MACHINE LEARNING INMERSION

ANDRÉ OMAR CHÁVEZ PANDURO

Perú

« Divide las dificultades que examinas en tantas partes como sea posible , para su mejor solución»

AGENDA

- Balanceo de Datos.
- Técnicas de UnderSampling y OverSampling.
- Smote, Tomek-Link, Clusters.

Balanceo de la Información

INTRODUCCIÓN

Información o «dataset» con múltiples clases es entendida como desbalanceada cuando las clases minoritarias están sub-representadas en oposición a la clase

PROBLEMÁTICA

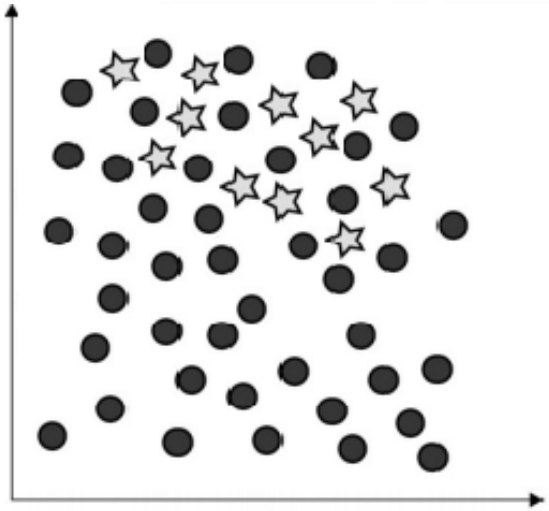
- Los algoritmos de clasificación funcionan pobremente en la clase minoritaria.
- El costo de mala clasificación en dicha clase suele ser mucho mayor que el resto.
- En muchos dominios del mundo real existe una clase dentro de la variable de estudio o target la cual acumula la gran mayoría de elementos.

ESTRATEGIAS DE BALANCEO

- Existen en la literatura muchas maneras o metodologías para balancear o equilibrar las clases, cada una dependiendo de la proporción de elementos en cada una de las clases y del tipo de problema que uno está abordando. Entre las más usadas y estudiadas tenemos:

✓ **UnderSampling.**

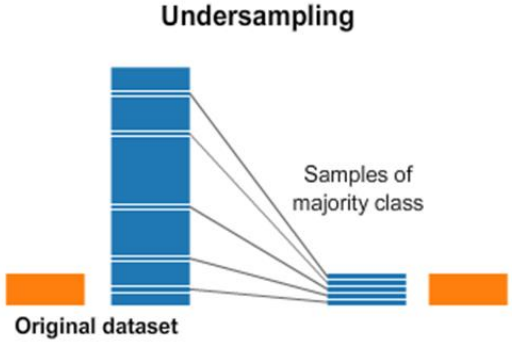
✓ **OverSamplig.**



ESTRATEGIAS DE BALANCEO

UNDERSAMPLING (RANDOM UNDERSAMPLING RUS)

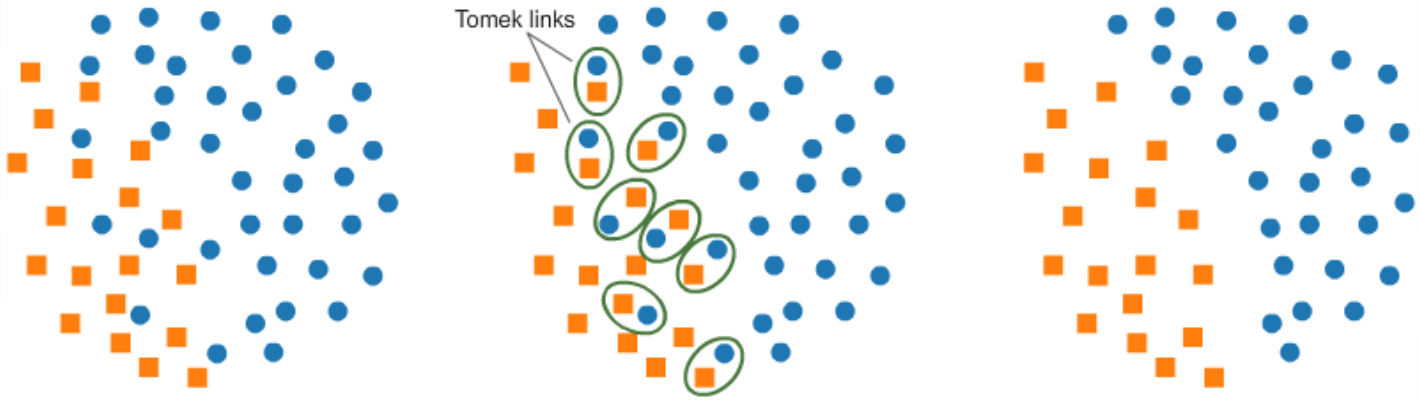
- Registros de la clase mayoritaria en el conjunto de entrenamiento se eliminan al azar hasta que la relación o proporción entre la clase minoritaria y mayoritaria se encuentre en el nivel deseado.



- **Desventaja** : Podrían eliminarse ejemplos potencialmente importantes para el proceso de modelado o aprendizaje.

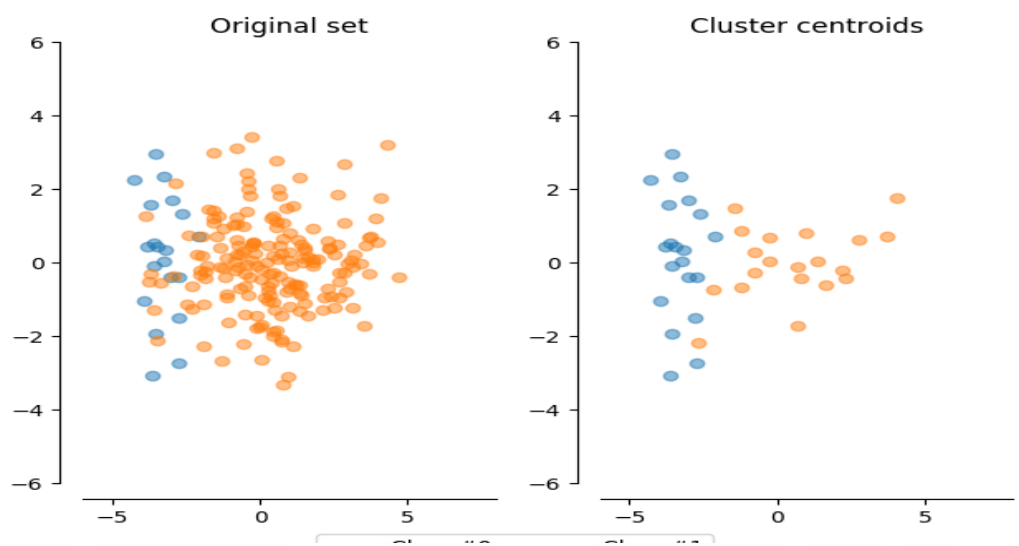
Tomek Link

- Elimina registros de la clase mayoritaria que se encuentren en la frontera de decisión de las clases.



Cluster Centroids

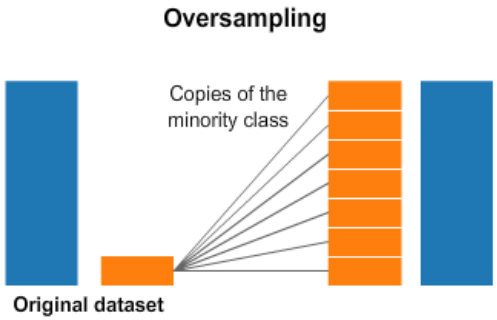
- Respecto a la clase mayoritaria , nos quedamos solamente con los puntos representativos o baricentros de las observaciones.



Estrategias de balanceo

OverSampling

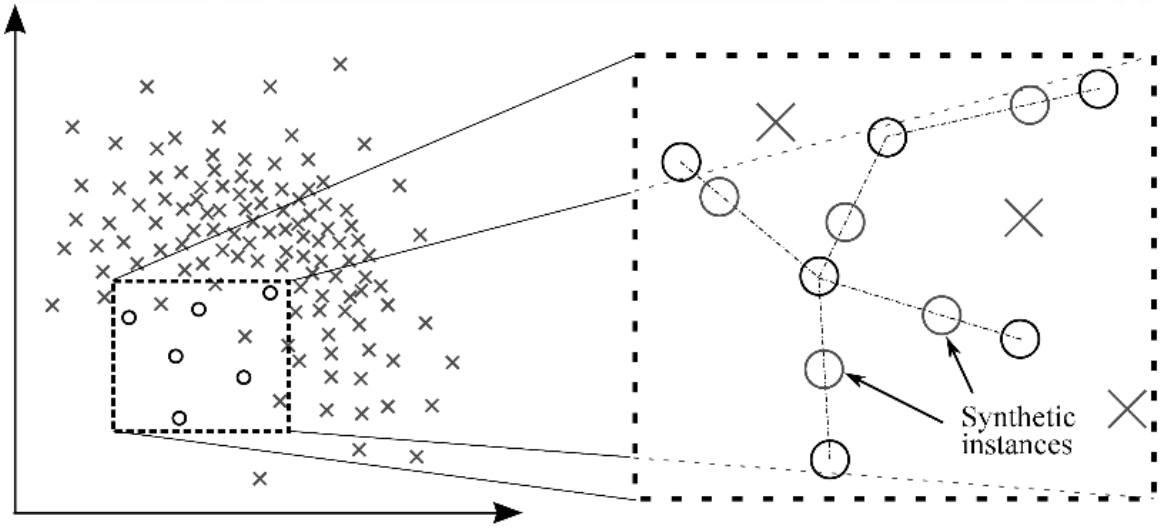
- Registros de la clase minoritaria en el conjunto de entrenamiento se replican o clonan al azar hasta que la relación o proporción entre la clase minoritaria y mayoritaria se encuentre en el nivel deseado.



- **Desventaja :** Podría generarse sobre-estimación o sub-estimación dependiendo de la aleatoriedad y es muy costoso computacionalmente.

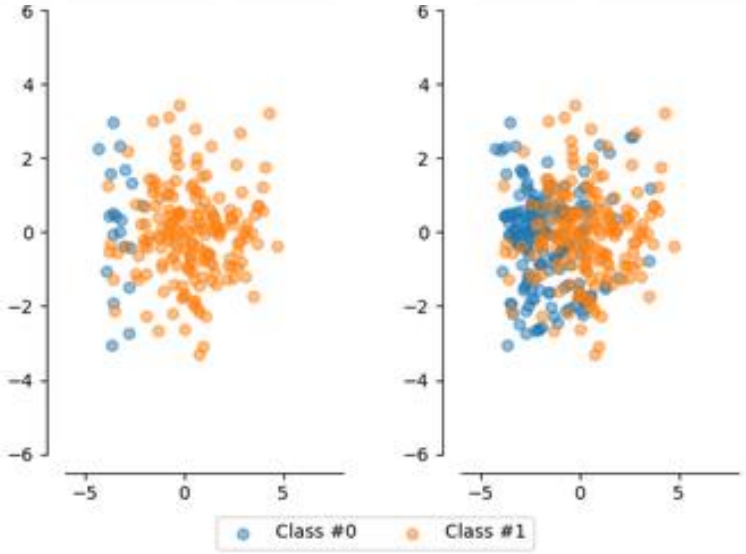
SMOTE (SYNTETIC MINORITY OVERSAMPLING TECHNIQUE)

- Para cada una de las instancias minoritarias se buscan las k instancias vecinas (más cercanas) de la misma clase y se crean N% instancias o elementos entre la línea que une la instancia original y cada una de sus vecinas.



RESAMPLING (UNDERSAMPLING Y OVERSAMPLING)

- Remuestra aleatoriamente del conjunto de datos. Permite disminuir la clase mayoritaria y generar réplicas o clones de la clase minoritaria de forma aleatoria.



CONCLUSIONES

- Es muy importante resolver problemas de desbalance en el procesamiento para clasificar correctamente a la clase minoritaria (De mayor relevancia).

- SMOTE y Resampling son las técnicas con mejor rendimiento, undersampling en algunos casos nos produce resultados irregulares.

- Algunos algoritmos de Machine Learning son menos sensibles que otros a problemas de muestras desbalanceadas, sin embargo siempre es conveniente balancear.

- Algunos algoritmos de Machine Learning incorporan dentro de la propia ejecución una metodología de balanceo de muestras.
- Ejemplo : Balanced Random Forest .

GRACIAS
POR SU ATENCIÓN