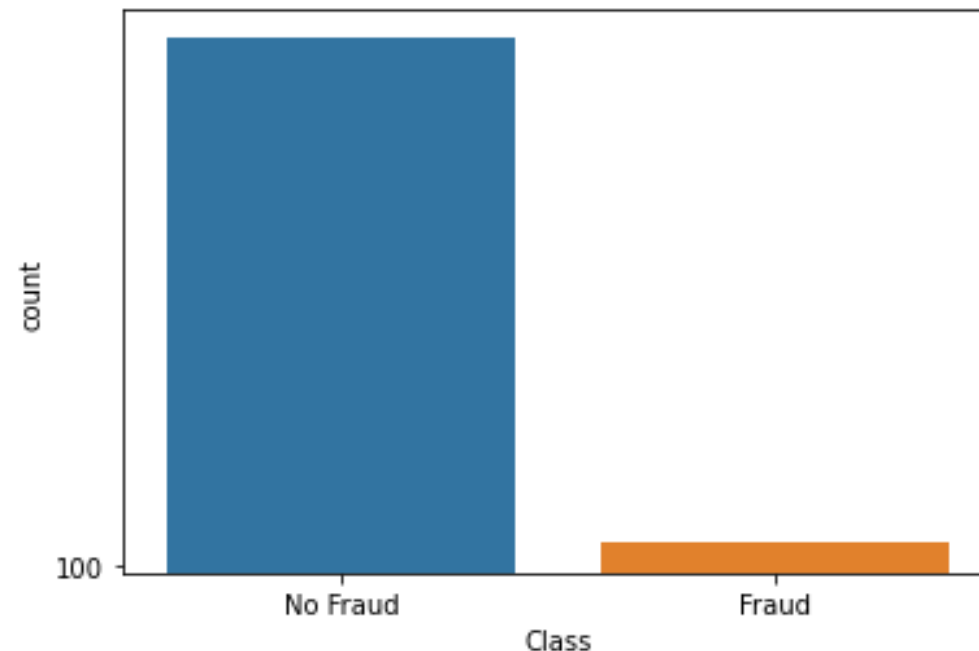# Machine Learning

Session 22 - T

## Data Imbalance in Machine Learning

**Ciência de Dados Aplicada**

**2023/2024**

# Data Imbalance

- **Common issue** in machine learning where class distribution in a dataset is highly skewed;

- **One class significantly outnumbers the others**;

- Real-world scenarios:
  - Fraud detection;
  - Medical diagnosis;
  - Text classification;
  - Imafe recognition;

# Data Imbalance

- Consequences of data imbalance:
  - Data imbalance can have a **huge impact on model performance**;
  - **Poor generalization for the minority class**.

- **Why?**

  - Machine learning models are typically designed to **optimize overall accuracy**, which means they tend to favor the majority class.

# Data Imbalance

- Model Performance on Imbalanced Datasets:
  - **High accuracy rate**, but **ineffective at minority class** identification and classification;

- Practical Implications:
  - In applications like fraud detection or medical diagnosis, may lead to:
    - **Undetected fraudulent transactions**
    - **Missed critical diagnoses**

- Adressing data imbalance:
  - **Rebalance dataset**;
  - **Adjust the model learning process**;
  - **Use specialized evaluation metrics** for imbalanced data performance.

# Approaches to Adress Data Imbalance

- Data-level methods:
  - **Oversampling**;
  - **Undersampling**;
  - Combined over and undersampling.
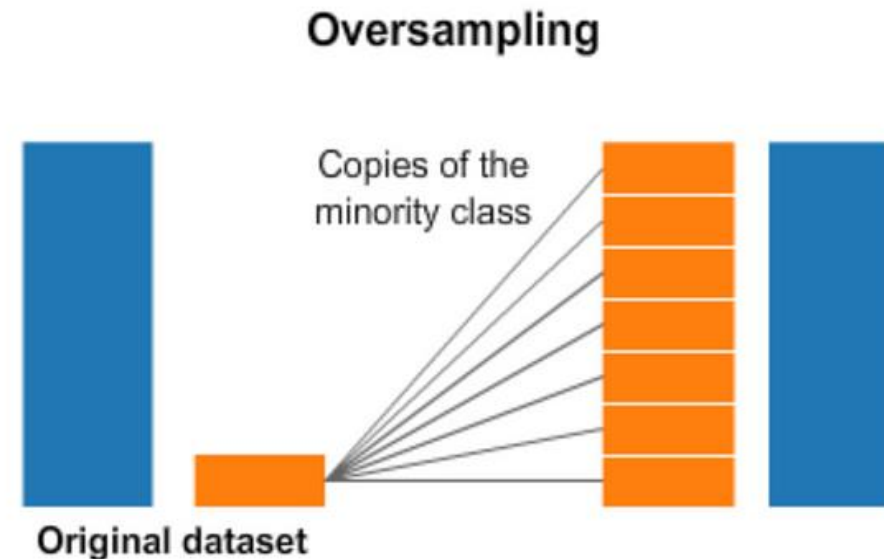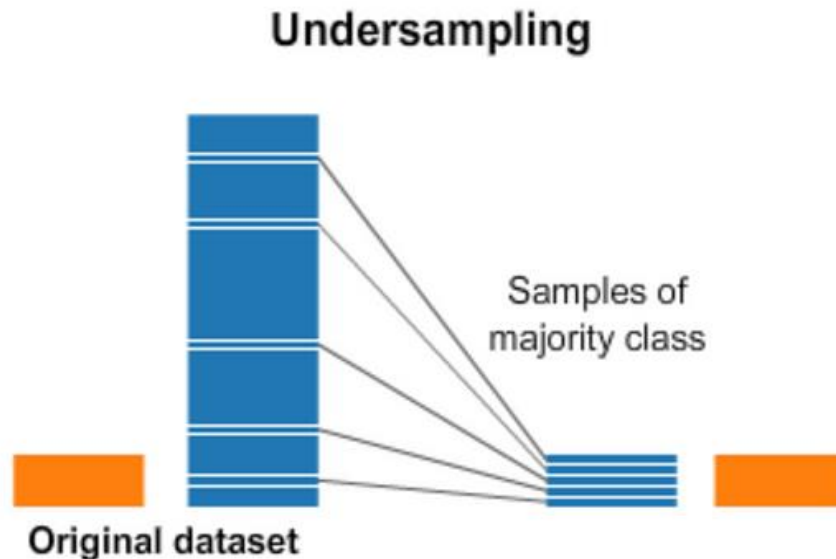
# Data-Level Methods

- **Resampling techniques** are a commonly used for adressing data imbalance;

- They modify the data by:
  - **Increasing the minority class** samples (oversampling);
  - **Decreasing the majority class** samples (undersampling).
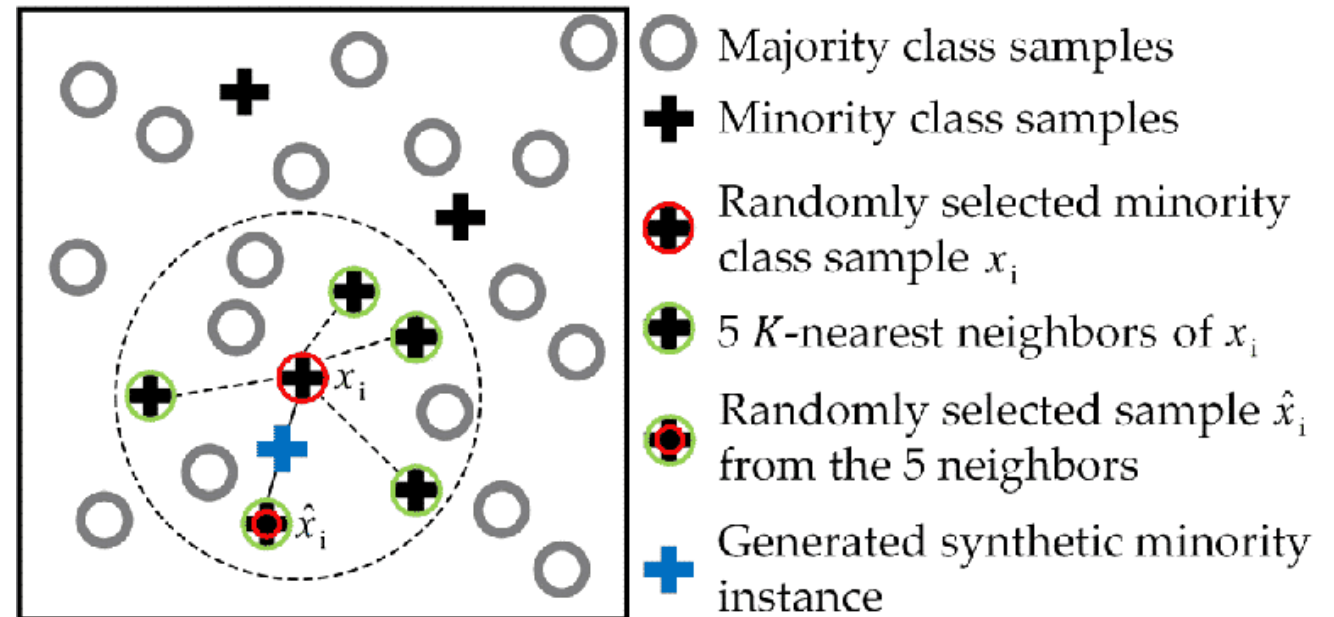
# Oversampling Techniques

- **Random oversampling**:
  - Increases the number of minority class samples by **randomly duplicating existinting minority class samples**;

  - Can **improve model performance on minority class**, but increases the risk of **overfitting** due to the repeated samples.
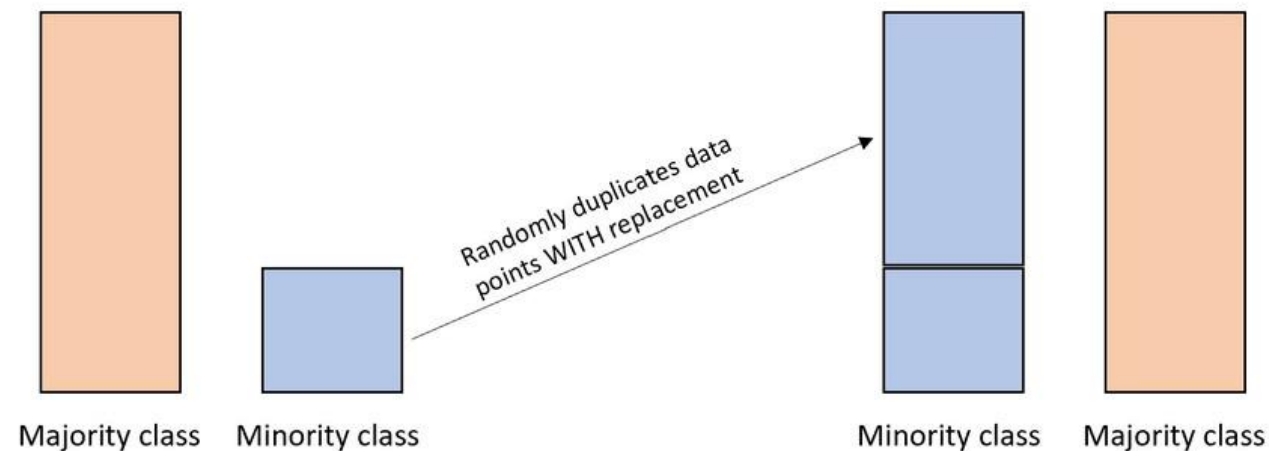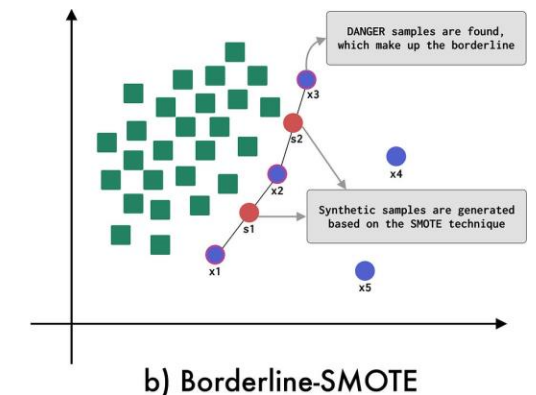
# Oversampling Techniques

- **SMOTE** (Synthetic Minority Oversampling Technique):
    - Increases minority class samples by creating **synthetic examples** (no duplicates);

    - New samples are generated by **interpolating between existing minority class** examples;

    - **Reduces risk of overfitting** (compared to random oversampling).



○ Majority class samples

✚ Minority class samples

⊕ Randomly selected minority class sample $x_i$

⊕ 5 $K$-nearest neighbors of $x_i$

⊕ Randomly selected sample $\hat{x}_i$ from the 5 neighbors

✚ Generated synthetic minority instance

# Oversampling Techniques

- SMOTE variations:
  - **SMOTEN**:
    - SMOTE for **nominal** (categorical) data.

  - **SMOTENC**:
    - SMOTE for **nominal and continuous** data.

  - **Borderline-SMOTE**:
    - Focus on **samples near the decision boundary**;
    - Aims to improve classification of difficult, borderline cases.



a) Class Imbalance



DANGER samples are found, which make up the borderline

Synthetic samples are generated based on the SMOTE technique

b) Borderline-SMOTE
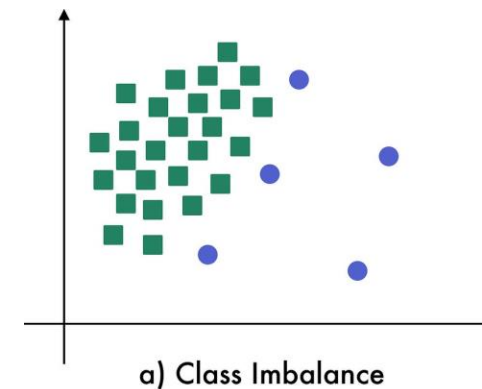
# Oversampling Techniques

- SMOTE variations:
  - **Kmeans-SMOTE**:
    - Combines k-means with SMOTE;
    - **Clusters the dataset** into K clusters using K-Means;
    - Applies SMOTE within each cluster to generate synthetic minority class samples;
    - Aims to create more **diverse and representative synthetic samples**.



Input data

Find k = 3 clusters and compute imbalance ratio (IR)

$IR_1 = \frac{5}{2}$

$IR_2 = \frac{1}{48}$

$IR_3 = \frac{14}{1}$

× Minority Sample
• Majority Sample
- - Decision Boundary
+ Generated Sample

Use SMOTE to oversample clusters with IR > 1, generating more samples in sparse clusters

$IR_2 = \frac{1}{48}$

Oversampled data rectifies decision boundary
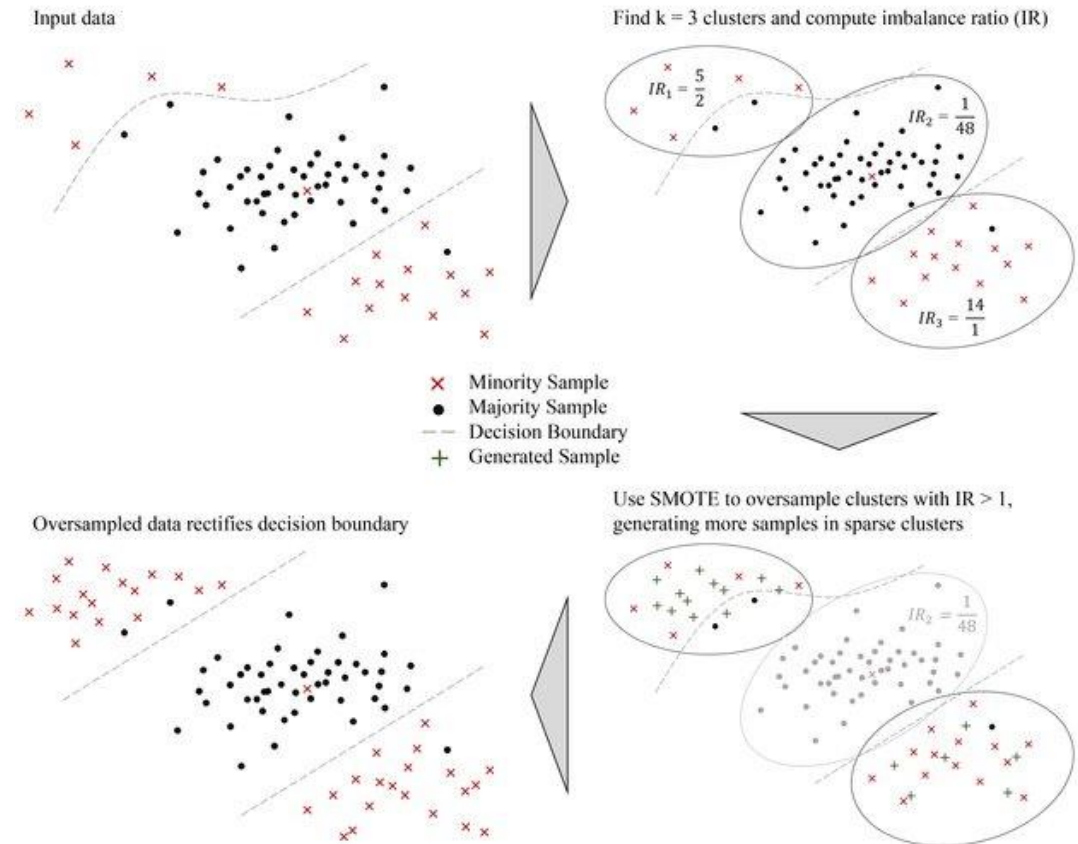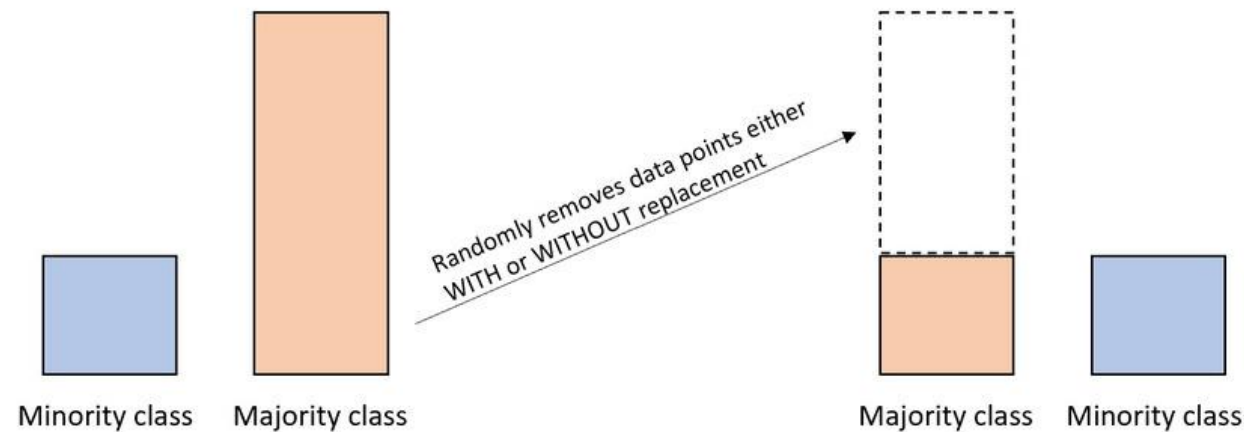
# Oversampling Techniques

- SMOTE variations:
    - **svm-SMOTE**:
        - Combines **SVMs** with **SMOTE**;
        - Uses SVM to identify the **decision boundary** between classes;
        - Generates **synthetic minority class samples near the SVM decision boundary**;
        - Focuses on **difficult-to-classify samples**.



○ Majority sample
● Minority sample
● Minority SV
○ Majority SV
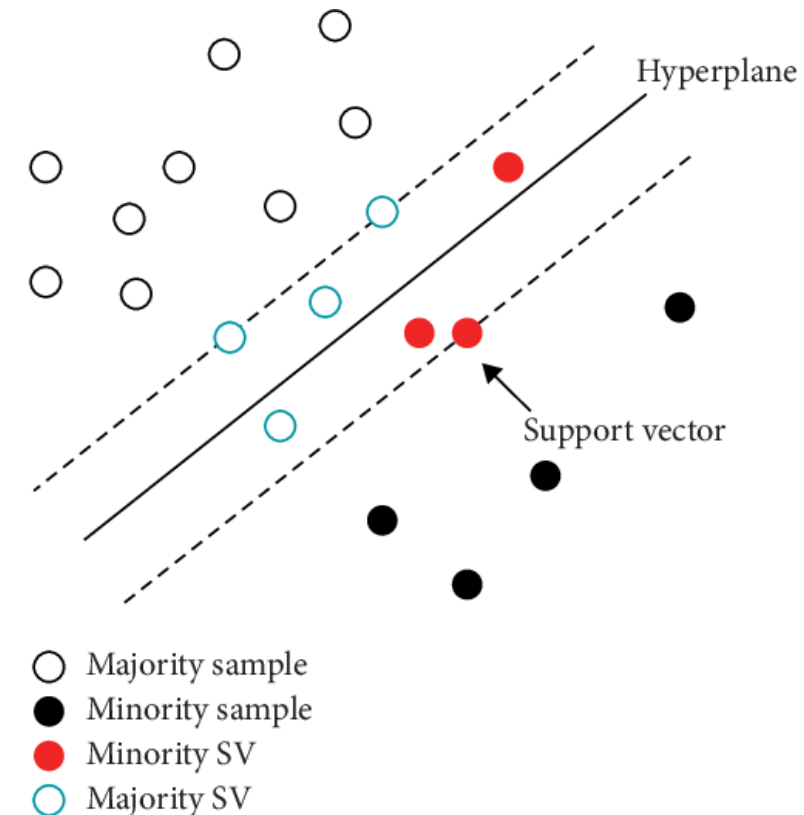
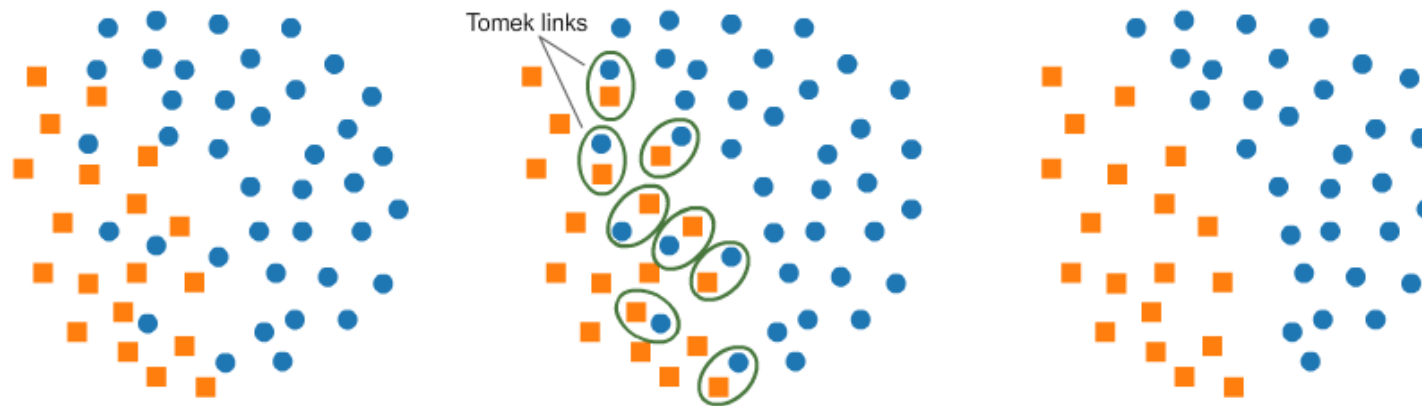# Undersampling Techniques

- **Random undersampling**:
  - Balances data by **randomly removing samples from the majority class**;

  - **Simple and quick**;

  - Risk of **losing important information** from the data.

# Undersampling Techniques

- **Tomek Links**:
  - Identifies and removes **Tomek links**, which are **pairs of nearest neighbors from different classes**;
  - Aims to **remove borderline examples** to clarify class boundaries;
  - Helps to **reduce class overlap** and improve classifier performance.

# Undersampling Techniques

- **Cluster Centroids Undersampling**:
  - Replaces majority class with **cluster centroids**;
  - Aims to **retain important information** while reducing majority class size;
  - Balances class distribution and **minimizes information loss**.

- **Condensed Nearest Neighbor Undersampling:**
  - Undersampling technique based on **nearest neighbor classification**;
  - Iteratively **selects samples that correctly classify others**;
  - **Retains representative samples** while **removing redundant ones**;
  - Helps reduce dataset size while preserving classification accuracy.

# Undersampling Techniques

- Others:

    - Edited Nearest Neighbours

    - AllKNN

    - NearMiss

    - One Sided Selection

    - Etc...

# Combining Under and Oversampling Techniques

- Using both under and oversampling may help mitigate the drawbacks of each technique while leveraging their advantages.

- **SMOTEENN** - SMOTE with Edited Nearest Neighbors (ENN):
  - Generates synthetic minority class samples (SMOTE) and removes majority class examples misclassified by a KNN classifier (Edited Nearest Neighbors).

- **SMOTETomek**- SMOTE with Tomek Links:
  - Generates synthetic minority class samples (SMOTE) and removes Tomek links.

# Approaches to Adress Data Imbalance

- **Algorithm-level techniques**:
    - Some models can inherently better deal with class imbalance (e.g. RandomForests, AdaBoost, etc.);
    - Building ensembles of multiple models;

# Approaches to Adress Data Imbalance

- **Algorithm-level techniques**:
    - Some models can inherently better deal with class imbalance (e.g. **RandomForests**, **AdaBoost**, etc.);
    - Building **ensembles** of multiple models;
    - Assigning **different misclassification costs to classes** can encourage the model to focus on the minority class;
    - **Threshold adjustment** to control the trade-off between precision and recall.

- **Evaluation metrics**:
    - When working with imblanced data, it is important to use appropriate evaluation metrics (e.g. **imbalanced accuracy, precision, recall, f1-score, AUC-ROC, Matthew's correlation coefficient**).

# Resources

- He, H., & Ma, Y. (2013). Imbalanced Learning: Foundations, algorithms, and applications (H. He & Y. Ma, Eds.; 1st ed.). Wiley-IEEE Press.

- Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from imbalanced data sets (1st ed.). Springer.

- https://imbalanced-learn.org/stable/