🍄 Informe del Proyecto de Machine Learning: Clasificación de Comestibilidad de Setas

1. Resumen Ejecutivo y Objetivo del Proyecto

El objetivo principal de este proyecto fue crear un modelo de clasificación altamente preciso para predecir si una seta es **comestible** o **venenosa** (un problema de clasificación binaria). El análisis se basó en un conjunto de datos con 23 características categóricas.

Tras la limpieza y el preprocesamiento de los datos, el modelo **Random Forest** (un modelo de ensamble) logró una **separación perfecta** con una precisión del

$$100.00\%$$

$\mathbf{100.00\%}$. Las técnicas de análisis exploratorio como **PCA** y **K-Means** se utilizaron para entender la estructura y complejidad de los datos.

---

2. Preparación de los Datos (Fase de Limpieza)

Antes de aplicar cualquier algoritmo, es esencial limpiar y organizar los datos para asegurar su calidad y que los modelos puedan procesarlos correctamente.

Competencia	Acción Realizada	Importancia
Limpieza de Datos	Los **valores faltantes** (marcados con '?') se rellenaron con el valor más común (la moda). Se **eliminó** una columna ('p.2') que solo contenía un valor, ya que no aportaba información útil.	Garantiza que los datos sean completos y que cada característica contribuya información única.
Preprocesamiento	Todos los códigos categóricos (ej. 'e' para comestible) se convirtieron en **columnas numéricas** utilizando **One-Hot Encoding** y **Label Encoding**.	Los modelos de Machine Learning solo trabajan con números. Esta conversión amplió el conjunto inicial de 22 columnas a **116 características numéricas**.
División de Datos	Los datos se dividieron en un **Conjunto de Entrenamiento (67%)** y un **Conjunto de Prueba (33%)**.	El modelo se entrena con el primer conjunto y se evalúa con el segundo (datos no vistos) para medir su rendimiento en el mundo real.

---

3. Exploración de la Estructura y Complejidad de los Datos

Con 116 características, los datos son complejos. Utilizamos el **Análisis de Componentes Principales (PCA)** y **Clustering (K-Means)** para obtener perspectivas sobre su estructura.

A. Reducción de Dimensionalidad (PCA)

PCA nos ayuda a encontrar las combinaciones de características más importantes que capturan la mayor variación en los datos.

- **Visualización:** Redujimos los datos a solo **dos componentes** para crear un gráfico de dispersión 2D , coloreado por el objetivo.
    - **Conclusión:** El gráfico mostró regiones extensas donde los dos tipos de setas estaban **perfectamente separados**, indicando que la tarea de clasificación sería sencilla.
- **Retención de Información:** Verificamos cuántos componentes son necesarios para mantener el
- 95%
- **95% de la información original**.
    - **Conclusión:** Se necesitaron **109 de las 116 componentes**. Esto indica que, aunque hay muchas características, la mayoría son **necesarias** para describir completamente la varianza del conjunto de datos.

B. Clustering No Supervisado (K-Means)

Utilizamos el algoritmo **K-Means** (aprendizaje no supervisado) para ver cómo los datos se agrupan de forma natural, sin usar las etiquetas 'comestible' o 'venenosa'.

- **Clústeres Óptimos:** El **Método del Codo (Elbow Method)** sugirió que
- $K=2$
- **$K=2$** clústeres era la mejor opción.
- **Conclusión:** Al comparar los clústeres de K-Means con las etiquetas reales (edible/venenosa), la división fue muy **mezclada** (
- $\approx 50/50$
- $\approx 50/50$ en cada clúster) .
- **Implicación:** Esto demuestra que los factores que hacen que una seta sea venenosa **no siempre coinciden** con los factores que hacen que sus características parezcan similares. El modelo supervisado encuentra límites complejos que la simple agrupación por semejanza no puede detectar.

---

4. Modelado Supervisado y Resultados Finales

Elegimos el modelo **Random Forest** (Bosque Aleatorio)—un método de ensamble conocido por su precisión—para la clasificación.

Competencia	Modelo Utilizado	Resultado
Modelo de Clasificación	**Random Forest Classifier** (entrenado con características estandarizadas)	**Precisión en Pruebas:** 1.0000 **1.0000 (100%)**
Verificación de Rendimiento	Se reentrenó el modelo usando los 109 **109 componentes** de PCA.	**Precisión en Pruebas:** 1.0000 **1.0000 (100%)**

Conclusión Final

El modelo **Random Forest** alcanzó una **precisión perfecta** (

100%

100%) en la distinción entre setas comestibles y venenosas. Esto confirma que existe una **señal fuerte y limpia** en el conjunto de datos que permite una separación impecable. El modelo es extremadamente fiable para esta tarea de clasificación.

Todos los gráficos y artefactos producidos durante este análisis se guardan en los directorios `results/figures/` y `results/reports/`, asegurando que el trabajo sea completamente reproducible.