CLASIFICACIÓN DE HONGOS

Hurtado Medina, Isaac

Ingeniería Mecatrónica; Universidad EIA; Envigado

Este proyecto tiene como objetivo construir modelos predictivos capaces de determinar si un es comestible o venenoso basado en sus características físicas.

---

1. Introducción

El presente informe documenta el desarrollo de un proyecto de clasificación binaria utilizando el conjunto de datos 'Mushroom Dataset' del repositorio UCI Machine Learning. El objetivo principal fue construir modelos predictivos capaces de clasificar hongos como comestibles ('e') o venenosos ('p') en base a sus características morfológicas.

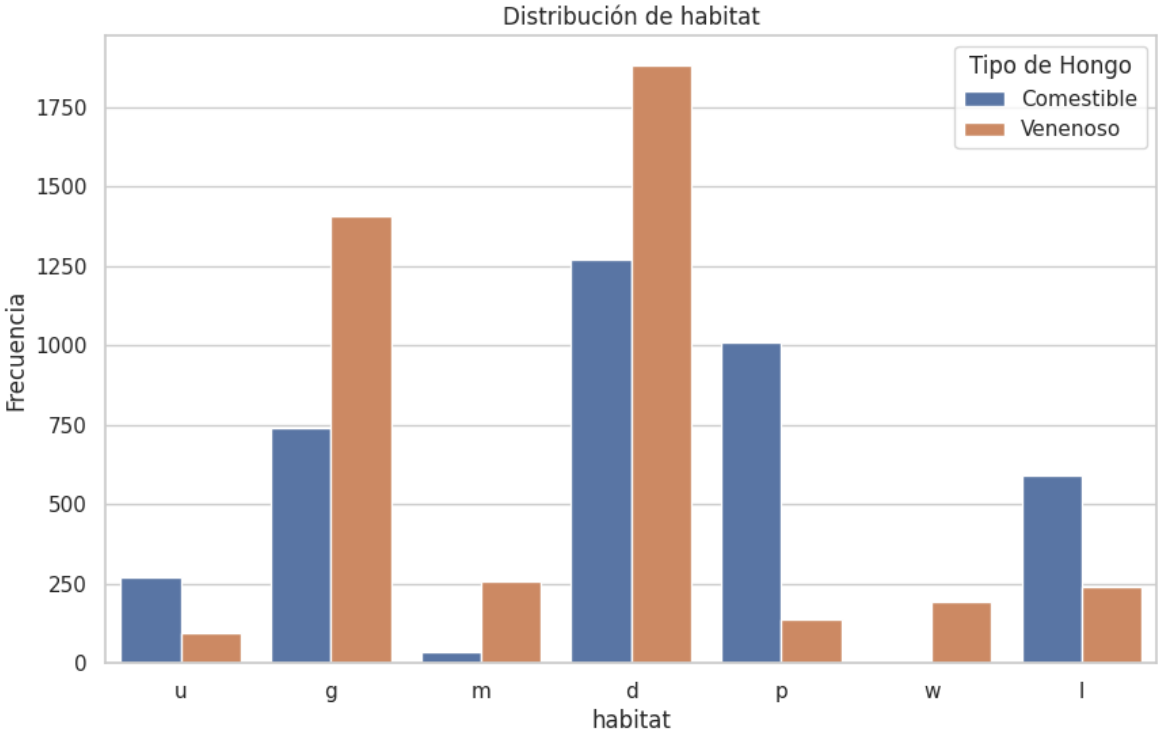
Este tipo de problema tiene una aplicación práctica evidente: identificar si un hongo encontrado en la naturaleza es seguro para el consumo humano o no. Se seleccionaron dos algoritmos de clasificación ampliamente utilizados: Árbol de Decisión y Random Forest. Ambos modelos fueron entrenados y evaluados utilizando técnicas estándar de aprendizaje supervisado, considerando métricas como la precisión, la sensibilidad, la especificidad y el coeficiente Kappa de Cohen.

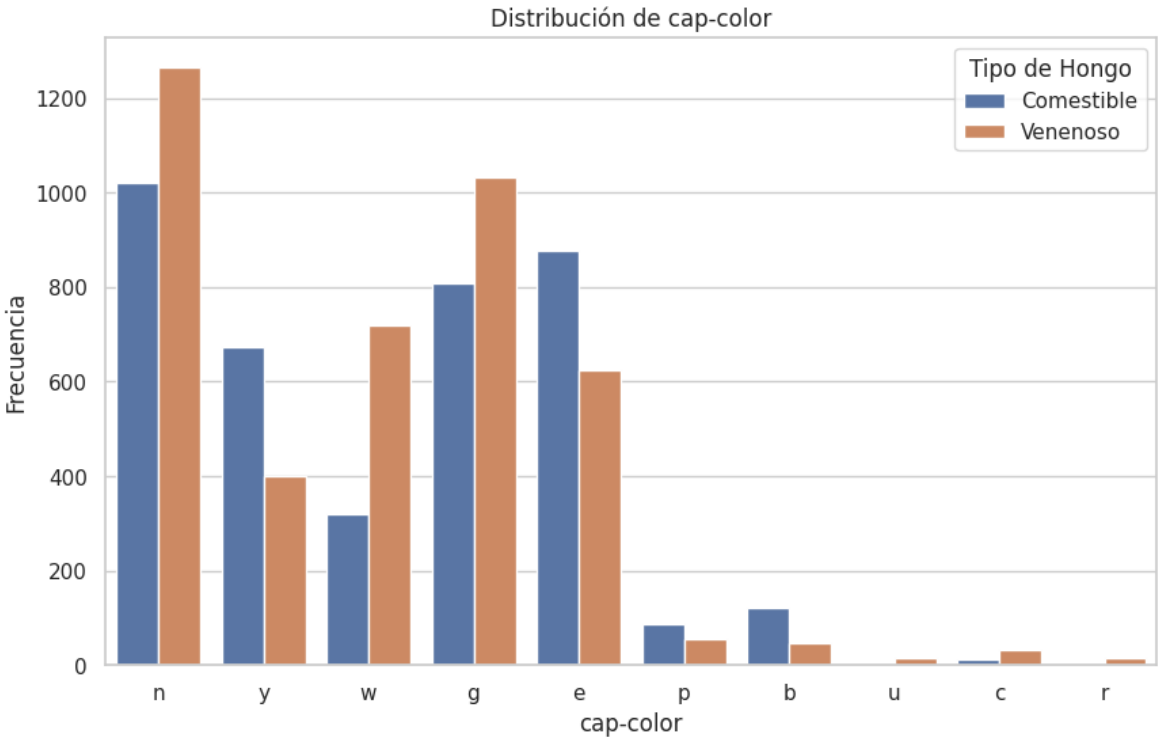
A continuación, se presentan los resultados obtenidos y las conclusiones derivadas del análisis comparativo de los modelos.

2. Desarrollo

Se realizó una exploración exhaustiva del dataset de hongos, incluyendo la visualización de la distribución de características categóricas y el análisis estadístico descriptivo para

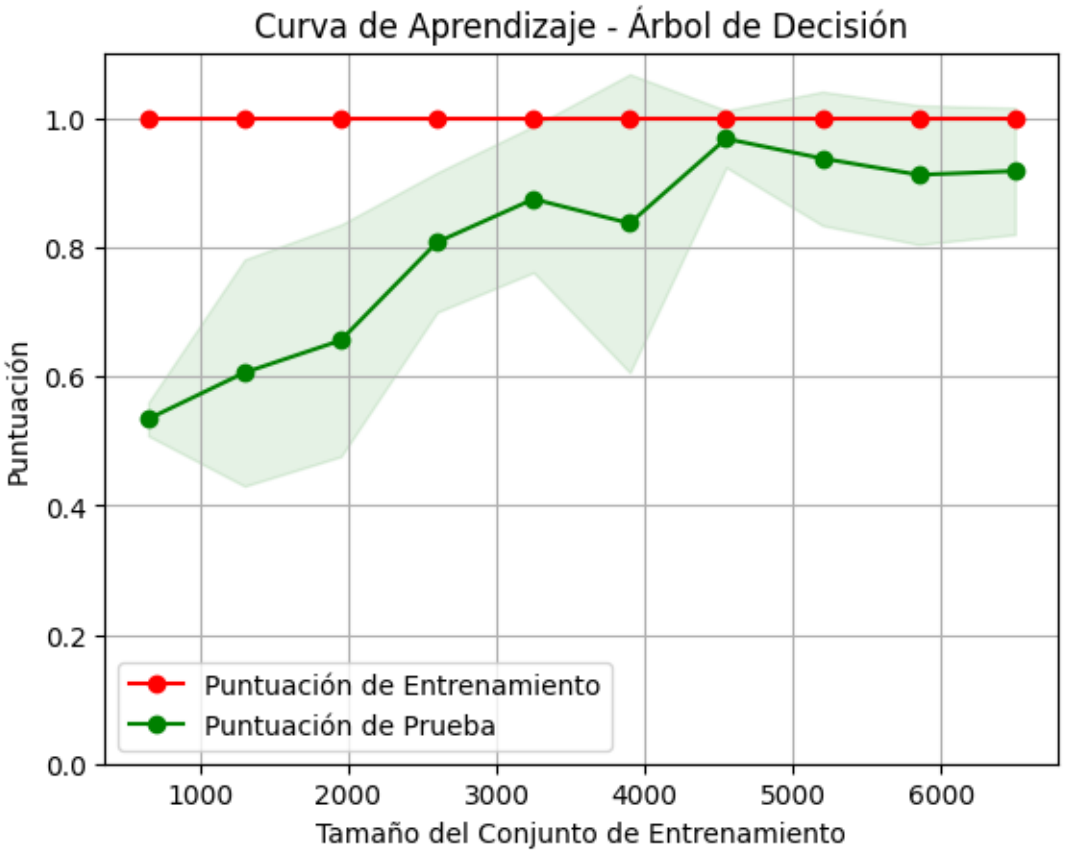
comprender mejor la naturaleza y calidad de los datos, también se realizaron graficas para visualizar la correlación entre las variables, se presentan algunas graficas a continuación:

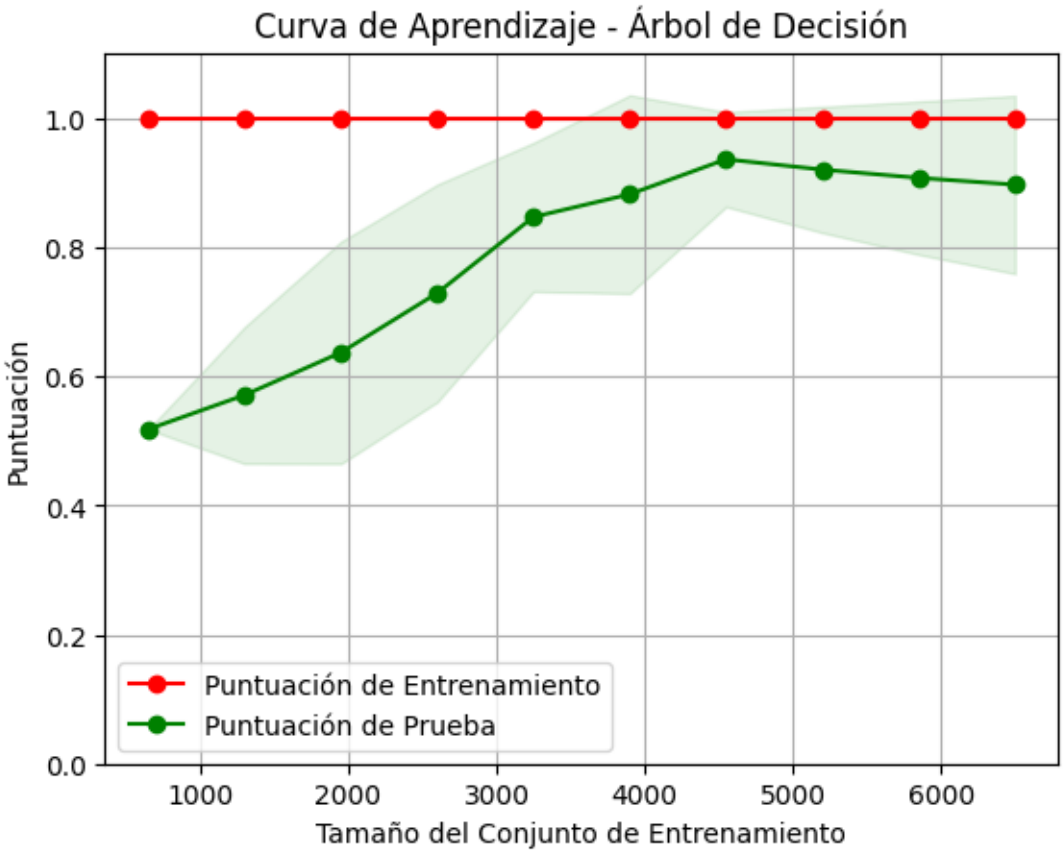




Posteriormente, se llevó a cabo el preprocesado del dataset, que incluyó la simulación y la imputación de datos faltantes mediante la imputación con la moda, así como la codificación one-hot de las variables categóricas para preparar los datos para los modelos de clasificación.

Seguido se dividieron los datos en conjuntos de entrenamiento y prueba (70%-30%), se entrenaron los modelos, y se realizaron predicciones sobre el conjunto de prueba. Se realizaron las curvas de aprendizaje, para evidenciar la capacidad de los modelos para generalizar y detectar la clase correcta de los hongos.


Curva de Aprendizaje - Árbol de Decisión
Puntuación
1.0
0.8
0.6
0.4
0.2
0.0
1000
2000
3000
4000
5000
6000
Tamaño del Conjunto de Entrenamiento
Puntuación de Entrenamiento
Puntuación de Prueba



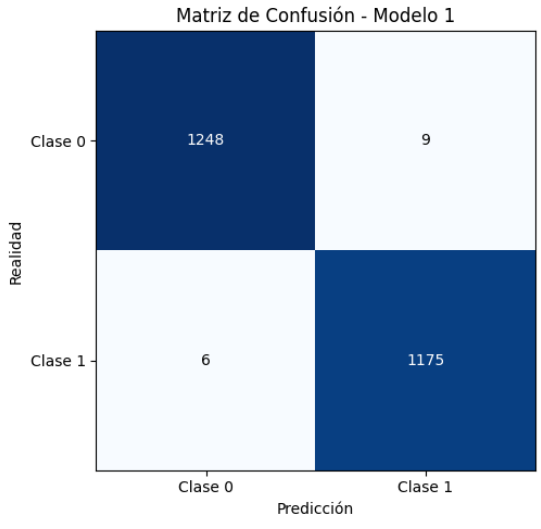
3. Resultados

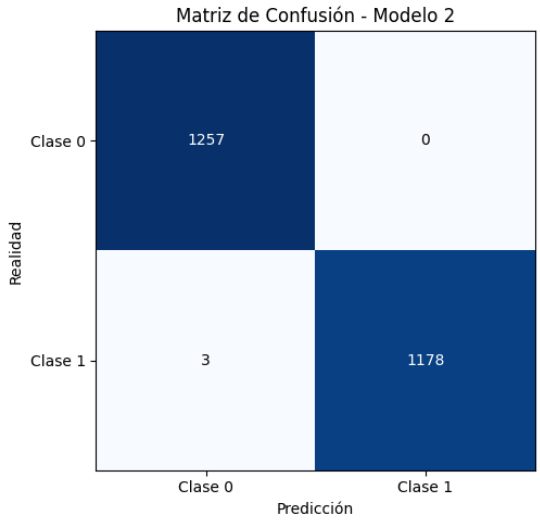
Los modelos se evaluaron sobre un conjunto de prueba compuesto por 2438 instancias. Se utilizaron matrices de confusión, reportes de clasificación, análisis de concordancia y coeficiente de Kappa de Cohen para determinar el rendimiento de cada modelo.

```
Reporte de Clasificación - Modelo 1 (Árbol de Decisión):
              precision    recall  f1-score   support

           e       1.00      0.99      0.99      1257
           p       0.99      0.99      0.99      1181

    accuracy                           0.99      2438
   macro avg       0.99      0.99      0.99      2438
weighted avg       0.99      0.99      0.99      2438


Reporte de Clasificación - Modelo 2 (Random Forest):
              precision    recall  f1-score   support

           e       1.00      1.00      1.00      1257
           p       1.00      1.00      1.00      1181

    accuracy                           1.00      2438
   macro avg       1.00      1.00      1.00      2438
weighted avg       1.00      1.00      1.00      2438
```

```
Análisis de Concordancia - Modelo 1 (Árbol de Decisión):
Precisión: 0.99, Sensibilidad: 0.99, Especificidad: 0.99

Análisis de Concordancia - Modelo 2 (Random Forest):
Precisión: 1.00, Sensibilidad: 1.00, Especificidad: 1.00
```

Coeficiente de Kappa de Cohen entre los modelos: 0.99





4. **Conclusiones**

Ambos modelos evaluados demostraron un rendimiento sobresaliente al abordar el problema de clasificación. Sin embargo, el modelo Random Forest destacó al alcanzar una precisión perfecta del 100% en los datos de prueba, lo que lo posiciona como la mejor opción entre los dos. Por su parte, el Árbol de Decisión también presentó un desempeño muy alto, con una precisión del 99%, aunque cometió 15 errores de clasificación, lo que sugiere una ligera desventaja frente al Random Forest en términos de exactitud.

A pesar de esta diferencia, la concordancia entre ambos modelos fue extremadamente alta, evidenciada por un coeficiente Kappa de 0.99, lo cual indica una fuerte consistencia en sus predicciones. Este nivel de acuerdo refleja que ambos algoritmos son capaces de captar patrones similares en los datos, aunque Random Forest lo hace con mayor precisión.

Durante el análisis, se identificaron características particularmente relevantes para la clasificación, como el olor, la forma del sombrero y el color de las branquias. Estas variables mostraron una alta correlación con las clases del conjunto de datos, lo que permitió distinguir con eficacia entre hongos comestibles y venenosos.

En general, el proyecto demuestra la aplicabilidad de los algoritmos de aprendizaje automático supervisado en problemas de clasificación dentro del ámbito biológico, especialmente en contextos donde la seguridad humana está en juego. Por ello, dada su precisión y robustez, se recomienda el uso del modelo Random Forest para implementaciones reales, como herramientas de campo para recolectores de hongos o aplicaciones móviles de identificación.