





# Hands-on Imbalanced Classification

Master in Data Science | 19 MAY 2021

Nicola Procopio

# Summary

**1# Healthware Overview**

**2# Data Science Team**

Scenario e obiettivi

**3# DTx Projects and Data Science**

**4# Hand-On Imbalanced Classification**

**5# Q&A**

01

# Healthware

A portrait of Roberto Ascione, CEO & Founder, smiling. He has dark hair and a beard, wearing a dark blazer over a light blue button-down shirt. The background is a neutral, light gray.

**In our vision digital  
technology &  
innovation are the  
driving forces behind  
the transformation in  
healthcare, leading to a  
world of increasingly  
relevant, human-sized,  
solutions to health  
challenges**

Roberto Ascione, CEO & Founder

# Born digital, forward looking, fully integrated

Advisory  
Marketing  
Medical  
Media  
Technology

Digital Health  
DTx R&D

Healthware works at the intersection of industry digital transformation and digital health by providing a novel, integrated solution to existing and emerging stakeholders combining marketing, communications, technology capabilities with innovation consultancy and a corporate venturing arm

Publishing  
Education  
Events

Venture Building



healthware<sup>®</sup>

INTOUCH



GROUP<sup>®</sup>

# The Largest, Most Respected, Independent Player In the Industry

New York | Boston | Kansas City | Chicago | San Diego | San Francisco | London | Barcelona | Cologne | Milan | Rome | Salerno | Rende | Helsinki | Mumbai



1,300+

in-house associates  
across 15 offices



50+

life science  
clients



Top 20

creative healthcare  
communications  
agencies  
worldwide



5 times in 6 years

agency of the year  
by  
Med Ad News,  
MM&M and PM360



150+

brands currently  
represented

# Focus on Data Science Team





**The full-service healthcare  
agency of Healthware Group**

We play at the intersection of science, creativity, boundless curiosity, and our understanding of human needs. That's how we design transformational healthcare experiences that engage, simplify and empower people's lives.

We are digital natives and multi-talented coders, connected and passionate to learn and innovate.

Our mission is to design and develop successful solutions and digital products.

# The Sila Valley

Healthware Data Science Team

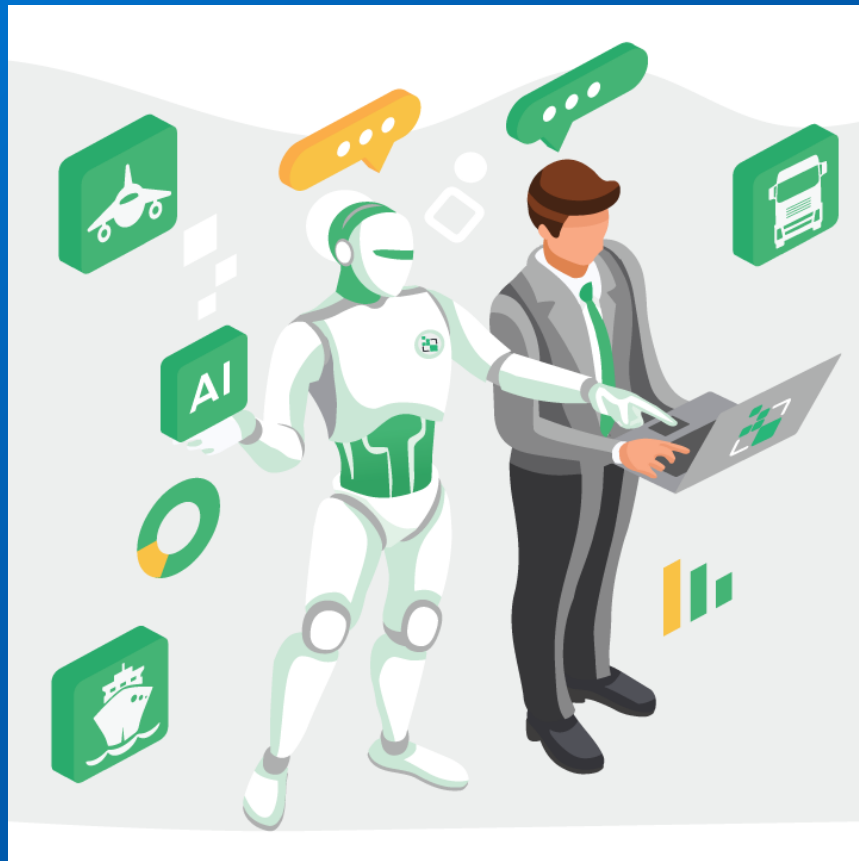
- **The Data Science team of Healthware covers with its expertise the entire design process:** Data Ingestion, Data Analysis and Analytics, Algorithms (NLP, ML/AI, DeepLearning, Statistics), Data Visualization.
- The team has deep expertise not only on models and algorithms, but also on architectures: Big Data and Cloud in particular.
- **The Healthware Data Science Team, is located in a district of ICT particularly focused on Artificial Intelligence:** Cosenza is a very stimulating environment due to the presence of **Universities and Research Centers, Startups and Communities**, other companies in the Artificial Intelligence sector.
- The medical campus of the **Magna Graecia University** offers, among other research lines, also a **research center in neuroscience and medical science.**



# Our Philosophy

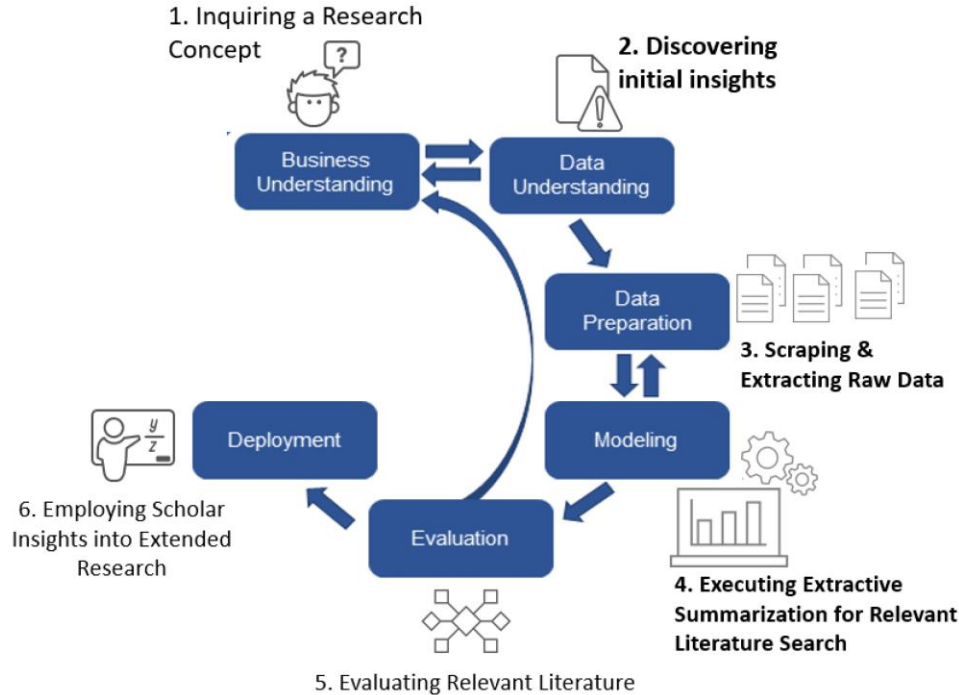
## Healthware Data Science Team

- **From Big to Smart Data.** Giving meaning to data is, therefore, the element that distinguishes us.
- **From Artificial to Augmented Intelligence.** Big data, NLP, machine learning, neural networks to support doctors to improve the quality of life.
- **Explainable Artificial Intelligence (XAI).** XAI can improve the user experience of a product or service by helping end users trust that the AI is making good decisions.



# CRISP - DM

Methodology for prototype



## CRISP – DM Framework

Methodology  
**C**ross Industry **S**tandard **P**rocess  
for **D**ata **M**ining

An **open standard process model**  
that describes common approaches  
used by data mining experts.

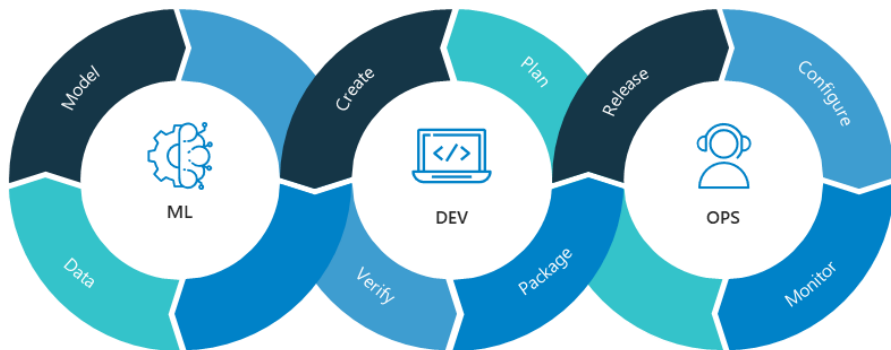
# MLOps

## Machine Learning in production

MLOps looks to increase automation and improve the quality of production ML while also focusing on business and regulatory requirements.

The predicted growth in machine learning includes an estimated doubling of ML pilots and implementations from 2017 to 2018, and again from 2018 to 2020.

In 2018, after having one presentation about ML productionization from Google, MLOps began to gain traction as a solution that can address the complexity and growth of machine learning in businesses.



### Some Tools

- MLflow
- Jira and Confluence
- Kubeflow
- Amazon Sagemaker
- MLLeap

# Who I am

Data Science Team



**Nicola Procopio**

Senior Data Scientist

## Contacts



[nicola.procopio@healthwareinternational.com](mailto:nicola.procopio@healthwareinternational.com)



<https://it.linkedin.com/in/nicolaprocopio>



<https://github.com/nickprock>



<https://www.slideshare.net/NicolaProcopio>

## Education

*Master in Applied Statistics for Economy and Finance.*

UNIVERSITÀ  
DELLA CALABRIA



## Background



## Community



healthware  
NEXT-GEN HEALTH CONSULTANCY



# DTx Projects and Data Science



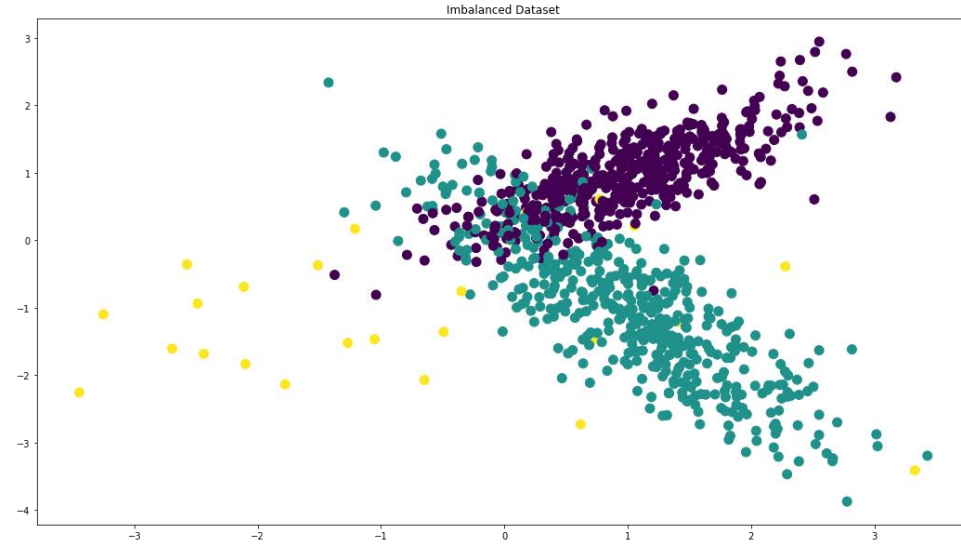
*"Our intelligence is what makes us human, and AI is an extension of that quality."*

*-Yann LeCun-*

# Hands-on Imbalanced Classification

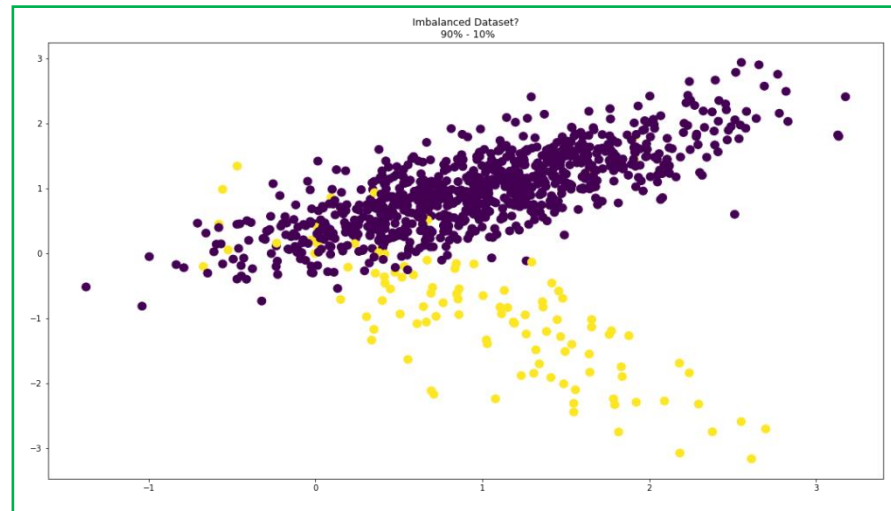
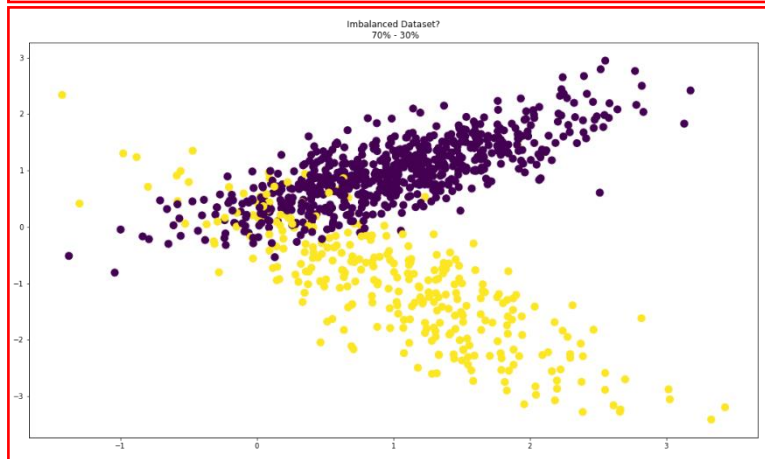
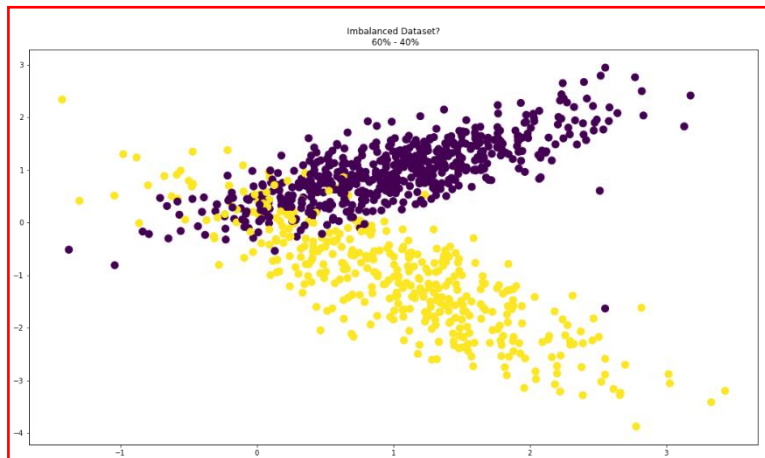
# What's Imbalanced Classification?

# The Problem



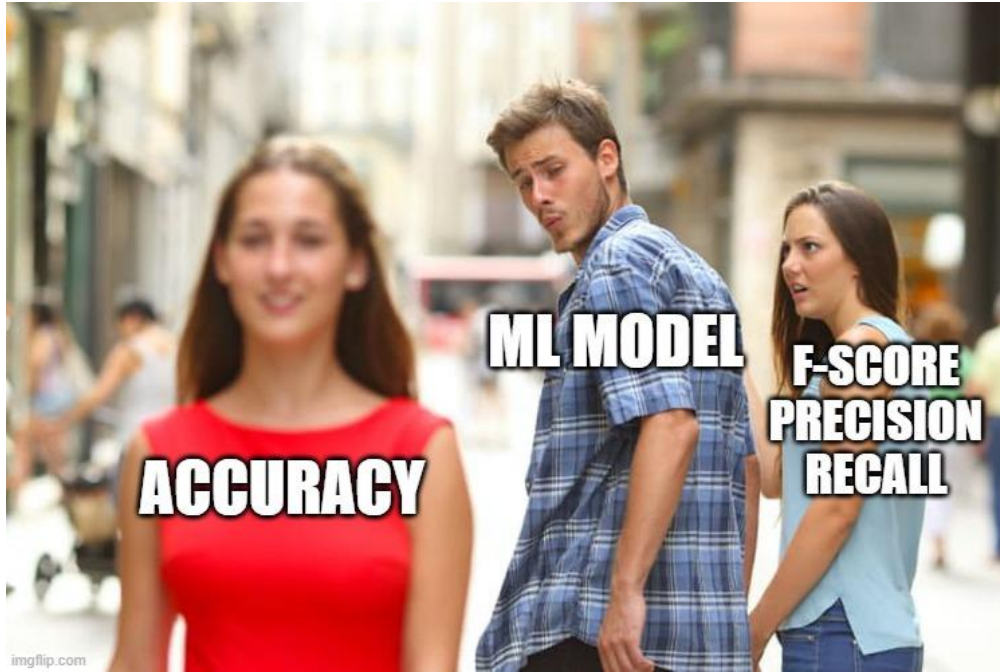
- We refer to ***imbalanced classification*** when a class (or more than one) is much less present than the others.
- It's a common problem in real world datasets.
- This bias in the training dataset can influence many machine learning algorithms.
- Some tasks:
  - Fraud Detection
  - Predictive Maintenance
  - Anomaly Detection

# When is a dataset Imbalanced?





# Why classification fails?



- Machine learning models are built under the hypothesis of a dataset with balanced classes
- Classical metrics to optimize the models are focused on the majority class
- The “*default*” probability **threshold** may not represent an optimal interpretation of the predicted probabilities.

# Metrics

# Classic Threshold Metrics

	Predicted label class 1	Predicted label class 2
True label class 1	<b>correct</b> true positive for class 1	<b>wrong</b> false positive for class 2
True label class 2	<b>wrong</b> false positive for class 1	<b>correct</b> true positive for class 2

$$\text{accuracy} = \frac{\text{orange} + \text{blue}}{\text{orange} + \text{yellow} + \text{blue} + \text{green}}$$
$$\begin{array}{l} \text{class 1 precision} = \frac{\text{orange}}{\text{orange} + \text{yellow}} \\ \text{class 2 precision} = \frac{\text{blue}}{\text{blue} + \text{green}} \end{array}$$
$$\begin{array}{l} \text{class 1 recall} = \frac{\text{orange}}{\text{orange} + \text{green}} \\ \text{class 2 recall} = \frac{\text{blue}}{\text{blue} + \text{yellow}} \end{array}$$

- **Accuracy:** the most intuitive performance indicator and it is simply a ratio of correctly predicted observation to the total observations.
- **Precision:** the ratio of correctly predicted positive observations to the total predicted positive observations.
- **Recall (Sensitivity):** is the ratio of correctly predicted positive observations to the total of observations in actual class.
- **Specificity:** is the ratio of correctly predicted negative observations to the total really negative observations. How good a test is at avoiding false alarms.

# Focus On

G-Mean, F-Measure, Brier Score

$$\text{G-Mean} = \sqrt{(\text{Sensitivity} \times \text{Specificity})}$$

The  **$F_\beta$  measure** is an abstraction of the F-measure where the balance of precision and recall in the calculation of the harmonic mean is controlled by a coefficient called beta. Like precision and recall, a poor F-Measure score is 0.0 and a best or perfect F-Measure score is 1.0.

$$\text{BrierScore} = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2$$

The **Geometric Mean** is a metric that measures the balance between classification performances on both the majority and minority classes. A low G-Mean is an indication of a poor performance in the classification of the positive cases even if the negative cases are correctly classified as such.

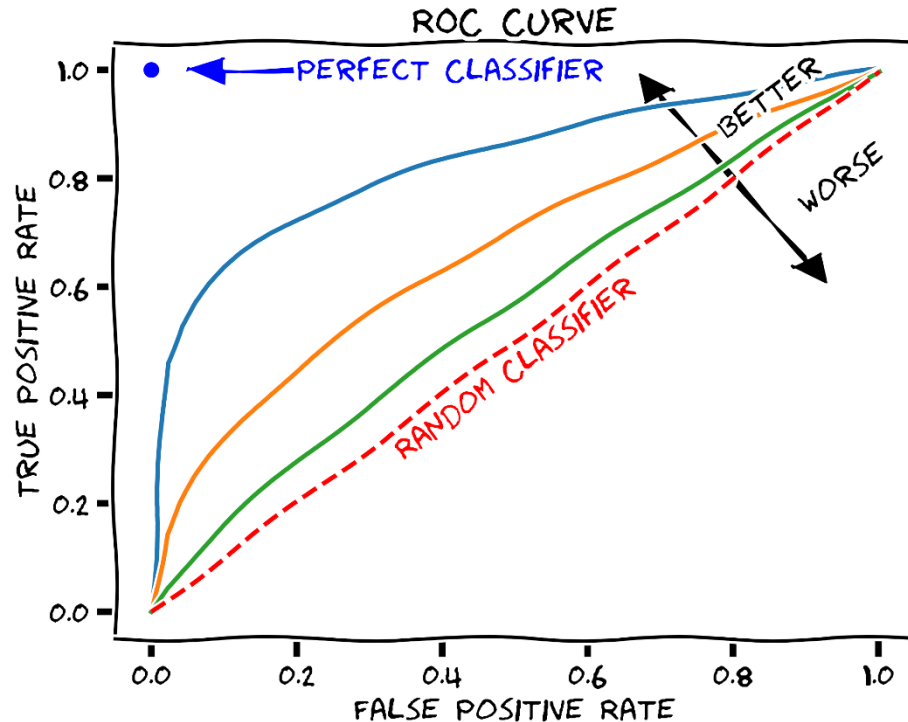
$$F_\beta = (1 + \beta^2) \times \frac{\text{Precision} \times \text{Recall}}{(\beta^2 \times \text{Precision}) + \text{Recall}}$$

The **Brier score** calculates the mean squared error between predicted probabilities and the expected values.

The score summarizes the magnitude of the error in the probability forecasts.

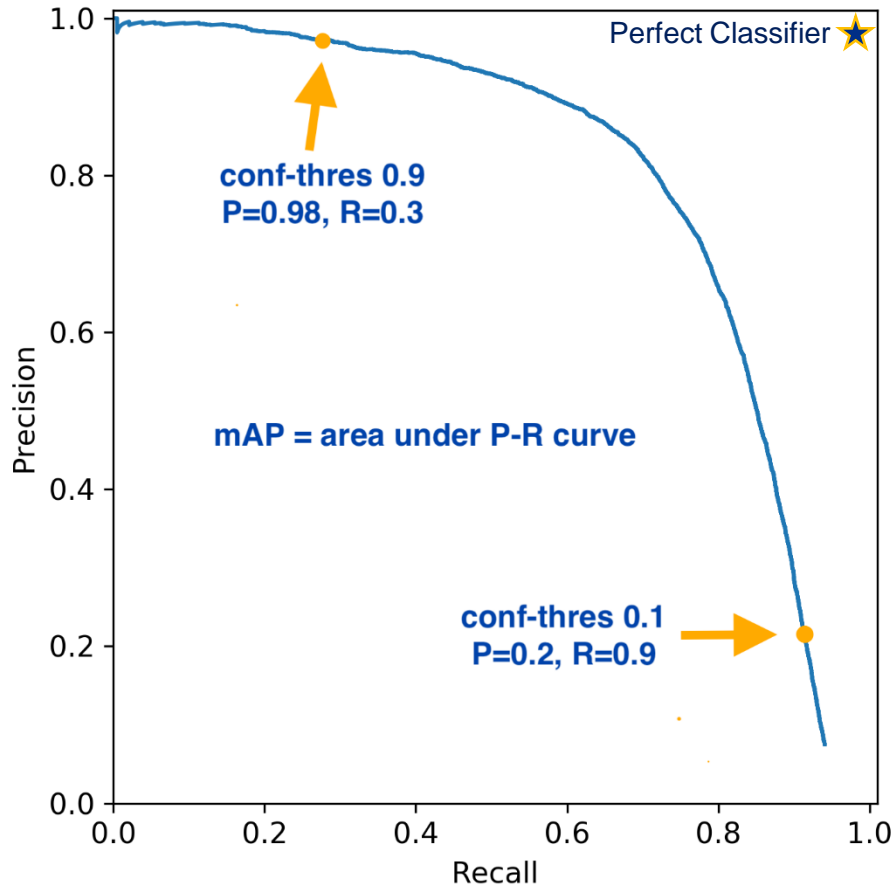
The error score is always between 0.0 and 1.0, where a model with perfect skill has a score of 0.0.

# ROC AUC



- **ROC curve** is a plot that summarizes the performance of a binary classification model on the positive class.
  - $TP\ rate = TP/(TP+FN)$
  - $FP\ rate = FP/(FP+TN)$
- Ideally, we want the fraction of correct positive class predictions to be 1 (top of the plot) and the fraction of incorrect negative class predictions to be 0 (left of the plot).
- Very useful for threshold-moving task, ideally, we want the threshold associates at the point [0,1], or that maximizes the area under the curve.

# PRC AUC

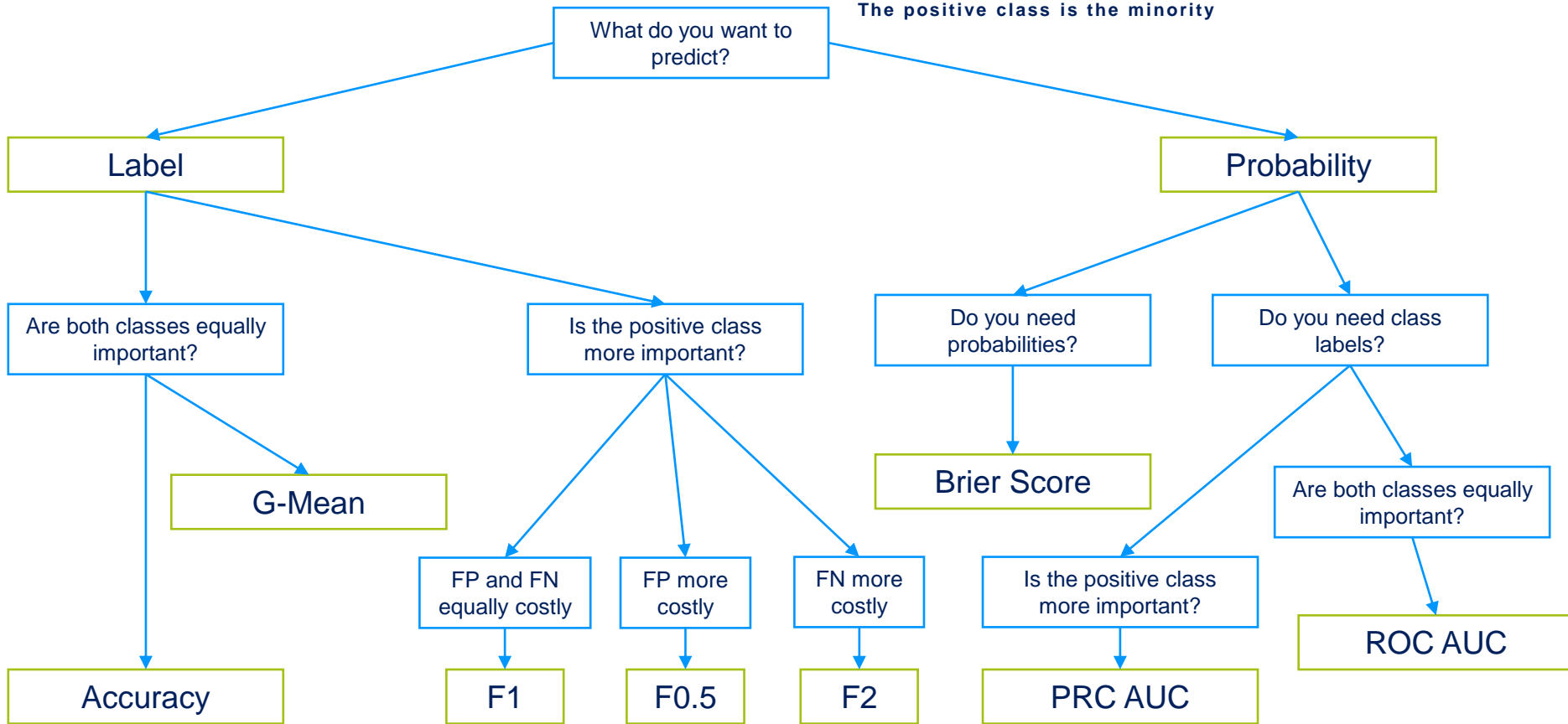


- A precision-recall curve is a plot of the precision and the recall for different probability thresholds.
- A model with perfect skill is depicted as a point at a coordinate of (1,1).
- A no-skill classifier will be a horizontal line on the plot with a precision that is proportional to the number of positive examples in the dataset.
  - In Imbalanced dataset the positive examples are the minority class.
- Very useful in health problems or in predictive maintenance tasks.



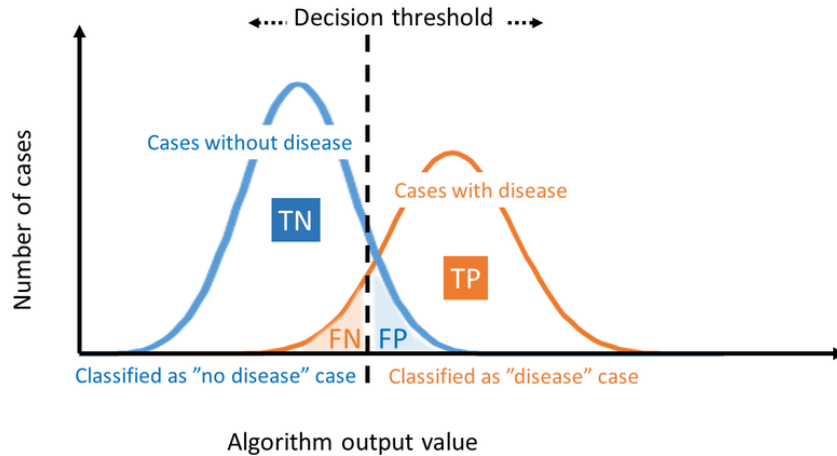
# Metrics for I.C.

The positive class is the minority



# Threshold-Moving

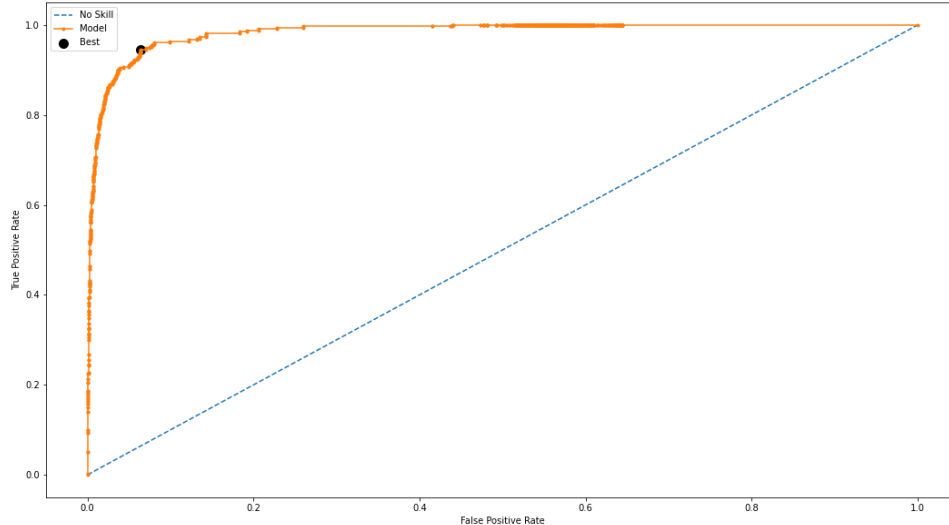
# Gentle Introduction



- The decision for converting a predicted probability or scoring into a class label is governed by a parameter referred to as the “**decision threshold**”
- The ML models are created for balanced dataset, the default threshold for binary classification usually is 0.5
- Every problem needs its decision threshold.
- Steps:
  - Train the model
  - Predict probabilities on Test set
    - Convert probabilities to class labels using Thresholds
    - Evaluate class labels
    - Choose the best Threshold
  - Use adopted Threshold on new data

# Using ROC Curve

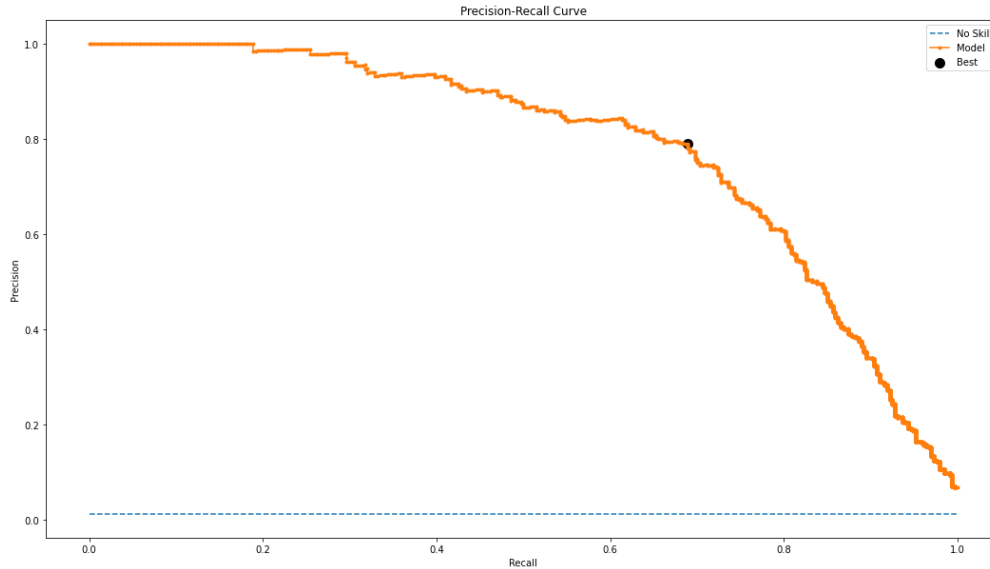
## Threshold-Moving



- There are many ways to find the Threshold with the best balance between False Positive Rates and True Positive Rates
- In this example we use the G-Mean but another simple method is the [Youden's J statistic](#)
- Calculate the G-Mean for each Threshold and locate the index with the largest score.
- Use that index to find the best Threshold

# Using PRC Curve

## Threshold-Moving

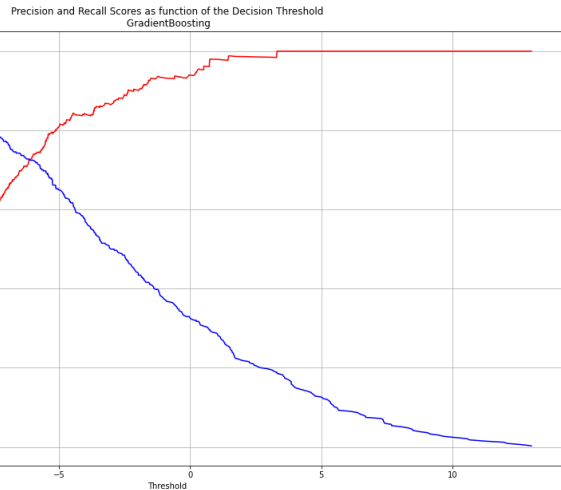
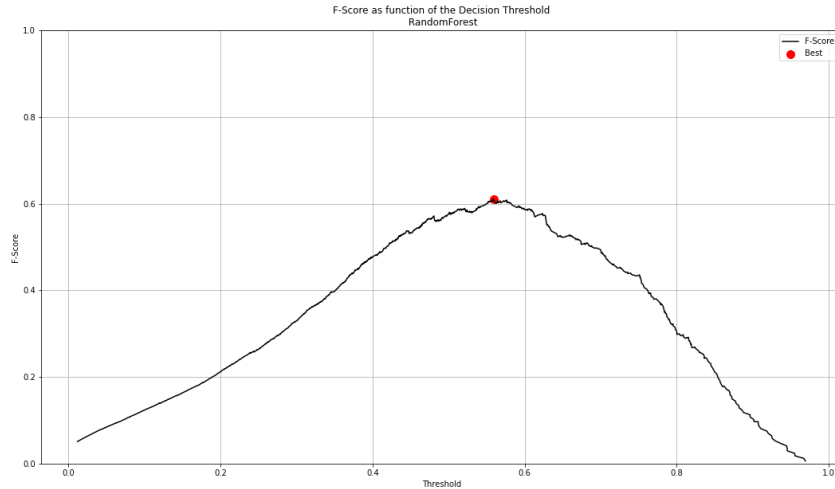


- The PRC **focuses on** the performance of a classifier on **the positive class**.
- In this example we use the F-Measure to find the optimal balance between precision and recall
- As for the G-Mean, also in this case we calculate the metrics for each threshold and find the index with the largest score.
- Use that index to find the best Threshold

# Other Methods

## Threshold-Moving

- *F-Score as function of the Decision Threshold*



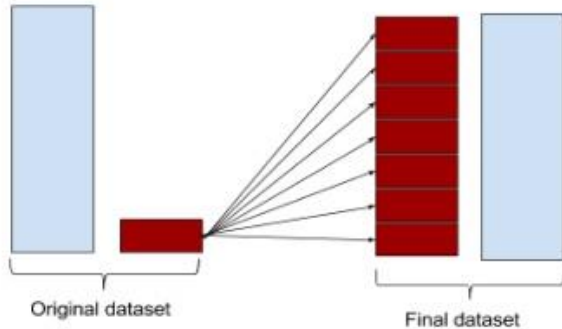
- *Precision and Recall as function of the Decision Threshold*



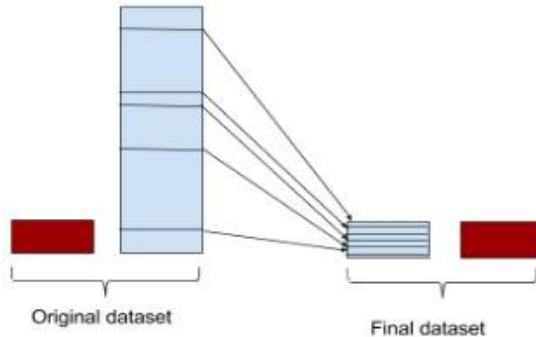
# Oversampling and Undersampling

# Resampling

**Oversampling** minority class



**Undersampling** majority class



- One approach to addressing the problem of class imbalance is to resample the training dataset.
- There are many methods to resample the dataset, everyone introduce or remove informations
- If we want to resize the minority class we use the **oversample**
- If we want to rebalance deleting some examples in the majority class we use the **undersample**
- [We use the Imblearn library](#)

# Random Resampling

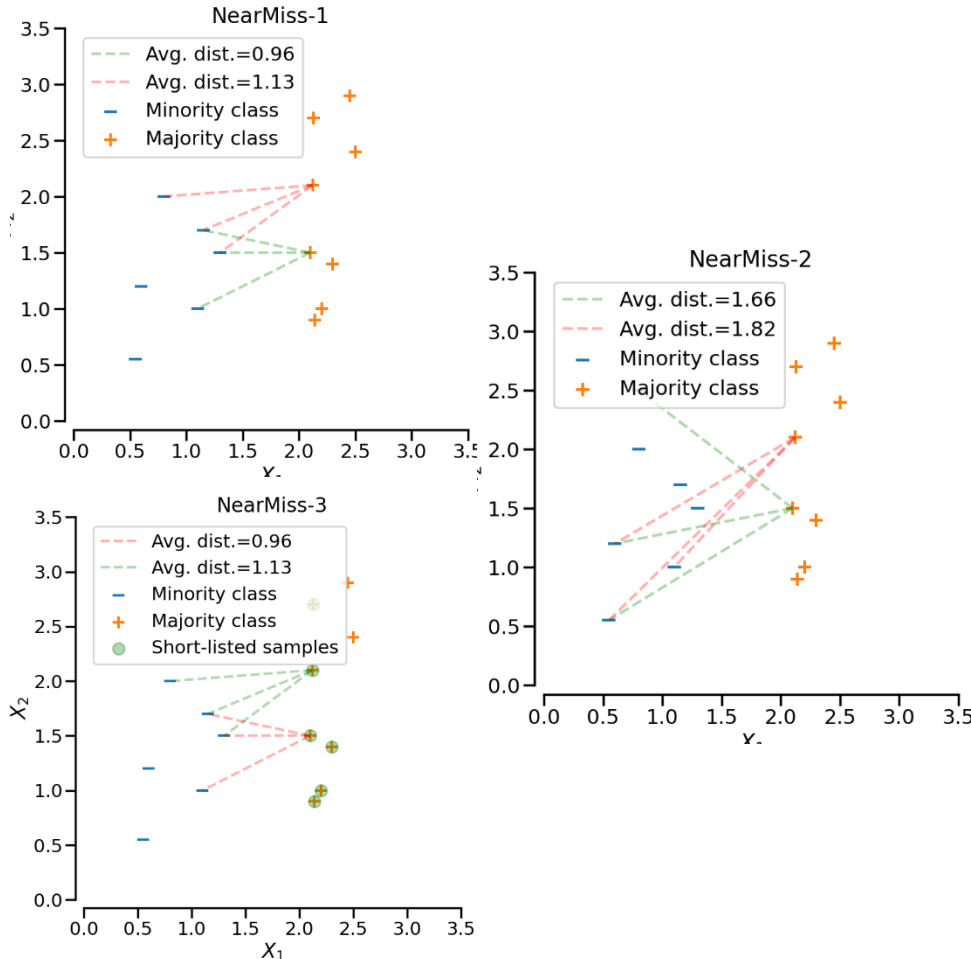


- *Random Undersampling:*  
Randomly delete examples in the majority class
  - **Advantage:** improves computation time.
  - **Disadvantage:** important information could be deleted.  
The sample may not be representative.
- *Random Oversampling:*  
Randomly duplicate examples in the minority class
  - **Advantage:** there is no loss of information
  - **Disadvantage:** overfitting

# NearMiss

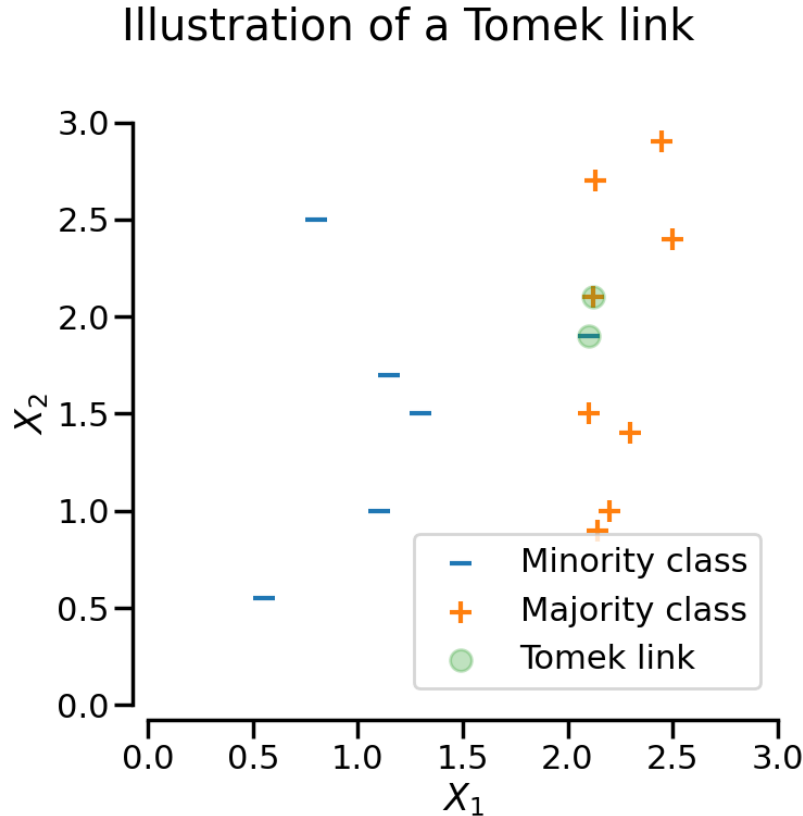
## Undersampling

- *Based on KNN*
- Implements 3 different types of heuristic.
  - **NearMiss-1** selects the majority class samples for which the average distance to the  $N$  closest samples of the minority class is the smallest.
  - **NearMiss-2** selects the majority class samples for which the average distance to the  $N$  farthest samples of the minority class is the smallest.
  - **NearMiss-3** is a 2-steps algorithm. First, for each minority class sample, their  $M$  nearest-neighbors will be kept. Then, the majority class samples selected are the one for which the average distance to the  $N$  nearest-neighbors is the largest.



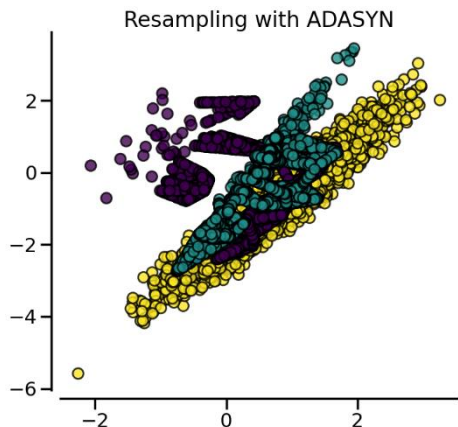
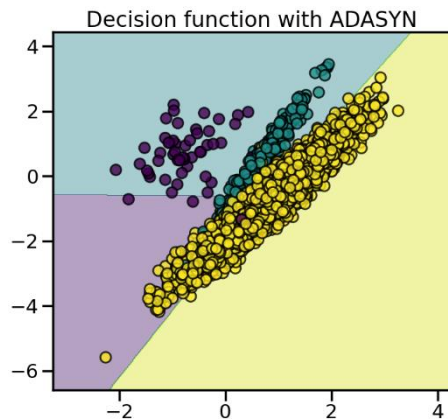
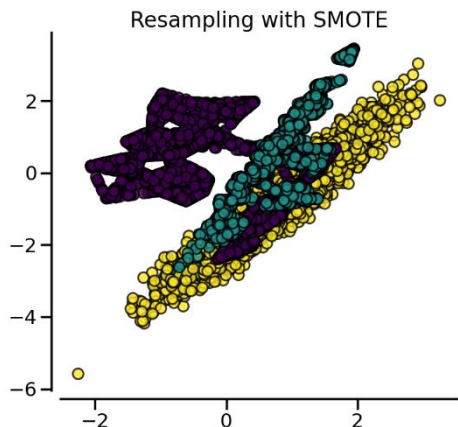
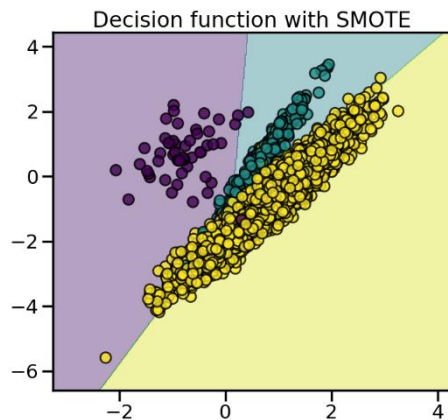
# Tomek Link (T-Link)

## Undersampling



- Based on KNN
- Tomek Link exist if the two samples of different classes are the nearest neighbors of each other in their class.
- If any two examples are T-Link then one of these examples is a noise or otherwise both examples are located on the boundary of the classes.
- The observations from the majority class are removed.

## Particularities of over-sampling with SMOTE and ADASYN



# SMOTE e ADASYN

## Oversampling

- *Based on KNN*
- Generate new samples in by interpolation.
- **ADASYN** focuses on generating samples next to the original samples which are wrongly classified using a k-Nearest Neighbors classifier.
- **SMOTE** will not make any distinction between easy and hard samples to be classified using the nearest neighbors rule.

# Other Methods

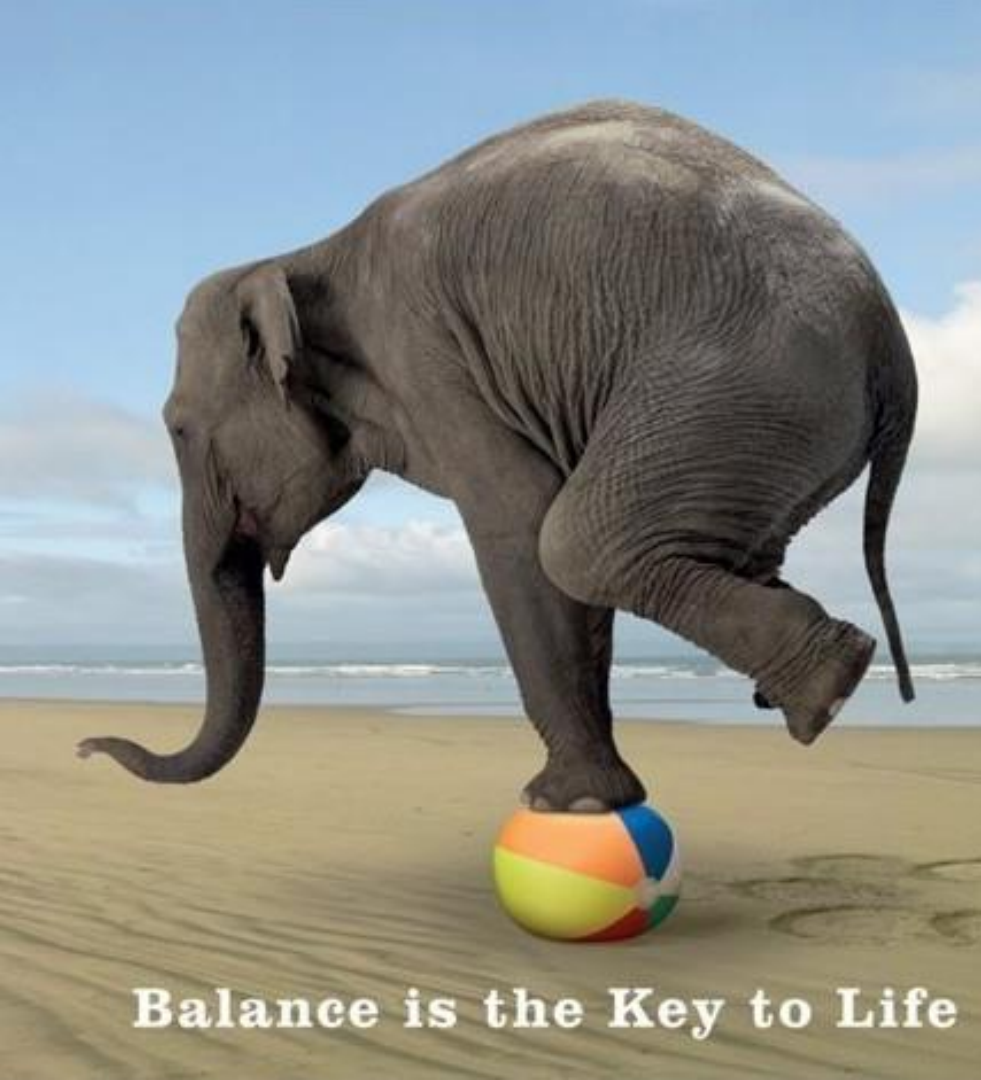


- Undersampling:
  - Edited Nearest Neighbors (ENN) and RENN
  - One Side Selection
  - Neighbourhood Cleaning Rule
  - AIKNN
- Oversampling:
  - SMOTE Variants:
    - Bordeline SMOTE
    - SVM SMOTE
    - K-Means SMOTE
  - GAN-Based
  - Cluster Based Oversampling



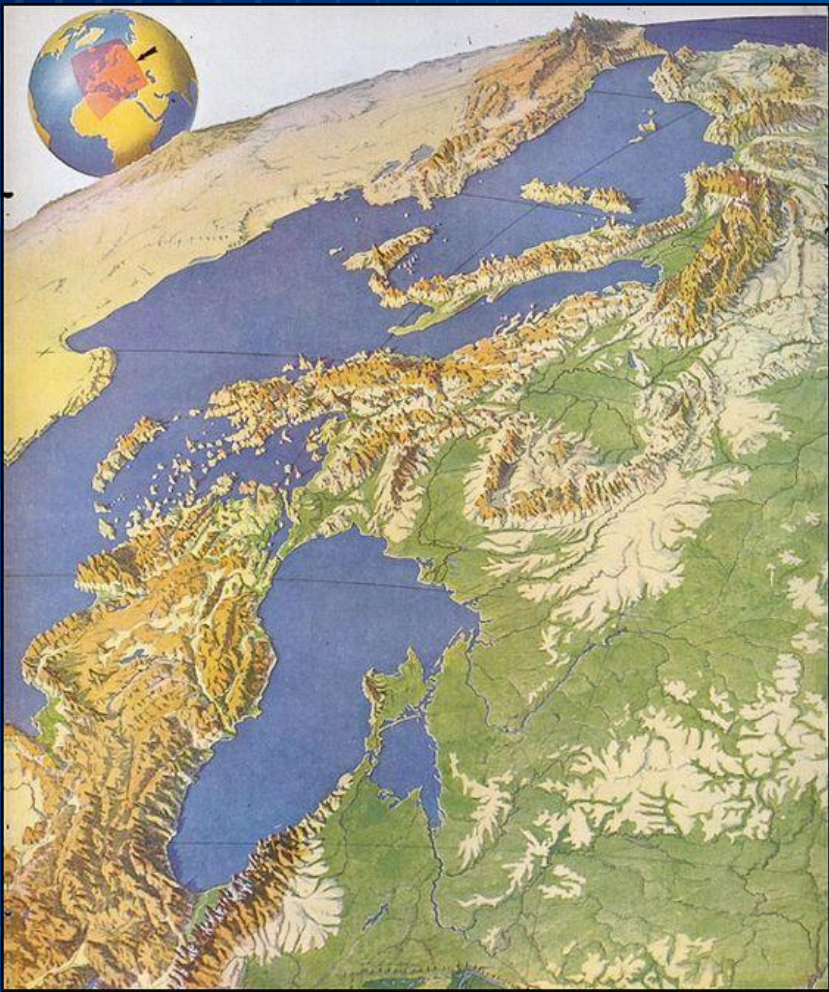
# Class weight





# Balance weights

- Similar to oversampling but introduce repetition of samples associated with the minority classes.
- The idea is to weigh the loss computed for different samples differently based on whether they belong to the majority or the minority classes.
- There are different ways to apply this method in scikit-learn



*If none of this works,  
change your point of view!*

# Q&A

*«I suppose it is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail.»*

*-Abraham Maslow – 1966-*

# Thank you

@healthware\_intl | @healthware\_ita  
linkedin.com/company/healthware-international  
facebook.com/healthwareintl  
instagram.com/healthware\_intl  
[www.healthwaregroup.com](http://www.healthwaregroup.com)

healthware<sup>■</sup>  
NEXT-GEN HEALTH CONSULTANCY