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

Jordi Blasco Lozano

Resumen

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** porque me parece muy interesante y contiene: 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 las 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 como el MLE FULL.

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

Validación Cruzada 5-fold Estratificada: Se aplica validación cruzada 5-fold estratificada sobre todo el conjunto de datos. En cada fold, el conjunto se divide en 80% entrenamiento y 20% test, manteniendo las proporciones de clases. El proceso se repite 5 veces, de modo que cada partición actúa como test una vez. Para cada modelo se obtienen 5 métricas independientes, cuya media y desviación estándar reflejan el rendimiento y la estabilidad del modelo.
Optimización de Hiperparámetros: En los modelos que requieren selección de hiperparámetros, se evalúan todos los valores candidatos en cada uno de los 5 folds. Se calcula el F1-macro promedio de cada hiperparámetro y se selecciona el que maximiza esta métrica.
Métricas: F1-macro (métrica principal), accuracy (para comparación general) y matrices de confusión (para analizar errores por clase).

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

Las Tables 1 and 2 muestran la media y desviación estándar de Accuracy y F1-macro obtenidas en los para cada dataset.

Table 1. Resultados en dataset desbalanceado (CV 5-fold)

	ACCURACY		F1-MACRO	
MODELO	MEAN	STD	MEAN	STD
Modelos Paramétricos				
NAIVE BAYES	0.96	0.06	0.90	0.14
MLE FULL	0.81	0.04	0.58	0.05
Modelos No Paramétricos				
HISTOGRAM BAYES	0.46	0.06	0.17	0.10
PARZEN ($h = 0.5$)	0.98	0.02	0.91	0.12
K-NN DENSITY ($k = 1$)	0.98	0.02	0.90	0.12
K-NN RULE ($k = 1$)	0.98	0.02	0.91	0.12

Table 2. Resultados en dataset balanceado (CV 5-fold)

	ACCURACY		F1-MACRO	
MODELO	MEAN	STD	MEAN	STD
Modelos Paramétricos				
NAIVE BAYES	0.88	0.07	0.87	0.08
MLE FULL	0.89	0.05	0.89	0.05
Modelos No Paramétricos				
HISTOGRAM BAYES	0.15	0.01	0.06	0.02
PARZEN ($h = 0.5$)	0.93	0.03	0.93	0.03
K-NN DENSITY ($k = 1$)	0.93	0.03	0.93	0.04
K-NN RULE ($k = 3$)	0.92	0.05	0.91	0.05

5.1. Análisis detallado por modelo

5.1.1. Naive Bayes Gaussiano

Desbalanceado: Acc = 0.96 $\pm$ 0.06, F1 = 0.90 $\pm$ 0.14; **Balanceado:** Acc = 0.88 $\pm$ 0.07, F1 = 0.87 $\pm$ 0.08

Naive Bayes obtiene excelente rendimiento en el dataset desbalanceado (F1=0.90), demostrando robustez ante clases minoritarias gracias a estimar solo $2d$ parámetros por clase ($\mu_{y,i}$ y $\sigma^2_{y,i}$). Con el dataset balanceado, el rendimiento baja ligeramente (F1=0.87), pero la desviación se reduce notablemente (0.14→0.08), indicando mayor estabilidad al disponer de más datos por clase. La matriz de confusión revela que las principales confusiones ocurren entre reptiles-anfibios (13 casos) y anfibios-reptiles (11 casos), reflejando el solapamiento real de características binarias entre estas clases. A pesar de la asunción "naive" de independencia, el modelo captura eficazmente las distribuciones de clase, siendo especialmente preciso con mamíferos, aves e invertebrados (recall $\approx$ 1.00).

5.1.2. MLE Multivariante (Full Bayesian Gaussian)

Desbalanceado: Acc = 0.81 $\pm$ 0.04, F1 = 0.58 $\pm$ 0.05; **Balanceado:** Acc = 0.89 $\pm$ 0.05, F1 = 0.89 $\pm$ 0.05

Este modelo exhibe la mejora más dramática con el balanceo (F1: 0.58 $\rightarrow$ 0.89, +53%). En el dataset desbalanceado, la matriz de confusión revela colapso total en clases minoritarias: 5 reptiles clasificados como mamíferos, 4 anfibios como mamíferos, resultando en precision cero para estas clases. La causa es que estimar la covarianza completa 16×16 (136 parámetros únicos por clase) con solo 3-5 ejemplos produce matrices mal condicionadas, generando predicciones inestables. Con el balanceo (41 ejemplos/clase), el modelo estima correctamente las correlaciones entre características y alcanza F1=0.89, igualando a Naive Bayes y demostrando que capturar dependencias completas es valioso cuando hay datos suficientes.

5.1.3. Histogram Bayes

Desbalanceado: Acc = 0.46 $\pm$ 0.06, F1 = 0.17 $\pm$ 0.10; **Balanceado:** Acc = 0.15 $\pm$ 0.01, F1 = 0.06 $\pm$ 0.02

Este modelo sufre el peor rendimiento, demostrando brutalmente la maldición de la dimensionalidad. Con $2^{16} = 65,536$ bins posibles pero solo 287 muestras (41/clase), la mayoría de bins quedan vacíos, asignando $p(\mathbf{x}|y) = 0$ a patrones no vistos y forzando predicciones por defecto. La matriz de confusión balanceada muestra el colapso: 283/287 muestras clasificadas como mamíferos (recall=1.00 mamífero, 0.00 resto). El accuracy paradójicamente empeora con el balanceo (0.46→0.15) porque en desbalanceado, predecir "mamífero" siempre acierta 40% del tiempo, mientras que con clases equiprobables solo acierta 14%. El F1-macro permanece consistentemente bajo (0.17→0.06) porque evalúa todas las clases por igual. Los histogramas requieren $N \propto b^d$ muestras; con $d = 16$ y $N = 287$, el método es fundamentalmente inadecuado.

5.1.4. Parzen Windows

Desbalanceado: Acc = 0.98 $\pm$ 0.02, F1 = 0.91 $\pm$ 0.12; **Balanceado:** Acc = 0.93 $\pm$ 0.03, F1 = 0.93 $\pm$ 0.03

Parzen Windows alcanza el mejor rendimiento global, especialmente tras balanceo (F1=0.93), con bandwidth óptimo $h = 0.5$ que proporciona suavizado continuo sin sobreajuste. En el dataset desbalanceado, obtiene F1=0.91 con alta varianza (std=0.12), reflejando sensibilidad a la composición de los folds con clases minoritarias. El balanceo estabiliza dramáticamente el modelo (std: 0.12→0.03), demostrando que el kernel Gaussiano funciona óptimamente con datos equilibrados. La matriz de confusión muestra confusiones concentradas en reptiles-anfibios (6+4=10 casos) e insectos dispersos entre múltiples clases, coherentes con solapamiento de características binarias. A diferencia de histogramas, el suavizado continuo mitiga la maldición de dimensionalidad prestando densidad de vecinos cercanos sin crear discontinuidades artificiales. La búsqueda de hiperparámetros revela que h pequeños (0.05-0.5) mantienen F1$\approx$0.93, mientras que h grandes (1.5-2.0) degradan a F1$\approx$0.60, confirmando importancia de localidad.

5.1.5. k-NN Density Bayes

Desbalanceado: Acc = 0.98 $\pm$ 0.02, F1 = 0.90 $\pm$ 0.12; **Balanceado:** Acc = 0.93 $\pm$ 0.03, F1 = 0.93 $\pm$ 0.04

Con $k = 1$ óptimo, este modelo iguala prácticamente a Parzen Windows (F1=0.93 vs 0.93 balanceado), sugiriendo que en este dataset la asignación de densidad local adaptativa funciona óptimamente con el vecino más cercano. En desbalanceado, obtiene F1=0.90 con alta varianza (0.12), similar a Parzen, reflejando sensibilidad

a clases minoritarias. El balanceo estabiliza el modelo (std: 0.12→0.04) y mejora F1 (0.90→0.93). La búsqueda de k revela degradación con valores mayores: $k = 3$ produce F1=0.92, $k5$ baja a F1 (0.83-0.88), confirmando que promediar muchos vecinos diluye información local crítica. La matriz de confusión muestra confusiones reptiles-anfibios (1+10=11 casos), coherentes con Parzen. Aunque el método estima densidades explícitamente mediante $\hat{p}(\mathbf{x}|y) = k/(NV_k(\mathbf{x}))$, con $k = 1$ se comporta esencialmente como asignación directa, explicando su equivalencia con k-NN Rule.

5.1.6. k-NN Rule

Desbalanceado: Acc = 0.98 ± 0.02, F1 = 0.91 ± 0.12; **Balanceado:** Acc = 0.92 ± 0.05, F1 = 0.91 ± 0.05

Este método clásico obtiene resultados excelentes, especialmente en desbalanceado (F1=0.91), con k óptimo variando según balanceo: $k = 1$ en desbalanceado, $k = 3$ en balanceado. La simplicidad del método (asignación directa por votación mayoritaria sin estimar densidades) resulta ventajosa: con $k = 1$ en desbalanceado, iguala a Parzen y k-NN Density, confirmando que la estructura espacial está bien definida. En balanceado, $k = 3$ aporta ligero suavizado (F1=0.91 vs 0.93 de Parzen), con matriz de confusión mostrando confusiones reptiles-anfibios (2+11=13 casos) coherentes con otros métodos. La búsqueda revela que k grandes degradan rendimiento: $k = 11$ baja a F1≈0.51-0.89, confirmando que promediar muchos vecinos borra fronteras de decisión locales. La alta varianza en balanceado (std=0.05 vs 0.03 de Parzen) sugiere mayor sensibilidad a la composición de folds, pero el rendimiento global permanece competitivo.

6. Discusión y Conclusiones

Comparación general de enfoques: Los modelos paramétricos muestran comportamientos complementarios: Naive Bayes es robusto con clases minoritarias (F1=0.90 desbalanceado) estimando solo $2d$ parámetros por clase, mientras MLE Full colapsa en desbalanceado (F1=0.58) pero alcanza F1=0.89 con balanceo, igualando a NB. Esto confirma el equilibrio sesgo-varianza: modelos simples son robustos con datos escasos; complejos necesitan más muestras pero capturan correlaciones valiosas. Entre no paramétricos, Histogram Bayes falla completamente (F1=0.06-0.17) por maldición de dimensionalidad: $2^{16} = 65,536$ bins con 287 muestras resultan en colapso predictivo. Los tres métodos de vecinos cercanos (Parzen, k-NN Density, k-NN Rule) logran rendimiento excepcional (F1=0.91-0.93 balanceado) mediante suavizado continuo o asignación directa. Parzen con $h = 0.5$ obtiene máximo F1=0.93 y mínima varianza (std=0.03), confirmándose como mejor modelo global. k-NN Density con $k = 1$ iguala a Parzen (F1=0.93), mientras k-NN Rule con $k = 3$ queda ligeramente atrás (F1=0.91), sugiriendo que votación mayoritaria es menos precisa que estimación de densidad en este problema.

Impacto del desbalance y balanceo: El desbalance original (meanIR=4.84) expone brutalmente debilidades de modelos complejos: MLE Full colapsa con clases minoritarias (F1=0.58), generando precision=0.00 para reptiles y anfibios al no poder estimar covarianzas con 3-5 ejemplos. El balanceo sintético ($101 \rightarrow 287$ instancias) transforma MLE Full: F1=$0.58 \rightarrow 0.89$ (+53%), alcanzando rendimiento competitivo. Histogram Bayes muestra paradoja accuracy: 0.46→0.15 (empeora), pero F1 permanece consistente en colapso (0.17→0.06), confirmando que predice clase mayoritaria por defecto. Los métodos de vecinos cercanos mejoran moderadamente con balanceo (Parzen: 0.91→0.93) pero exhiben mejora crítica en estabilidad: desviaciones estándar reducen dramáticamente (Parzen: 0.12→0.03, k-NN Density: 0.12→0.04), revelando que con 41 ejemplos/clase las estimaciones locales son robustas. Naive Bayes se mantiene estable (F1: 0.90→0.87) con reducción de varianza (0.14→0.08), confirmando su robustez intrínseca.

Métricas: El contraste accuracy vs F1-macro es crucial: Histogram obtiene accuracy=0.46 pero F1=0.17, prediciendo predominantemente "mamífero". El accuracy engaña porque predecir clase mayoritaria acierta 40% en desbalanceado; F1-macro revela el colapso real al promediar todas las clases equitativamente.

Conclusiones finales: Este trabajo confirma empíricamente tres principios del aprendizaje automático: (1) el equilibrio entre complejidad del modelo y cantidad de datos (MLE Full: F1=0.58→0.89 con balanceo, +53%; Naive Bayes mantiene F1≈0.87-0.90 consistentemente); (2) la maldición de dimensionalidad (Histogram colapsa con F1=0.06-0.17; métodos de vecinos cercanos alcanzan F1=0.91-0.93); y (3) la importancia crítica de estabilidad (balanceo reduce desviaciones: Parzen 0.12→0.03, k-NN Density 0.12→0.04). **Parzen Windows con $h = 0.5$** emerge como mejor modelo (F1=0.93±0.03 balanceado), combinando alto rendimiento y mínima varianza. **Lecciones clave:** validación cruzada estratificada optimiza hiperparámetros críticos ($h = 0.5$ Parzen, $k = 1$ k-NN Density, $k = 3$ k-NN Rule); F1-macro es métrica esencial en multiclase desbalanceado (Histogram: accuracy=0.46 pero F1=0.17); la asunción "naive" es sorprendentemente efectiva (F1=0.87-0.90) sugiriendo independencia aproximada entre características binarias; confusiones reptiles-anfibios (10-13 casos) son biológicamente coherentes dado solapamiento de atributos (huevos, respira, venenoso); y el balanceo sintético mediante investigación profunda produce datos válidos que estabilizan modelos complejos.