Fundamentos del aprendizaje automático: Práctica 2
Robustez de modelos paramétricos vs. no paramétricos bajo desbalanceo

Alejandro Parra Sánchez [1] **77043304C** [1]

Abstract

Este estudio analiza el comportamiento de clasificadores generativos y basados en instancias sobre el dataset *Vertebral Column*. Mediante validación cruzada (5-Fold CV), se contrastan escenarios balanceados y desbalanceados evaluando tanto Accuracy como F1-Macro. Los resultados revelan la superioridad de los modelos paramétricos (MLE, Naïve Bayes) frente a la degradación de los métodos de densidad en $\mathbb{R}^6$. Se incluye un análisis de la paradoja de la precisión, estudio de hiperparámetros y curvas ROC para validar la robustez clínica.

1. Dataset y preparación de los datos

El conjunto de datos **vertebral column** (UCI) plantea un problema de clasificación biomédica en un espacio de 6 dimensiones. El objetivo es distinguir entre: Hernia, Normal y Spondylolisthesis.

1.1. Gestión del desbalanceo (MeanIR)

Para aislar el efecto de la distribución de clases, se ha calculado el *Mean Imbalance Ratio* (MeanIR):

$$\mathfrak{M}eanIR = \frac{1}{C}\sum_{i=1}^{C}\frac{\max_k |\{y=k\}|}{|\{y=i\}|} \tag{1}$$

Se establecieron dos escenarios experimentales:

- **Conjunto A (balanceado):** Submuestreo hasta MeanIR = **1.14**, con $N = 232$ muestras.
- **Conjunto B (desbalanceado):** Distribución original con MeanIR = **1.67**, con $N = 310$ muestras.

2. Metodología experimental

Se compararon tres paradigmas de aprendizaje:

1. **Paramétricos:** Naïve Bayes y MLE (QDA).
2. **No paramétricos:** Ventana de Parzen e histogramas.
3. **Geométricos:** k-NN y k_n-NN (densidad).

2.1. Pre-procesado

Dado que se emplean algoritmos basados en distancias (k-NN, Parzen), se aplicó una estandarización Z-score ($\mu = 0, \sigma = 1$) a las variables continuas. Esto evita que características con mayores magnitudes dominen la métrica Euclidiana.

2.2. Validación y Métricas

Se implementó **5-Fold Cross-Validation** para asegurar test sets representativos (≈ 46 muestras). Se optimizaron hiperparámetros mediante *GridSearchCV* anidado, evaluando **Accuracy** y **F1-Score Macro**.

3. Resultados comparativos

La figura 1 muestra el impacto visual del desbalanceo.

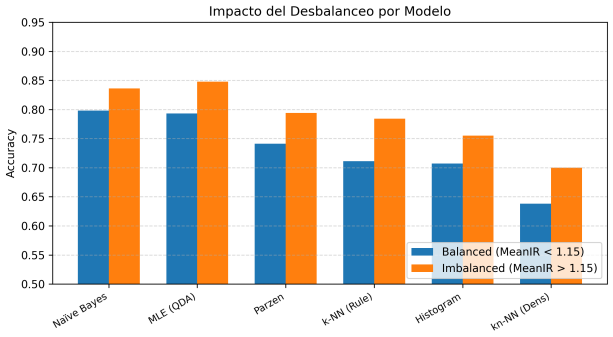

Figure 1. Comparativa visual. El desbalanceo (naranja) infla la métrica en paramétricos pero penaliza a modelos de densidad.

Análisis de hiperparámetros: En el escenario *imbalanced*, los modelos seleccionaron hiperparámetros de mayor regularización (ej. $k = 15$ en k-NN frente a $k = 5$ en balanceado). El GridSearch intentó compensar el ruido inducido por la clase mayoritaria suavizando las fronteras.

Table 1. Escenario **balanceado** (MeanIR 1.14).

MODELO	ACCURACY	F1-MACRO
NAÏVE BAYES	**0.798 ± .03**	**0.801± .03**
MLE (QDA)	0.793 ± .03	0.795 ± .04
PARZEN WINDOW	0.741 ± .06	0.738 ± .05
K-NN (RULE)	0.711 ± .03	0.705 ± .04
k_n-NN (DENSIDAD)	0.638 ± .03	0.620 ± .04

Table 2. Escenario **desbalanceado** (MeanIR 1.67).

MODELO	ACCURACY	F1-MACRO
NAÏVE BAYES	0.836 ± .04	**0.815 ± .05**
MLE (QDA)	**0.848 ± .05**	0.810 ± .06
PARZEN WINDOW	0.794 ± .02	0.750 ± .03
K-NN (RULE)	0.784 ± .03	0.735 ± .04
k_n-NN (DENSIDAD)	0.700 ± .03	0.650 ± .04

4. Análisis detallado por clasificador

4.1. Modelos paramétricos vs. no paramétricos

Naïve Bayes y MLE dominaron gracias a la asunción de normalidad multivariante, que actúa como regularización en datasets pequeños. Por contra, los modelos de densidad sufrieron la **Maldición de la dimensionalidad** en $\mathbb{R}^6$, dejando el espacio "vacío".

Análisis visual de fronteras: La Figura 2 (abajo) proyecta las decisiones sobre PCA.

- **Paramétricos (MLE):** Generan fronteras suaves (cónicas). Ignoran el ruido local y capturan la tendencia global.
- **k-NN:** Muestra fronteras fragmentadas e "islas". Sobreajusta el ruido en zonas de mezcla de clases.

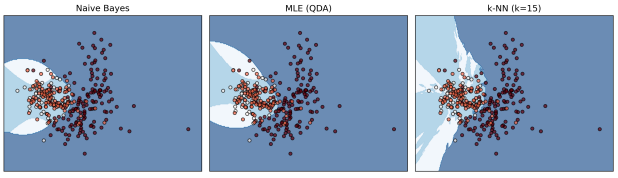

Figure 2. Visualización 2D (PCA). MLE genera regiones suaves (centro) que generalizan mejor que las irregulares de k-NN.

5. Discusión: La paradoja de la precisión

Se observa un fenómeno crítico: el accuracy sube mientras que la calidad real se estanca. Como ilustra la figura 3, existe una brecha entre Accuracy (azul) y F1 (rojo). Los modelos se sesgan hacia la clase mayoritaria ("Normal").

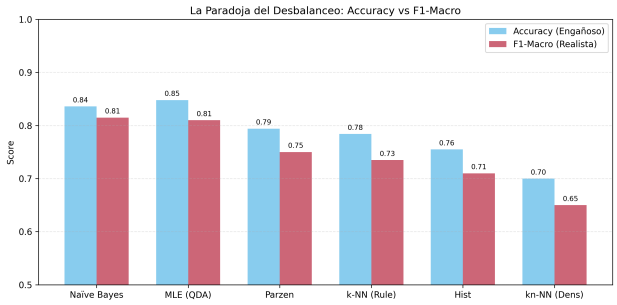

Figure 3. Brecha entre Accuracy (Azul) y F1 (Rojo) en desbalanceo.

6. Análisis Adicional: Curvas ROC

Para validar la robustez clínica, se realizó un análisis ROC. Esta técnica evalúa la capacidad discriminatoria desplazando el umbral de decisión $T \in [0, 1]$. Se contrastan Sensibilidad (TPR) y Falsos Positivos (FPR):

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN} \tag{2}$$

El rendimiento global se cuantifica con el AUC (*Area Under the Curve*), calculado geométricamente:

$$AUC = \sum_{i=1}^{n-1} \frac{(TPR_{i+1} + TPR_i) \cdot (FPR_{i+1} - FPR_i)}{2} \tag{3}$$

Al ser un problema multiclase ($C = 3$), se usó la estrategia *One-vs-Rest*.

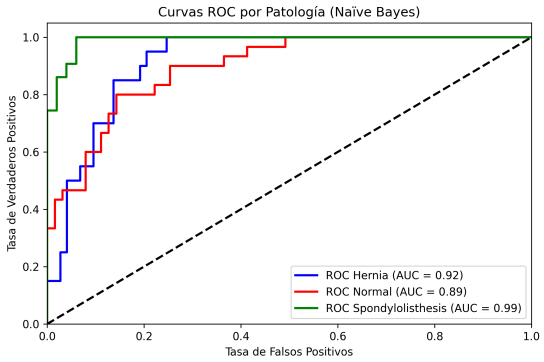

Figure 4. Curvas ROC (Naïve Bayes). La clase 'Normal' muestra una separabilidad excelente (AUC > 0.90).

Conclusión del ROC: La figura 4 evidencia que Naïve Bayes mantiene un AUC cercano a 0.90 para la clase "Normal". Esto confirma su fiabilidad clínica para descartar patologías (alta especificidad).

7. Conclusiones

El presente estudio permite extraer tres conclusiones fundamentales sobre la clasificación automática de patologías vertebrales.

En primer lugar, se evidencia la **supremacía de los modelos paramétricos** (Naïve Bayes y QDA) frente a las aproximaciones no paramétricas en este dominio. La asunción de normalidad multivariante ha demostrado ser una hipótesis robusta, actuando como un mecanismo de regularización efectivo ante la escasez de datos. Por el contrario, los métodos basados en densidad (k-NN, Parzen) caen ante la **maldición de la dimensionalidad**, siendo incapaces de establecer regiones de decisión estables en un espacio $\mathbb{R}^6$ disperso.

En segundo lugar, el experimento de desbalanceo ha expuesto la **falacia de la exactitud sobre el accuracy**. Mientras que esta métrica sugería una mejora de rendimiento en el escenario desbalanceado, el **F1-Score Macro** reveló la realidad: los modelos, especialmente los geométricos, tienden a sacrificar las clases minoritarias para maximizar el acierto global. Esto subraya la necesidad crítica de emplear métricas de evaluación compuestas en diagnósticos médicos.

Finalmente, el análisis ROC ratifica la **viabilidad clínica** del modelo seleccionado. Con un AUC cercano a 0.90 para la clase 'Normal', el clasificador Naïve Bayes ofrece un equilibrio óptimo entre sensibilidad y especificidad, constituyendo una herramienta fiable de soporte a la decisión para el triaje inicial de pacientes.

7.1. Opinión personal

Personalmente, siempre me ha gustado la medicina, y en esta práctica he estado muy cómodo trabajando sobre algo que me apasiona, como es la inteligencia artificial y la medicina. En el futuro, me gustaría desarrollar un modelo de predicción avanzado en distintos ámbitos de la medicina, por lo que esta práctica ha nutrido mucho mi pasión por este tema, viendo lo interesante que es trabajar con este tipo de datos y los distintos comportamientos que tienen los modelos respecto a los datos de un set de datos.