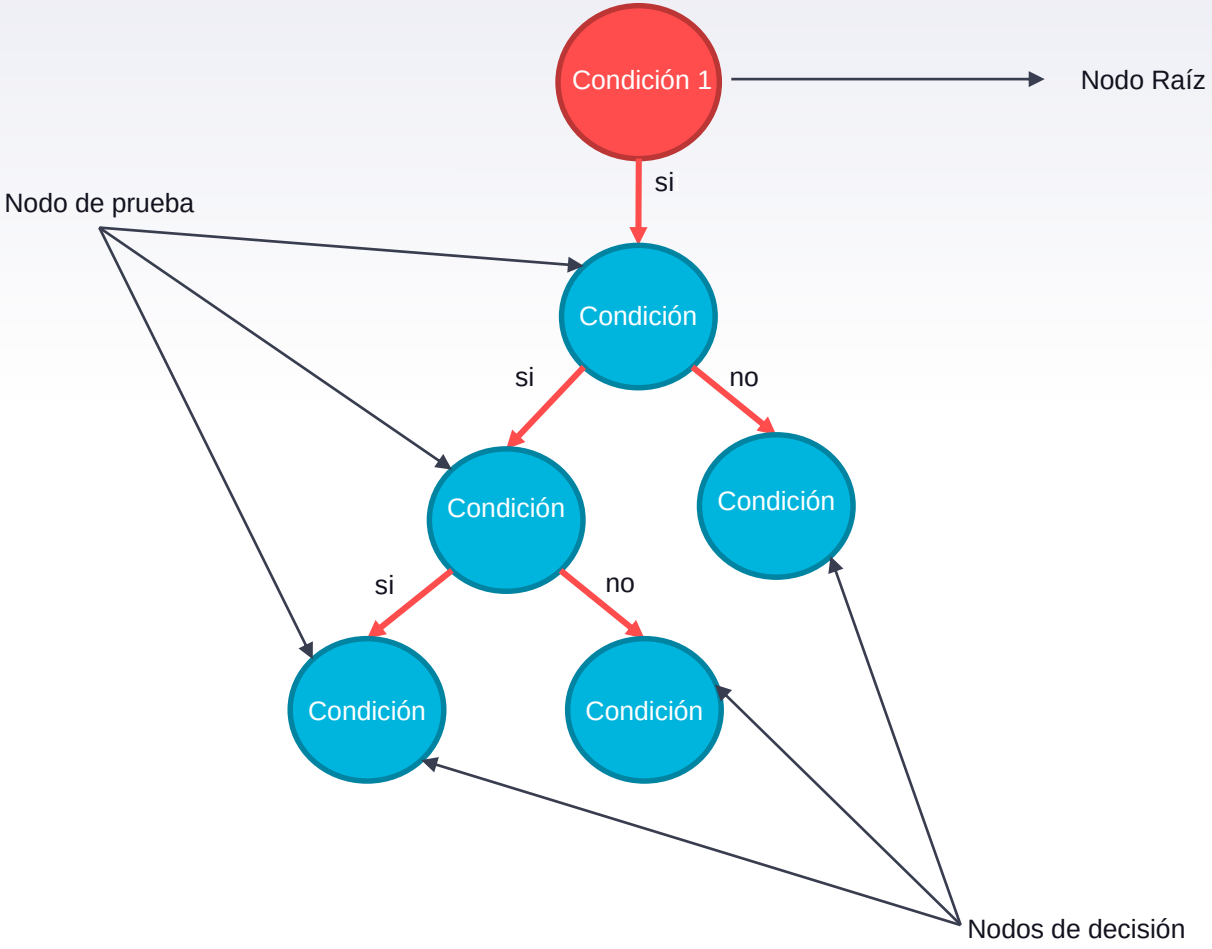
MACHINE LEARNING

Árbol de decisión

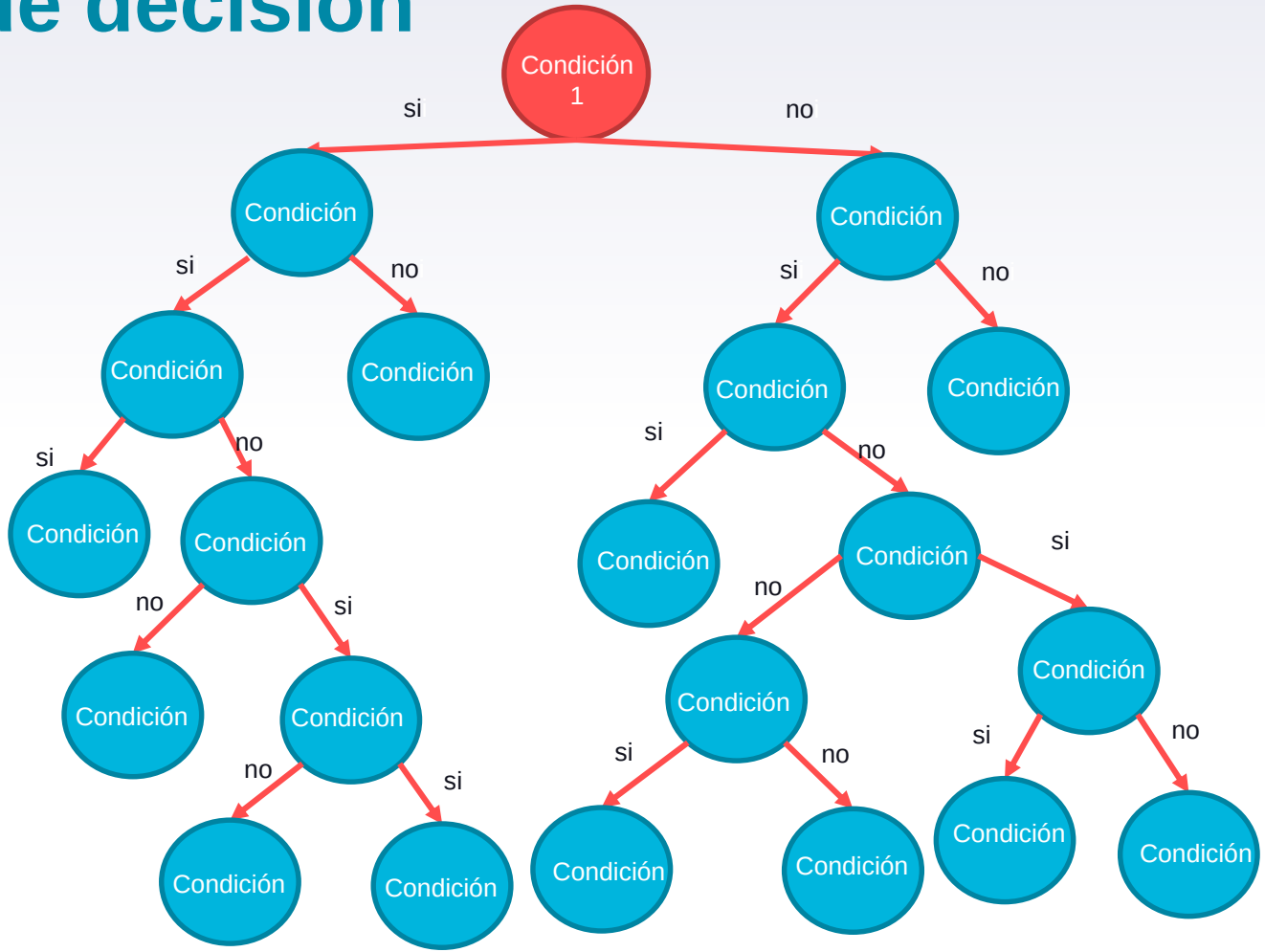
Puede ser fácilmente visible para que un humano pueda entender lo que está sucediendo

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

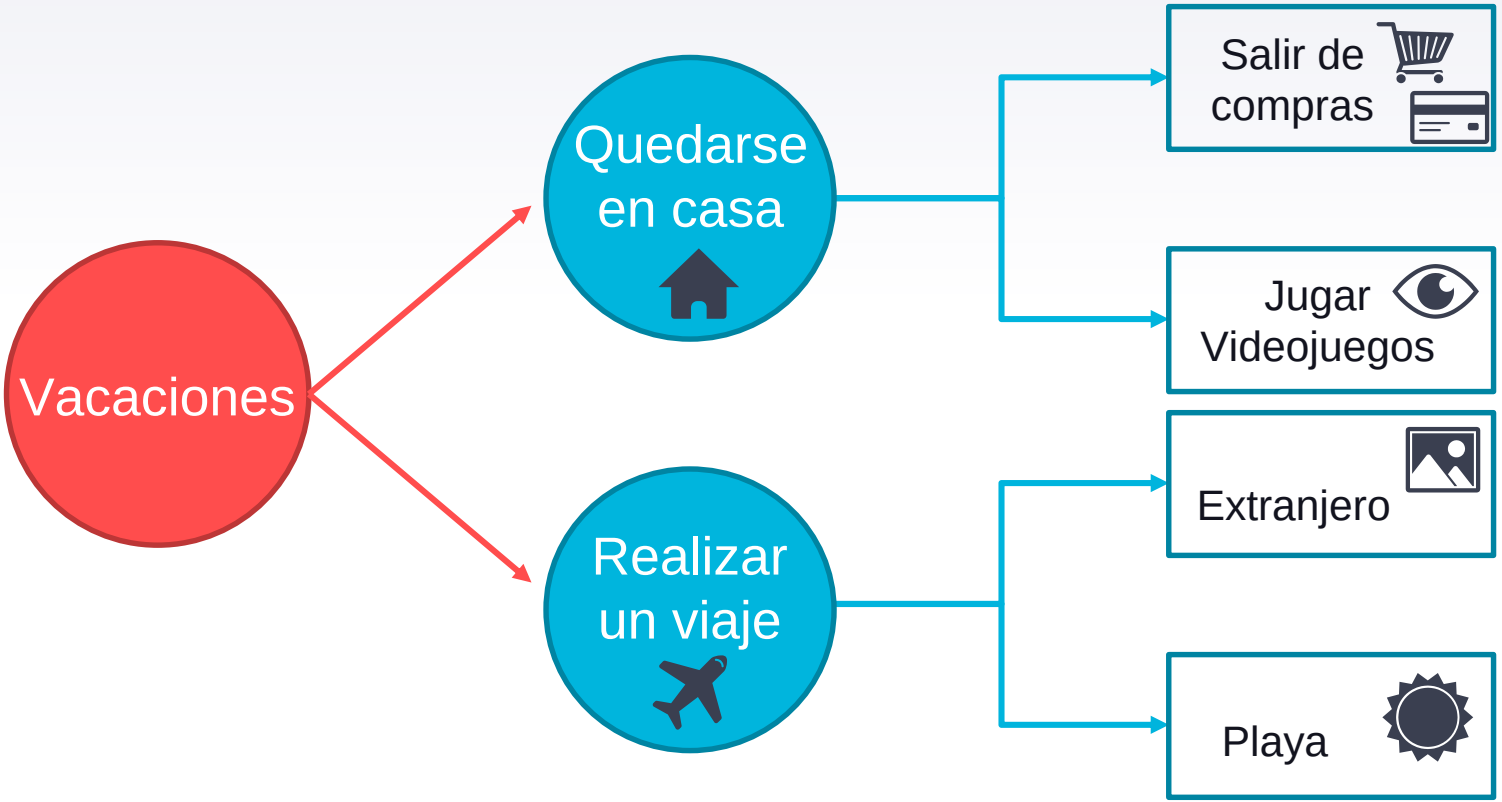
Árbol de decisión



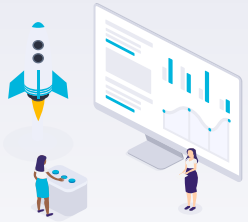
Árbol de decisión



EJEMPLO


Vacaciones
Quedarse en casa
Realizar un viaje
Salir de compras
Jugar Videojuegos
Extranjero
Playa

Árbol de decisión



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

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

Árbol de decisión

Ganancia de información

Cuando usamos un nodo de 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 a la raíz del nodo.

Utilizar la ganancia de información para elegir que atributo etiquetar cada nodo con cual.

Construir recursivamente cada subárbol en el subconjunto de instancias de capacitación que se clasificarían

Árbol de decisión

Índice de Gini

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

Árbol de decisión

Ventajas

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

Requiere menos procesamiento de datos por parte del usuario, por ejemplo, no es necesario normalizar una columna.

Se puede utilizar para ingeniería de características, como las predicciones 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.

Árbol de decisión

Desventajas

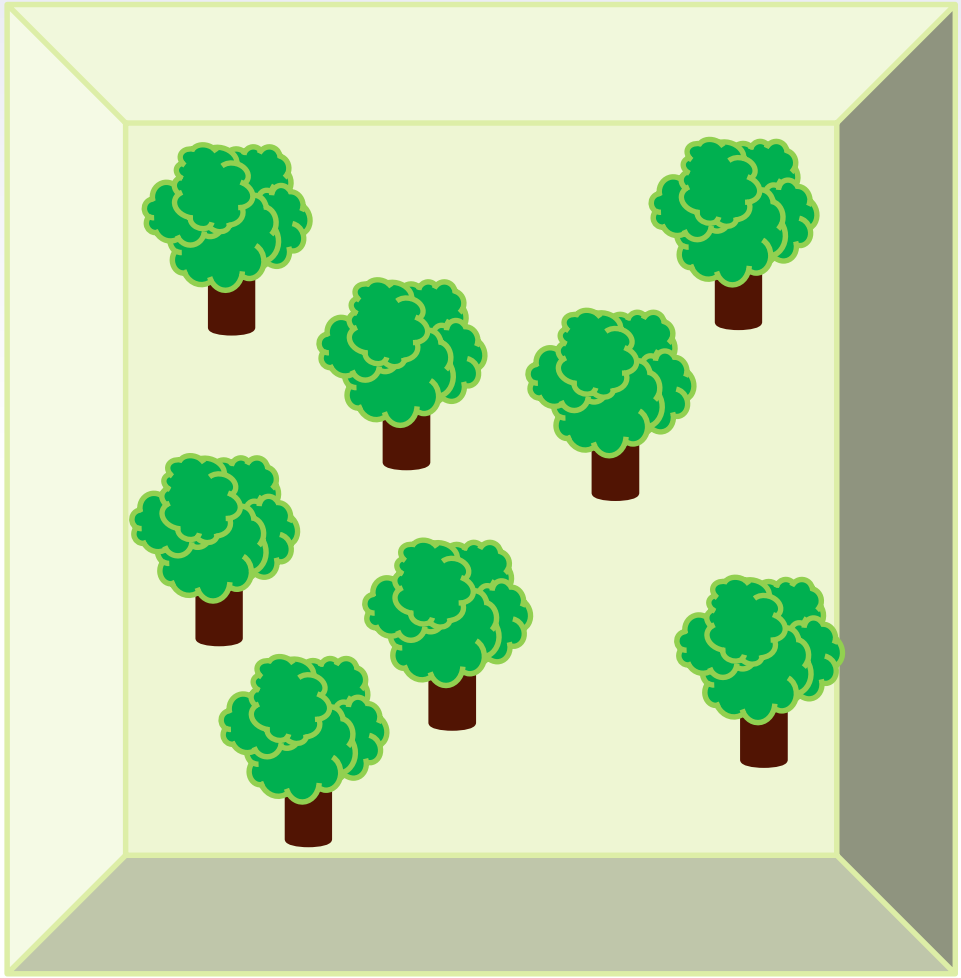
Datos sensibles al ruido, puede sobredimensionar los datos ruidosos.

La pequeña variación de los datos puede dar lugar a un árbol de decisión diferente.

Esta sesgado con un conjunto de datos de desequilibrio, por lo que se recomienda equilibrar el conjunto de datos antes de crear el árbol de decisión.

2

Bosque aleatorio



Bosque aleatorio

Es un algoritmo de machine Learning flexible y fácil de usar que produce, incluso sin ajuste de parámetro, un gran resultado la mayor parte de tiempo,

Clasificación

Regresión

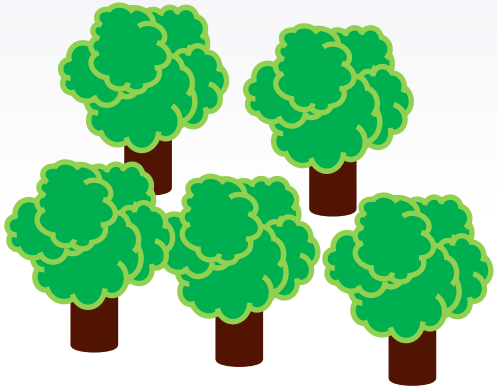
Bosque aleatorio Regresión

Crea un bosque y lo hace de alguna manera aleatorio

Crea múltiples arboles de decisión y los combina para obtener una predicción más precisa y estable

Al crecer los árboles busca la mejor característica entre un subconjunto aleatorio de características

Se puede hacer que los arboles sean más aleatorios, usando umbrales aleatorios para cada función



Bosque aleatorio Regresión

Árbol de decisión regresión

Formula un conjunto de reglas a las características de entrenamiento que se utilizaran para hacer las predicciones.

Los árboles de decisión son muy profundos pueden sufrir de sobreajuste

Bosque aleatorio regresión

Selecciona al azar las observaciones y características para construir varios árboles de decisión y luego promedia los resultados.

Evitan el exceso de adaptación la mayor parte de tiempo, creando subconjuntos aleatorios de características

Bosque aleatorio Regresión

Ventajas

Es un algoritmo muy útil y fácil de usar ya que los parámetros predeterminados a menudo producen un buen resultado de predicción

Si hay suficientes árboles en el bosque, el algoritmo no se adaptara al modelo, evitando el sobreajuste.

Desventajas

Una gran cantidad de árboles puede hacer que el algoritmo sea lento e ineficiente para las predicciones en tiempo real

Es una herramienta de modelado predictivo y no una herramienta descriptiva

Bosque aleatorio Regresión

Es un gran algoritmo para entrenar temprano en el proceso de desarrollo del modelo para ver cómo se desempeña y es difícil de construir un mal modelo con este algoritmo debido a su simplicidad.