Predicción de Clase de Animal mediante Clasificadores Bayesianos, Estimadores No Paramétricos y k-NN

Jordi Blasco Lozano

Abstract

Este trabajo aborda la práctica 2 en la asignatura de Fundamentos del Aprendizaje Automático empleando el dataset Zoo de UCI, aplicando seis algoritmos de clasificación: Naive Bayes Gaussiano, MLE Multivariante, Histogram Bayes, Parzen Windows, k-NN Density Bayes y k-NN Rule. Se analiza su rendimiento en una tarea real de clasificación multiclase (7 clases) con características binarias y clases minoritarias.

1. Introducción y justificación del dataset

He seleccionado el **dataset Zoo** de UCI porque cumple con los requerimientos del enunciado: clasificación multiclase (7 clases: mamífero, ave, reptil, pez, anfibio, invertebrado, insecto), 16 atributos binarios interpretables, y tamaño adecuado (101 instancias originales) para observar comportamientos de métodos paramétricos y no paramétricos.

Dado el desbalance original (41 mamíferos vs 4 anfibios y un **meanIR** de 4.84), generé sintéticamente nuevos animales mediante investigación profunda usando IA añadiendo nuevos animales con sus características, alcanzando 287 instancias con 41 ejemplares por clase. Esto permite evaluar cómo el balanceo afecta el rendimiento de cada modelo, especialmente aquellos sensibles a datos escasos (MLE).

2. Dataset y Análisis Exploratorio

Dataset original: 101 instancias, 17 atributos (15 binarios + 1 numérico + 1 clase), 7 clases desbalanceadas (mamíferos: 41, aves: 20, reptiles: 5, peces: 13, anfibios: 4, invertebrados: 8, insectos: 10).
Dataset balanceado: 287 instancias con 41 ejemplares por clase, generados sintéticamente mediante investigación de características específicas de cada tipo de animal.

Los 17 atributos incluyen **15 características binarias**: pelo, plumas, huevos, leche, vuela, acuático, depredador, dientes, columna vertebral, respira, venenoso, aletas, cola, doméstico, tamaño grande; **1 atributo numérico**: patas; y **1 atributo de tipo clase** (valores 1-7 codificados como los grupos de animales anteriores).

3. Metodología: Particionado y Validación

División Train-Test (80%-20%): División estratificada manteniendo proporciones de clases (original: 80 train / 21 test; balanceado: 229 train / 58 test).

Validación Cruzada: 5-fold estratificada solo en entrenamiento para optimizar hiperparámetros (bandwidth en Parzen, k en k-NN).

Métricas: F1-macro (métrica principal por desbalance), accuracy, precision, matrices de confusión.

4. Modelos Implementados

Seis clasificadores: (1) Naive Bayes Gaussiano, (2) MLE Multivariante con covarianza completa, (3) Histogram Bayes, (4) Parzen Windows con kernel Gaussiano, (5) k-NN Density Bayes, (6) k-NN Rule.

5. Resultados

He evaluado los seis modelos en el conjunto de prueba de 21 muestras. Las Tables 1 and 2 muestran los resultados de validación cruzada y test respectivamente, separando modelos paramétricos (1-2) de no paramétricos (3-6).

Table 1. Validación cruzada 5-fold en entrenamiento

	ORIGINAL		BALANCEADO	
MODELO	F1	STD	F1	STD
Modelos Paramétricos				
NAIVE BAYES	0.85	0.14	0.84	0.04
MLE FULL	0.53	0.10	0.87	0.01
Modelos No Paramétricos				
HISTOGRAM BAYES	0.25	0.13	0.06	0.02
PARZEN (H=0.1)	0.86	–	0.94	–
K-NN DENSITY (K=11)	0.57	–	0.70	–
K-NN RULE (K=1)	0.83	–	0.94	–

Table 2. Resultados en conjunto de prueba

	ORIGINAL		BALANCEADO	
MODELO	ACC	F1	ACC	F1
Modelos Paramétricos				
NAIVE BAYES	1.00	1.00	0.91	0.91
MLE FULL	0.71	0.46	0.95	0.95
Modelos No Paramétricos				
HISTOGRAM BAYES	0.38	0.08	0.16	0.07
PARZEN (H=0.1)	1.00	1.00	0.95	0.95
K-NN DENSITY (K=11)	0.48	0.57	0.74	0.68
K-NN RULE (K=1)	1.00	1.00	0.95	0.95

5.1. Análisis detallado por modelo

5.1.1. Naive Bayes Gaussiano

Original: Accuracy = 1.0, F1-macro = 1.0
Balanceado: Accuracy = 0.91, F1-macro = 0.91

En el dataset original desbalanceado, Naive Bayes logra clasificación perfecta (21/21 muestras correctas). Sin embargo, con el dataset balanceado (287 muestras, 41 por clase), el rendimiento baja ligeramente a 91%, revelando que la perfección inicial se debía en parte al pequeño tamaño del conjunto de prueba. A pesar de la asunción "naive" de independencia entre características, el modelo funciona excepcionalmente bien en ambos casos. La matriz de confusión muestra que las principales confusiones ocurren entre reptiles-anfibios y anfibios-insectos, lo cual es razonable dado que comparten características físicas similares. El modelo estima solo d parámetros por clase ($\mu_{y,i}$ y $\sigma^2_{y,i}$ para cada característica), lo que lo hace robusto incluso con clases minoritarias.

5.1.2. MLE Multivariante (Full Bayesian Gaussian)

Original: Accuracy = 0.71, F1-macro = 0.46
Balanceado: Accuracy = 0.95, F1-macro = 0.95

Este modelo muestra la mejora más dramática al balancear los datos. En el dataset original, falla significativamente con clases minoritarias (reptiles, anfibios, invertebrados), clasificándolas erróneamente como mamíferos. La causa es que con solo 3-6 ejemplos en entrenamiento, la estimación de la matriz de covarianza completa 16×16 (136 parámetros únicos por clase) es imposible, resultando en matrices singulares o mal condicionadas. Con el dataset balanceado (41 ejemplos por clase), el modelo puede estimar correctamente las covarianzas y alcanza 95% de rendimiento, igualando a los mejores modelos. Esto demuestra que los modelos más complejos **necesitan más datos** para funcionar correctamente. La covarianza completa captura las correlaciones entre características, pero requiere $O(d^2)$ muestras por clase.

5.1.3. Histogram Bayes

Original: Accuracy = 0.38, F1-macro = 0.08
Balanceado: Accuracy = 0.16, F1-macro = 0.07

Este modelo obtiene el peor rendimiento de todos, evidenciando la maldición de la dimensionalidad. Con 16 características binarias existen $2^{16} = 65,536$ bins posibles, pero solo 287 muestras en el dataset balanceado (y con menos muestras aún en el desbalanceado). Dado que los histogramas requieren $N \propto b^d$ muestras, la mayoría de bins permanecerán vacíos, asignando $p(\mathbf{x}|y) = 0$ a regiones no vistas, forzando al modelo a predecir sistemáticamente la clase mayoritaria. El accuracy empeora al balancear (0.380.16) porque en el dataset original predecía "mamífero" siempre, y esta clase supone casi un 40 por ciento de las muestras. Mientras el F1-macro se mantiene bajo (0.07 aprox) en ambos casos porque evalúa cada clase individualmente. Es por esto por lo que deberíamos de tener más en cuenta el F1 en vez del accuracy al estudiar clasificaciones.

5.1.4. Parzen Windows

Original: Accuracy = 1.0, F1-macro = 1.0
Balanceado: Accuracy = 0.95, F1-macro = 0.95

Parzen windows con kernel Gaussiano mantiene excelente rendimiento en ambos escenarios. Con bandwidth $h = 0.1$ óptimo (obtenido por validación cruzada en el entrenamiento), el modelo produce estimaciones de densidad continuas y suaves que funcionan bien incluso con datos escasos. A diferencia de histogramas, el suavizado del kernel Gaussiano compensa la falta de muestras, evitando las discontinuidades artificiales de los bins. En el dataset balanceado, obtiene el mejor F1-macro (0.95), superior a Naive Bayes (0.91), confirmando que el bandwidth optimizado generaliza correctamente al capturar características conjuntas que NB asume como independientes, también es por esto por lo que las confusiones ocurren principalmente entre clases con características solapadas (anfibios-reptiles), lo cual es esperado. Este método mitiga efectivamente la maldición de la dimensionalidad mediante un leve suavizado continuo. Obtenemos un 1 en el test del dataset desbalanceado gracias a haber elegido el suavizado correcto y al tener solo 21 ejemplos. En nuestro caso, una ventana pequeña pero no mínima, "prestando" densidad de vecinos cercanos sin manchar clusters separados, es clave para obtener la máxima precisión.

5.1.5. k-NN Density Bayes

Original: Accuracy = 0.48, F1-macro = 0.57
Balanceado: Accuracy = 0.74, F1-macro = 0.68

Rendimiento mediocre que mejora con el balanceo pero sigue siendo inferior a otros métodos. Con $k = 11$ óptimo

en ambos datasets, los 11 vecinos más cercanos diluyen la señal local, resultando en estimación de densidad por volumen local inestable. El modelo adapta el volumen $V_k(\mathbf{x})$ localmente (pequeño en regiones densas, grande en dispersas), pero esta estrategia resulta problemática con clases que tienen distribuciones espaciales dispersas. En el dataset original, confunde masivamente mamíferos como reptiles. Con el balanceo mejora, pero las confusiones persisten entre reptiles-peces y insectos distribuidos entre múltiples clases. Comparado con Parzen (que fija bandwidth globalmente), la adaptación local de volumen es menos efectiva en este dataset. La estimación explícita de densidad añade complejidad sin beneficio claro.

5.1.6. k-NN Rule

Original: Accuracy = 1.0, F1-macro = 1.0
Balanceado: Accuracy = 0.95, F1-macro = 0.95

El método clásico de k-NN mantiene excelente rendimiento con $k = 1$ óptimo en ambos datasets (vecino más cercano). En el dataset original logra clasificación perfecta, indicando que las clases están bien separadas espacialmente. Con el dataset balanceado, mantiene 95% de accuracy, igualando a MLE Full y Parzen. Con $k = 1$, el método simplemente asigna la clase del vecino más cercano, lo cual es óptimo cuando los datos están bien separados y tienen poco ruido. La simplicidad del método (asignación directa basada en vecinos sin calcular densidades probabilísticas) resulta ventajosa: menos pasos intermedios significan menos fuentes de error. Las confusiones coinciden exactamente con las de Parzen (anfibios-reptiles), confirmando que ambos métodos capturan correctamente la estructura espacial. Valores mayores de k promediarían en exceso los vecinos cercanos, borrando detalles locales innecesariamente.

6. Discusión y Conclusiones

6.1. Comparación general de enfoques

Los modelos paramétricos muestran comportamientos opuestos: Naive Bayes es robusto con clases minoritarias (F1=0.85 original) al estimar solo $2d$ parámetros por clase, mientras MLE Full colapsa (F1=0.46) requiriendo $O(d^2)$ parámetros. Con datos suficientes (41 ejemplares/clase), MLE Full alcanza 95% superando a NB (91%), confirmando el equilibrio entre sesgo y varianza: modelos simples son robustos pero limitados; complejos necesitan más datos pero alcanzan mayor rendimiento.

Entre no paramétricos, Histogram Bayes falla completamente (F1=0.07-0.08) por la maldición de dimensionalidad: $2^{16} = 65,536$ bins con solo 287 muestras resulta en predicciones por defecto. Parzen y k-NN Rule logran 95-100% evitando particiones rígidas mediante suavizado continuo ($h = 0.1$) o asignación directa ($k = 1$). k-NN Density (F1=0.57-0.68) queda en medio, sugiriendo que estimar densidades explícitamente añade complejidad sin beneficio en alta dimensión.

6.2. Impacto del desbalance y balanceo

El desbalance original (meanIR=4.84) expone brutalmente debilidades de modelos complejos: MLE Full no puede estimar covarianzas con 3-5 ejemplos/clase, clasificando anfibios como mamíferos. El balanceo sintético (101287 instancias) transforma MLE Full dramáticamente: F1 = $0.46 \rightarrow 0.95$ (+107%), pasando de peor a mejor modelo. Paradójicamente, Histogram empeora (accuracy $0.38 \rightarrow 0.16$) al no poder explotar el sesgo mayoritario, confirmando su inadecuación fundamental. Los mejores modelos originales (NB, Parzen, k-NN Rule con F1=1.0) bajan ligeramente a 91-95%, pero esto refleja que al aumentar el conjunto de datos trabajamos en un test más realista (58 vs 21 muestras) que revela verdadera capacidad de generalización.

6.3. Métricas y lecciones prácticas

El contraste accuracy vs F1-macro es crucial: Histogram obtiene accuracy = 0.38 pero F1 = 0.08 en desbalanceado, prediciendo siempre "mamífero" (al ser clase mayoritaria acierta 8 ejemplos de 21, de aquí la "gran" precisión). El accuracy premia este sesgo pero F1-macro lo penaliza al promediar todas las clases. Esto confirma que **accuracy es engañoso en multiclase desbalanceado**, siendo F1-macro la métrica correcta para evaluar poder discriminativo real.

6.4. Conclusiones finales

Este trabajo confirma empíricamente tres principios del aprendizaje automático: (1) el trade-off complejidad-datos (modelos con $O(d^2)$ parámetros necesitan más muestras), (2) la maldición de dimensionalidad (particionado rígido colapsa, suavizado continuo mitiga), y (3) la importancia del balanceo para revelar limitaciones sin sesgos artificiales. Los tres mejores modelos (MLE Full, Parzen, k-NN Rule con F1=0.95) triunfan por razones distintas: MLE captura correlaciones completas, Parzen suaviza óptimamente, k-NN Rule aprovecha separación espacial. **Las lecciones clave son:** validar hiperparámetros con CV estratificada es crucial ($h = 0.1$, $k = 1$ optimizados); usar F1-macro en multiclase desbalanceado; la asunción "naive" es sorprendentemente efectiva (F1=0.91) sugiriendo correlaciones débiles entre clases; y el balanceo sintético mediante características reales es válido (pero hay que comprender también que puede haber confusiones entre anfibios y reptiles pero son biológicamente coherentes).