INFORME EJECUTIVO

Clustering y Análisis PCA en Dataset de Setas

Aplicación de Técnicas de Machine Learning No Supervisado

Fecha: Mayo 2025

Proyecto: Workshop Clustering y PCA

Equipo: Polina Pavlova y Pepe Ruiz

Dataset: Mushroom Classification (8,124 observaciones)

Objetivo: Evaluar la efectividad del clustering no supervisado para identificar setas comestibles vs. venenosas

---

1. RESUMEN EJECUTIVO

El análisis implementó técnicas avanzadas de machine learning no supervisado sobre un dataset de 8,124 setas con 22 características morfológicas relevantes. Los resultados demuestran que **el clustering K-Means logró una separación casi perfecta** entre setas comestibles y venenosas **sin utilizar etiquetas previas**, alcanzando una pureza del **99.8%** para setas venenosas y **84.1%** para comestibles.

Hallazgos Clave

- **Separabilidad natural excepcional:** Las características morfológicas permiten distinción automática de toxicidad
- **Reducción dimensional ultra-eficiente:** PCA reveló que solo **5 de 95 variables** contienen 99.4% de la información discriminativa
- **Clustering superior al esperado:** K-Means identificó patrones biológicos reales con **90.1% de precisión global**

---

2. METODOLOGÍA Y PROCESAMIENTO DE DATOS

2.1 Preparación del Dataset

- **Volumen:** 8,124 setas × 22 variables categóricas (eliminada `veil-type` por falta de variabilidad)
- **Calidad:** Sin valores nulos, dataset completo
- **Encoding:** Transformación a **95 variables binarias** mediante One-Hot Encoding
- **Valores faltantes:** 2,480 casos (30.5%) con valores faltantes en `stalk-root` tratados como categoría independiente "missing" por ser MNAR (omitidos no por azar)

2.2 División de Datos

- **Entrenamiento:** 5,443 observaciones (67%)
- **Test:** 2,681 observaciones (33%)
- **Balance de clases:** 52% comestibles vs. 48% venenosas (bien balanceado)

---

3. ANÁLISIS DE COMPONENTES PRINCIPALES (PCA)

3.1 Reducción de Dimensionalidad

El análisis PCA reveló una **estructura dimensional ultra-eficiente**:

Componentes	Precisión Clasificación	Reducción Dimensional
2 componentes	**92.58%**	97.9% reducción
5 componentes	**99.37%**	**94.7% reducción**
10 componentes	**99.89%**	89.5% reducción
15 componentes	99.93%	84.2% reducción
20 componentes	99.96%	78.9% reducción

3.2 Visualización 2D

- **PC1: 18.3%** de varianza explicada
- **PC2: 12.8%** de varianza explicada

- **Total: 31.1%** de información original en 2 dimensiones
- **Resultado:** Separación visual clara entre clases, confirmando estructura natural agrupable

---

4. ANÁLISIS DE CLUSTERING

4.1 Selección de K Óptimo

Aplicación del **método del codo** para determinar número óptimo de clusters:

K	Inercia (WCSS)	Reducción
1	48,083	-
2	**40,034**	**-8,049**
3	34,369	-5,665
4	30,735	-3,634
5	28,726	-2,009

Decisión: K=2 muestra la mayor caída inicial, coincidiendo con estructura biológica real (binario: comestible o no)

4.2 Resultados K-Means (K=2)

Cluster	Comestibles	Venenosas	Pureza	Interpretación
Cluster 0	**5**	**2,080**	**99.8%**	Casi exclusivamente venenosas
Cluster 1	**2,825**	**533**	**84.1%**	Mayormente comestibles

Precisión global del clustering: 90.1% (4,905 clasificaciones correctas de 5,443 total)

4.3 Validación Visual

La comparación entre clusters K-Means y clases reales en espacio PCA mostró **coincidencia espacial extraordinaria**, confirmando que el algoritmo detectó la estructura biológica natural sin conocimiento previo.

---

5. COMPARACIÓN DE RENDIMIENTO

5.1 Métodos Supervisados vs. No Supervisados

Método	Precisión	Información Requerida	Aplicabilidad
Random Forest	**100.00%**	Etiquetas conocidas	Clasificación supervisada
K-Means	**90.1%**	Sin etiquetas	Exploración automática

5.2 Valor del Clustering

- **Descubrimiento automático** de patrones sin conocimiento previo
- **Identificación de setas peligrosas** con **99.8%** de precisión
- **Aplicabilidad en campo** para especies no catalogadas previamente
- **Eficiencia computacional** usando solo 5 componentes PCA

---

6. CONCLUSIONES Y RECOMENDACIONES

6.1 Conclusiones Técnicas

1. **Estructura dimensional ultra-simple:** Solo **5 componentes PCA** capturan 99.4% de la información discriminativa
2. **Separabilidad natural excepcional:** Las características morfológicas permiten distinción automática de toxicidad con 90% de precisión
3. **Clustering altamente efectivo:** K-Means detectó patrones biológicos reales, especialmente efectivo para identificar setas venenosas (99.8% precisión)

6.2 Aplicaciones Prácticas

- **Herramienta de campo:** Sistema de alerta temprana con 99.8% precisión para setas peligrosas
- **Clasificación preliminar:** Reducción de 95 características a solo 5 componentes principales
- **Investigación botánica:** Descubrimiento automático de patrones en especies no clasificadas

6.3 Recomendaciones

1. **Implementar PCA con 5 componentes:** Punto óptimo entre eficiencia (95% reducción) y precisión (99.4%)
2. **Sistema híbrido:** Clustering para exploración inicial + Random Forest para confirmación final
3. **Desarrollo de aplicación móvil:** Herramienta portátil usando los 5 componentes principales identificados

---

7. ESPECIFICACIONES TÉCNICAS

Herramientas utilizadas:

- Python 3.12 con scikit-learn, pandas, numpy
- Técnicas: PCA, K-Means, Random Forest, One-Hot Encoding
- Visualización: matplotlib, seaborn

Métricas de evaluación:

- Inercia (WCSS) para selección de K óptimo
- Pureza de clusters y precisión global para evaluación de calidad
- Varianza explicada para validación de PCA

Este análisis demuestra que el machine learning no supervisado es altamente efectivo para problemas críticos de seguridad alimentaria, logrando 90% de precisión general y 99.8% para identificación de setas venenosas sin conocimiento previo de toxicidad.