Predicción de quiebra bancaria taiwanesa mediante Clasificadores Bayesianos, Estimadores No Paramétricos y k-NN

Jordi Blasco Lozano 74527208D

3 de noviembre de 2025

Resumen

He abordado la práctica 2 seleccionando el dataset "Taiwanese Bankruptcy Prediction" de UCI, aplicando los métodos estudiados en la asignatura. Explico cada proceso (preprocesado, partición, modelado, optimización y análisis de resultados) usando terminaciones verbales en primera persona, incluyendo los códigos implementados para que cada apartado sea reproducible y comprensible.

1 Introducción

He decidido realizar la práctica usando el dataset "Taiwanese Bankruptcy Prediction"porque corresponde a un problema clásico de clasificación binaria con fuerte desbalance y datos financieros reales. Este dataset presenta características que lo hacen especialmente interesante para evaluar clasificadores:

- **Alta dimensionalidad:** 95 características financieras que describen la salud económica de empresas
- **Desbalance severo:** Solo el 3.23 % de las empresas están en bancarrota, lo que representa un reto significativo para los algoritmos de clasificación
- **Relevancia práctica:** La predicción de quiebras empresariales tiene aplicaciones directas en análisis de riesgo crediticio y decisiones de inversión
- **Datos reales:** Proviene de empresas taiwanesas reales, lo que añade complejidad y ruido natural

El objetivo principal es comparar el rendimiento de diferentes enfoques de clasificación (paramétricos vs. no paramétricos) bajo condiciones de desbalance extremo, poniendo especial énfasis en la correcta aplicación de técnicas de validación cruzada para evitar resultados optimistas y engañosos.

2 Dataset y Preprocesado

He descargado el dataset desde el repositorio UCI Machine Learning. El conjunto contiene 6819 empresas taiwanesas con 95 características financieras cada una. Tras el análisis exploratorio he identificado las siguientes propiedades:

2.1 Características del Dataset

- **Tamaño:** 6819 ejemplos con 96 columnas (95 features + 1 target)
- **Tipos de datos:** 93 variables float64 (continuas) y 3 int64 (flags binarios)
- **Valores faltantes:** Ninguno (100 % de datos completos)
- **Distribución de clases:** 220 bancarrotas (3.23 %) vs 6599 no bancarrotas (96.77 %)
- **Ratio de desbalance:** Aproximadamente 30:1 (clase mayoritaria:minoritaria)

Este desbalance extremo es un desafío fundamental que afectará significativamente al entrenamiento y evaluación de los modelos. Los clasificadores tienden naturalmente a favorecer la clase mayoritaria, lo que puede llevar a modelos con alta accuracy pero pobre capacidad de detectar bancarrotas.

2.2 Preprocesamiento

No he realizado normalización ni estandarización en esta fase inicial porque tanto Naive Bayes como k-NN de scikit-learn pueden manejar variables en escalas diferentes. Sin embargo, esto es un punto de mejora potencial, especialmente para k-NN que es sensible a la escala de las variables.

El código para cargar y explorar el dataset incluye la configuración para evitar warnings en Windows relacionados con la detección de núcleos de CPU por parte de joblib:

```
import pandas as pd
import numpy as np
import os
import warnings
from sklearn.model_selection import StratifiedKFold, GridSearchCV, cross_val_predict
from sklearn.naive_bayes import GaussianNB
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

# Silenciar warnings de joblib en Windows
os.environ['LOKY_MAX_CPU_COUNT'] = '4'
warnings.filterwarnings('ignore', category=UserWarning, module='joblib')

# Cargar datos
df = pd.read_csv('data.csv')
print(df.info())
print(df['Bankrupt?'].value_counts())
print(f"\nPorcentaje de bancarrotas: {df['Bankrupt?'].mean() * 100:.2f}%")
```