

Clasificación con datos desbalanceados

Deep Learning y Series de tiempo



Marco Teran
Universidad Sergio Arboleda

2023

Contenido

1 Introducción

2 ¿Qué son los problemas de clasificación de Clases desequilibradas?

- El desequilibrio de clases

3 Métricas y matriz de confusión

4 Estrategias para el manejo de Datos desbalanceados

- Ajuste de Parámetros del modelo
- Reequilibrar el Dataset
- Muestras sintéticas

Introducción

Introducción

- El desequilibrio entre las clases en los conjuntos de datos es común
- Es poco habitual encontrar conjuntos de datos que estén bien equilibrados

¿Cómo manejo una clase con pocas muestras?

- La respuesta obvia es **recopilar más datos**, pero a veces no es posible conseguir más muestras de las clases minoritarias, como en casos de salud.
- Habría que tener también cuidado con las **métricas**, que pueden proveer información **sesgada**

**¿Qué son los problemas de clasificación
de Clases desequilibradas?**

¿Qué son los problemas de clasificación de Clases desequilibradas?

Problema de clasificación de clases desequilibradas

Se presenta cuando un conjunto de datos de entrenamiento contiene una clase minoritaria con muy pocas muestras en comparación con las demás clases

- Problemas de clasificación como la detección de spam
- Áreas como la salud, donde los conjuntos de datos pueden tener miles de registros de pacientes negativos y pocos casos positivos de la enfermedad que se desea clasificar

¿Qué son los problemas de clasificación de Clases desequilibradas?

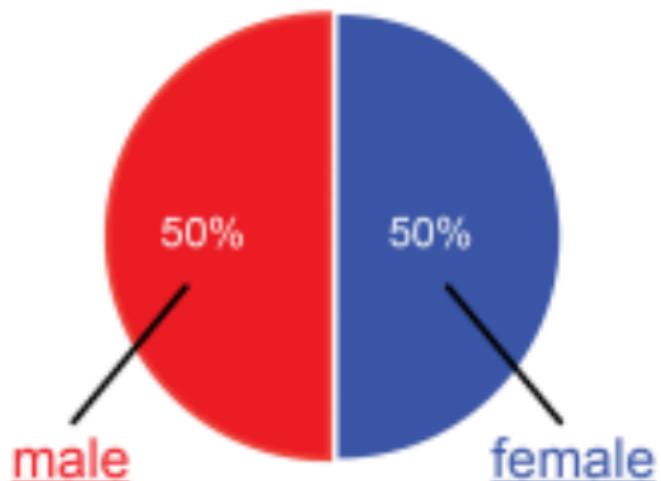
El problema de equilibrio corresponde a la diferencia en el número de muestras en las diferentes clases.

El desequilibrio de clases

El desequilibrio es común

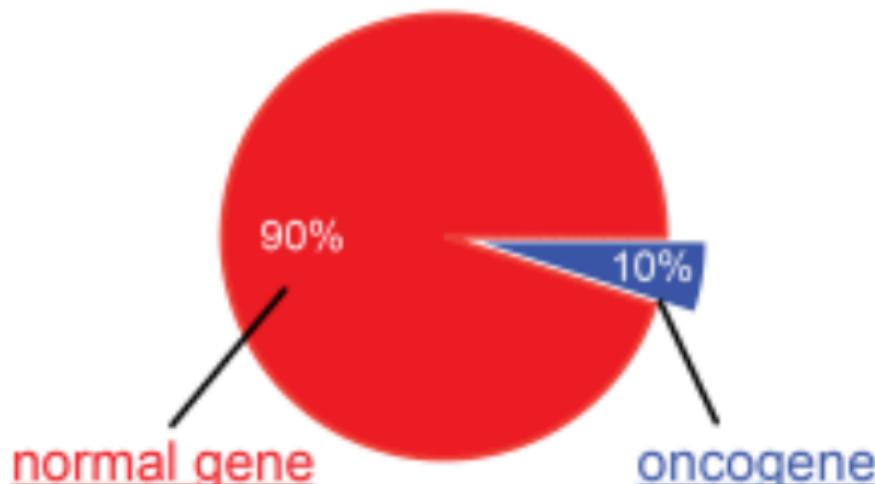
- La mayoría de los conjuntos de datos de clasificación no tienen exactamente el mismo número de instancias en cada clase
- Una pequeña diferencia a menudo no importa.
- Hay problemas donde el desequilibrio de clases no solo es común, sino que se espera

Example of balanced and imbalanced data



Negatives ≈ Positives

Balanced



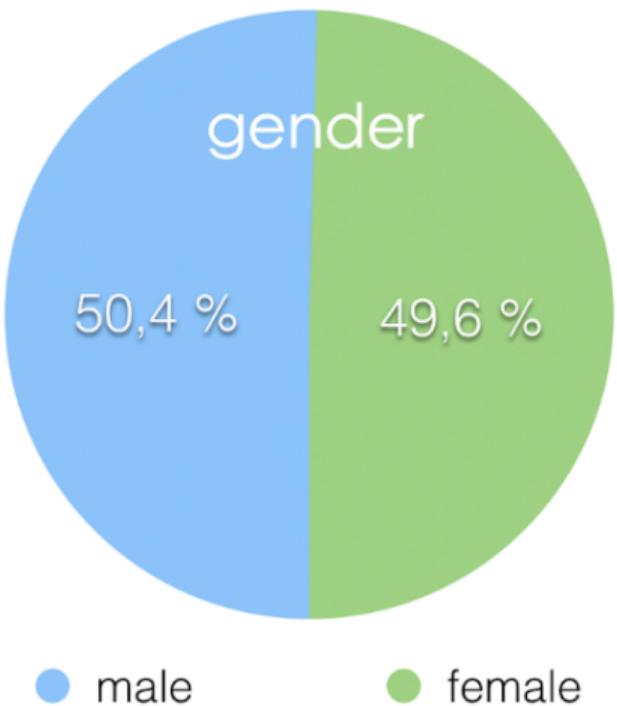
Negatives > Positives

Imbalanced

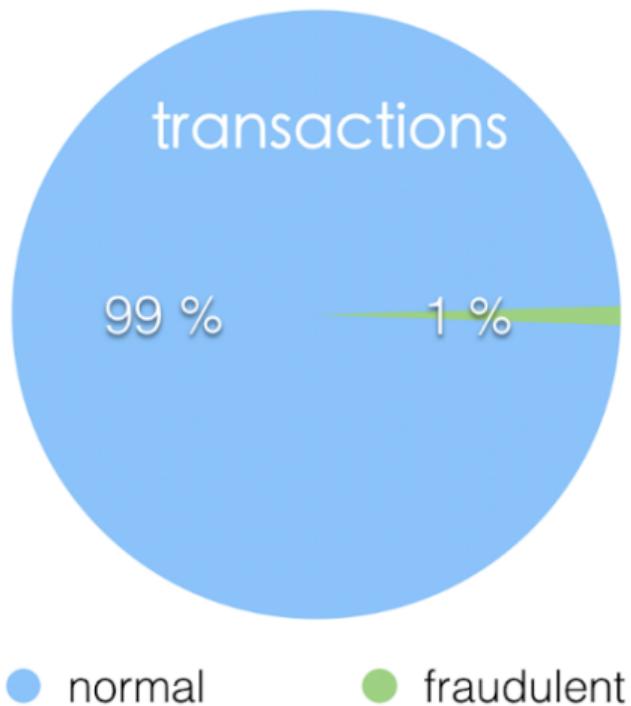
El desequilibrio es común

- **Transacciones fraudulentas:** La gran mayoría de las transacciones estarán en la clase **No-Fraude** y una pequeña minoría estarán en la clase **Fraude**
- Cuando hay un desequilibrio de clases modesto como **4:1** en el ejemplo anterior, puede causar problemas.

Balanced Dataset



Unbalanced Dataset



El desequilibrio de clases

- **Clase Mayoritaria** Más de la mitad de los ejemplos pertenecen a esta clase, a menudo el caso negativo o normal.
- **Clase Minoritaria** Menos de la mitad de los ejemplos pertenecen a esta clase, a menudo el caso positivo o anormal.

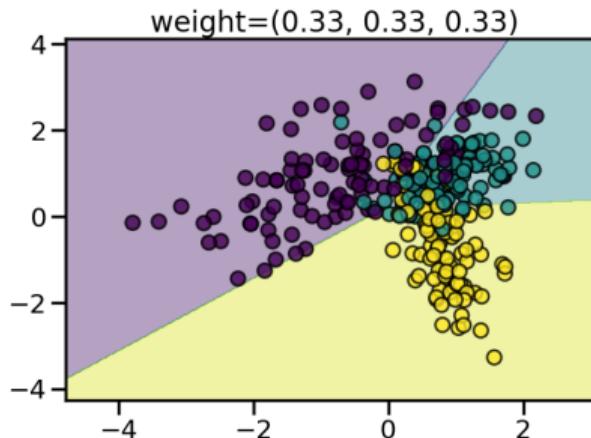
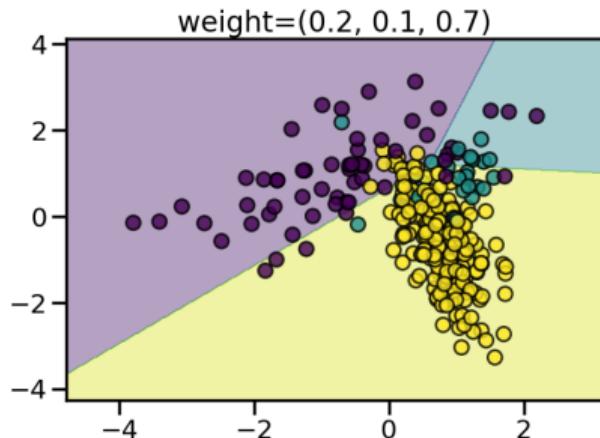
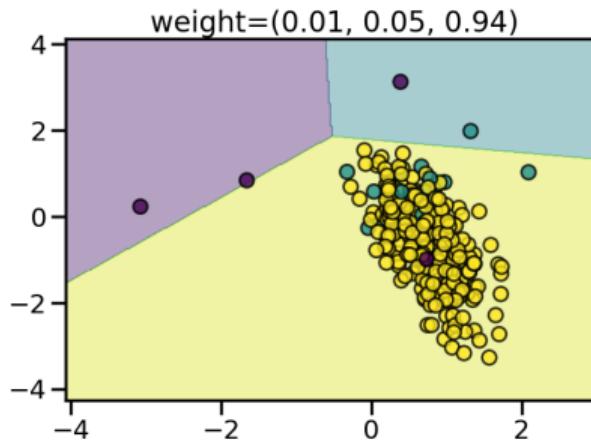
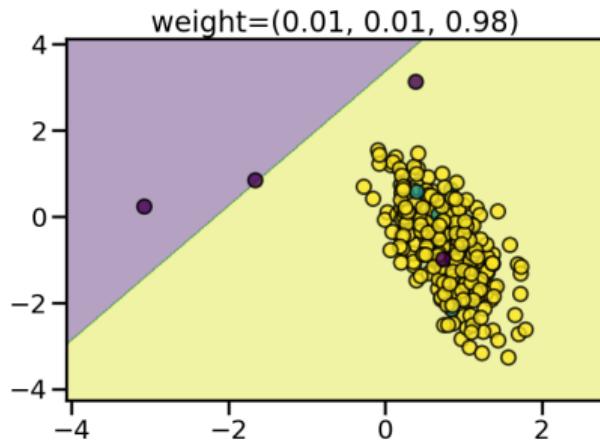
El desequilibrio de clases

- **Desequilibrio Ligero** Donde la distribución de ejemplos es desigual en una pequeña cantidad en el conjunto de datos de entrenamiento (por ejemplo, 4:6)
- **Desequilibrio Grave** Donde la distribución de ejemplos es desigual en una gran cantidad en el conjunto de datos de entrenamiento (por ejemplo, 1:100 o más)

¿Qué son los problemas de clasificación de Clases desequilibradas?

Se ilustra el efecto de entrenar un clasificador SVM lineal con diferentes niveles de equilibrio de clases.

Decision function of LogisticRegression



¿Qué son los problemas de clasificación de Clases desequilibradas?

- La función de decisión de la SVM lineal varía significativamente dependiendo del **grado de desequilibrio en los datos**
- Cuanto mayor es el desequilibrio, más favorece la función de decisión a la **clase mayoritaria**

Métricas y matriz de confusión

Métricas y matriz de confusión

Matriz de confusión

Desglose de predicciones en una tabla que muestra predicciones correctas (la diagonal) y los tipos de predicciones incorrectas realizadas (a qué clases se asignaron predicciones incorrectas).

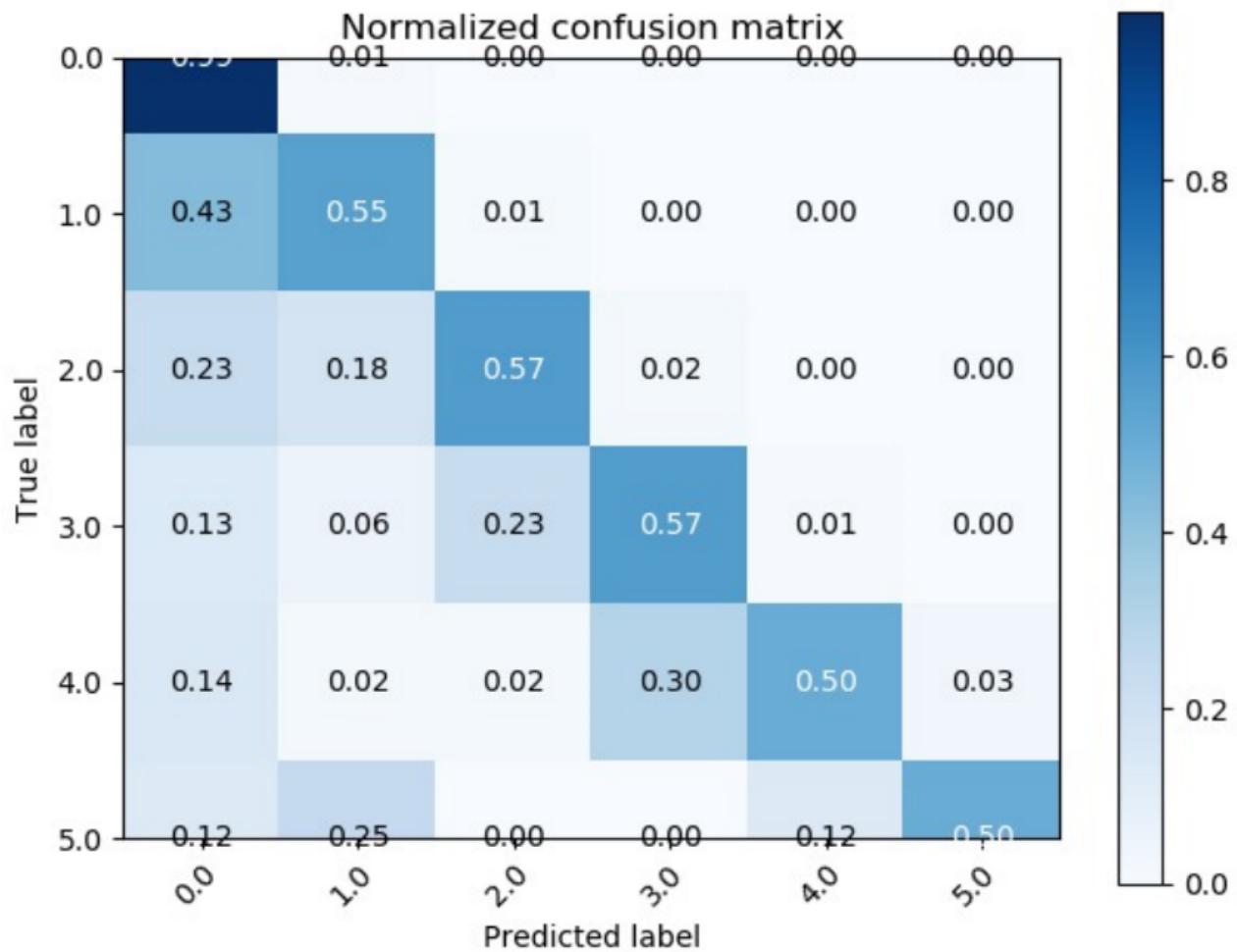
| | Predicción Clase 1 | Predicción Clase 2 |
|-----------------------|--------------------------------------|--------------------------------------|
| Valor real Clase 1 | Aciertos True Positive Clase 1 | Fallos False Positive Clase 2 |
| Valor real Clase 2 | Fallos False Positive Clase 1 | Aciertos True Positive Clase 2 |

Métricas y matriz de confusión

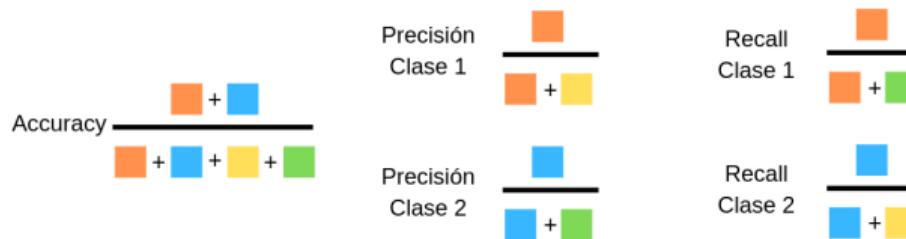
- Evaluación de la efectividad de un modelo que se entrena con datos **desbalanceados**, no se debe medir solo la cantidad de aciertos que se tienen en la clase mayoritaria, ya que esto puede dar una falsa sensación de que el modelo funciona correctamente.
- Se deben utilizar herramientas como la **matriz de confusión**, las métricas de precisión y **recall** para comprender mejor las salidas del modelo.
- De la **matriz de confusión** salen nuevas métricas: precisión y recall

Métricas y matriz de confusión

- **Precisión** Una medida de la exactitud de un clasificador
- **Accuracy** Número total de predicciones correctas dividido por el número total de predicciones realizadas
- **Precisión de una clase** Fiabilidad del modelo para identificar correctamente si un punto pertenece a dicha clase
- **Recall** Una medida de la exhaustividad de un clasificador. Capacidad del modelo para detectar correctamente dicha clase
- **F1 Score** Promedio ponderado de la Precisión y el Recall
$$(2 \times \text{Precisión} \times \text{Recall}) / (\text{Precisión} + \text{Recall})$$
, y combina ambas métricas en una sola.

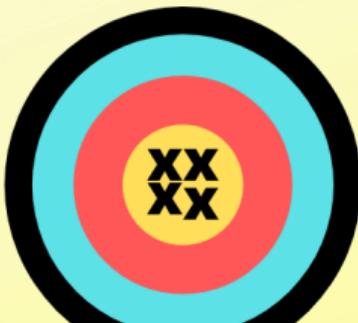


| | Predicción Clase 1 | Predicción Clase 2 |
|-----------------------|--------------------------------------|--------------------------------------|
| Valor real Clase 1 | Aciertos True Positive Clase 1 | Fallos False Positive Clase 2 |
| Valor real Clase 2 | Fallos False Positive Clase 1 | Aciertos True Positive Clase 2 |

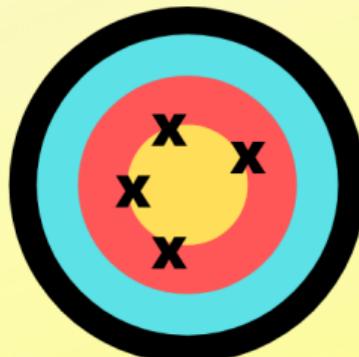


Accuracy and Precision

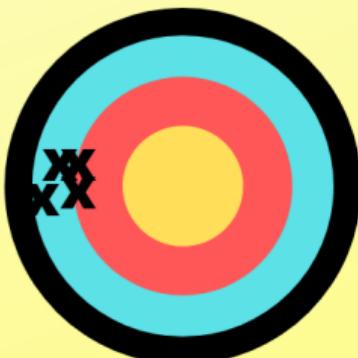
Accurate
Precise



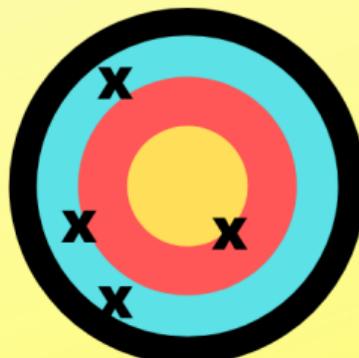
Accurate
Not Precise

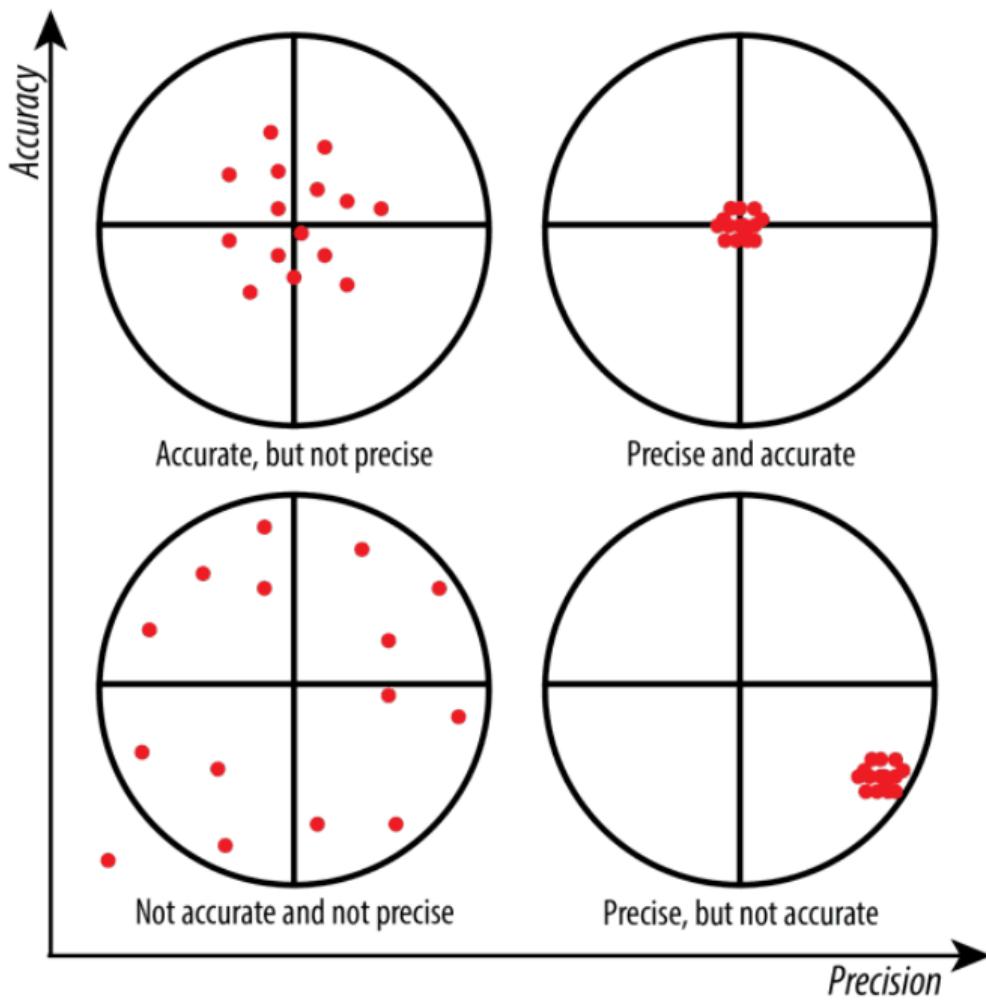


Not Accurate
Precise



Not Accurate
Not Precise





| | | true class | | total |
|-----------------|-------------|-------------------------|-------------------------|---------------|
| | | EFR | LFR | |
| predicted class | EFR | True Positives (TP) | False Positives (FP) | predicted EFR |
| | LFR | False Negatives (FN) | True Negatives (TN) | predicted LFR |
| | true EFR | | true LFR | |

$$PR = \frac{TP}{TP+FP}$$

$$RE = \frac{TP}{TP+FN}$$

$$CA = \frac{TP+TN}{TP+TN+FP+FN}$$

$$F_1 = \frac{2TP}{2TP+FP+FN}$$

Ejemplo

| | Predicción Gato | Predicción Perro |
|---------------------|--------------------|---------------------|
| Valor real Gato | Aciertos 990 | 0 |
| Valor real Perro | Fallos 10 | 0 |

| | Predicción Gato | Predicción Perro |
|---------------------|--------------------|---------------------|
| Valor real Gato | Aciertos 990 | 0 |
| Valor real Perro | Fallos 10 | 0 |

Accuracy $\frac{990 + 0}{990 + 0 + 10 + 0}$

| | | | |
|----------------------|------------------------|-------------------|-----------------------|
| Precisión Clase 1 | $\frac{990}{990 + 10}$ | Recall Clase 1 | $\frac{990}{990 + 0}$ |
| Precisión Clase 2 | $\frac{0}{0 + 0}$ | Recall Clase 2 | $\frac{0}{0 + 10}$ |

Estrategias para el manejo de Datos desbalanceados

Estrategias para el manejo de Datos desbalanceados

En el ámbito de la tecnología de la inteligencia artificial, Machine Learning y Deep Learning, existen diversas estrategias para manejar datos desbalanceados:

- **Ajuste de Parámetros del modelo**
- **Modificar el Dataset**
- **Muestras artificiales**
- **Balanced Ensemble Methods**

Ajuste de Parámetros del modelo

Ajuste de Parámetros del modelo

Ajustar los parámetros o métricas del propio algoritmo:

- Intentar equilibrar la clase minoritaria y penalizar la clase mayoritaria durante el entrenamiento
 - Por ejemplo, en **árboles**, se puede ajustar el peso
 - Mientras que en **logistic regression**, se puede utilizar el parámetro **class_weight="balanced"**
 - En redes neuronales, por ejemplo, se puede ajustar la métrica de Loss para que penalice a las clases mayoritarias.

Reequilibrar el Dataset

Reequilibrar el Dataset

Puedes cambiar el conjunto de datos que usas para construir tu modelo predictivo para tener datos más equilibrados

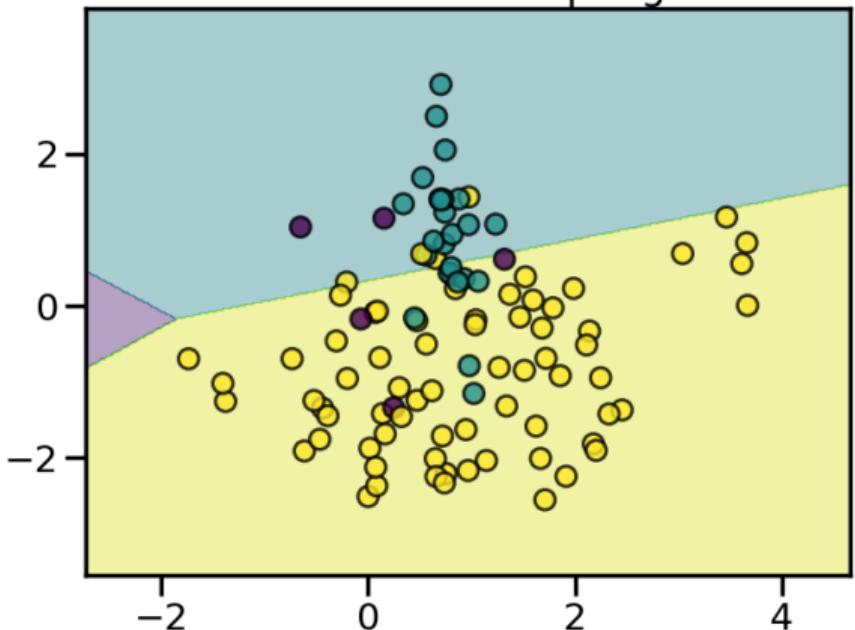
- Se puede eliminar instancias de la clase sobre-representada, llamado **sub-muestreo**
- Eliminando muestras de la clase mayoritaria para reducir su cantidad e intentar equilibrar la situación
- **Consecuencia:** Peligroso, ya que se podrían **prescindir** de *muestras importantes* que brinden información y empeorar el modelo.
- Para seleccionar qué muestras eliminar se debería seguir algún criterio

Over-sampling (random over-sampling)

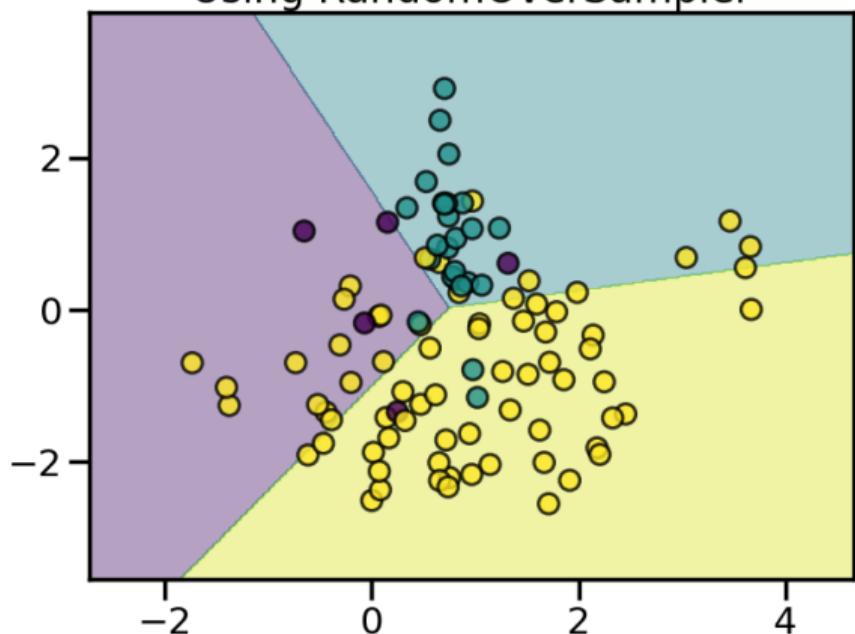
- Generar nuevas muestras seleccionando aleatoriamente con reemplazo entre las muestras disponibles.

Decision function of LogisticRegression

Without resampling



Using RandomOverSampler



Over-sampling

- Como resultado, la clase mayoritaria no domina las demás clases durante el proceso de entrenamiento. En consecuencia, todas las clases están representadas por la función de decisión.

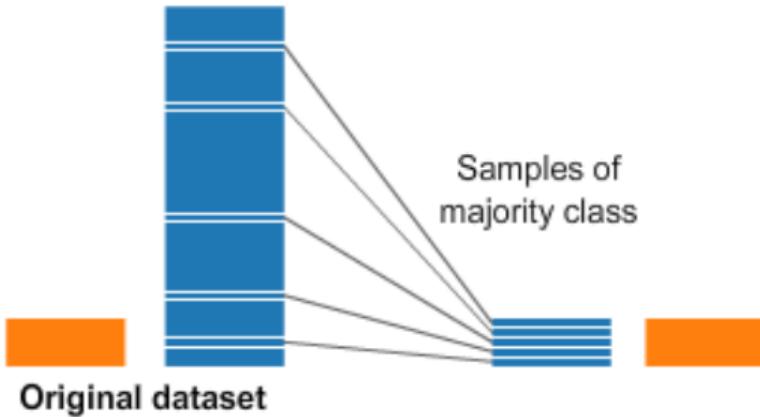
Reequilibrar el Dataset

- Agregar nuevas filas con los mismos valores de las clases minoritarias (redundancia)
- Se puede agregar copias de instancias de la clase subrepresentada llamada **sobre-muestreo** (o muestreo con reemplazo más formalmente)
- Estos enfoques suelen ser muy fáciles de implementar y rápidos de ejecutar. Son un excelente punto de partida

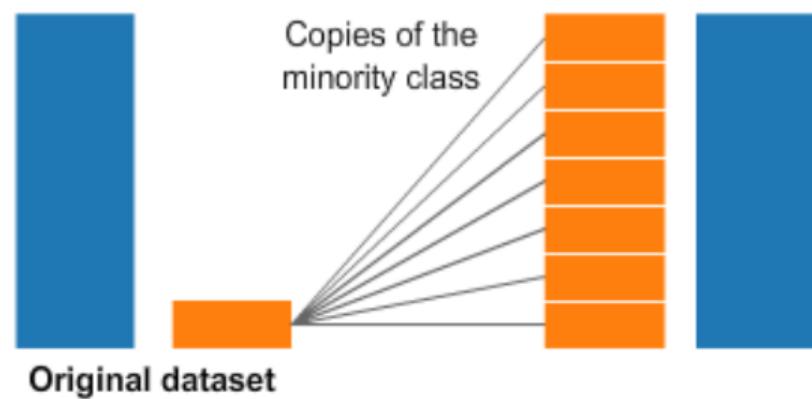
Reequilibrar el Dataset

Estos enfoques suelen ser muy fáciles de implementar y rápidos de ejecutar. Son un excelente punto de partida

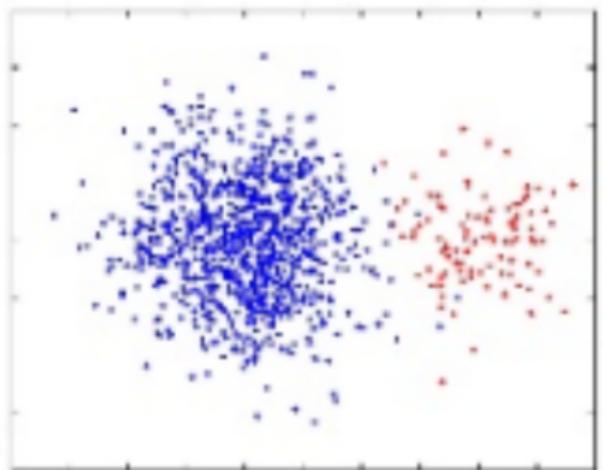
Undersampling



Oversampling

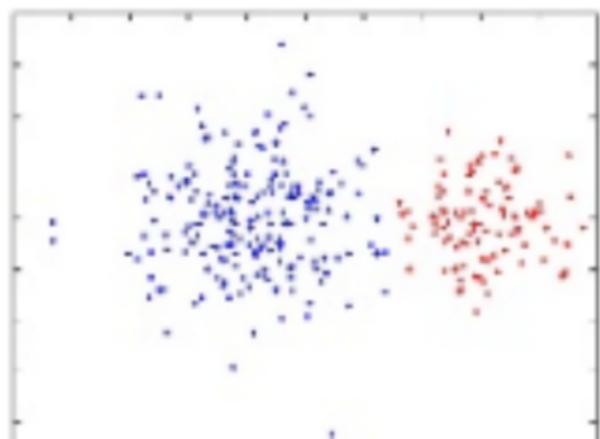


Sampling: Rebalancing the dataset

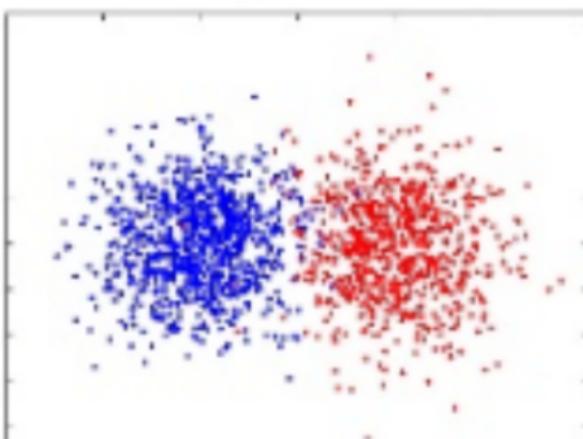


Imbalanced Data

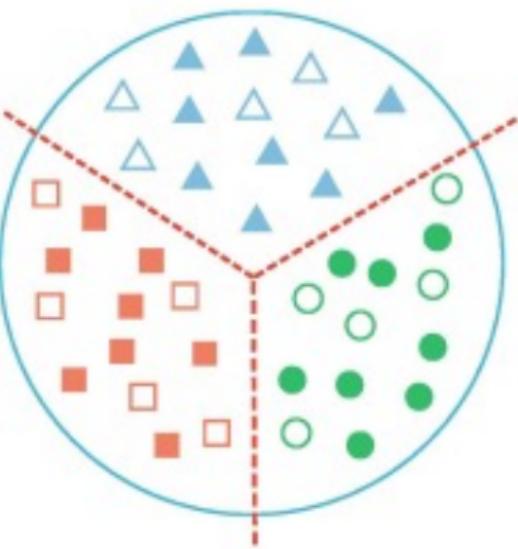
Under-sampling



Over-sampling



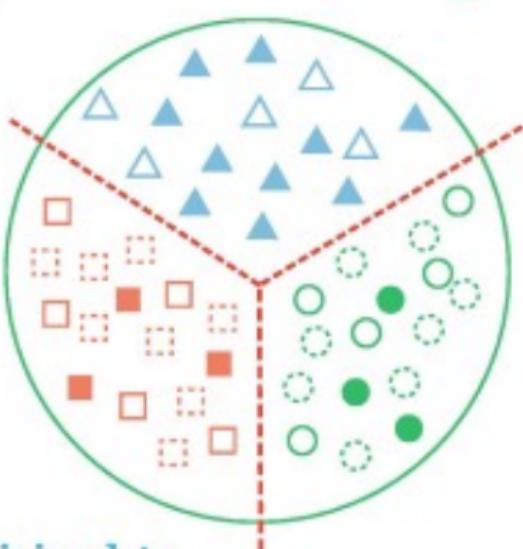
Balanced Data



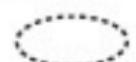
Imbalanced Data



Synthetic Over-Sampling



Classifier



Misclassified

△ ■ ● Training data

△ □ ○ Testing data

□ ○ Synthetic data

Reequilibrar el Dataset

Algunas reglas generales:

- Considera probar sub-muestreo cuando tienes muchos datos (decenas o cientos de miles de instancias o más).
- Considera probar sobre-muestreo cuando no tienes muchos datos (decenas de miles de registros o menos).
- Considera probar esquemas de muestreo aleatorio y no aleatorio (por ejemplo, **estratificado**).
- Considera probar diferentes radios de muestreo (por ejemplo, no tienes que apuntar a una relación 1:1 en un problema de clasificación binaria, prueba otros radios).

Muestras sintéticas

Del random over-sampling al SMOTE y ADASYN

Aparte del muestreo aleatorio con reemplazo, existen dos métodos populares para sobremuestrear clases minoritarias:

- Técnica de Sobremuestreo Minoritario Sintético (SMOTE)
- El método de muestreo sintético adaptativo (ADASYN)

Muestras artificiales

- Esta alternativa es la creación de muestras sintéticas utilizando diversos algoritmos que intentan seguir la tendencia del grupo minoritario
- Según el método, se podrían mejorar los resultados
- Existe un riesgo en la creación de muestras sintéticas, ya que se podría alterar la distribución "natural" de esa clase y confundir al modelo en su clasificación.

Muestras sintéticas

- Una forma simple de generar muestras sintéticas es muestrear aleatoriamente los atributos de las instancias en la clase minoritaria
- Puedes muestrearlos empíricamente dentro de tu conjunto de datos o puedes utilizar un método como el de **Naive Bayes** que puede muestrear cada atributo de forma independiente cuando se ejecuta en reversa
- Tendrás más y diferentes datos, pero las relaciones no lineales entre los atributos pueden no ser preservadas
- Existen algoritmos sistemáticos que puedes utilizar para generar muestras sintéticas

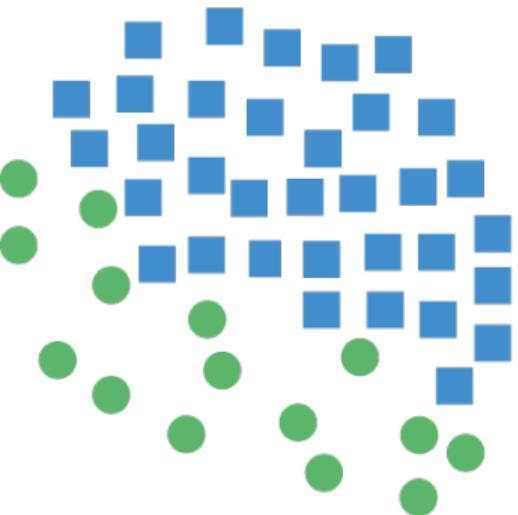
SMOTE: Definición

- SMOTE (**S**ynthetic **M**inority **O**versampling **T**echnique) es una técnica de sobremuestreo que se utiliza para equilibrar conjuntos de datos desbalanceados en clasificación binaria.
- SMOTE crea nuevas instancias sintéticas de la clase minoritaria, tomando como base las instancias existentes y combinándolas con vecinos sintéticos generados.
- SMOTE es una técnica determinista y depende de dos parámetros: el número de vecinos cercanos a utilizar para generar instancias sintéticas y el factor de sobremuestreo que indica la cantidad de nuevas instancias a generar.
- El objetivo de SMOTE es aumentar la cantidad de datos de la clase minoritaria sin duplicar instancias existentes, lo que puede mejorar el rendimiento del modelo de clasificación en la predicción de la clase minoritaria.

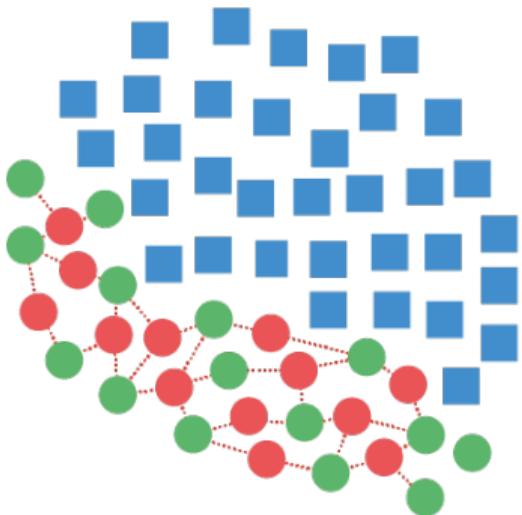
SMOTE (Técnica de Sobremuestreo Sintético para Clases Minoritarias)

- Funciona creando muestras sintéticas de la clase minoritaria en lugar de crear copias
- El algoritmo selecciona dos o más instancias similares (utilizando una medida de distancia) y perturba una instancia un atributo a la vez por una cantidad aleatoria dentro de la diferencia con las instancias vecinas
- SMOTE no es adecuado para conjuntos de datos con clases extremadamente desequilibradas o en presencia de ruido en los datos.

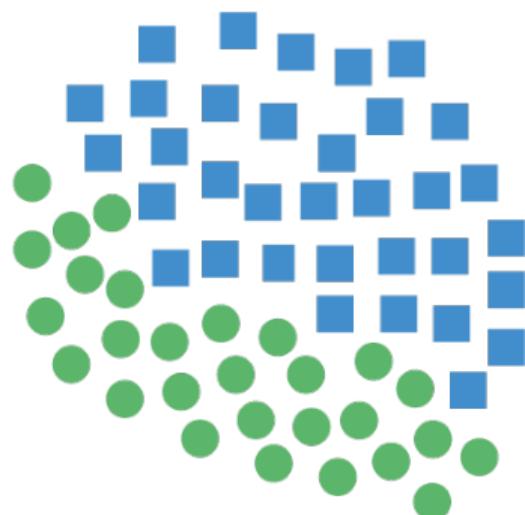
Synthetic Minority Oversampling Technique



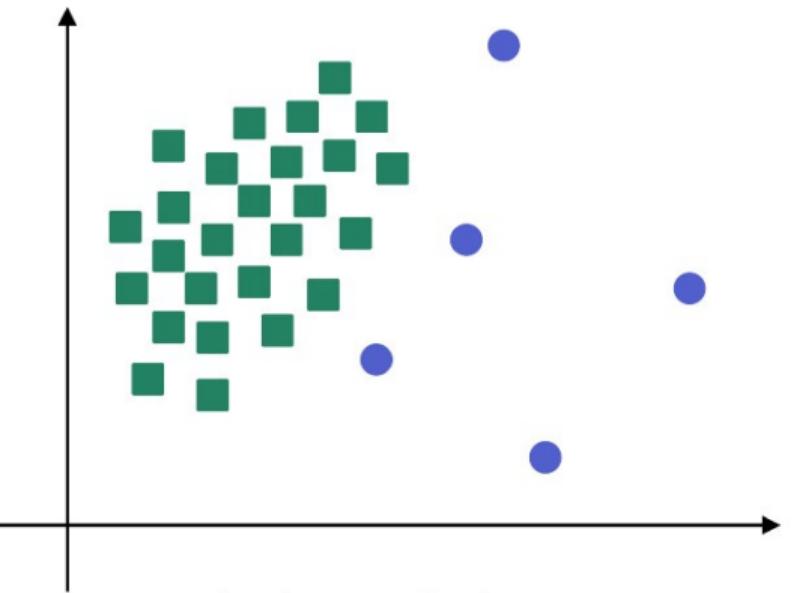
Original Dataset



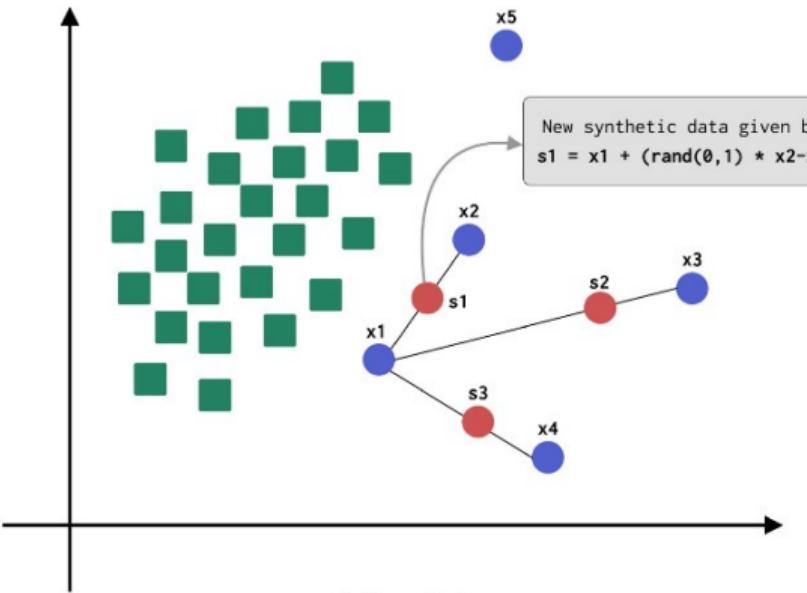
Generating Samples



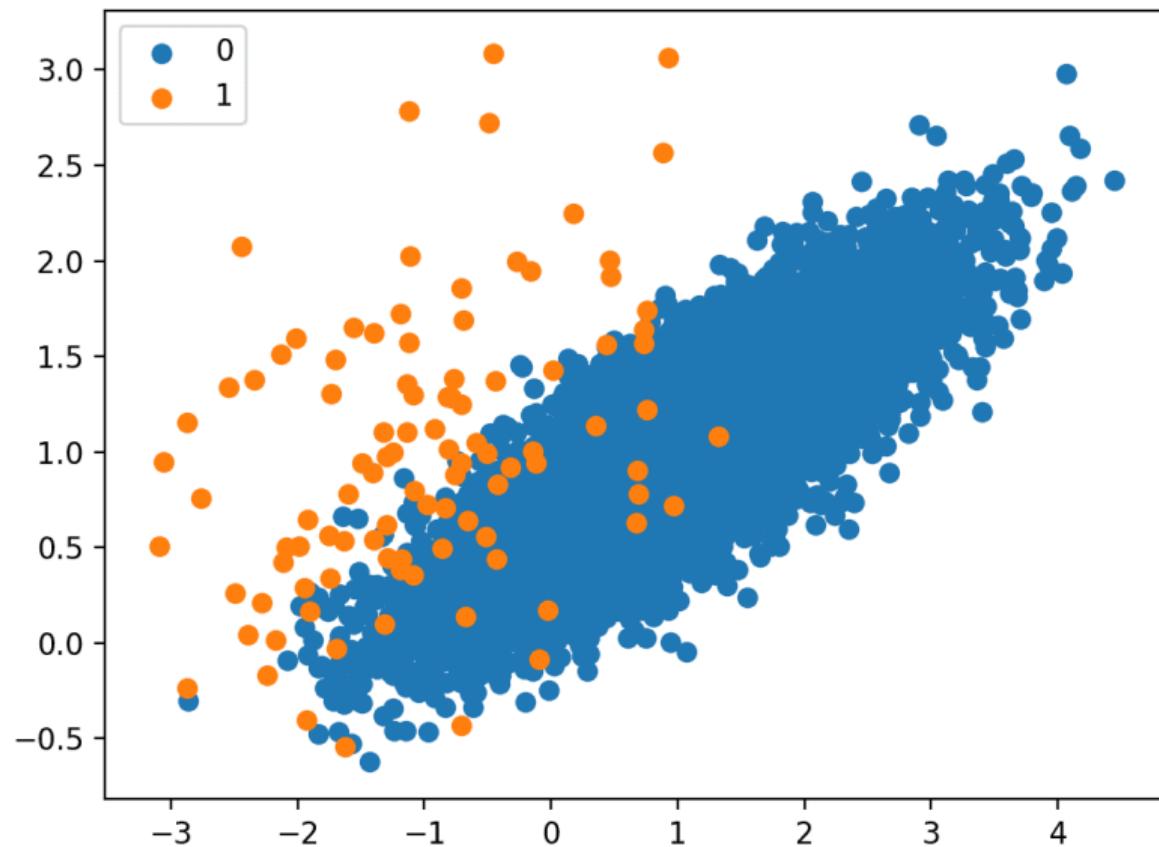
Resampled Dataset

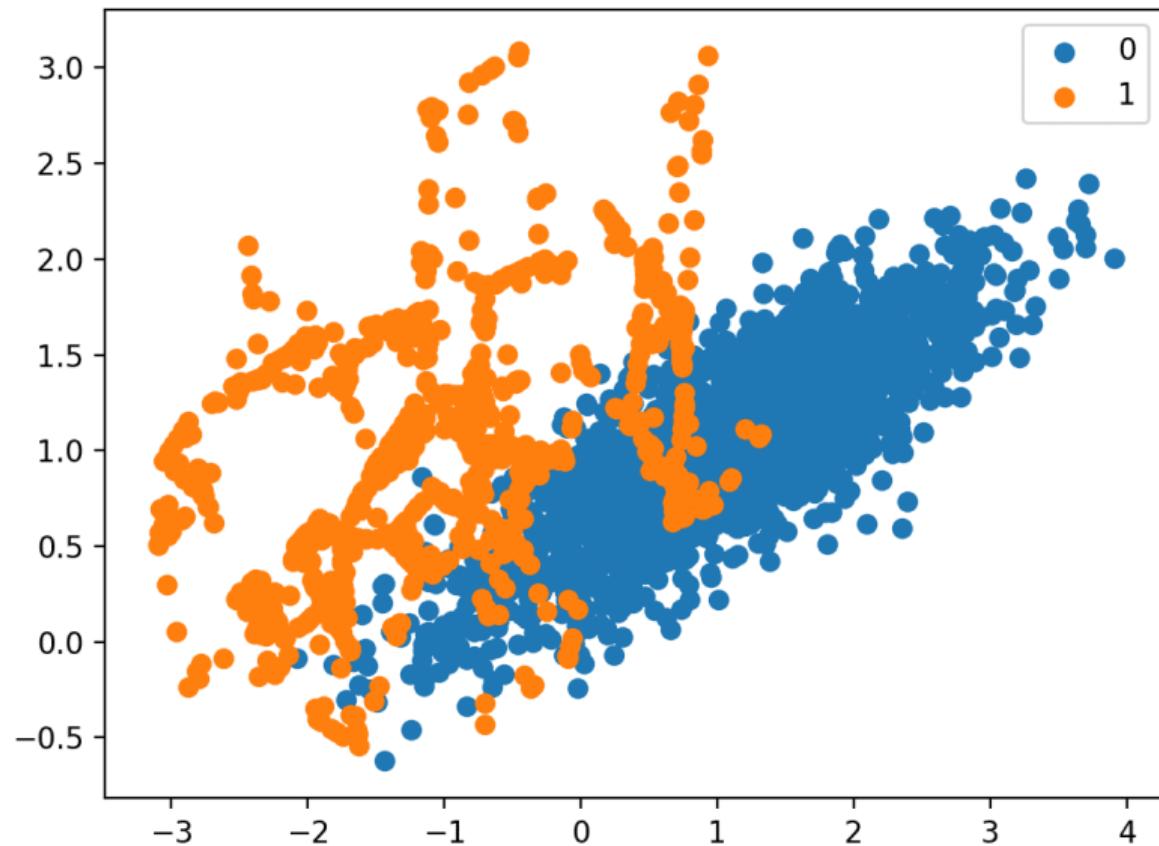


a) Class Imbalance

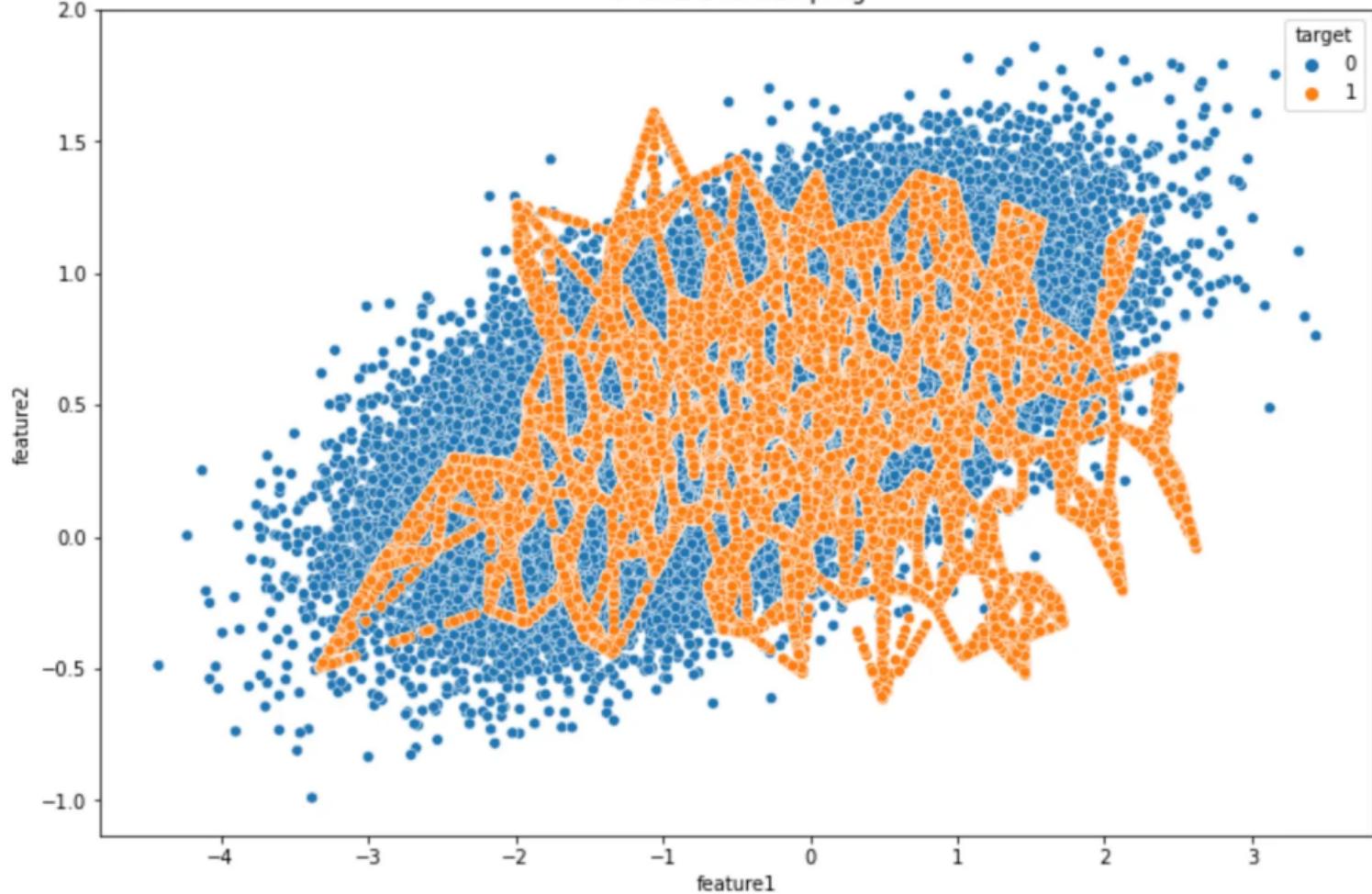


b) SMOTE



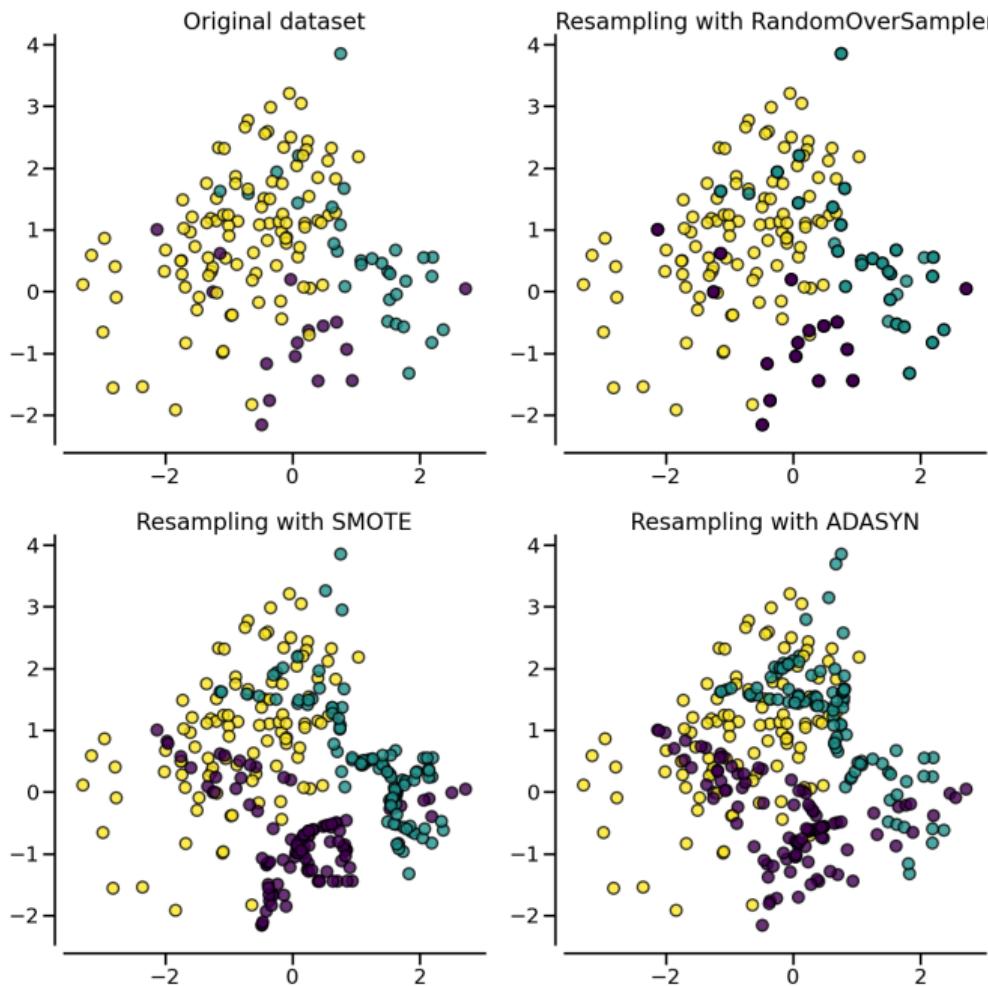


SMOTE Over Sampling



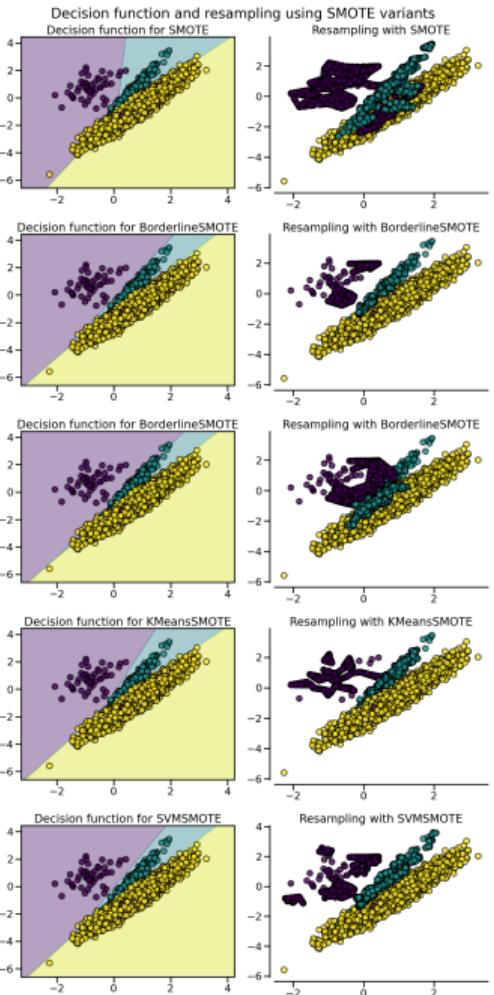
Del random over-sampling al SMOTE y ADASYN

La figura a continuación ilustra la diferencia principal de los diferentes métodos de sobremuestreo.



Variantes de SMOTE

- SMOTE podría conectar valores atípicos y valores típicos
- ADASYN podría enfocarse únicamente en valores atípicos, lo que en ambos casos podría llevar a una función de decisión subóptima.



Balanced Ensemble Methods

- Esta estrategia que aprovecha las ventajas de hacer ensamble de métodos, es decir, entrenar diversos modelos y entre todos obtener el resultado final (por ejemplo, "votando"), pero asegurándose de tomar muestras de entrenamiento equilibradas.

¡Muchas gracias por su atención!

¿Preguntas?



Contacto: Marco Teran
webpage: marcoteran.github.io/
e-mail: marco.teran@usa.edu.co

