MACHINE LEARNING

Tecnológico de Monterrey

# ARBOLES DE DECISION Y BOSQUES ALEATORIOS

Miguel Padrón Vences A01362804
Marce Martínez y Martínez A01381548
Roberto Arturo Gómez Mercado A01365947

# Árbol de Decisión

Un árbol de decisión es un modelo predictivo que divide el espacio de los predictores agrupando observaciones con valores similares para la variable respuesta o dependiente.

Son útiles para entender la estructura de un conjunto de datos y sirven para resolver problemas tanto de clasificación, como de regresión.

# MÉTODO

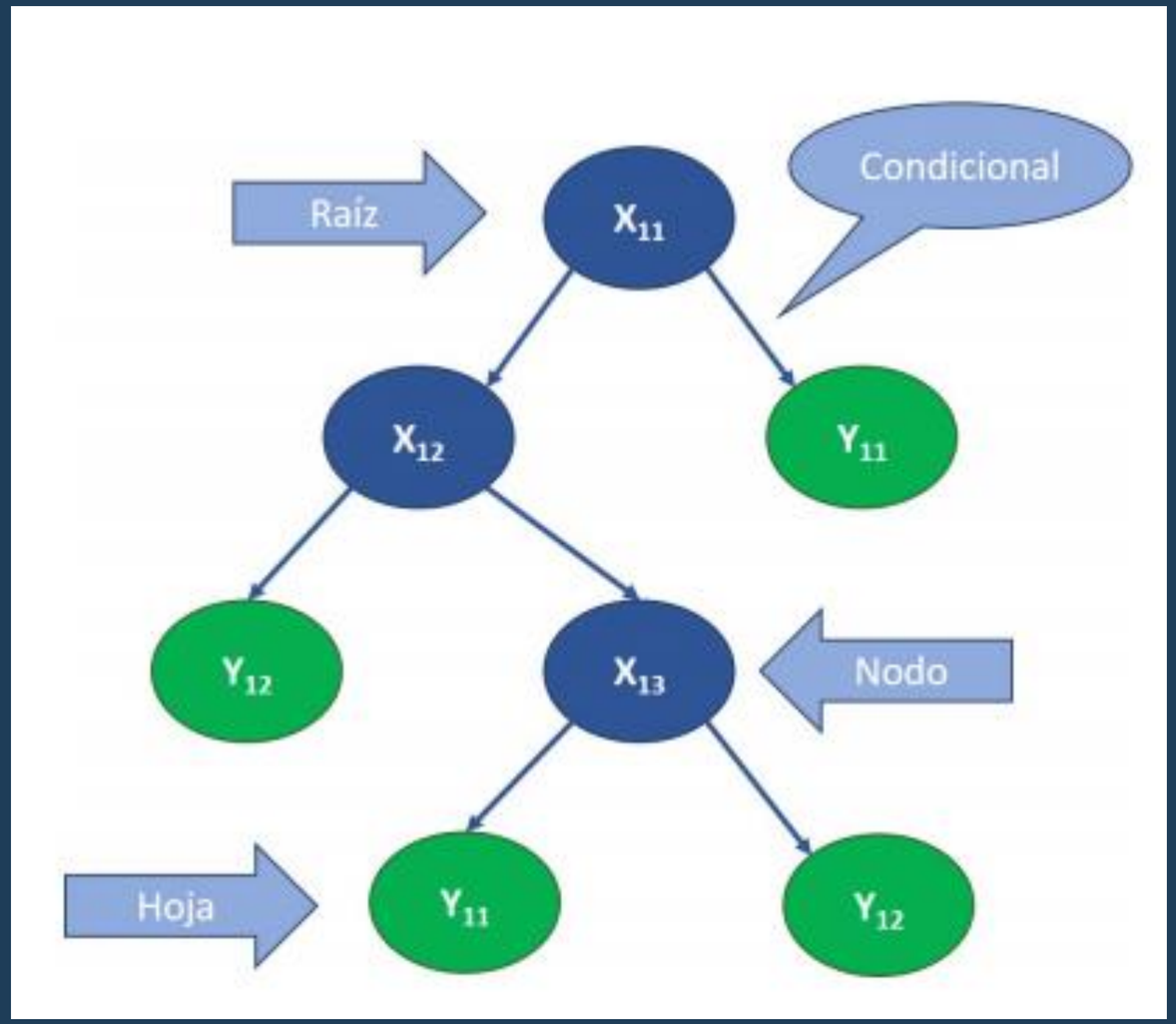
El tipo de problema a resolver dependerá de la variable a predecir:

- Variable continua: problema de **regresión**.
- Variable discreta: problema de **clasificación.**

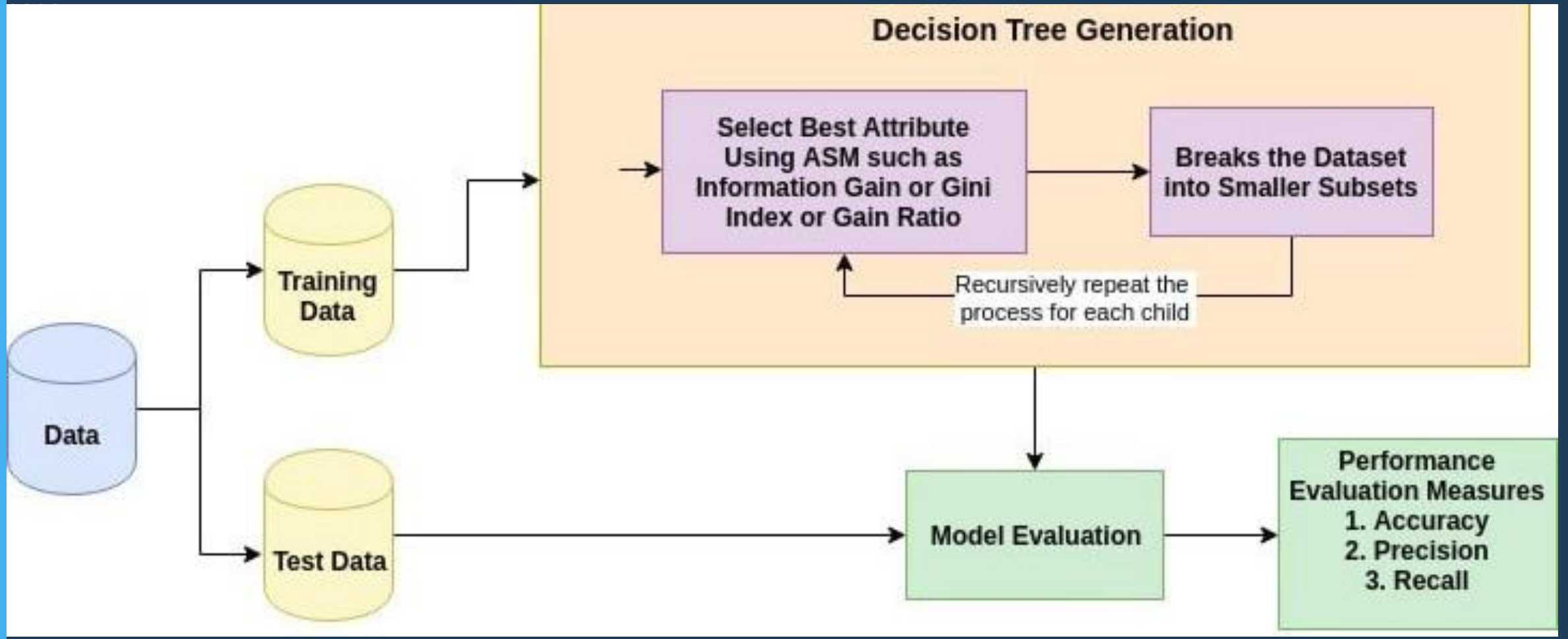
# ESTRUCTURA

El árbol tiene los siguientes elemetos:

- Raíz - Inicio
- Condicional - Condición de cambio
- Nodo - Decisiones
- Hojas - Clasificación

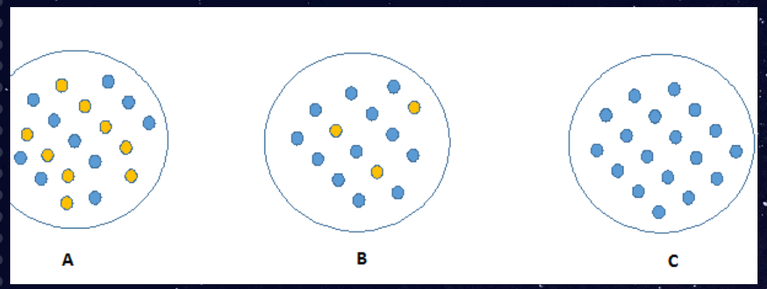


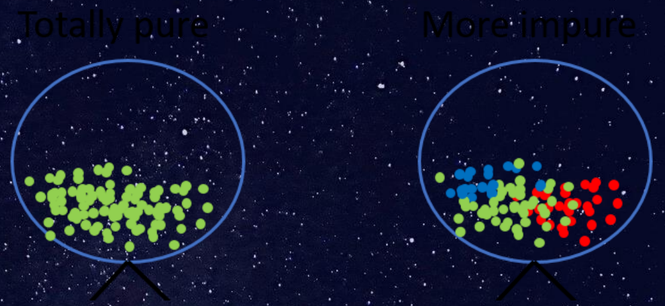
# PASOS

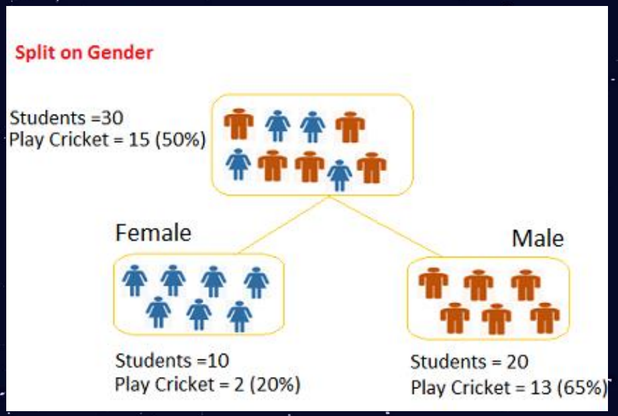



*Nota: Pensando y/o omitiendo los pasos anteriores del CRISP-DM.

# SELECCION DE INFORMACION




## GANANCIA INFORMACION

La ganancia de información es una propiedad estadística que mide qué tan bien un atributo dado separa los ejemplos de entrenamiento de acuerdo con sus clasificación objetivo.

## ENTROPÍA

El concepto de entropía, que mide la impureza del conjunto de entrada. En teoría de la información, se refiere a la impureza en un grupo de ejemplos. La ganancia de información es una disminución de la entropía.

## GINI --> METODO

Si se seleccionan dos datos de la población al azar, estos dos deben ser de la misma clase, y si la población es pura, la probabilidad de que esto suceda es 1.
Funciona con la variable objetivo discreta “Éxito” o “Fracaso”.

# VENTAJAS ARBOL

## FACILIDAD

Son fáciles de construir, interpretar y visualizar.

## ELIJE VARIABLES

Selecciona las variables más importantes y en su creación no siempre se hace uso de todos los predictores.

## PREDICCIONES TEMPRANAS

Si faltan datos no podremos recorrer el árbol hasta un nodo terminal, pero sí podemos hacer predicciones promediando las hojas del sub-árbol que alcancemos.

## VARIABLES SIN RELACION

Permiten relaciones no lineales entre las variables explicativas y la variable dependiente.

## TODO TIPO VARIABLES

Sirven tanto para variables dependientes cualitativas como cuantitativas, como para variables predictoras o independientes numéricas y categóricas. Además, no necesita variables dummys, aunque a veces mejoran el modelo.

# DESVENTAJAS ARBOL

## SOBREAJUSTE

Tienden al sobreajuste u overfitting de los datos, por lo que el modelo al predecir nuevos casos no estima con el mismo índice de acierto.

## OUTLIERS

Creando árboles con ramas muy profundas que no predicen bien para nuevos casos. Se deben eliminar dichos outliers.

## INEFICIENTES

No suelen ser muy eficientes con modelos de regresión.

## SESGO

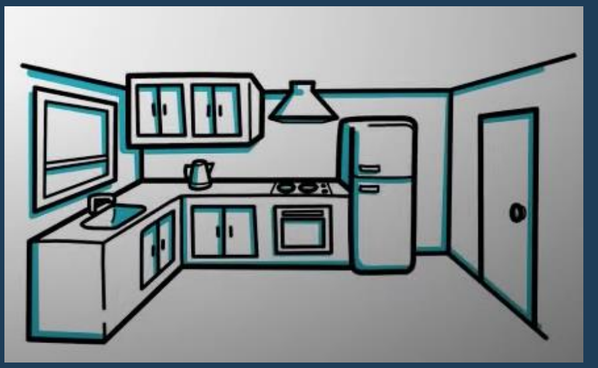
Se pueden crear árboles sesgados si una de las clases es más numerosa que otra.

## PÉRDIDA DE INFORMACIÓN

Se pierde información cuando se utilizan para categorizar una variable numérica continua.

# APLICACIONES


RENDIMIENTO DE COMBUSTIBLE



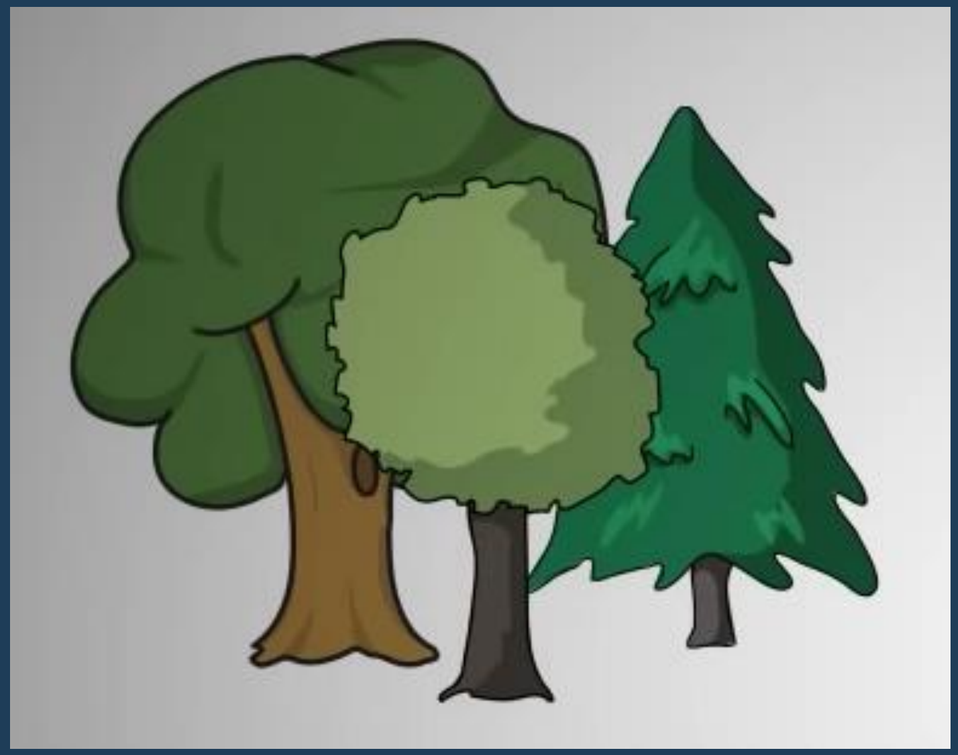
Precio de un inmueble.

Renovación del seguro de vida.

PREDECIR QUETIPO DE ENFERMEDAD

# BOSQUES ALEATORIOS

Si se aplica de manera iterativa (n veces) el algoritmo de árboles de decisión con diferentes parámetros sobre los mismos datos, se obtiene un bosque aleatorio de decisión.



# MÉTODO

## CREAR BOOTSTRAPPED DATASET

Seleccionar aleatoriamente datos para crear un dataset nuevo. (No se seleccionan todos los datos y se pueden repetir.)

## ÁRBOL DE DECISIÓN

Con base en el Bootstrapped dataset, se crea un árbol de decisión.

## ITERACIÓN

El proceso se repite *n* veces hasta crear *n* número de árboles.

# Ventajas

- Se pueden utilizar en ambos métodos, clasificación y regresión.
- Maneja los valores perdidos y mantiene la precisión con la falta de datos.
- No sobreajusta el modelo.
- Manejar grandes conjuntos de datos con mayor dimensionalidad.
- Genera predicciones más robustas.

# Desventajas

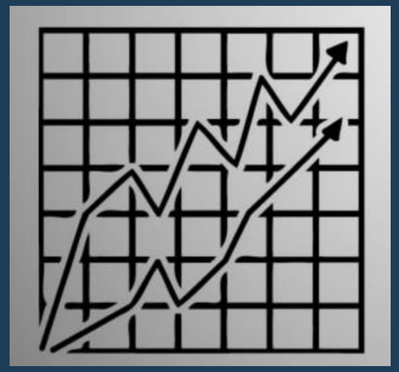
- No funciona tan bien con los problemas de regresión. No predice más allá del primer dataset.
- No se tiene mucho control sobre lo que realiza el modelo.
- Ya que crea muchos clasificadores, al momento de buscar el mejor de éstos, se vuelve una tarea tediosa.

# APLICACIONES

Sector bancario

e-commerce

Sector Salud



Bolsa de valores

# Fuentes


BELLOSTA, C. (2018). R PARA PROFESIONALES DE LOS DATOS: UNA INTRODUCCIÓN. HTTPS://WWW.DATANALYTICS.COM/LIBRO_R/ARBOLES-DE-DECISION.HTML

VILLALBA, F. (2018). CAPÍTULO 8 BOSQUES ALEATORIOS DE DECISIÓN | APRENDIZAJE SUPERVISADO EN R. HTTPS://FERVILBER.GITHUB.IO/APRENDIZAJE-SUPERVISADO-EN-R/BOSQUES.HTML

KANJEE, R. (2017). RANDOM FOREST - FUN AND EASY MACHINE LEARNING [VIDEO]. RETRIEVED FROM HTTPS://WWW.YOUTUBE.COM/WATCH?V=D_2LKHMJCFY