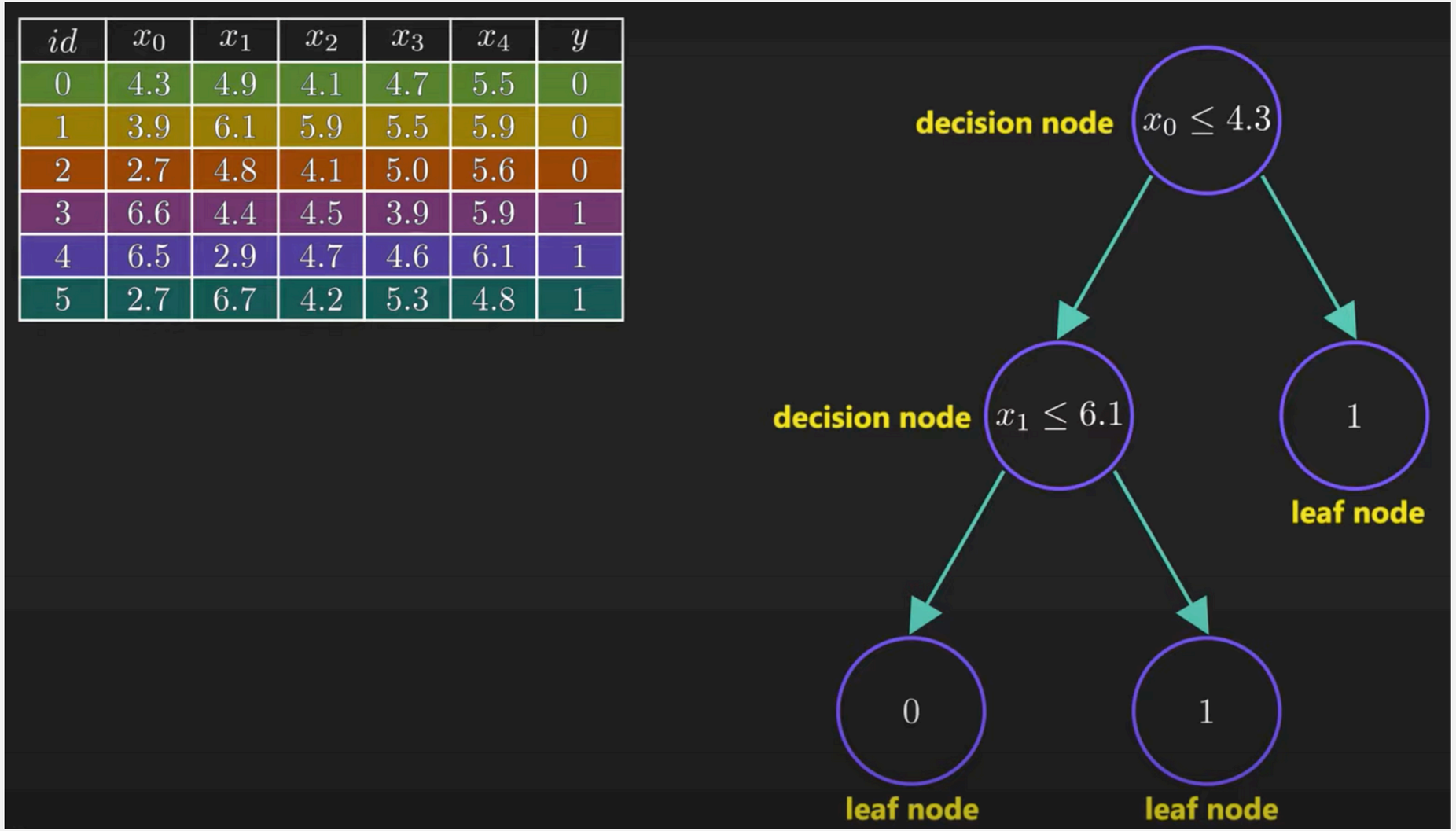
- **The goal of this presentation is to share our improvements to the Random Forest algorithm;**

- **We'll explain how the algorithm works, explore the challenges of small and imbalanced datasets, and compare our results step by step;**

- **Our aim is to provide clear and valuable insights into this topic.**

2.1. Random Forests - Base Learners

id	x_0	x_1	x_2	x_3	x_4	y
0	4.3	4.9	4.1	4.7	5.5	0
1	3.9	6.1	5.9	5.5	5.9	0
2	2.7	4.8	4.1	5.0	5.6	0
3	6.6	4.4	4.5	3.9	5.9	1
4	6.5	2.9	4.7	4.6	6.1	1
5	2.7	6.7	4.2	5.3	4.8	1



“There is wisdom in crowds”

Image Source: https://www.youtube.com/watch?v=v6VJ2RO66Ag&ab_channel=NormalizedNerd

2.2. Random Forests - Motivation

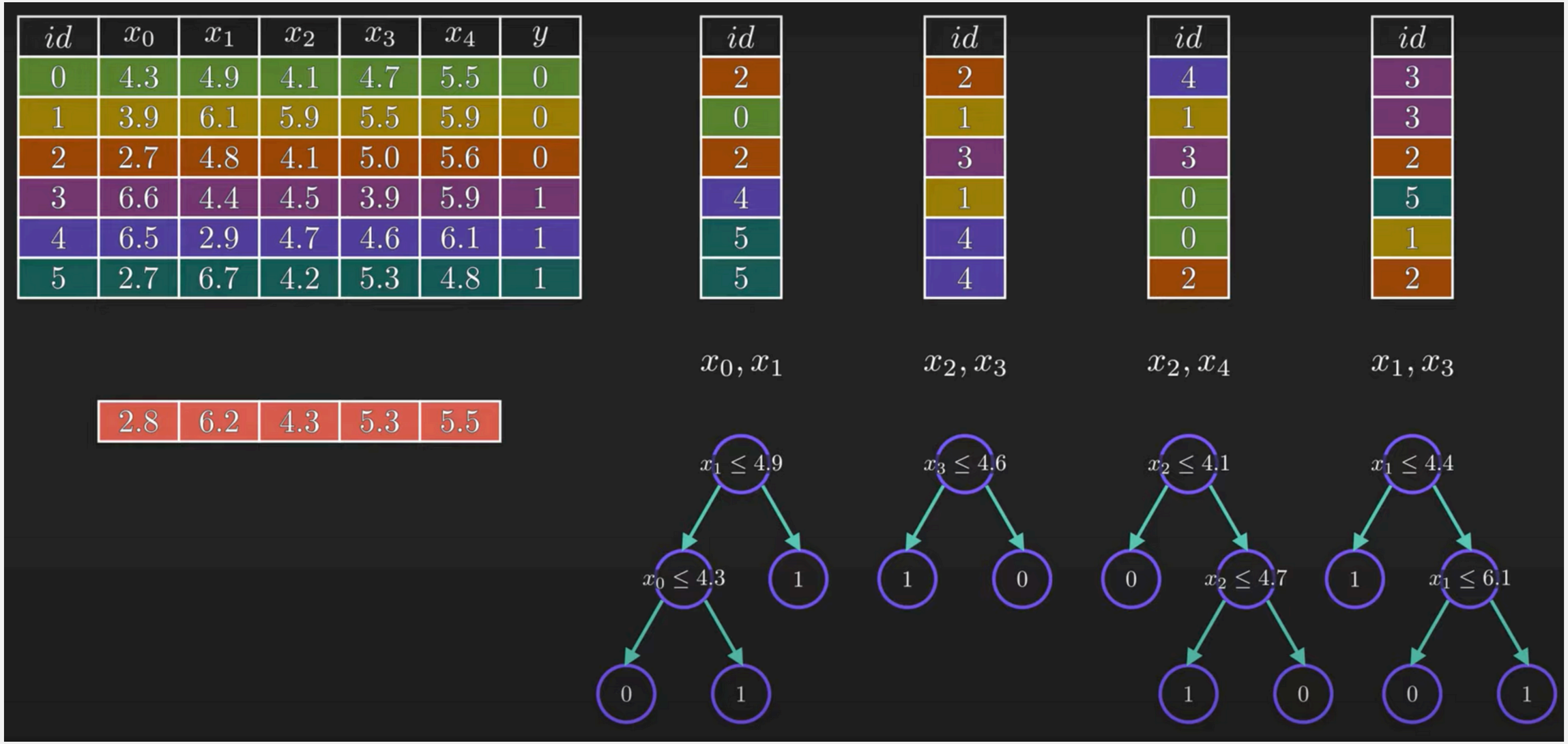


Image Source: https://www.youtube.com/watch?v=v6VJ2RO66Ag&ab_channel=NormalizedNerd

2.3. Random Forests - Explanation

2.3.1. Bootstrap

Reduces bias and assures nor big or important chunks of data are left behind/ignored

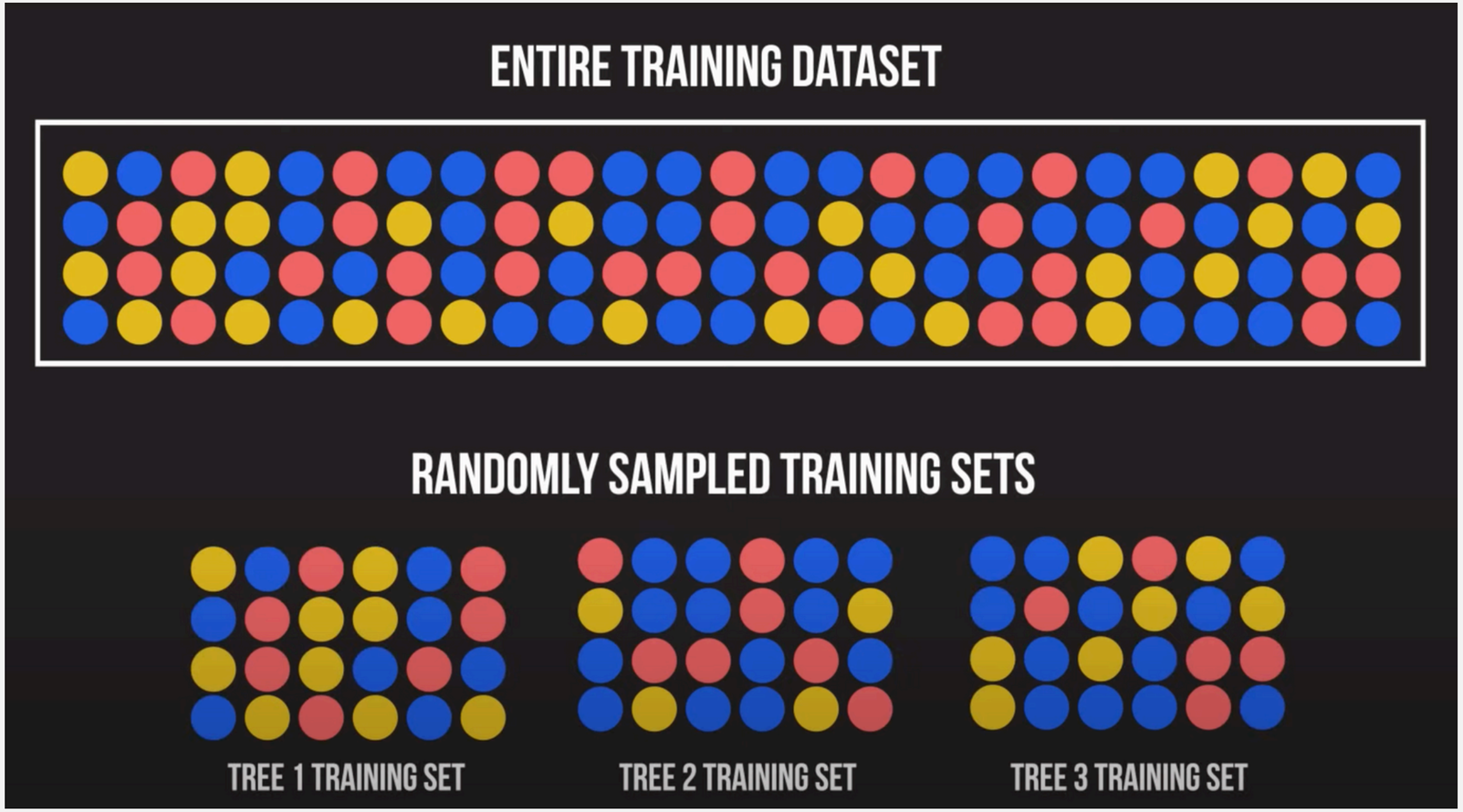


Image Source: https://www.youtube.com/watch?v=cIbj0WuK41w&ab_channel=Econoscent

2.3. Random Forests - Explanation

2.3.2. Random Feature Selection

Reduces overfitting by making sure our trees disagree, producing different predictions;

It is okay to have bad predictions as they will balance out in the final result

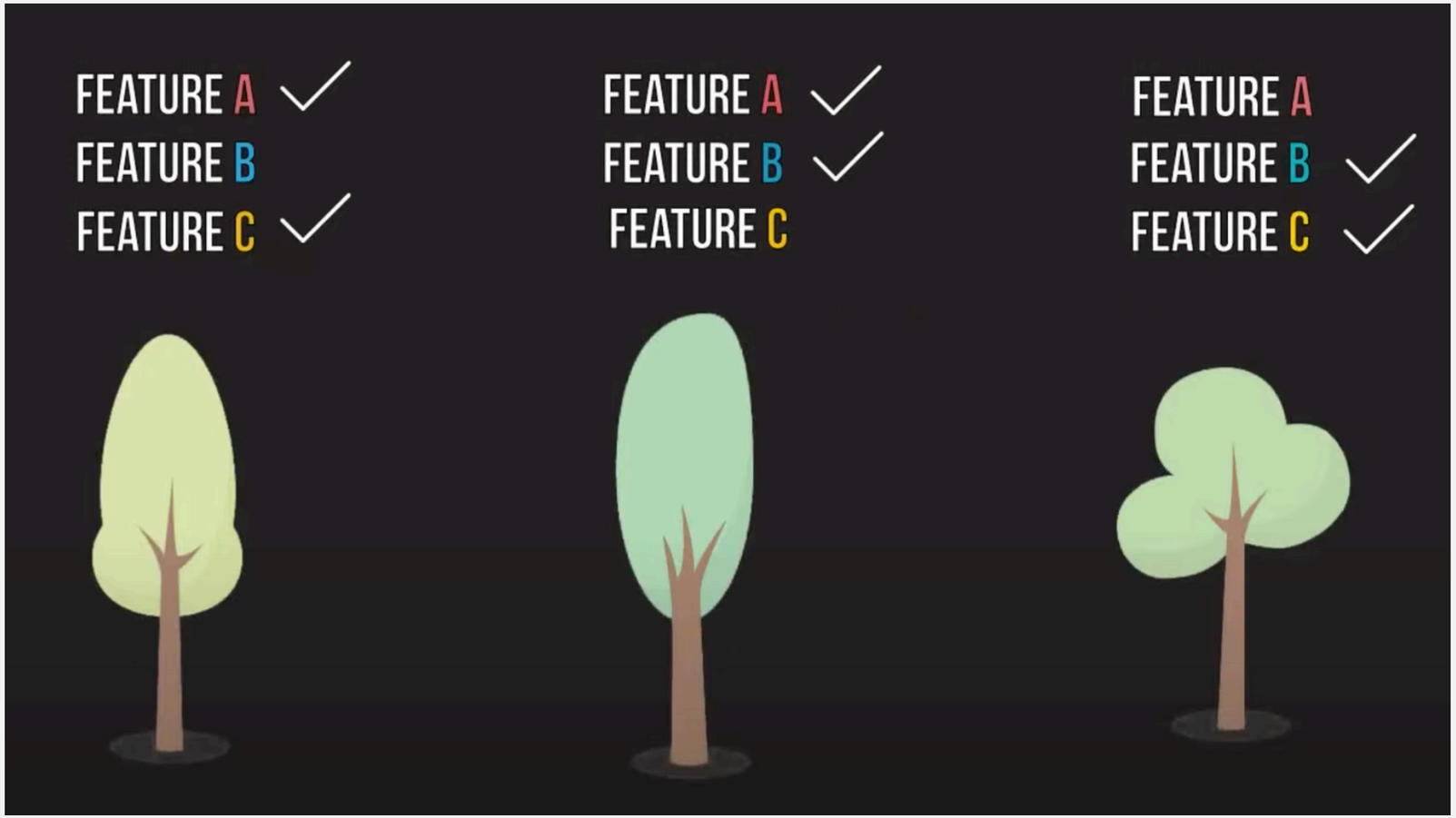
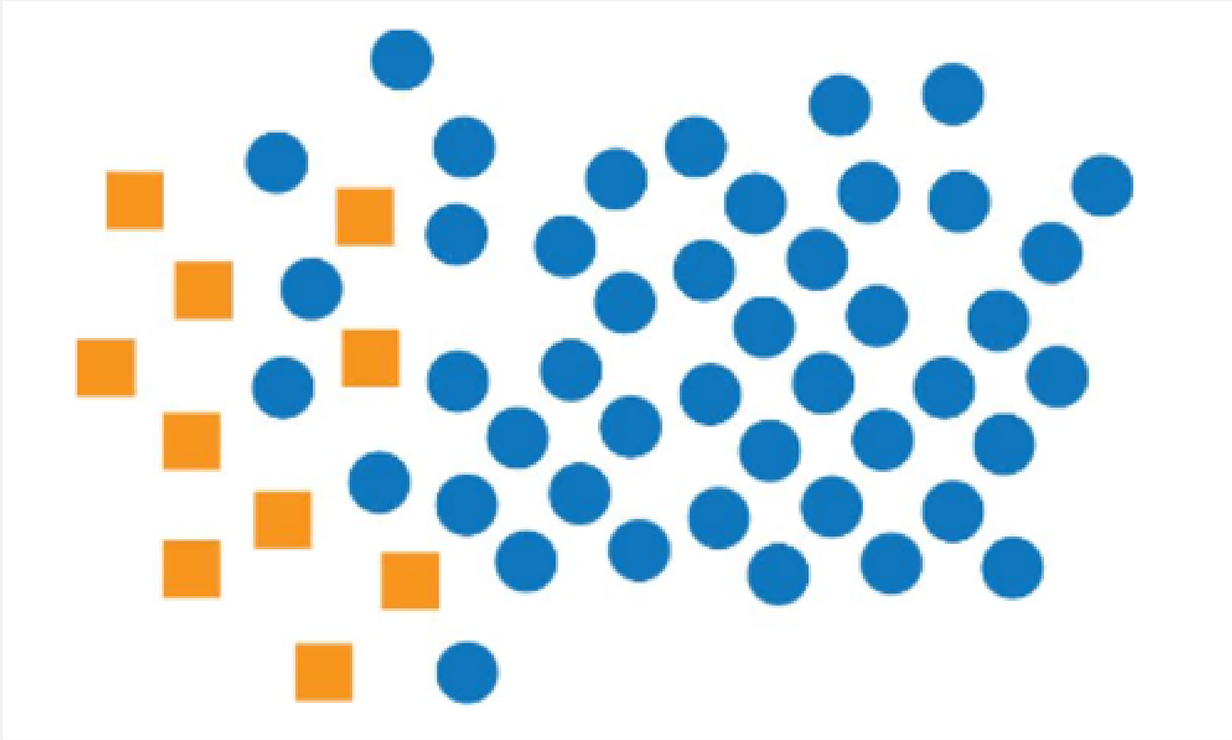


Image Source: https://www.youtube.com/watch?v=cIbj0WuK41w&ab_channel=Econoscent

2.4. Random Forests - Issues

2.4.1. Imbalanced Datasets

Imbalanced data can make the Random Forest prune to error, since our trees can train on samples where the minority class is practically negligible



2.4. Random Forests - Issues

2.4.2. Dataset Size

How small is the minority data?

Is it enough information to run predictions?

3. Performance Metrics

3.1. Accuracy

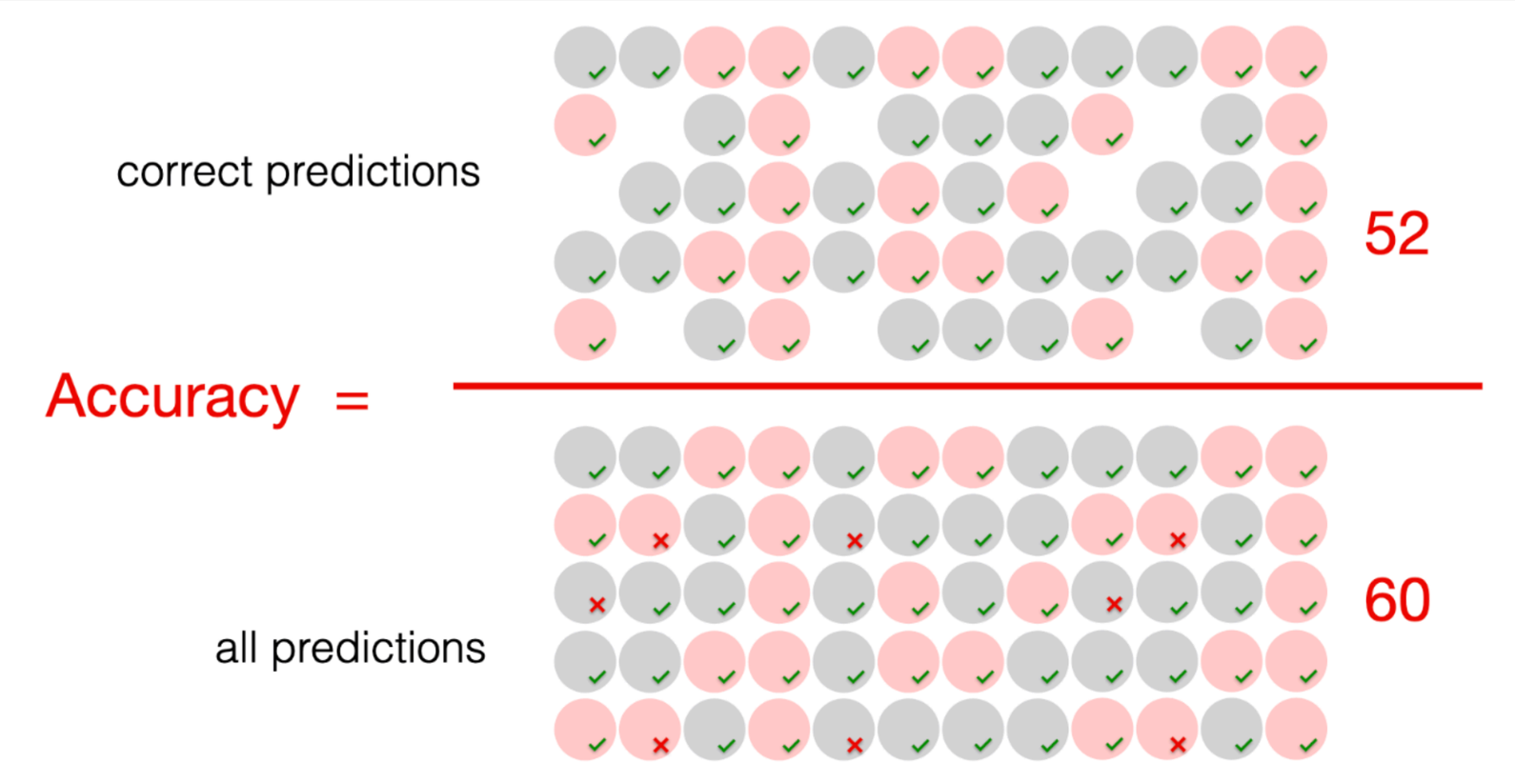


Image Source: https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall

3. Performance Metrics

3.2. Recall

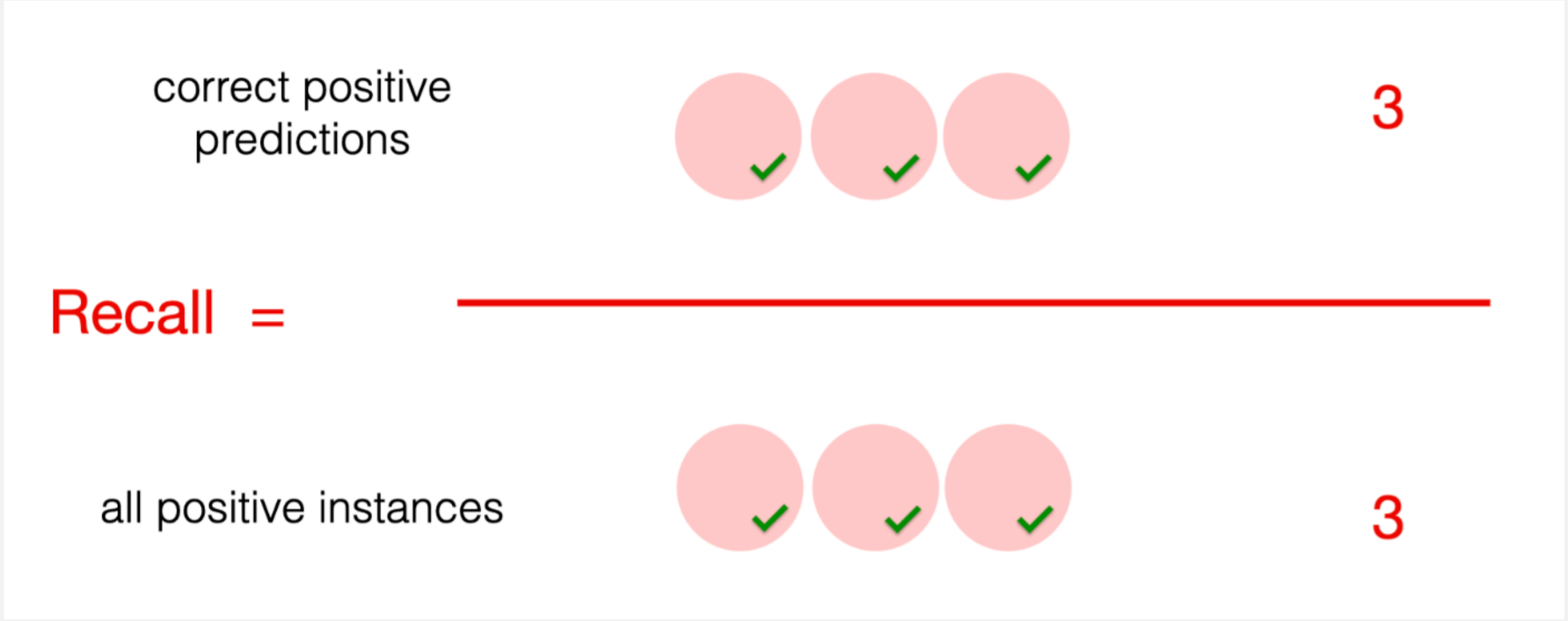


Image Source: https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall

4. SMOTE - Synthetic Minority Oversampling Technique

Oversampling Technique;

Creates synthetic instances;

The point is not obtaining more information about the minority class, but to assure the predictions are not dominated by the majority class;

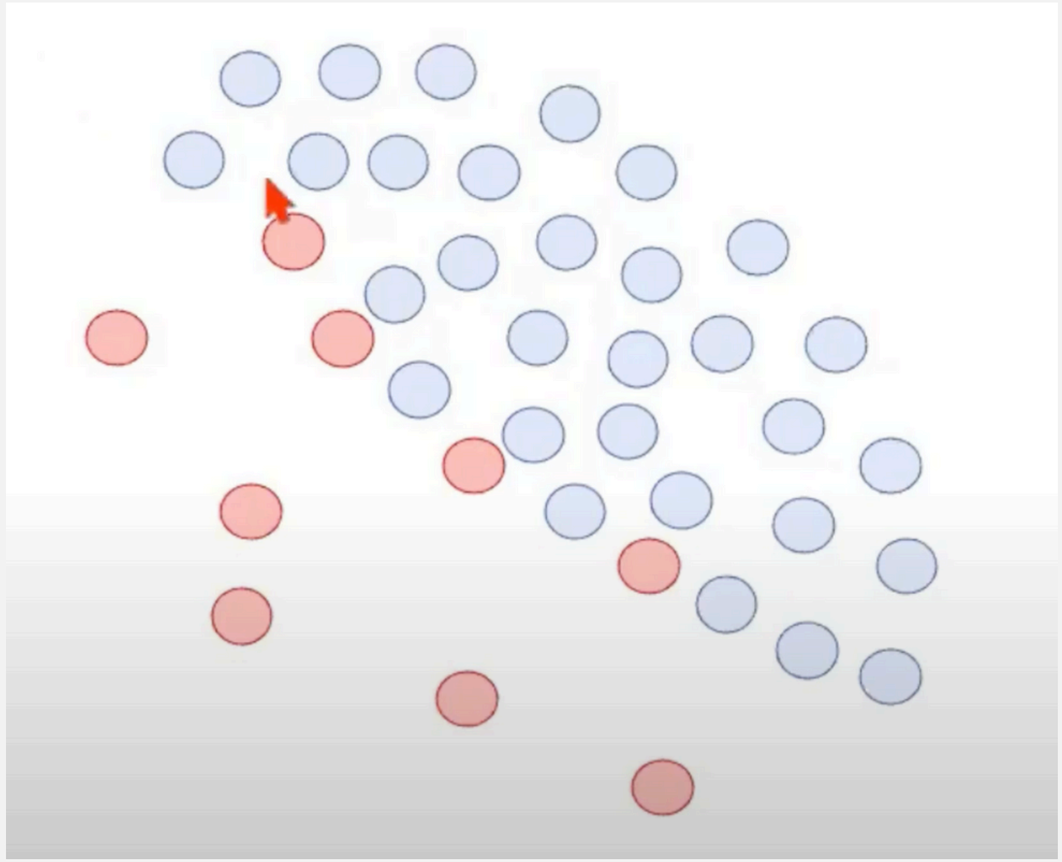


Image Source: https://www.youtube.com/watch?v=NvrfSntsm7w&ab_channel=TravisTang

4. SMOTE

Calculates the distances to k-nearest neighbours (pre-defined parameter)

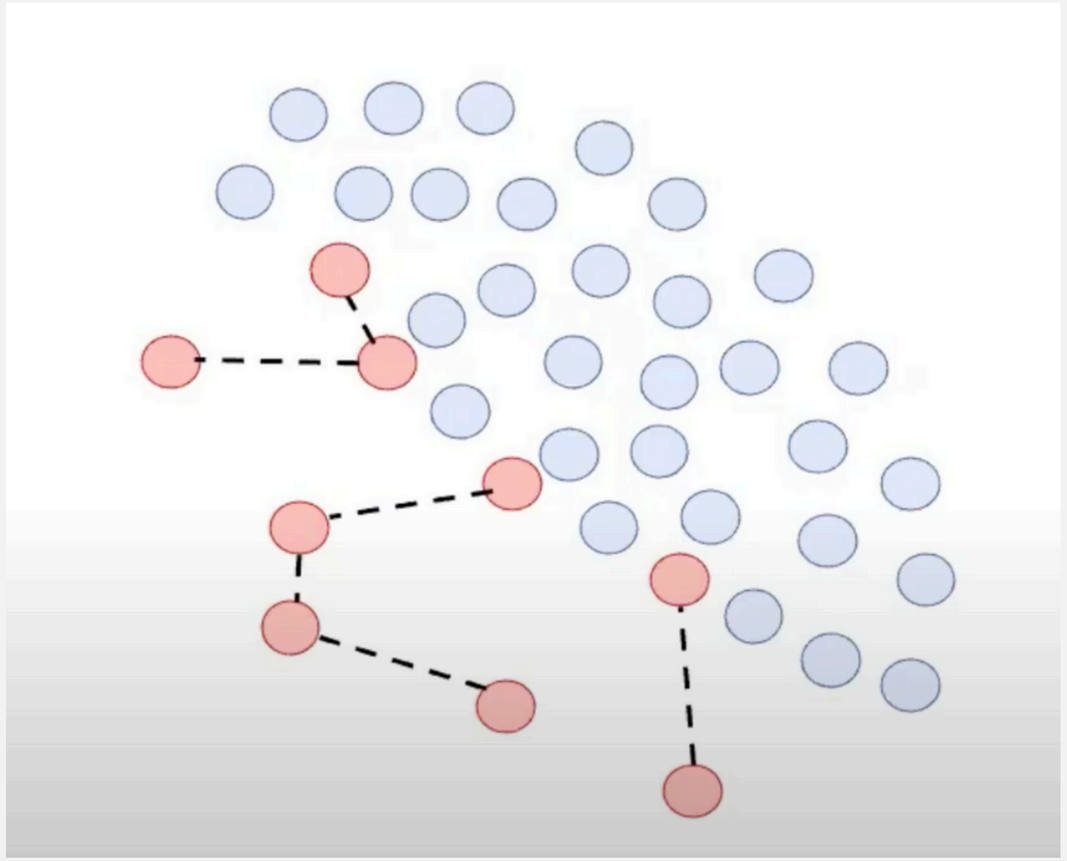


Image Source: https://www.youtube.com/watch?v=NvrfSntsm7w&ab_channel=TravisTang

4. SMOTE

Creates new data points (as many as we pre-define) along the links calculated previously

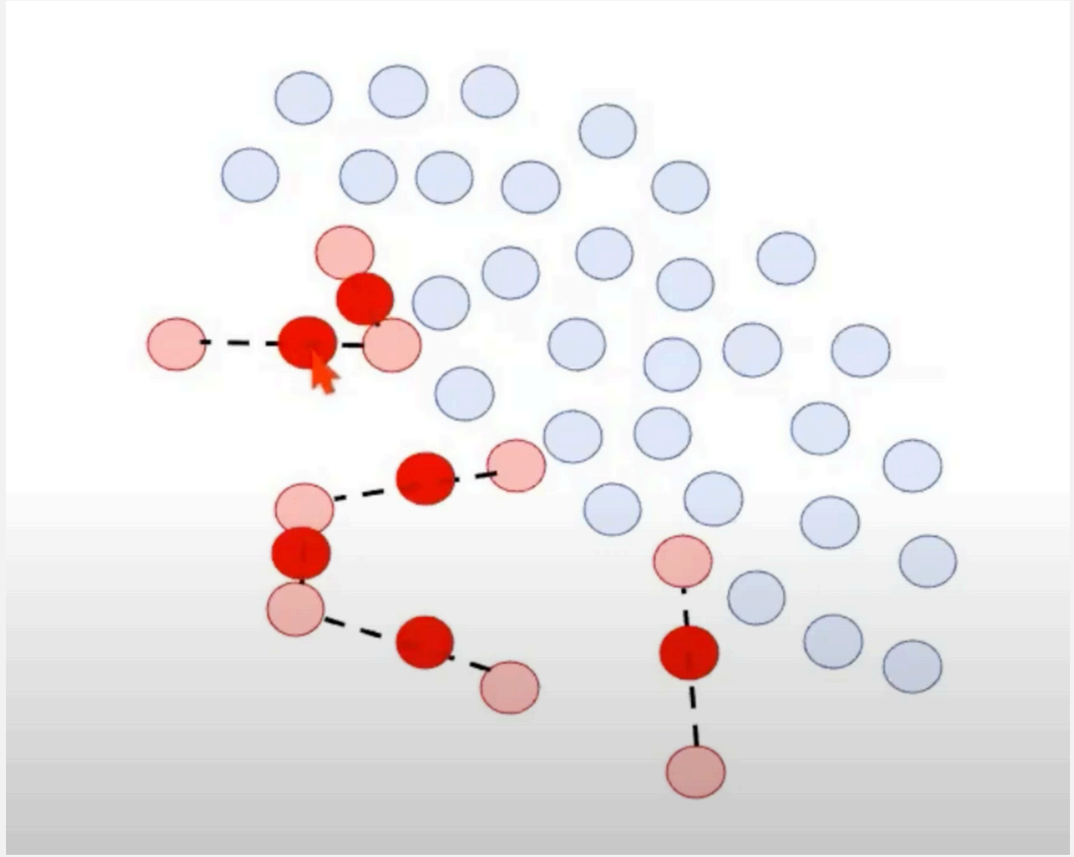
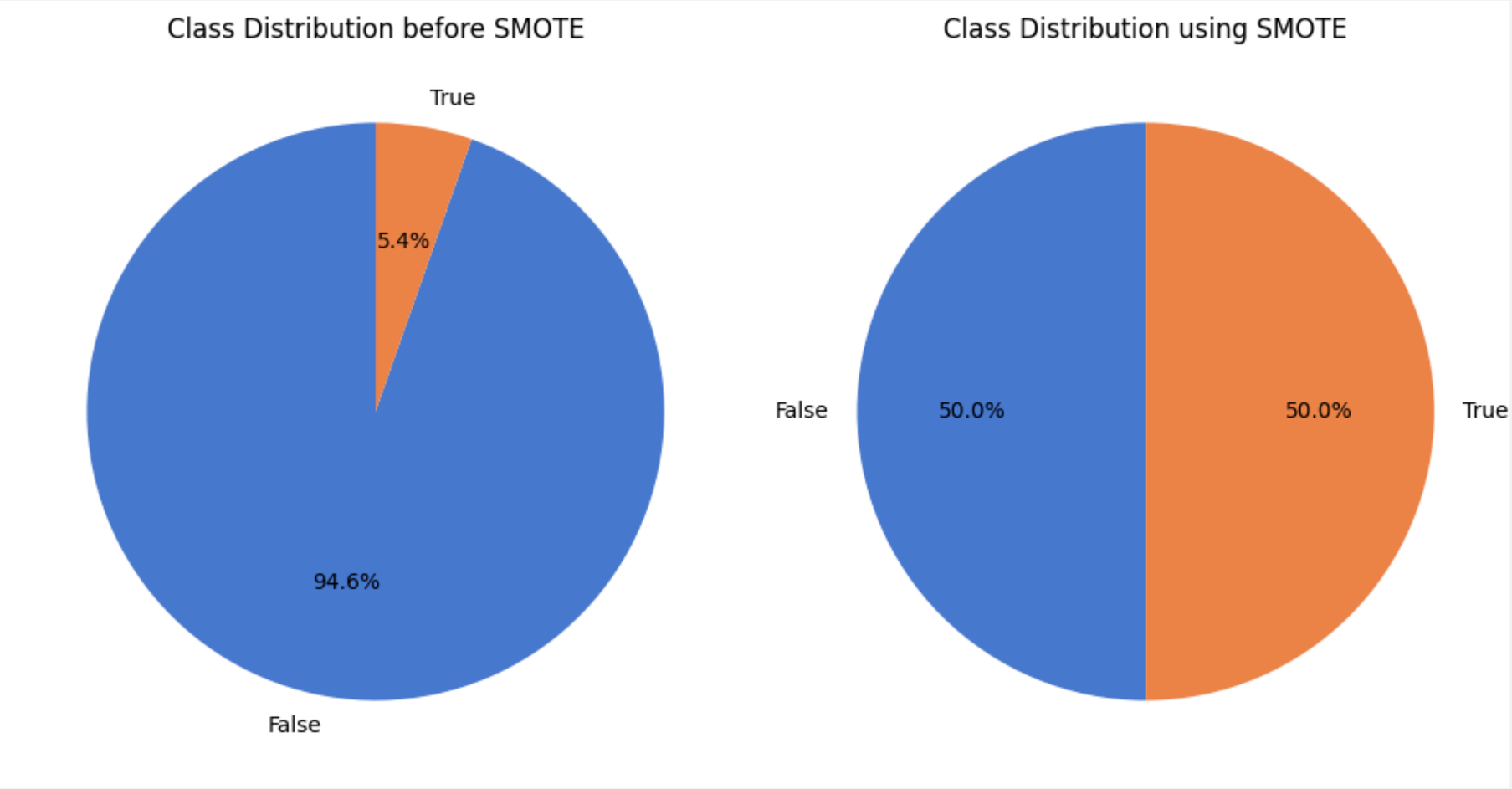


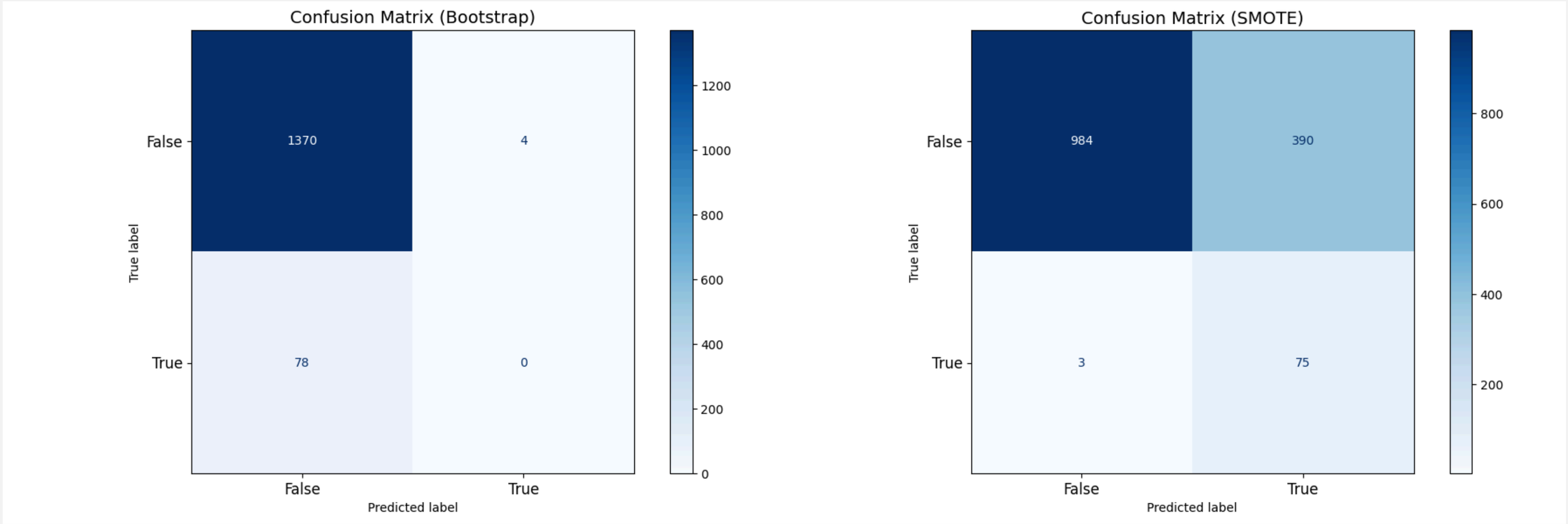
Image Source: https://www.youtube.com/watch?v=NvrfSntsm7w&ab_channel=TravisTang

4. SMOTE

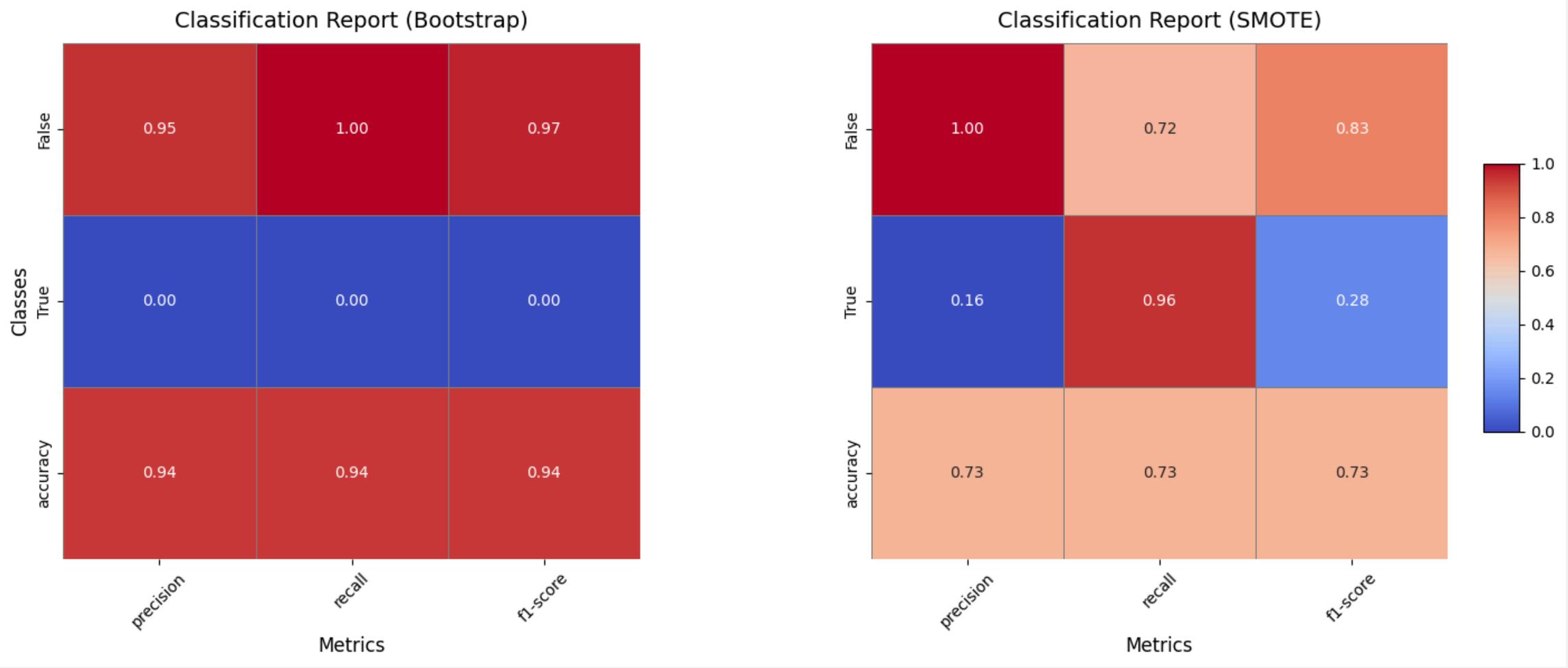
Using this technique allowed us to have balanced proportions between the minority and the majority class



5. Result Analysis - Comparing Model Performance



5. Result Analysis - Performance Metrics



6. Conclusions

- **Depending on the problem at hand, we should focus on maximizing the most relevant metrics for that specific context;**

- **SMOTE, although it "creates" synthetic data, is a reliable oversampling method, even if it is not the most efficient in terms of resource usage;**

- **Random Forest models trained on imbalanced datasets struggle to generalize effectively, as demonstrated by the lower ROC-AUC scores.**

U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Random Forests on Small Imbalanced Datasets

Introduction to Machine Learning and Knowledge Extraction - MDSE

Group 07 - Bruno Fernandes e Hugo Abelheira