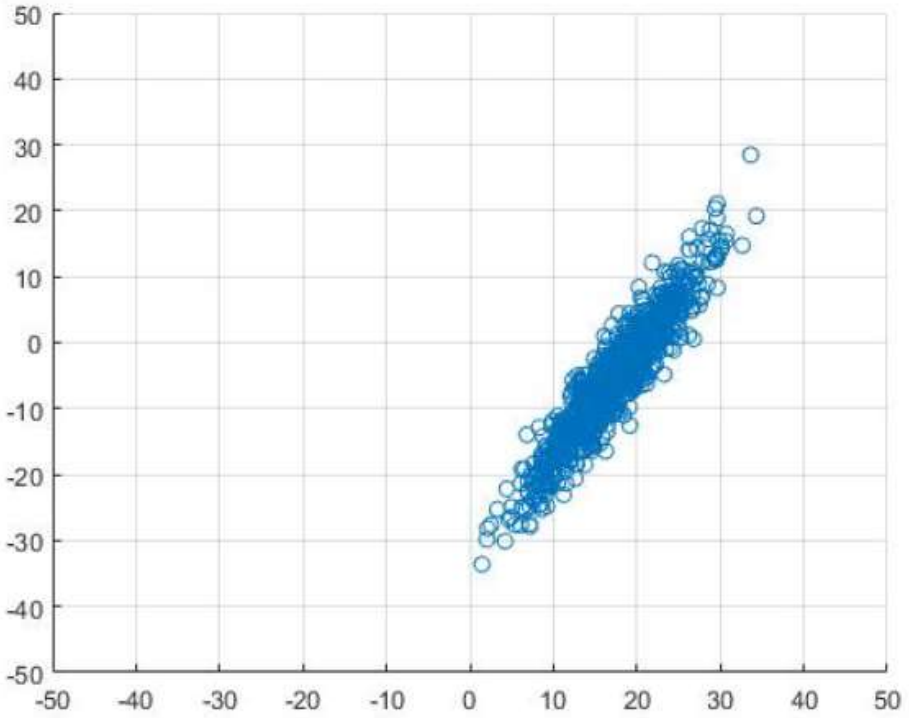
Temas Varios

PCA

Deep Dive

Covarianza +



- $\Sigma = \begin{bmatrix} 26.8 & 42.4 \\ 42.4 & 76.3 \end{bmatrix}$
- $\mu = \begin{bmatrix} 17 \\ -5 \end{bmatrix}$
- Covarianza positiva

- La covarianza nos explica la relación de una variable vs otra variable. Y la medida de “dispersión” de cada una de las variables

Matriz de Covarianza

- La matriz de covarianza tiene propiedades particulares
  - Diagonal
  - Cuadrada
  - Simétrica.
- Podemos descomponerla

- $\Sigma = \begin{bmatrix} var\ (x_1) & cov(x_1, x_2) \\ cov(x_2, x_1) & var\ (x_2) \end{bmatrix}$

Descomposición de la matriz de covarianza

- Descomponer una matriz

$$\Sigma = U \cdot D \cdot U^{-1}$$

- Donde D es una matriz diagonal compuesta de los valores propios y U son los vectores propios en columnas

$$U \cdot D \cdot U^{-1} = \begin{bmatrix} \vdots & \vdots & \vdots \\ \overrightarrow{u_1} & \overrightarrow{u_2} & \overrightarrow{u_3} \\ \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \begin{bmatrix} \vdots & \vdots & \vdots \\ \overrightarrow{u_1} & \overrightarrow{u_2} & \overrightarrow{u_3} \\ \vdots & \vdots & \vdots \end{bmatrix}^{-1}$$

Por propiedad, U es una matriz ortogonal (90°) y unitaria

$$U \cdot U^{-1} = U \cdot U^{T} = 1$$

Transformación Lineal

- https://shad.io/MatVis/

PCA vs Autoencoder

PCA

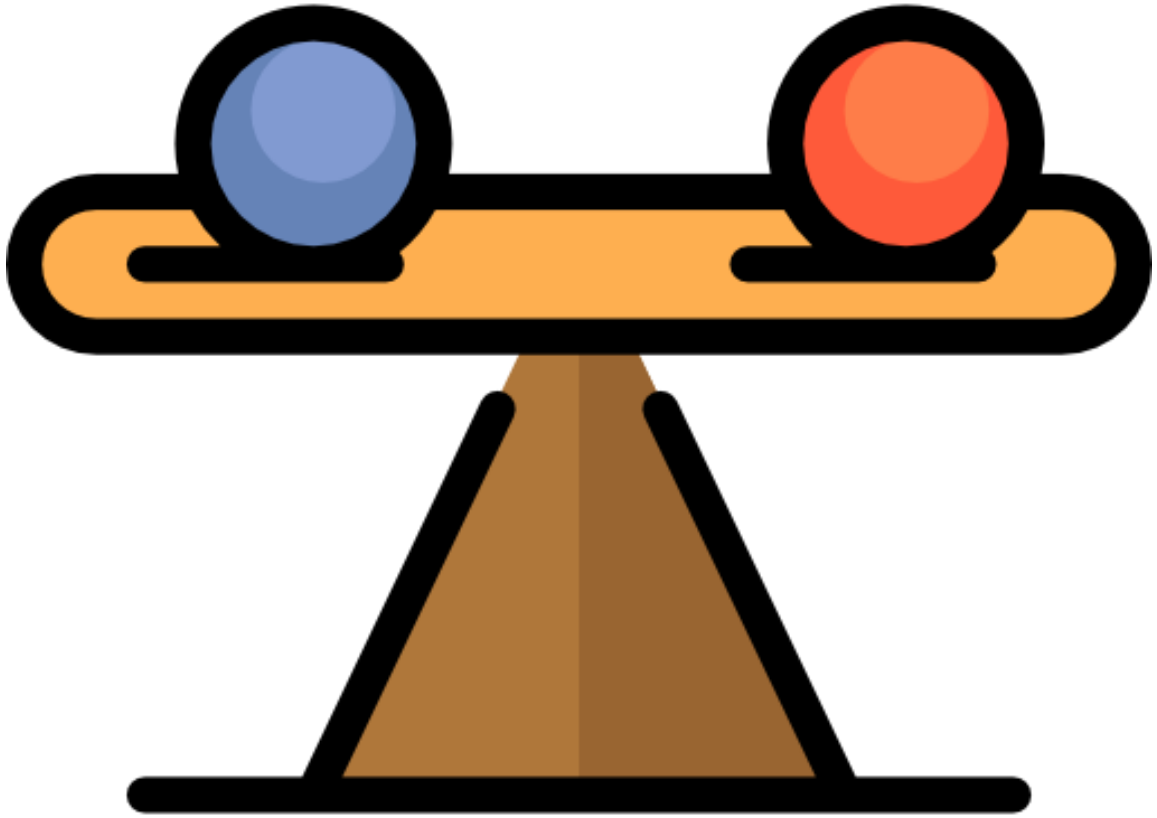
- Componentes ortogonales
- Componentes ordenados de mayor a menor
- Transformaciones lineales

Autoencoder

- Compomentes no necesariamente ortogonales
- Componentes con importancia similar
- Transformaciones lineales y no lineales

Balanceo de Datos

Para Clasificación



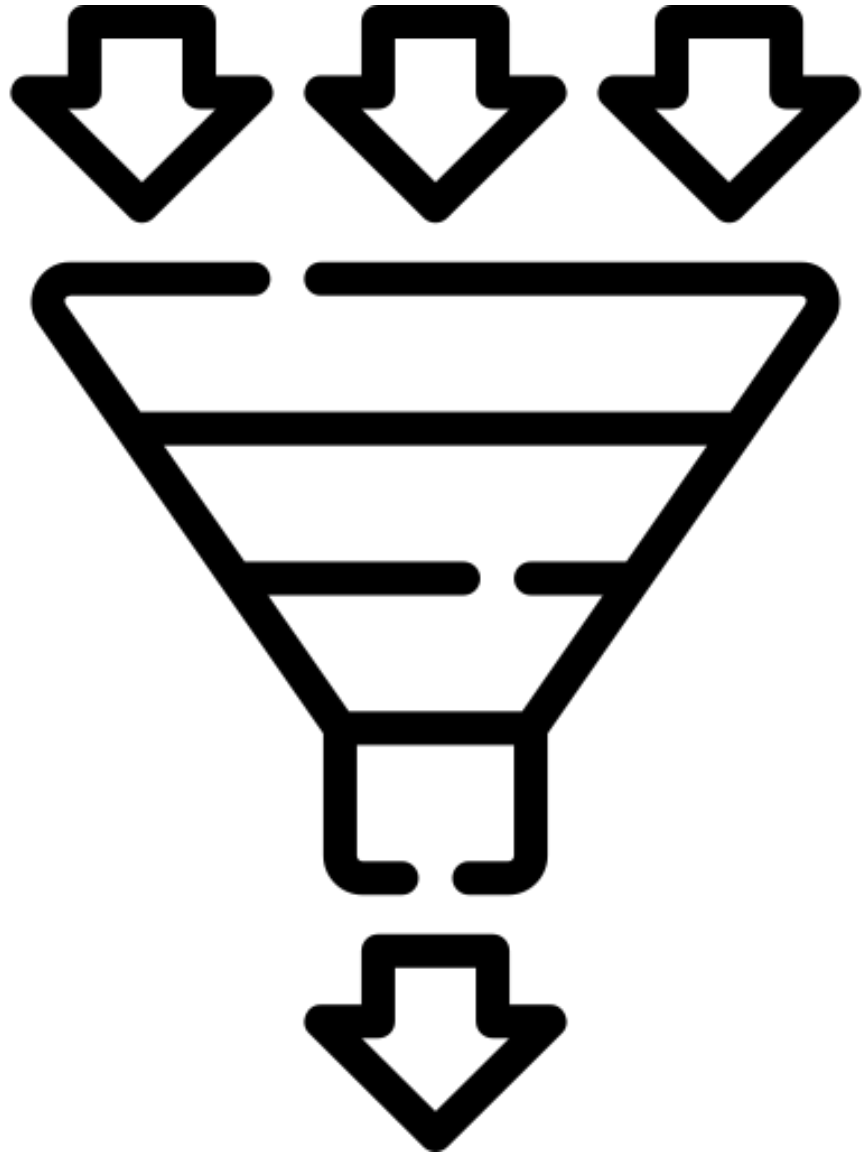
Balanceo

- Por lo general, los modelos de clasificación requieren de clases balanceadas para evitar sesgos.

- El desbalance de clases no solo es común, es incluso esperado.
  - Casos de fraudes en transacciones
  - Enfermedades
  - Situaciones anómalas

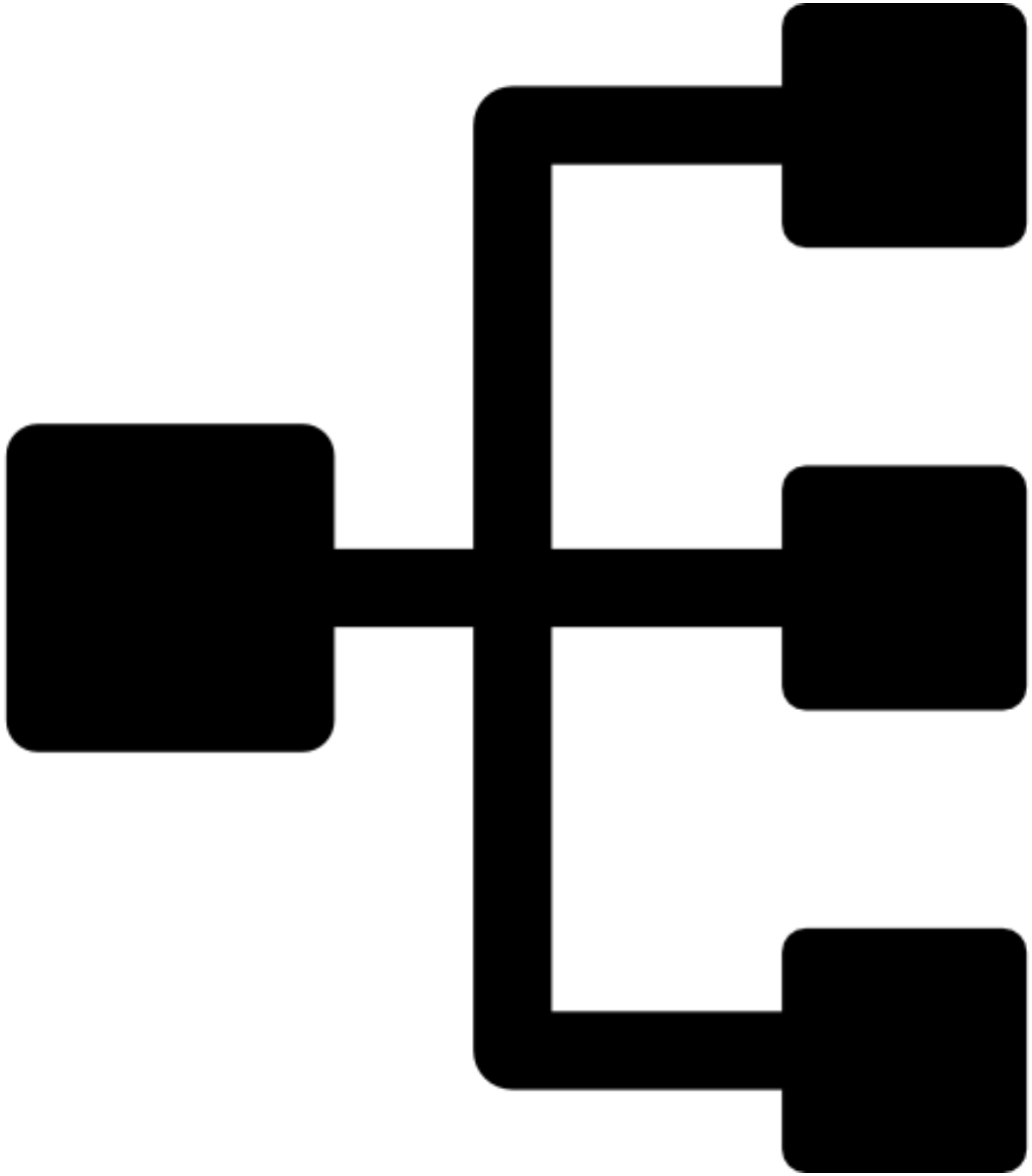
Reducción de data

- Aleatoriamente se reduce el número de ítems en la clase mayoritaria hasta que esté cercana al número de datos en la clase minoritaria
- Random Undersampling



Duplicar data

- Aleatoriamente, se toman muestras de la clase minoritaria y se duplica el item
- Random oversampling



SMOTE (Synthetic Minority Oversampling Technique)

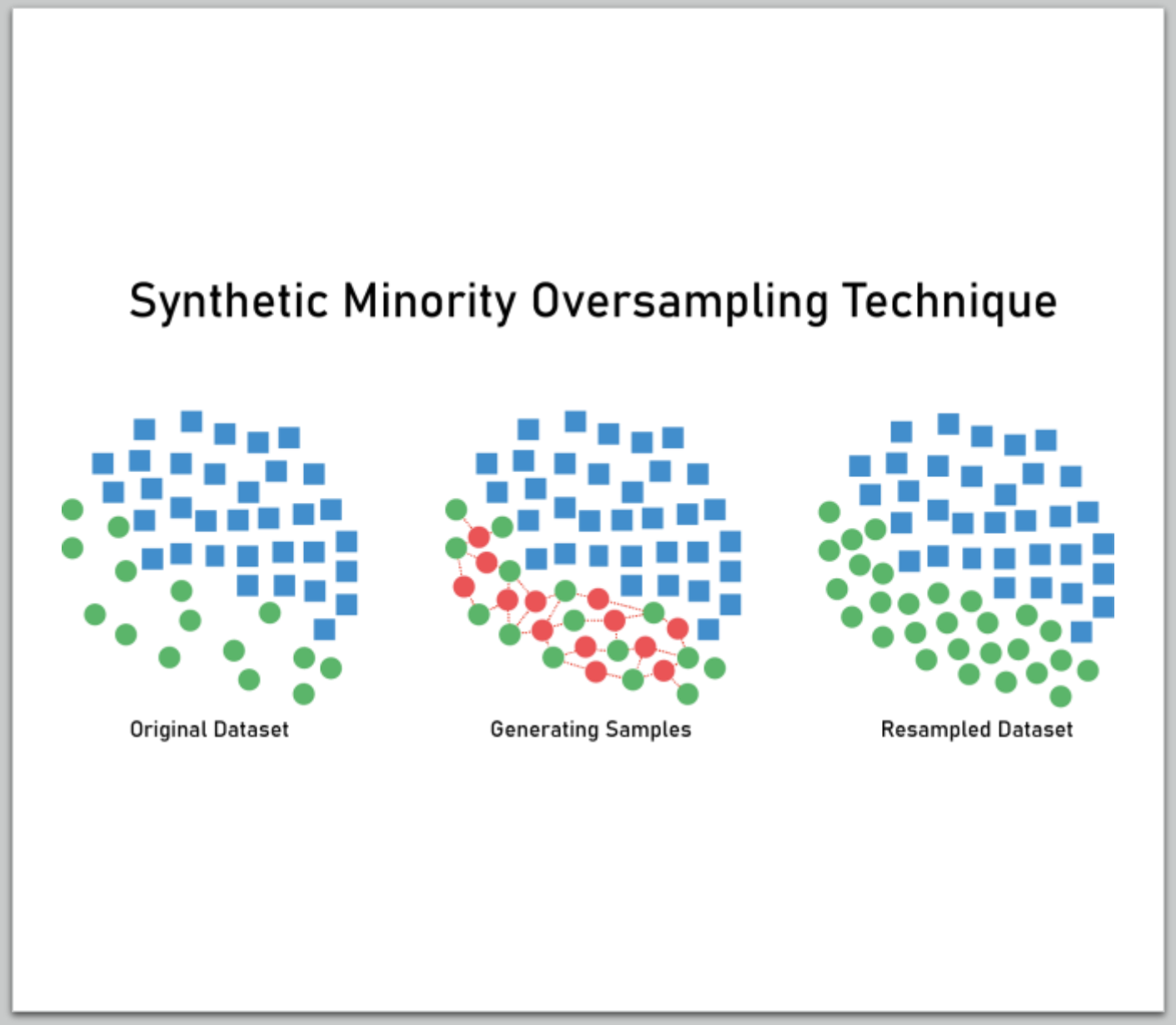
- Desarrollado en el 2002, (https://arxiv.org/abs/1106.1813)
- Genera aleatoriamente data de la clase minoritaria basado en la distancia entre ellas.
- La los puntos generada son ligeramente diferente de la data original

SMOTE

1. Elegir un punto aleatorio de la clase minoritaria
2. Calcular la distancia entre el punto y los k vecinos mas cercanos
3. Multiplicar la diferencia con un numero aleatorio entre 0 y 1 y agregarlo al dataset con la etiqueta minoritaria.
4. Repetir hasta tener una proporción más adecuada.

SMOTE

- Funciona porque la data sintética es generada cercana a la data real, agregando más información que una simple duplicación



Tomek Links

- https://doi.org/10.1109/TSMC.1976.4309452
- 1976 Condensed Nearest Neighbors

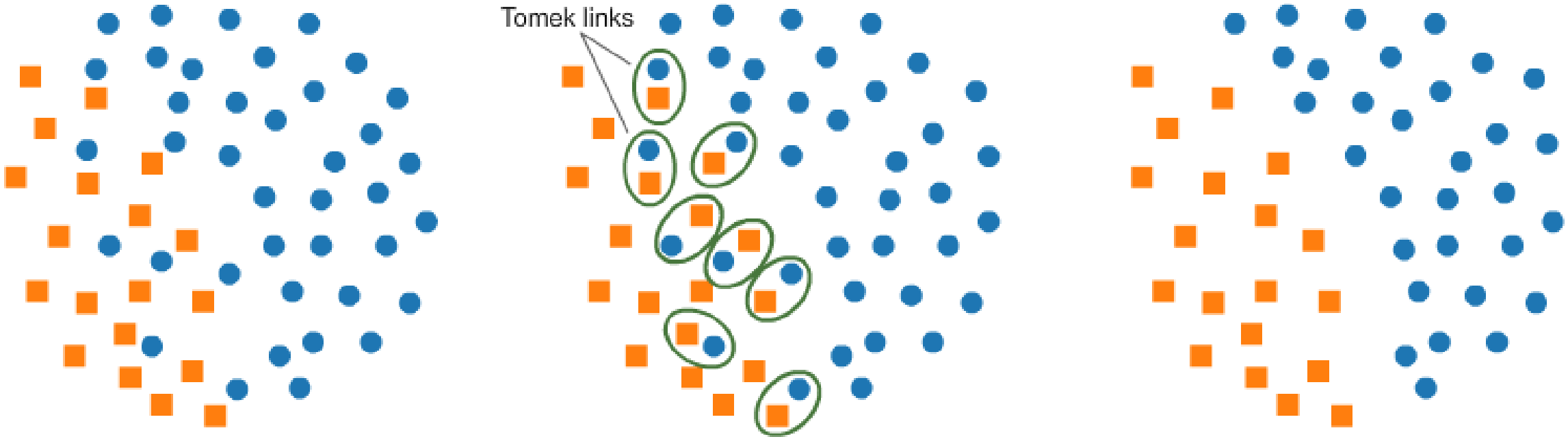
- Elimina puntos en la clase mayoritaria que se encuentren cercanos a la data minoritaria.

Tomek Links

- Seleccionamos dos puntos A y B, con las siguientes condiciones

- El vecino más cercano del punto A, es B
- El vecino más cercano del punto B, es A
- A y B pertenecen a diferentes clases

- El par A, B es el Tomek link.
- Eliminar el punto del Tomek link que pertenece a la case mayoritaria.


Tomek links

SMOTE-Tomek Links

- Combina la habilidad de generar data del SMOTE, y la habilidad de eliminar del Tomek.
- https://doi.org/10.1145/1007730.1007735

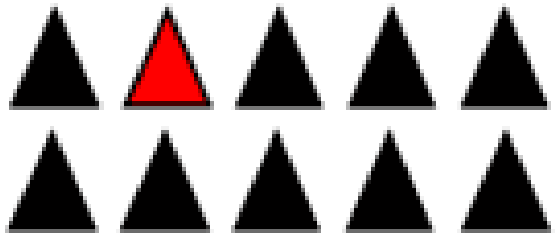
- Ejecuta primero SMOTE y luego Tomek Links

SMOTE-Tomek Links

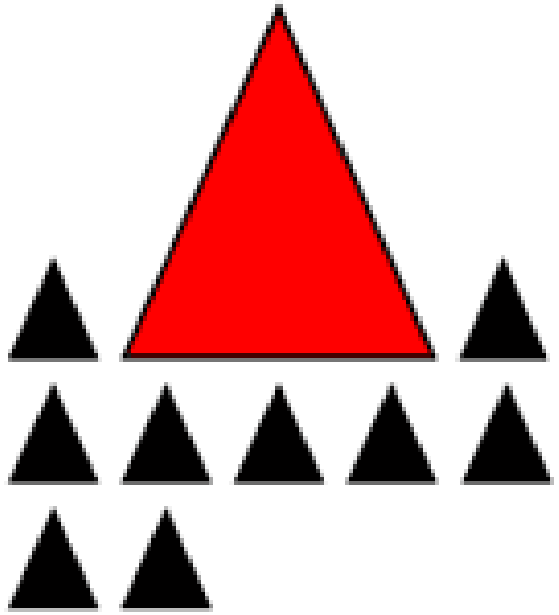
1) (SMOTE) Elegir un punto aleatorio de la clase minoritaria
2) Calcular la distancia entre el punto y los k vecinos más cercanos
3) Multiplicad la diferencia entre [0,1] y agregar la nueva data
4) Repetir hasta tener la proporción adecuada.
5) (Tomek) Elegir un punto aleatorio de la clase mayoritaria
6) Si es cercano a la clase minoritaria, crear el Tomek link y eliminar

Agregar un costo a la clase minoritaria

- Diseñar la función de costo de tal forma que equivocarse en la clase minoritaria tiene un mayor peso.



Data



Cost