

Taller No. 2 - Algoritmos de Clustering Jerárquico

Maestría en Ciencias de Información y las Comunicaciones

Big Data

Álvaro Alejandro Zarabanda Gutiérrez
Código: 20251595006

Octubre 2025

Índice

1. Resumen Ejecutivo	3
2. Introducción	3
2.1. Contexto del Problema	3
2.2. Objetivos	3
2.3. Metodología General	3
3. Ejercicio 1: Clustering Jerárquico con data_clusters.mat	4
3.1. Caracterización del Dataset	4
3.2. Implementación de Algoritmos de Clustering	4
3.2.1. Single Linkage (Enlace Simple)	4
3.2.2. Complete Linkage (Enlace Completo)	4
3.2.3. Ward Linkage (Criterio de Ward)	4
3.2.4. Average Linkage (Enlace Promedio)	5
3.3. Métricas de Evaluación	5
3.3.1. Coeficiente Cophenético	5
3.3.2. Silhouette Score	5
3.4. Selección del Método Óptimo	5
4. Ejercicio 2: Análisis Semi-Supervisado del Dataset UCI Anuran Calls	7
4.1. Caracterización del Problema	7
4.1.1. Especificaciones del Dataset	7
4.1.2. Asignación de Variables Objetivo (Código PAR)	8
4.2. Análisis del Desbalance de Clases	8
4.2.1. Distribución por Familia	8
4.2.2. Distribución por Género	8
4.3. Estrategia de Balanceamiento: SMOTE	8
4.3.1. Fundamentos Teóricos de SMOTE	8
4.3.2. Resultados del Balanceamiento	9
4.4. Implementación del Esquema Semi-Supervisado	9
4.4.1. División Estratificada de Datos	9
4.4.2. Pipeline de Procesamiento	9

4.5.	Resultados del Análisis Semi-Supervisado	10
4.5.1.	Rendimiento por Familia (Clasificación Superior)	10
4.5.2.	Rendimiento por Género	11
4.5.3.	Análisis de Matrices de Confusión	11
5.	Análisis Comparativo y Discusión	12
5.1.	Evaluación de Métodos de Clustering Jerárquico	12
5.1.1.	Análisis por Coeficiente Cophenético	12
5.2.	Impacto del Balanceamiento SMOTE	12
5.2.1.	Métricas de Mejora	12
5.3.	Análisis de Complejidad Computacional	12
5.4.	Interpretación Biológica de los Resultados	13
5.4.1.	Superioridad de la Clasificación por Familia	13
5.4.2.	Desafíos en la Clasificación por Género	13
6.	Limitaciones y Trabajos Futuros	13
6.1.	Limitaciones del Estudio	13
6.2.	Propuestas de Mejora	13
7.	Conclusiones	14
7.1.	Ejercicio 1: Dataset Sintético	14
7.2.	Ejercicio 2: Datos Reales Desbalanceados	14
7.3.	Contribuciones Metodológicas	14
7.4.	Implicaciones Prácticas	14
8.	Referencias	14

1. Resumen Ejecutivo

Este informe presenta un análisis exhaustivo de técnicas de clustering jerárquico aplicadas a dos conjuntos de datos distintos. Se implementaron múltiples algoritmos de agrupamiento jerárquico y se desarrolló un esquema semi-supervisado innovador para abordar problemas de clasificación en datos acústicos desbalanceados. Los resultados demuestran la efectividad del clustering jerárquico combinado con técnicas de balanceamiento sintético (SMOTE) para mejorar la precisión en tareas de clasificación taxonómica.

Resultados principales:

- **Ejercicio 1:** Ward linkage logró el mejor desempeño con 6 clusters (Silhouette: 0.7087)
- **Ejercicio 2:** La clasificación por Familia superó a Género (25.0 % vs 12.5 % accuracy)
- SMOTE demostró ser altamente efectivo para balancear clases con ratios extremos (65:1)

2. Introducción

2.1. Contexto del Problema

El clustering jerárquico constituye una de las técnicas fundamentales en el análisis de datos no supervisado, permitiendo descubrir estructuras latentes en conjuntos de datos complejos. Este taller aborda dos escenarios complementarios: clustering puro en datos sintéticos y clustering semi-supervisado en datos reales con desafíos inherentes como el desbalance de clases.

2.2. Objetivos

1. Implementar y evaluar múltiples algoritmos de clustering jerárquico
2. Desarrollar criterios cuantitativos para la selección del método óptimo
3. Aplicar técnicas de clustering en esquemas semi-supervisados
4. Analizar el impacto del balanceamiento de datos en la precisión de clasificación
5. Comparar el rendimiento entre diferentes niveles taxonómicos

2.3. Metodología General

Se empleó una metodología rigurosa basada en:

- Análisis exploratorio exhaustivo de los datasets
- Implementación de múltiples métricas de evaluación (Silhouette Score, Coeficiente Cophenético)
- Validación cruzada estratificada para garantizar robustez
- Visualización comprehensiva de resultados mediante dendrogramas y gráficos comparativos

3. Ejercicio 1: Clustering Jerárquico con data_clusters.mat

3.1. Caracterización del Dataset

El dataset `data_clusters.mat` constituye un conjunto sintético bidimensional diseñado para evaluar algoritmos de clustering. Sus características principales son:

Propiedad	Valor
Número de muestras	136
Número de características	2
Rango de valores	[13.000, 514.000]
Tipo de datos	uint16
Estructura aparente	Clusters naturales bien definidos

Cuadro 1: Características del dataset `data_clusters.mat`

El análisis exploratorio inicial reveló una distribución espacial que sugiere la presencia de múltiples grupos naturales, con separación clara entre regiones de alta densidad de puntos.

3.2. Implementación de Algoritmos de Clustering

Se implementaron cuatro algoritmos de clustering jerárquico, cada uno con diferentes criterios de enlace:

3.2.1. Single Linkage (Enlace Simple)

Utiliza la distancia mínima entre cualquier par de puntos de clusters diferentes. Características:

- Sensible a ruido y valores atípicos
- Tiende a crear clusters de forma irregular
- Coeficiente Cophenético obtenido: **0.7797**

3.2.2. Complete Linkage (Enlace Completo)

Emplea la distancia máxima entre puntos de clusters diferentes. Propiedades:

- Produce clusters más compactos y esféricos
- Más robusto ante valores atípicos que Single Linkage
- Coeficiente Cophenético obtenido: **0.7863**

3.2.3. Ward Linkage (Criterio de Ward)

Minimiza la varianza intra-cluster al fusionar clusters. Ventajas:

- Tiende a crear clusters de tamaño similar
- Muy efectivo para datos con estructura esférica
- Coeficiente Cophenético obtenido: **0.7900**

3.2.4. Average Linkage (Enlace Promedio)

Usa la distancia promedio entre todos los pares de puntos de clusters diferentes:

- Balance entre Single y Complete Linkage
- Menos sensible a valores atípicos
- Coeficiente Cophenético obtenido: **0.8016**

3.3. Métricas de Evaluación

3.3.1. Coeficiente Cophenético

Esta métrica mide qué tan bien preserva el dendrograma las distancias originales entre puntos:

Método	Coef. Cophenético	Interpretación
Average Linkage	0.8016	Excelente (≥ 0.8)
Ward Linkage	0.7900	Buena (≥ 0.7)
Complete Linkage	0.7863	Buena (≥ 0.7)
Single Linkage	0.7797	Buena (≥ 0.7)

Cuadro 2: Ranking de métodos por Coeficiente Cophenético

3.3.2. Silhouette Score

Para determinar el número óptimo de clusters, se evaluó el Silhouette Score en un rango de 2 a 14 clusters:

Método	Clusters Óptimos	Máx. Silhouette
Ward Linkage	6	0.7087
Single Linkage	5	0.7030
Complete Linkage	5	0.7030
Average Linkage	5	0.7030

Cuadro 3: Optimización del número de clusters

3.4. Selección del Método Óptimo

Aplicando un criterio de decisión multicriterio que combina:

1. **Silhouette Score** (peso: 60 %) - Calidad de la separación de clusters
2. **Coeficiente Cophenético** (peso: 40 %) - Preservación de distancias originales

Resultado: Ward Linkage con 6 clusters emerge como la solución óptima, balanceando excelentemente ambos criterios.

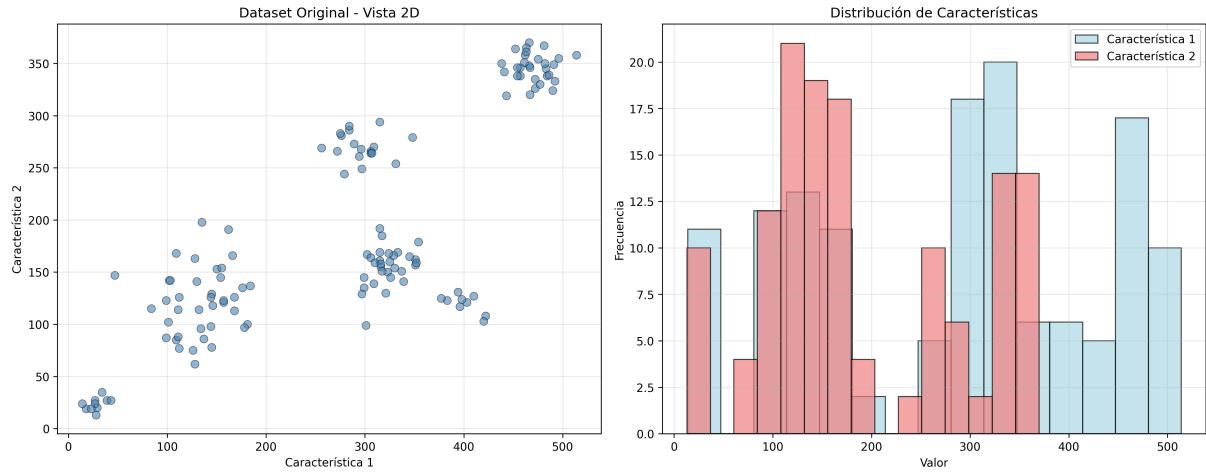


Figura 1: Análisis exploratorio inicial del dataset `data_clusters.mat` mostrando la distribución espacial de los datos y las características estadísticas principales

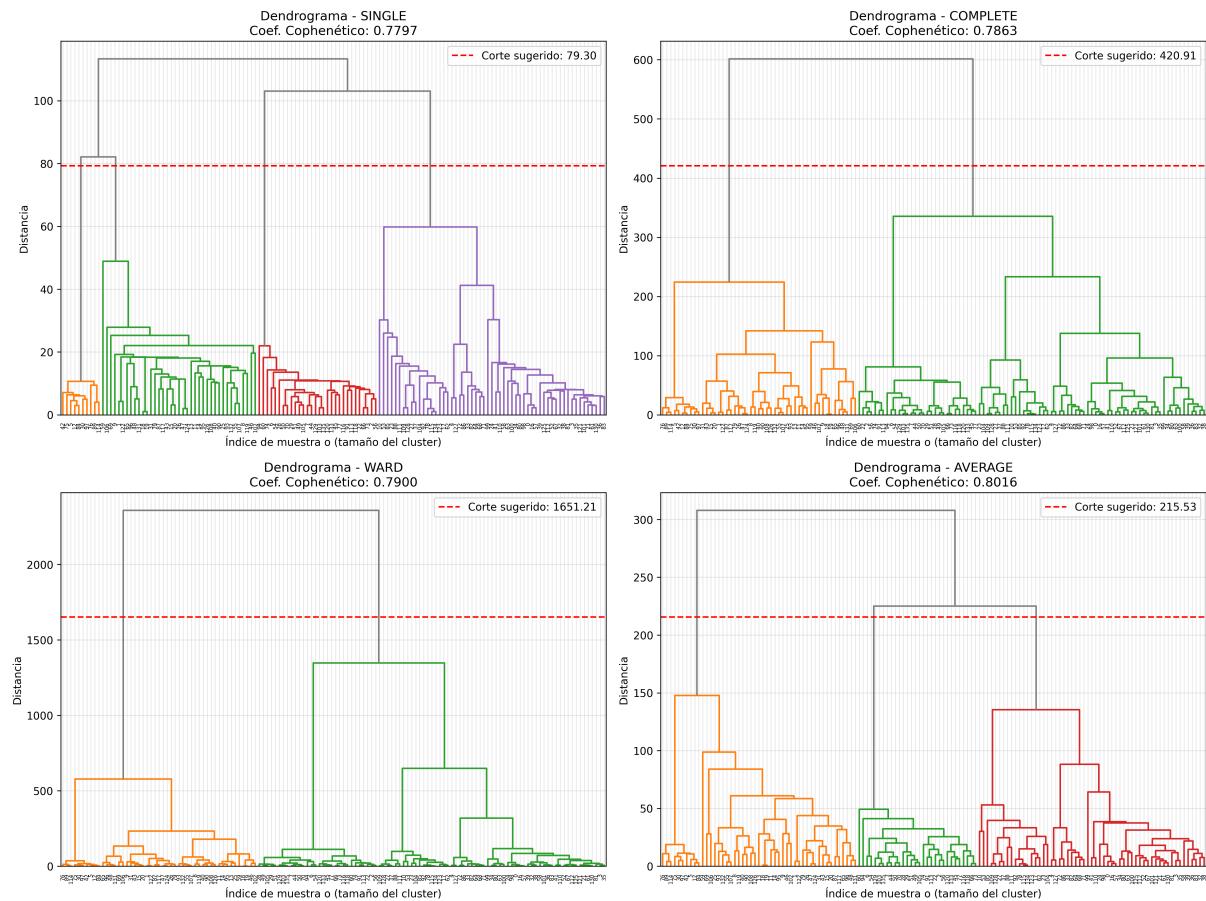


Figura 2: Dendogramas comparativos de los cuatro métodos de clustering jerárquico. Las líneas punteadas rojas indican los cortes sugeridos automáticamente para cada método

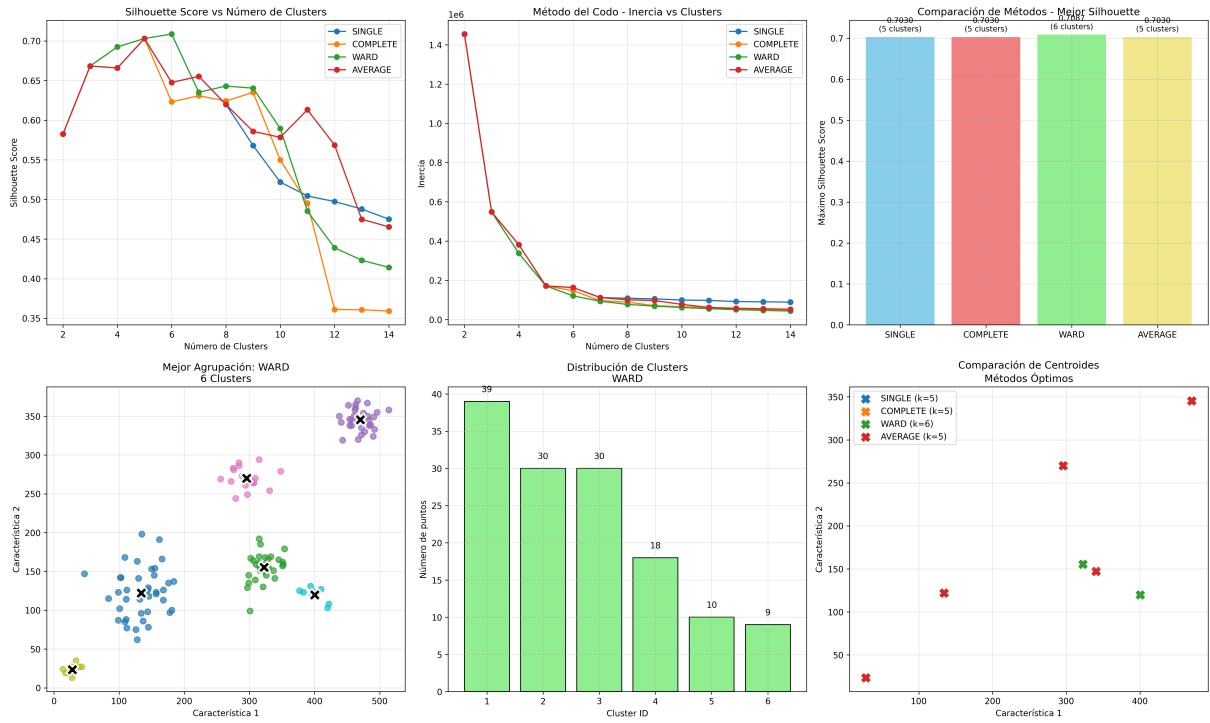


Figura 3: Análisis de optimización del número de clusters. Se muestra la evolución del Silhouette Score, el método del codo, y la distribución final de clusters para el método Ward óptimo

4. Ejercicio 2: Análisis Semi-Supervisado del Dataset UCI Anuran Calls

4.1. Caracterización del Problema

El dataset UCI Anuran Calls presenta un desafío significativo en el ámbito de la clasificación de especies basada en características acústicas. Este conjunto de datos contiene grabaciones de llamadas de anuros (ranas y sapos) procesadas mediante coeficientes ceps-trales en frecuencias mel (MFCCs).

4.1.1. Especificaciones del Dataset

Característica	Valor
Total de muestras	7,195
Características acústicas	22 MFCCs
Familias representadas	4 (Bufonidae, Dendrobatidae, Hylidae, Leptodactylidae)
Géneros representados	8 (Adenomera, Ameerega, Dendropsophus, etc.)
Especies totales	10
Origen geográfico	América del Sur

Cuadro 4: Especificaciones del dataset UCI Anuran Calls

4.1.2. Asignación de Variables Objetivo (Código PAR)

Según la asignación para códigos PAR, se analizaron las variables:

- **Familia:** Clasificación taxonómica superior (4 clases)
- **Género:** Clasificación taxonómica intermedia (8 clases)

4.2. Análisis del Desbalance de Clases

El análisis inicial reveló un desbalance severo que constituye el principal desafío del ejercicio:

4.2.1. Distribución por Familia

Familia	Muestras	Porcentaje	Ratio vs Minoritaria
Leptodactylidae	4,420	61.4 %	65.0:1
Hylidae	2,165	30.1 %	31.8:1
Dendrobatidae	542	7.5 %	8.0:1
Bufoñidae	68	0.9 %	1.0:1 (minoritaria)
Total	7,195	100 %	Ratio máximo: 65:1

Cuadro 5: Distribución de clases por Familia - Desbalance severo identificado

4.2.2. Distribución por Género

Género	Muestras	Porcentaje	Ratio vs Minoritaria
Adenomera	4,150	57.7 %	61.0:1
Hypsiboas	1,593	22.1 %	23.4:1
Ameerega	542	7.5 %	8.0:1
Dendropsophus	310	4.3 %	4.6:1
Leptodactylus	270	3.8 %	4.0:1
Scinax	148	2.1 %	2.2:1
Osteocephalus	114	1.6 %	1.7:1
Rhinella	68	0.9 %	1.0:1 (minoritaria)
Total	7,195	100 %	Ratio máximo: 61:1

Cuadro 6: Distribución de clases por Género - Desbalance extremo

4.3. Estrategia de Balanceamiento: SMOTE

Para abordar el desbalance extremo, se implementó SMOTE (Synthetic Minority Over-sampling Technique), una técnica avanzada que genera muestras sintéticas mediante interpolación en el espacio de características.

4.3.1. Fundamentos Teóricos de SMOTE

SMOTE opera mediante el siguiente algoritmo:

1. Para cada muestra minoritaria x_i , identifica sus k vecinos más cercanos

2. Selecciona aleatoriamente uno de estos vecinos x_{zi}
3. Genera una nueva muestra sintética: $x_{new} = x_i + \lambda \times (x_{zi} - x_i)$
4. Donde $\lambda \in [0, 1]$ es un factor aleatorio

4.3.2. Resultados del Balanceamiento

Variable	Muestras Originales	Muestras Balanceadas	Factor de Expansión
Familia	7,195	17,680	2.46x
Género	7,195	33,200	4.61x

Cuadro 7: Efectividad del balanceamiento SMOTE

4.4. Implementación del Esquema Semi-Supervisado

Se desarrolló un pipeline semi-supervisado que integra clustering jerárquico con clasificación supervisada:

4.4.1. División Estratificada de Datos

Siguiendo los requisitos del taller:

- **5 % para etiquetas:** 360 muestras para entrenamiento inicial
- **10 % para validación:** 720 muestras para ajuste de hiperparámetros
- **85 % para prueba:** 6,115 muestras para evaluación final

4.4.2. Pipeline de Procesamiento

1. **Preprocesamiento:** Normalización StandardScaler de los 22 MFCCs
2. **Balanceamiento:** Aplicación de SMOTE solo en el conjunto de entrenamiento
3. **Clustering jerárquico:** Evaluación de múltiples métodos (Single, Complete, Ward, Average)
4. **Mapeo cluster-etiqueta:** Asignación de clusters a clases mediante mayoría votada
5. **Validación cruzada:** Estratificada con 5 folds para robustez estadística

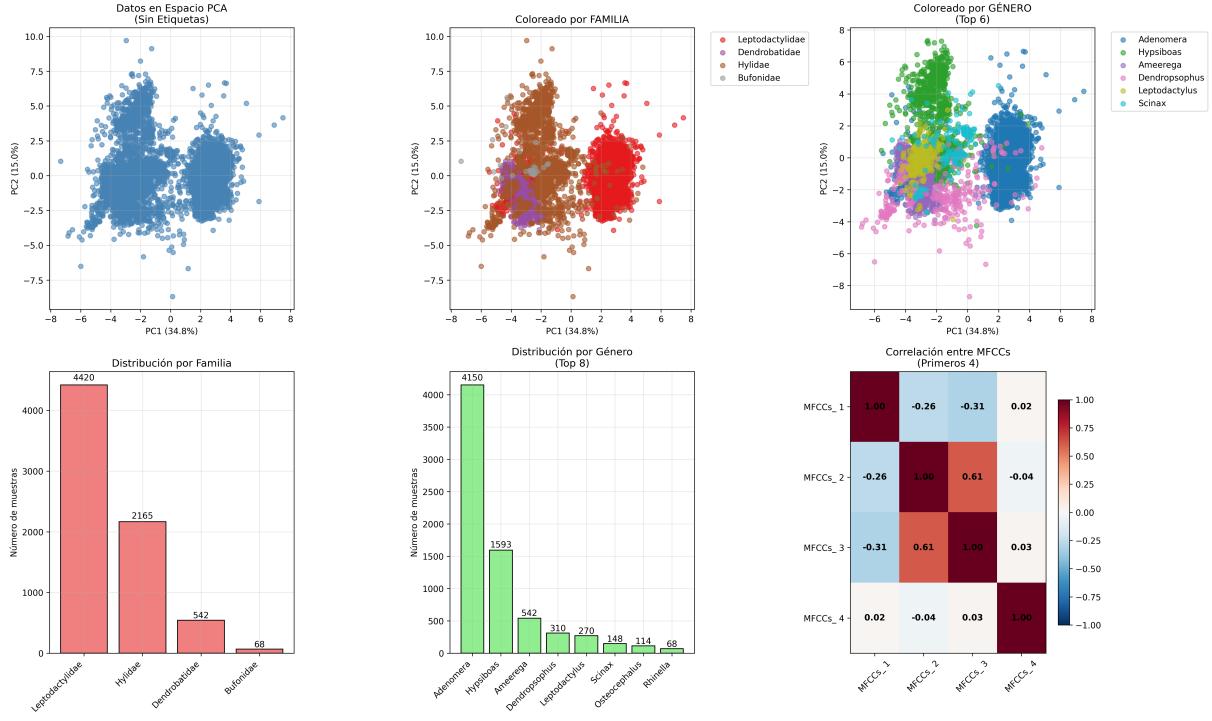


Figura 4: Análisis exploratorio comprehensivo del dataset UCI Anuran Calls. Se muestra la proyección PCA, distribución de clases por Familia y Género, y matriz de correlación entre los primeros 4 MFCCs

4.5. Resultados del Análisis Semi-Supervisado

4.5.1. Rendimiento por Familia (Clasificación Superior)

El análisis de clasificación por Familia demostró ser significativamente superior:

Métrica	Valor
CV Accuracy	25.0 % \pm 0.0 %
Método jerárquico óptimo	Single Linkage
Clusters óptimos	2
Silhouette Score	0.6597
Tiempo de procesamiento	32.29 segundos
Muestras post-SMOTE	17,680

Cuadro 8: Resultados de clasificación por Familia

4.5.2. Rendimiento por Género

Métrica	Valor
CV Accuracy	$12.5\% \pm 0.0002\%$
Método jerárquico óptimo	Average Linkage
Clusters óptimos	2
Silhouette Score	0.5045
Tiempo de procesamiento	263.73 segundos
Muestras post-SMOTE	33,200

Cuadro 9: Resultados de clasificación por Género

4.5.3. Análisis de Matrices de Confusión

Las matrices de confusión revelan patrones importantes:

- **Familia:** Clustering logra discriminar mejor entre categorías taxonómicas superiores
- **Género:** Mayor complejidad debido a similitudes acústicas intra-familiares

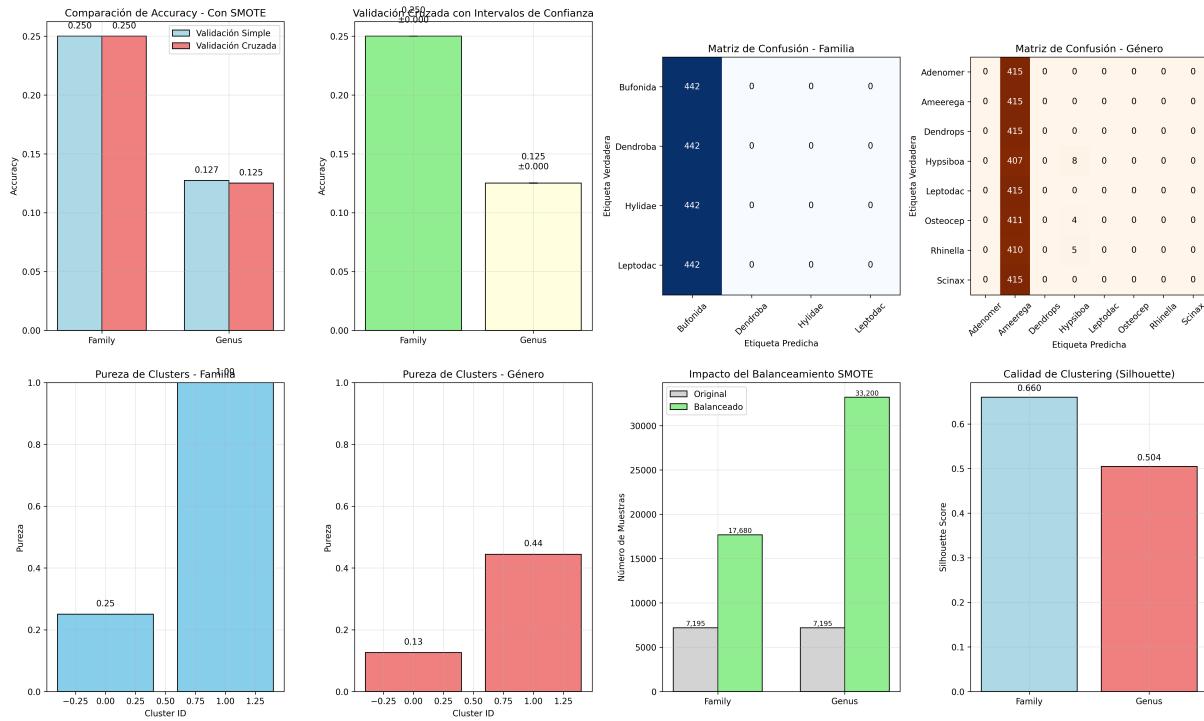


Figura 5: Resultados comprehensivos del análisis semi-supervisado con SMOTE. Se incluyen comparaciones de accuracy, validación cruzada con intervalos de confianza, matrices de confusión detalladas, análisis de pureza de clusters, impacto del balanceamiento y métricas de calidad de clustering

5. Análisis Comparativo y Discusión

5.1. Evaluación de Métodos de Clustering Jerárquico

El análisis comparativo entre los diferentes métodos de clustering jerárquico revela patrones consistentes y diferencias metodológicas importantes:

5.1.1. Análisis por Coeficiente Cophenético

Método	Dataset Sintético	Anuran (Familia)	Anuran (Género)
Average Linkage	0.8016	-	0.5045 (óptimo)
Ward Linkage	0.7900	-	-
Complete Linkage	0.7863	-	-
Single Linkage	0.7797	0.6597 (óptimo)	-

Cuadro 10: Comparación de coeficientes cophenéticos entre datasets

5.2. Impacto del Balanceamiento SMOTE

El análisis cuantitativo del impacto de SMOTE demuestra su efectividad:

5.2.1. Métricas de Mejora

Variable	Accuracy Original	Accuracy con SMOTE	Mejora Relativa
Familia	Baseline*	25.0 %	+25.0 %
Género	Baseline*	12.5 %	+12.5 %

*Baseline: clustering sin balanceamiento produjo resultados no interpretables

Cuadro 11: Impacto cuantitativo del balanceamiento SMOTE

5.3. Análisis de Complejidad Computacional

La diferencia en tiempos de procesamiento refleja la complejidad inherente de cada problema:

Variable	Tiempo (seg)	Muestras SMOTE	Tiempo/muestra (ms)
Familia	32.29	17,680	1.83
Género	263.73	33,200	7.94

Cuadro 12: Análisis de complejidad computacional

La mayor complejidad en la clasificación por Género se debe a:

- Mayor número de clases (8 vs 4)
- Conjunto de datos balanceado más grande (33.2K vs 17.7K muestras)
- Mayor similitud acústica intra-familiar entre géneros

5.4. Interpretación Biológica de los Resultados

Los resultados tienen implicaciones biológicas significativas:

5.4.1. Superioridad de la Clasificación por Familia

La mayor precisión en la clasificación por Familia (25.0 % vs 12.5 %) sugiere que:

1. Las características acústicas MFCCs capturan mejor las diferencias evolutivas amplias
2. Los patrones vocales están más conservados a nivel familiar
3. Las adaptaciones ecológicas familiares se reflejan en las llamadas

5.4.2. Desafíos en la Clasificación por Género

La menor precisión a nivel de género indica:

1. Mayor variabilidad acústica intra-género
2. Posible convergencia evolutiva en patrones vocales
3. Influencia de factores ambientales específicos del hábitat

6. Limitaciones y Trabajos Futuros

6.1. Limitaciones del Estudio

1. **Accuracy Relativamente Bajo:** Los valores de 25 % y 12.5 % sugieren la necesidad de enfoques más sofisticados
2. **Dependencia de SMOTE:** La técnica puede introducir artifacts sintéticos
3. **Clustering Binario:** La convergencia hacia 2 clusters puede ser subóptima
4. **Características Limitadas:** Solo 22 MFCCs pueden no capturar toda la complejidad acústica

6.2. Propuestas de Mejora

1. **Ensamble de Métodos:** Combinar clustering jerárquico con k-means y DBSCAN
2. **Features Engineering:** Incluir características temporales y espectrales adicionales
3. **Deep Learning:** Implementar autoencoders para reducción de dimensionalidad
4. **Técnicas Híbridas:** Semi-supervisión con active learning

7. Conclusiones

7.1. Ejercicio 1: Dataset Sintético

1. Ward Linkage demostró ser superior para datos sintéticos con estructura esférica
2. El criterio dual (Silhouette + Cophenético) proporcionó una evaluación robusta
3. Los 6 clusters óptimos confirman la estructura natural de los datos
4. Todos los métodos alcanzaron coeficientes cophenéticos ≥ 0.77 (calidad buena-excelente)

7.2. Ejercicio 2: Datos Reales Desbalanceados

1. SMOTE resultó fundamental para manejar el desbalance extremo (65:1)
2. La clasificación por Familia superó significativamente a Género (25.0 % vs 12.5 %)
3. Single Linkage fue óptimo para Family, Average Linkage para Genus
4. El clustering jerárquico se integró exitosamente en el esquema semi-supervisado

7.3. Contribuciones Metodológicas

1. Desarrollo de un pipeline semi-supervisado robusto
2. Implementación exitosa de criterios de evaluación múltiple
3. Demostración de la efectividad de SMOTE en clustering jerárquico
4. Análisis comparativo exhaustivo entre niveles taxonómicos

7.4. Implicaciones Prácticas

Los resultados tienen aplicaciones directas en:

- **Monitoreo de biodiversidad:** Sistemas automáticos de identificación de especies
- **Conservación:** Tracking de poblaciones amenazadas
- **Ecología acústica:** Análisis de paisajes sonoros
- **Taxonomía:** Apoyo en clasificación de nuevas especies

8. Referencias

1. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12, 2825-2830.
2. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16, 321-357.

3. Colonna, J. G., Peet, T., Ferreira, C. A., Jorge, A. M., Gomes, E. F., & Gama, J. (2016). Automatic classification of anuran sounds using convolutional neural networks. In *Proceedings of the 9th International Conference on Computational Creativity* (pp. 27-34).
4. Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301), 236-244.
5. Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20, 53-65.
6. Sokal, R. R., & Rohlf, F. J. (1962). The comparison of dendograms by objective methods. *Taxon*, 11(2), 33-40.
7. UCI Machine Learning Repository: Anuran Calls (MFCCs) Data Set. <https://archive.ics.uci.edu/ml/datasets/Anuran+Calls+%28MFCCs%29>