Tecnológico de Monterrey

Instituto Tecnológico de estudios superiores de Monterrey

Campus Estado de México

Departamento de Ingeniería

TC3006C

Inteligencia Artificial Avanzada

Grupo: 101

Profesor: Jorge Adolfo Ramírez Uresti

Momento de Retroalimentación

"*Implementación de una técnica de aprendizaje máquina sin el uso de un framework*"

Fecha: 5/09/2025

Alumno:
Santiago Villazón Ponce de León A01746396

Introducción

En el presente trabajo se implementó un algoritmo de aprendizaje supervisado basado en **árboles de decisión y ensambles** para la clasificación de datos categóricos. El objetivo principal fue evaluar el desempeño de este modelo en un dataset ampliamente utilizado en la literatura de *machine learning*: el conjunto de datos **Mushrooms**, disponible públicamente y con el cual se pueden realizar experimentos en problemas de clasificación binaria.

El dataset **Mushrooms** contiene observaciones de distintos hongos, cada uno descrito por una serie de atributos categóricos que incluyen características como el tipo de sombrero, color, olor, forma de la raíz, superficie del tallo, entre otros. El atributo de salida, denominado **class**, distingue si un hongo es **edible (e)**, es decir comestible, o **poisonous (p)**, venenoso. La clasificación correcta de hongos es una tarea de gran relevancia práctica, ya que en la vida real consumir un hongo venenoso podría tener consecuencias graves o incluso letales.

Para este proyecto, se utilizó un enfoque basado en **árboles de decisión con índice Gini como criterio de impureza**, y además se aplicó una técnica de **ensamble (bagging con submuestreo de atributos)**, lo cual permite mejorar la estabilidad del modelo y reducir la varianza al combinar varios árboles entrenados sobre subconjuntos aleatorios de datos y características.

El conjunto de datos fue dividido en dos partes: un **80% para entrenamiento** del modelo y un **20% para prueba**, de modo que fuera posible evaluar su capacidad de generalización sobre ejemplos no vistos durante el proceso de ajuste.

Configuración del modelo

Los parámetros del modelo se establecieron de la siguiente manera:

- **Máxima profundidad de los árboles (MAX_DEPTH):** 8
- **Mínimo de muestras para dividir un nodo (MIN_SAMPLES_SPLIT):** 15
- **Mínimo de muestras por hoja (MIN_SAMPLES_LEAF):** 10
- **Ganancia mínima para realizar un split (MIN_GAIN):** 1e-4
- **Número de árboles en el ensamble (N_TREES):** 35
- **Fracción de características seleccionadas en cada árbol (FEATURE_FRACTION):** 0.6
- **Fracción de datos utilizados en el muestreo bootstrap (BOOTSTRAP_FRACTION):** 1.0
- **Umbral fijo de clasificación (PRED_THRESHOLD):** 0.5

Estos valores permiten al modelo tener un balance entre complejidad y generalización. Por ejemplo, limitar la profundidad y el número de muestras en hojas evita que los árboles memoricen excesivamente los datos de entrenamiento. Por otra parte, el uso de varios árboles (ensamble) reduce la sensibilidad a las fluctuaciones de los datos y mejora la robustez del clasificador.

Resultados

Tras ejecutar el algoritmo sobre el dataset de hongos, se obtuvieron los siguientes resultados en el conjunto de prueba:

- **Umbral de decisión empleado:** 0.5
- **Matriz de confusión:**
  - Verdaderos Positivos (TP): 783
  - Verdaderos Negativos (TN): 842
  - Falsos Positivos (FP): 0
  - Falsos Negativos (FN): 0
- **Métricas principales en test:**
  - Exactitud (Accuracy): 1.0000
  - Precisión (Precision): 1.0000

La matriz de confusión muestra un desempeño perfecto: el modelo logró clasificar correctamente los 783 hongos venenosos y los 842 hongos comestibles del conjunto de prueba, sin incurrir en ningún error de predicción.

En consecuencia, las métricas de exactitud y precisión alcanzaron el valor máximo posible (1.0). La exactitud del 100% indica que todas las predicciones coincidieron con las etiquetas reales, mientras que la precisión del 100% significa que todos los ejemplos clasificados como venenosos eran efectivamente venenosos, sin falsos positivos.

Análisis de los resultados

Aunque a primera vista un rendimiento perfecto puede parecer demasiado bueno para ser verdad, en este caso es un resultado esperado debido a las características del dataset. El conjunto de datos de **Mushrooms** es conocido por ser **altamente separable**, es decir, que las variables categóricas disponibles contienen información suficiente para distinguir de manera clara entre hongos venenosos y comestibles.

Esto significa que existen patrones muy evidentes en los datos que permiten al modelo aprender reglas simples y efectivas para diferenciar ambas clases. Por ejemplo, ciertos valores de atributos como el olor (*odor*) o el color del sombrero (*cap-color*) son determinantes para identificar hongos venenosos con certeza.

Sin embargo, el hecho de obtener métricas perfectas también obliga a reflexionar sobre el riesgo de **overfitting**. El sobreajuste ocurre cuando un modelo se ajusta demasiado a los datos de entrenamiento, memorizando en lugar de generalizar. Generalmente, el overfitting se manifiesta cuando el rendimiento en entrenamiento es muy alto, pero disminuye en prueba. En este caso, tanto el entrenamiento como la prueba alcanzaron 100%, lo cual sugiere que el dataset está tan bien estructurado que prácticamente no hay ruido ni solapamiento entre clases.

Dicho de otra forma, no es que el modelo esté necesariamente sobreajustando, sino que el problema planteado es tan sencillo (por la claridad de las variables) que cualquier modelo basado en reglas lógicas tendrá un desempeño perfecto. Por esta razón, este dataset suele usarse como ejemplo didáctico, más que como un desafío realista en aprendizaje automático.

Conclusiones

El modelo implementado, basado en un ensamble de árboles de decisión con índice Gini, logró un desempeño perfecto al clasificar hongos como comestibles o venenosos en el dataset Mushrooms. Las métricas de exactitud y precisión fueron de 1.0000, y la matriz de confusión confirmó que no se cometió ningún error de predicción en el conjunto de prueba.

Estos resultados reflejan que el dataset de hongos es un caso particular donde la separación entre clases es clara y no existe ruido significativo. La información categórica disponible contiene patrones tan distintivos que permiten al modelo construir reglas deterministas para cada clase, alcanzando así un rendimiento perfecto.

No obstante, es importante remarcar que este tipo de desempeño no es común en problemas reales de clasificación. En contextos prácticos, siempre existe cierto nivel de incertidumbre, ruido y variabilidad que hacen imposible lograr métricas perfectas. Por ello, aunque el experimento es útil para validar la implementación del algoritmo y comprobar su correcto funcionamiento, no debe interpretarse como una garantía de que el modelo funcionará igual de bien en datasets más complejos y menos estructurados.

En conclusión, este ejercicio muestra la efectividad de los árboles de decisión y de las técnicas de ensamble al aplicarse en un conjunto de datos bien definido y categórico. Sin embargo, también deja claro que, para aplicaciones más realistas, será necesario considerar estrategias adicionales de regularización, validación cruzada y análisis de generalización para evitar caer en sobreajuste y obtener resultados confiables en producción.