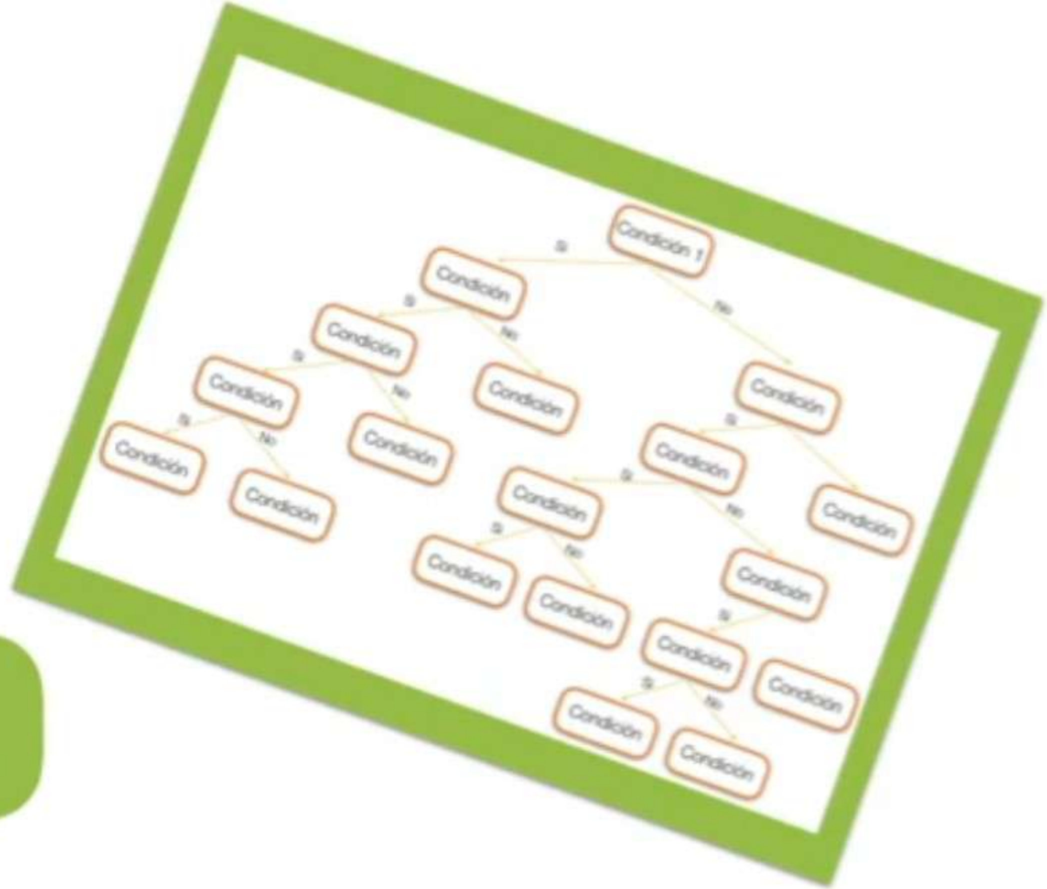
ÁRBOLES DE DECISIÓN CLASIFICACIÓN

TEORÍA



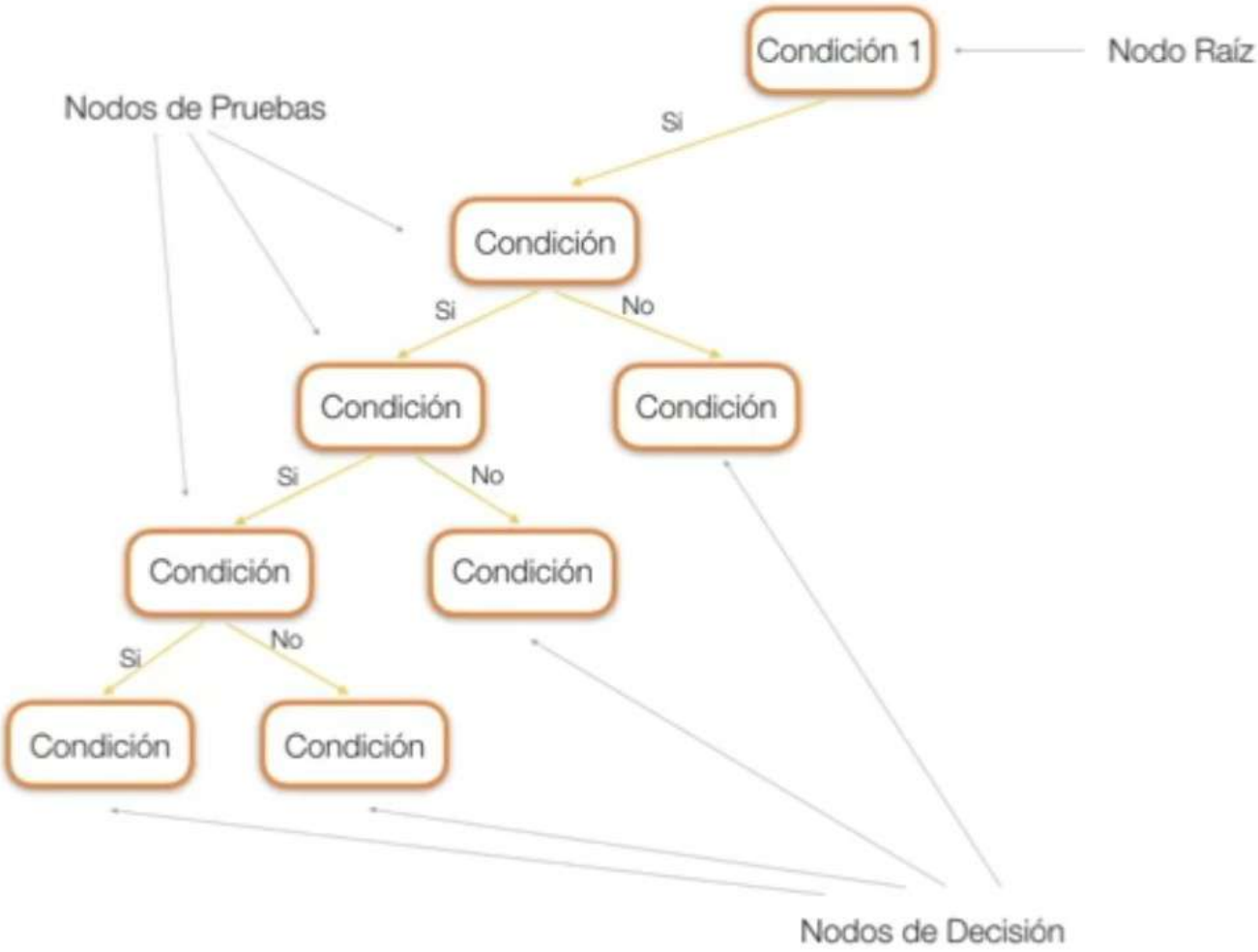
ÁRBOLES DE DECISIÓN CLASIFICACIÓN - TEORÍA

Árboles de Decisión

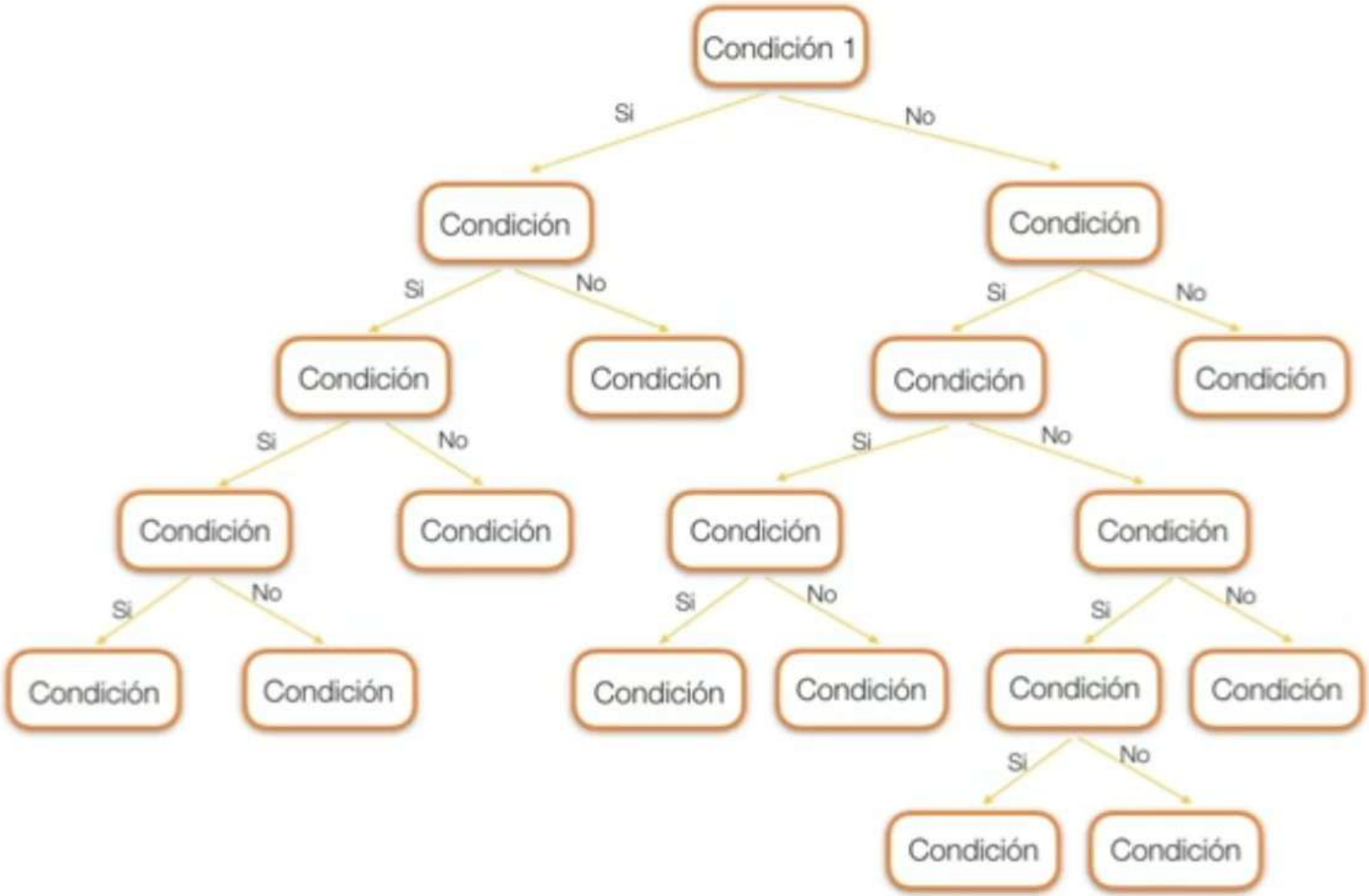
Puede ser fácilmente visible para que un humano pueda entender lo que está sucediendo

Imagina un diagrama de flujo, donde cada nivel es una pregunta con una respuesta de si o no. Eventualmente una respuesta te dará una solución al problema inicial

ÁRBOLES DE DECISIÓN CLASIFICACIÓN - TEORÍA



ÁRBOLES DE DECISIÓN CLASIFICACIÓN - TEORÍA



ÁRBOLES DE DECISIÓN CLASIFICACIÓN - TEORÍA

La medida de selección de atributos es una heurística para seleccionar el criterio de división que divide los datos de la mejor manera posible

Esta medida proporciona un rango a cada característica, explicando el conjunto de datos dado. El atributo de mejor puntuación se seleccionará como atributo de división

ÁRBOLES DE DECISIÓN CLASIFICACIÓN - TEORÍA

Ganancia de Información

Cuando usamos un nodo en un árbol de decisión para particionar las instancias de formación en subconjuntos más pequeños, la entropía cambia. La ganancia de información es una medida de este cambio en la entropía

- Comenzar con todas las instancias de formación asociadas al nodo raíz
- Utilizar la ganancia de información para elegir qué atributo etiquetar cada nodo con cual
- Construir cada subárbol en el subconjunto de instancias de capacitación que se clasificarían

Índice de Gini

Es una métrica para medir la frecuencia con la que un elemento elegido al azar sería identificado incorrectamente. Esto significa que se debe preferir un atributo con un índice de Gini más bajo

ÁRBOLES DE DECISIÓN CLASIFICACIÓN - TEORÍA

Ventajas

Los árboles de decisión son fáciles de interpretar y visualizar y pueden capturar fácilmente patrones no lineales

Requiere menos preprocesamiento de datos por parte del usuario, por ejemplo, no es necesario normalizar las columna

Se puede utilizar para ingeniería de características, como la predicción de valores perdidos, adecuada para la selección de variables

El árbol de decisión no tiene suposiciones sobre la distribución debido a la naturaleza no paramétrica del algoritmo

ÁRBOLES DE DECISIÓN CLASIFICACIÓN - TEORÍA

Desventajas

- Datos sensibles al ruido, puede sobredimensionar los datos ruidosos
- La pequeña variación en los datos puede dar lugar a un árbol de decisión diferente
- Están sesgados con un conjunto de datos de desequilibrio, por lo que se recomienda equilibrar el conjunto de datos antes de crear el árbol de decisión