Informe del Proyecto - Predicción de Especies de Iris

Juan Pablo Oviedo Herrera

Eddie Alejandro Arenas Vita

Jhon Jairo Moguea Caballero

Ciencia de datos

Luis Fernando Sánchez

Sena

Análisis y desarrollo de software

04/04/2025

Este informe presenta el desarrollo y los resultados de un modelo de clasificación basado en **Árboles de Decisión** para la predicción de especies de flores utilizando el conjunto de datos **Iris**. El objetivo principal es evaluar el desempeño del modelo y analizar su efectividad en la clasificación de las muestras.

1. Descripción del Dataset

El **conjunto de datos Iris** es uno de los más conocidos en el ámbito del aprendizaje automático. Fue introducido por el estadístico **Ronald Fisher** en 1936 y contiene **150 muestras** de tres especies diferentes de flores del género *Iris*:

- **Setosa**
- **Versicolor**
- **Virginica**

Cada muestra está representada por **cuatro características numéricas**, que son:

- **Longitud del sépalo (cm)**
- **Ancho del sépalo (cm)**
- **Longitud del pétalo (cm)**
- **Ancho del pétalo (cm)**

La variable objetivo del dataset es la **especie de la flor**, la cual será predicha en base a las cuatro características mencionadas.

El dataset se encuentra balanceado, es decir, hay **50 muestras** de cada una de las tres especies.

2. Modelo Utilizado: Árbol de Decisión

Para la clasificación de las especies, se utilizó un **Árbol de Decisión**, un algoritmo supervisado que organiza los datos en forma de estructura jerárquica para tomar decisiones en función de características clave.

Ventajas del Árbol de Decisión

- Fácil interpretación mediante diagramas visuales.
- Bajo costo computacional en comparación con otros modelos más complejos.
- No requiere normalización de datos ni transformación de variables.

Configuración del Modelo

El proceso de entrenamiento del modelo se realizó dividiendo los datos en dos partes:

- **70% (105 muestras) para entrenamiento**
- **30% (45 muestras) para prueba**

El modelo se ajustó utilizando el criterio de **índice de Gini**, que mide la pureza de los nodos en el árbol de decisión. Un valor de **0** significa que el nodo es puro (contiene solo una clase).

3. Resultados

Tras entrenar el modelo y evaluarlo en los datos de prueba, se obtuvo una **precisión del 100%**, lo que indica que el modelo logró clasificar correctamente todas las muestras del conjunto de prueba.

Métricas de Evaluación

El desempeño del modelo se evaluó utilizando las siguientes métricas:

- **Precisión (Precision):** Proporción de predicciones correctas entre todas las realizadas.
- **Recall:** Proporción de casos positivos detectados correctamente.
- **F1-score:** Media armónica entre precisión y recall, útil cuando las clases están desbalanceadas.

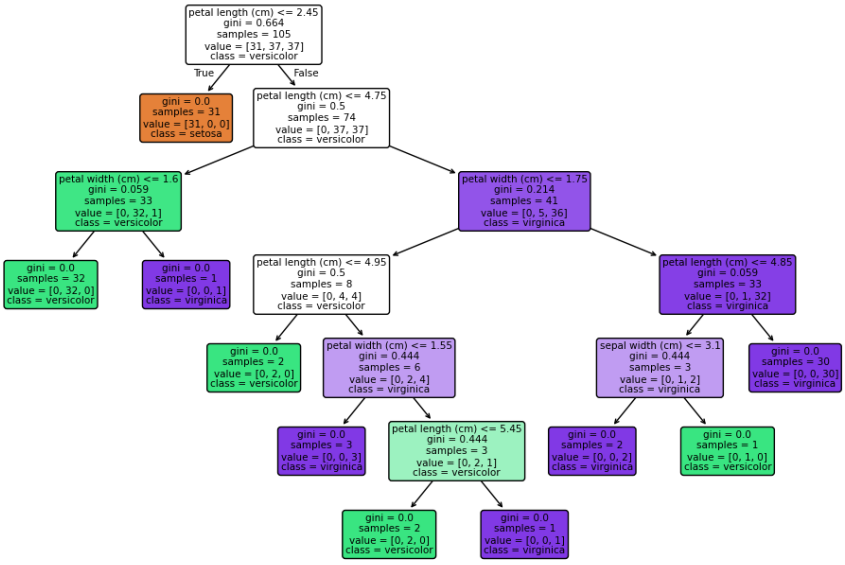
A continuación, se muestra el reporte de clasificación obtenido:

Clase	Precisión	Recall	F1-score	Soporte
Setosa	1.00	1.00	1.00	19
Versicolor	1.00	1.00	1.00	13
Virginica	1.00	1.00	1.00	13
Promedio	**1.00**	**1.00**	**1.00**	**45**

Estos resultados reflejan que el modelo no cometió **ningún error** al clasificar las especies en los datos de prueba.

Visualización del Árbol de Decisión

Se generó una representación gráfica del **Árbol de Decisión**, donde cada nodo representa una condición de separación basada en las características del dataset. El modelo segmenta los datos en función de valores específicos de **longitud y ancho de sépalos y pétalos**, dividiendo las especies con gran precisión.



4. Conclusión

El modelo de Árbol de Decisión resultó altamente efectivo para la clasificación del conjunto de datos **Iris**, logrando una precisión perfecta del **100%** en la predicción de las especies.

A pesar de su buen desempeño, es importante considerar que:

- El modelo puede haber sobreajustado los datos, ya que la precisión es perfecta.
- Árboles de decisión más complejos pueden ser más difíciles de interpretar y generalizar peor en datos nuevos.
- Se podría probar con otros modelos, como **Regresión Logística** o **Máquinas de Soporte Vectorial (SVM)**, para comparar su desempeño.

En general, el Árbol de Decisión es una herramienta valiosa para tareas de clasificación y proporciona una forma clara y visual de entender cómo se toman las decisiones en el modelo.