Informe Ejecutivo: Análisis de Clasificación de Hongos Comestibles y Venenosos

Resumen Ejecutivo

Se ha analizado el conjunto de datos "Mushroom Dataset" para desarrollar un modelo que clasifique hongos como comestibles o venenosos basado en sus características físicas. Los modelos de clasificación alcanzaron una precisión del 100%, y el análisis de clustering demostró que las características físicas de los hongos tienen una correlación natural con su comestibilidad (89% de precisión sin usar etiquetas).

Contexto y Objetivos

El dataset contiene 8,124 ejemplares de hongos categorizados como comestibles o venenosos, con 22 atributos categóricos que describen sus características físicas como color, forma, olor y hábitat. El objetivo principal fue:

- Identificar los atributos más predictivos de la comestibilidad
- Desarrollar un modelo preciso de clasificación
- Explorar patrones naturales en los datos sin usar las etiquetas

Metodología

1. ***Preprocesamiento de datos:***
    - ❖ Tratamiento de valores ausentes (30% en "stalk-root") como categoría válida
    - ❖ Eliminación de características sin variabilidad o redundantes
    - ❖ Codificación one-hot de variables categóricas
2. ***Reducción de dimensionalidad:***
    - ❖ Aplicación de PCA para visualización y compresión de datos
    - ❖ Evaluación de rendimiento con diferentes números de componentes
3. ***Modelado:***
    - ❖ Comparación de Random Forest, SVM y KNN
    - ❖ Clustering con K-means sin usar etiquetas

Hallazgos Clave

Análisis de Características

- Odor (olor) mostró la mayor correlación con la comestibilidad (0.5438)
- Spore-print-color fue el segundo indicador más importante (0.4707)
- Se identificaron y eliminaron variables altamente correlacionadas entre sí

Rendimiento de Modelos de Clasificación

- **Random Forest:** Precisión 100%
- **SVM:** Precisión cercana al 100%
- **KNN:** Precisión superior al 99%

Reducción de Dimensionalidad

- Con solo 36 componentes principales (de más de 100 variables originales) se logró precisión del 100%
- Reducción del 65% en dimensionalidad sin pérdida de rendimiento
- La visualización con 2 componentes principales ya mostraba clara separación entre clases

Clustering

- K-means con k=2 logró separar automáticamente las setas con 89.15% de precisión
- Los clusters formados naturalmente corresponden en gran medida a la división comestible/venenoso

Conclusiones e Implicaciones

1. Las características físicas de los hongos permiten una clasificación extremadamente precisa
2. Existe una correlación natural entre atributos físicos y comestibilidad
3. El olor es el indicador más fuerte de sí un hongo es comestible o venenoso
4. El modelo tiene potencial para aplicaciones prácticas en identificación de hongos

Recomendaciones

- Implementar un sistema automatizado de identificación de hongos basado en Random Forest
- Centrarse en capturar con precisión los atributos de mayor importancia (especialmente olor)
- Para contextos sin etiquetado, utilizar clustering como aproximación inicial seguida de validación experta
- Considerar la reducción a 36 componentes principales para futuros modelos, optimizando eficiencia computacional