MINERIA DE DATOS

TRABAJO FINAL: COMPANY BANKRUPTCY PREDICTION

El dataset elegido almacena los datos macroeconómicos de distintas compañías y nos indica si estas fueron a bancarrota o no. La idea detrás de nuestro trabajo consiste tratar los datos de forma que sean lo suficientemente útiles como para que una vez entrenado un modelo sobre ellos sea capaz de anticipar de forma consistente cuándo una empresa vaya a ir a bancarrota, lo que permitiría a la dirección hacer los cambios pertinentes para evitarlo. Consecuentemente, el recall será la principal métrica de interés.

PROBLEMAS DEL DATASET

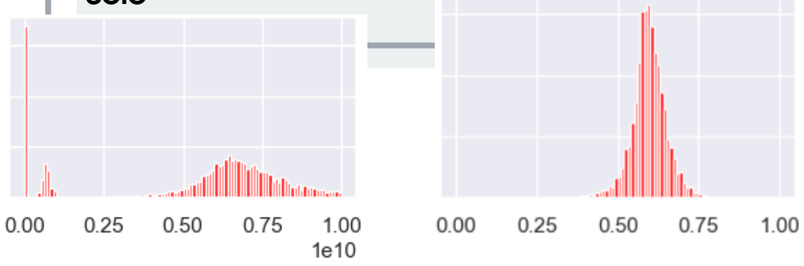
- Gran desbalanceo entre clases (6599 a 220) -> se tiende a predecir la clase mayoritaria y no se identifica correctamente la minoritaria 1
- Normalizaciones incompletas, datos mezclados -> reduce la calidad de los datos 2
- Outliers -> reducen la calidad de la normalización porque confina la mayoría de los datos a un rango menor 3
- Muchos atributos (95 variables) -> reduce velocidad de trabajo y calidad de resultados 4

PROCESAMIENTO Y PREDICCIÓN

Problema tratado: 2

Corrección de rangos

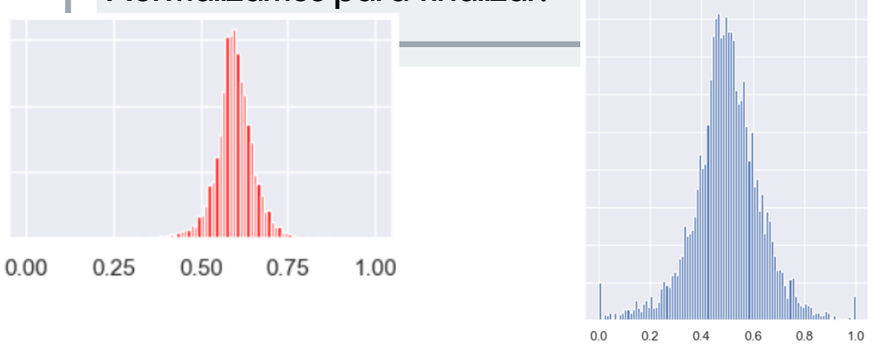
Asumimos que el rango de los valores mayores que 1 es similar a los que están normalizados. Normalizamos y combinamos ambos conjuntos en uno solo



Problema tratado: 3

Detección de outliers

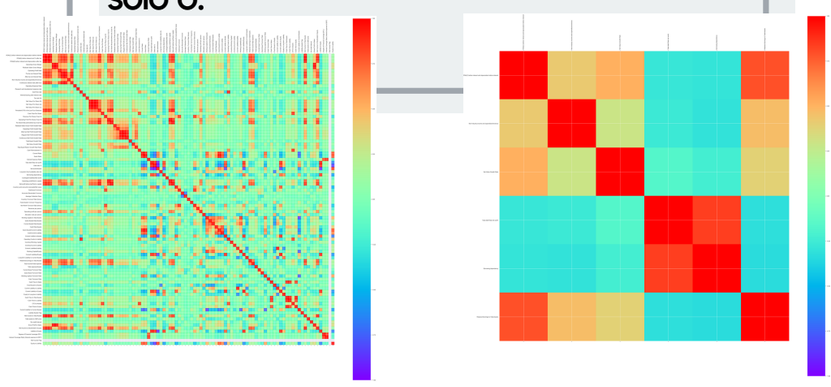
Aplicamos la detección de outliers por **IQR** y los asignamos a los extremos del nuevo rango para dejar constancia de que son valores alejados de la media. Normalizamos para finalizar.



Problema tratado: 4

Selección de variables

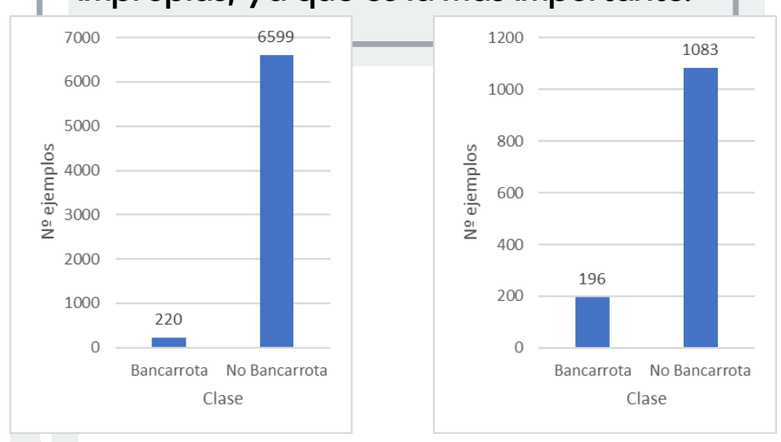
Aplicamos el filtro de correlación con el filtro **ANOVA** para hacer la selección de los atributos. Pasamos de 95 a tan solo 6.



Problema tratado: 1

Muestreo de datos

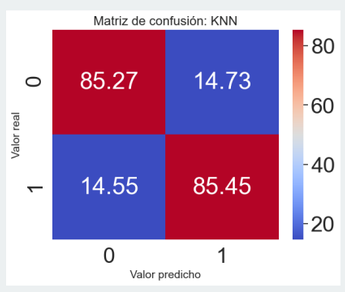
Aplicamos **RUS** (Random UnderSampling) para reducir el numero de ejemplos de la clase mayoritaria, y así evitar que en la minoritaria aparezcan características impropias, ya que es la mas importante.



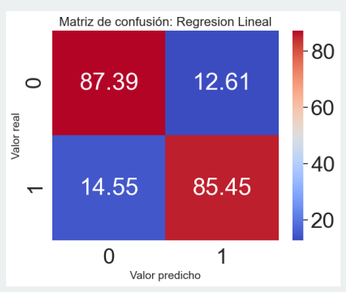
Entrenamiento y clasificación

Probaremos los siguientes clasificadores para observar su desempeño después de todo el proceso de tratado y limpieza de datos (mostramos el **recall** por clase):

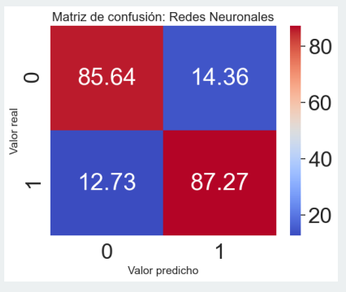
KNN



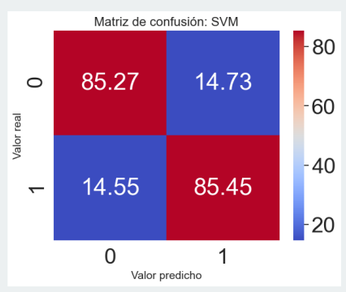
Regresión Lineal



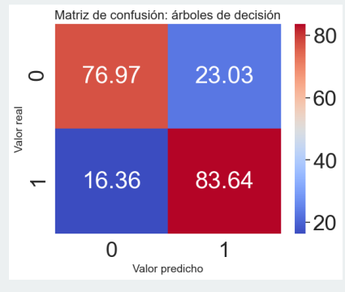
Redes Neuronales



SVMs



Árboles de decisión



1650ej clase 0 || 55ej clase 1

* A los datos no se les aplica el paso de la selección de variables por que el modelo lo hace de por sí

RESULTADOS

Se alcanza el objetivo de obtener un buen resultado en recall. Pasamos del 2.4% al 87% sobre la clase 1 (bancarrota), lo que significa que en caso de que la compañía este en riesgo de quiebra, nuestros modelos la identificaran en la mayoría de los casos. El principal problema es que debido a que siempre hay mas ejemplos de la clase 0, aunque tenga un recall alto, habrá muchos falsos positivos.

LINEAS FUTURAS

- Tratar el problemas del alto FPR
- Comprobar los resultados de todas las posibles combinatorias de procesamientos de datos para evitar perder combinaciones más óptimas
- Comprensión e interpretación de las variables de mano de un experto en economía.
- Diversificar los métodos de detección de oultiers y normalización para distribuciones no normales

IGOR VONS & AITOR URANGA