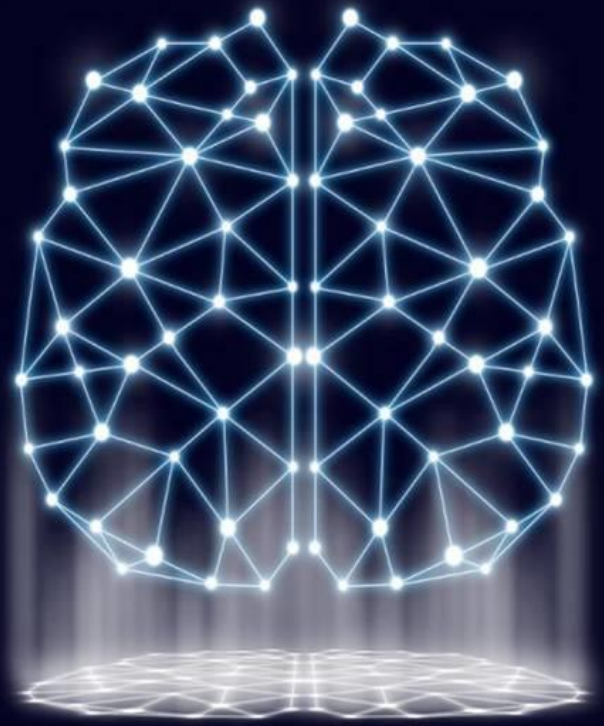
FULLSTACK



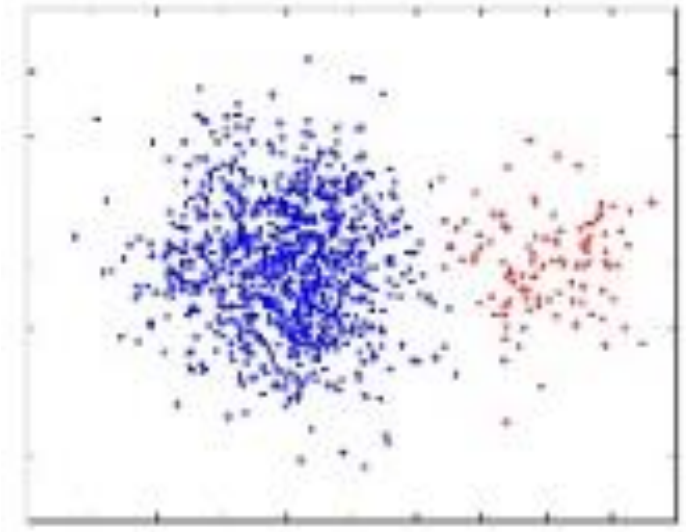
Curso de Especialización de Machine Learning con Python

TÉCNICAS DE BALANCEO DE DATOS

¿Qué es data desbalanceada?

Comúnmente hace referencia a la variable “Target”.

Def: Se da cuando la frecuencia de clases de la variable Target son muy distintas o muy desiguales.



Data desbalanceada

¿Qué consecuencias puede traer?

Al momento de entrenar un algoritmo de ML con el dataset desbalanceado, se puede originar un sesgo hacia una clase en particular de la variable target. Hacia la clase mayoritaria.

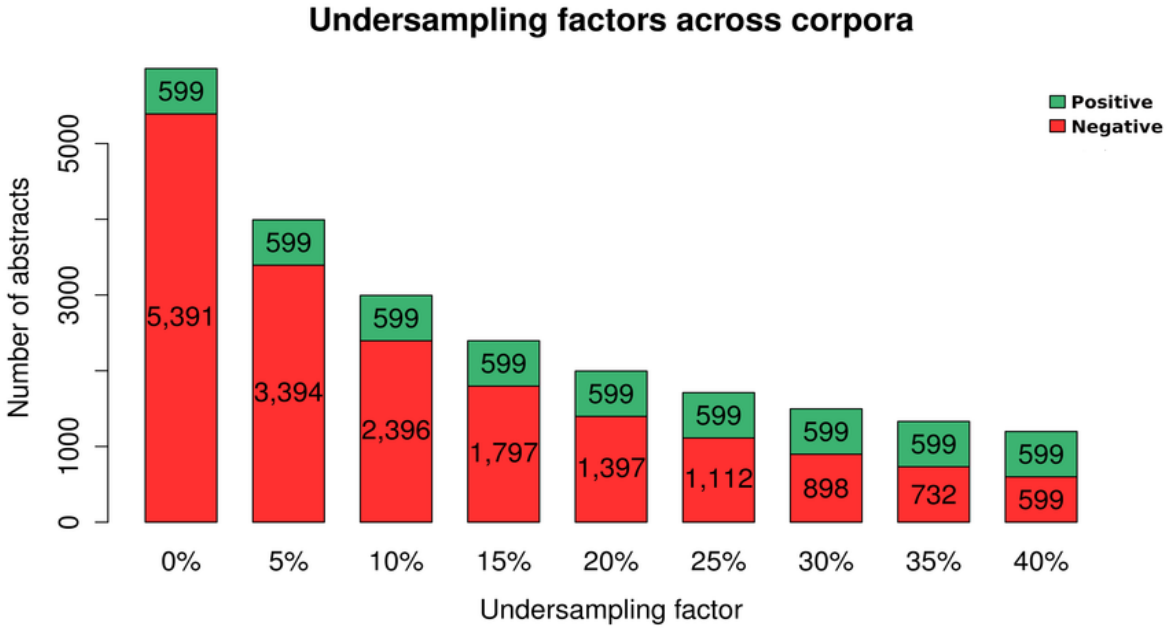
¿Cómo lo solucionamos?

Balanceando la data.

- Técnicas: Undersampling, Oversampling, SMOTE, a criterio propio, otras

UNDERSAMPLING

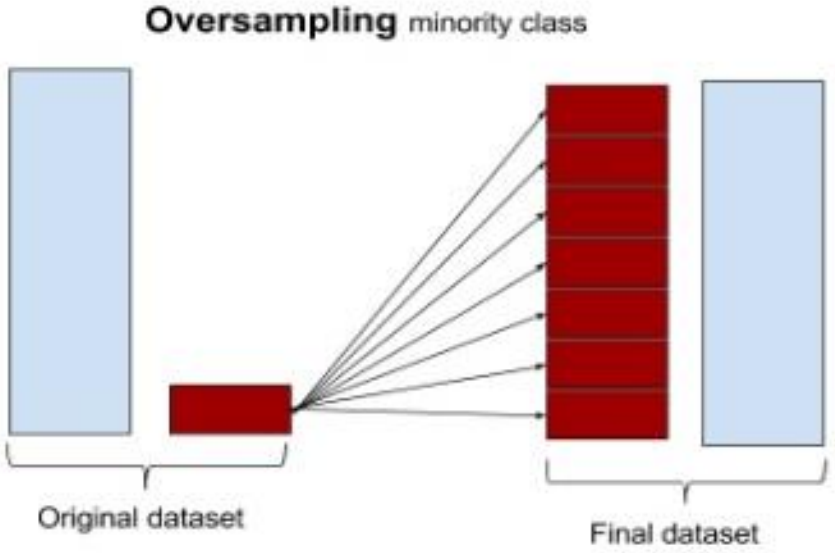
En esta técnica se busca reducir la cantidad de registros de la clase mayoritaria a la cantidad de registros de la clase minoritaria.



Mayor a menor → Undersampling

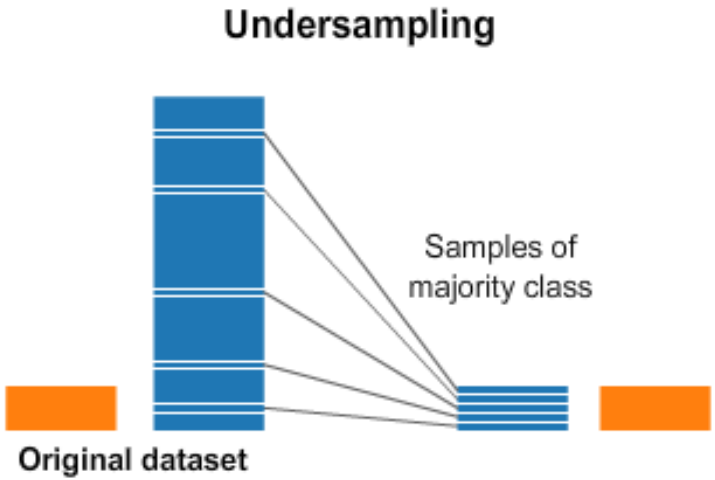
OVERSAMPLING

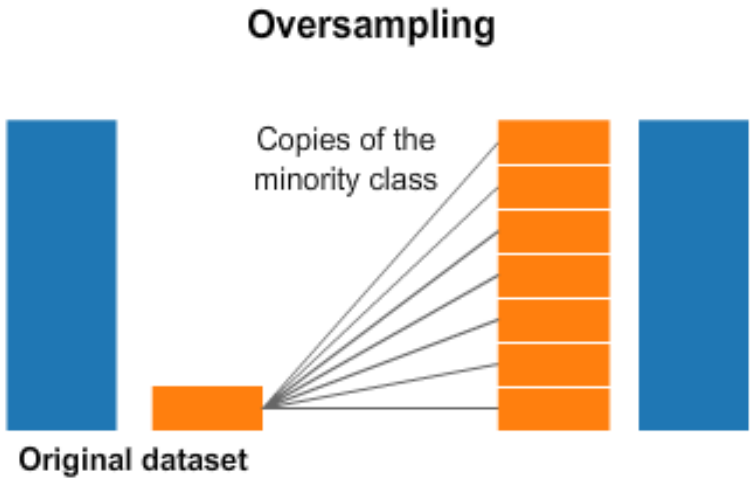
En esta técnica se busca incrementar la cantidad de registros de la clase minoritaria a la cantidad de registros de la clase mayoritaria.



Menor a mayor → Oversampling

RESUMEN


Undersampling
Samples of
majority class
Original dataset


Oversampling
Copies of the
minority class
Original dataset

SMOTE

Synthetic Minority Over-sampling Technique

Es una técnica de oversampling.

El objetivo es crear puntos sintéticos a partir de la data de la clase minoritaria.
Ejemplo: k = 5