Informe Ejecutivo

Análisis Exploratorio de Datos – Clasificación de Hongos

1. Objetivo del Análisis

Este informe tiene como finalidad documentar los hallazgos obtenidos a partir del análisis exploratorio de un conjunto de datos sobre hongos, con el objetivo de identificar patrones y atributos que permitan discriminar de manera confiable entre especies comestibles y venenosas. Se han empleado técnicas de análisis exploratorio y métodos de aprendizaje no supervisado, con miras a evaluar la viabilidad de futuros modelos predictivos aplicables a contextos de salud pública, agricultura o recolección silvestre.

2. Descripción del Conjunto de Datos

El estudio se basa en un conjunto de datos proveniente del **UCI Machine Learning Repository**, reconocido por su uso académico y experimental. Este conjunto está conformado por:

- **Total de instancias**: 8.124 registros.
- **Características**: 22 variables categóricas, sin atributos numéricos.
- **Variable objetivo**: `class`, con dos categorías codificadas:
    - `e`: comestible (edible).
    - `p`: venenoso (poisonous).

Cabe destacar que todas las características representan observaciones morfológicas, lo cual subraya la utilidad de este dataset para el desarrollo de modelos explicativos e interpretables.

3. Principales Hallazgos del Análisis Exploratorio de Datos (EDA)

a. Balance de Clases

Uno de los primeros aspectos analizados fue la distribución de clases. A diferencia de otros datasets biológicos, este se encuentra razonablemente equilibrado:

- Comestibles: 4.208 registros (52%).
- Venenosos: 3.916 registros (48%).

Esta distribución equitativa resulta favorable para tareas de modelado, ya que reduce el riesgo de sesgo hacia una clase dominante.

b. Valores Faltantes

No se identificaron valores nulos explícitos. No obstante, la variable `stalk-root` presenta un valor no estándar (`?`) que, en lugar de ser eliminado o imputado, se optó por tratar como una categoría independiente. Esta decisión se justifica dado que el signo podría representar una falta estructural de información en ciertas condiciones de recolección.

c. Características Relevantes

Durante el análisis se identificaron variables con alto potencial discriminativo. Particularmente:

- `odor`: ciertos olores como `foul`, `fishy` y `spicy` aparecen fuertemente asociados con hongos venenosos.
- `gill-color` y `gill-size`: atributos morfológicos que presentan clara diferenciación por clase.
- `spore-print-color`: el color de la impresión de esporas también muestra un patrón relevante para la clasificación.
- `ring-type` y `stalk-surface-below-ring`: aportan información complementaria para segmentar especies.

d. Visualizaciones

Se emplearon herramientas gráficas como histogramas y análisis de Cramér's V para evaluar la fuerza de asociación entre variables categóricas. Las visualizaciones revelaron relaciones consistentes y patrones claros, lo que refuerza la premisa de que ciertos atributos pueden utilizarse eficazmente en modelos de clasificación.

4. Preparación para el Modelado

Con el fin de facilitar el uso de algoritmos de aprendizaje automático, se procedió a convertir todas las variables categóricas mediante **Label Encoding**. Esta transformación fue suficiente para los modelos utilizados, sin requerir imputación de datos ni escalado adicional.

5. Modelos de Aprendizaje Automático

Aunque el enfoque inicial fue exploratorio, se llevaron a cabo pruebas con algoritmos de agrupamiento (K-Means) y árboles de decisión como aproximación al modelado predictivo. Los resultados preliminares son prometedores: sin necesidad de etiquetas supervisadas, los algoritmos fueron capaces de generar agrupaciones alineadas con la variable objetivo, lo cual sugiere que la estructura interna del dataset es altamente informativa.

6. Conclusiones

A la luz del análisis efectuado, se concluye que el conjunto de datos en cuestión posee atributos suficientemente diferenciadores como para sustentar un modelo de clasificación robusto. Variables como `odor`, `gill-color` y `spore-print-color` emergen como pilares fundamentales para la discriminación entre hongos comestibles y venenosos.

Además, la calidad y equilibrio del dataset lo convierten en una excelente base para proyectos educativos, así como para desarrollos aplicados en entornos agrícolas, forestales o de salud ambiental.