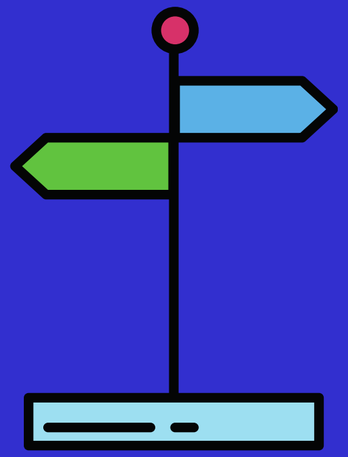
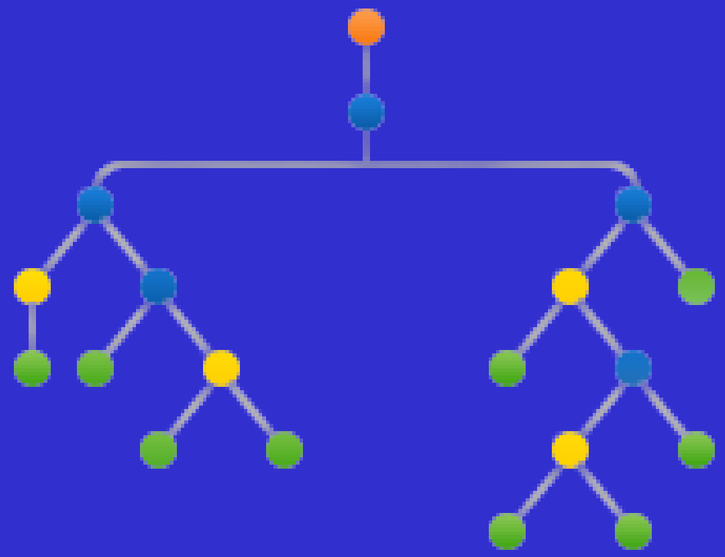
ÁRBOL DE DECISIÓN Y BOSQUES ALEATORIOS

Camila Nieto A00819174
Laura López A01363564
Daniela Gómez A01364823
RicardoTapia A01364132

Árbol de decisión

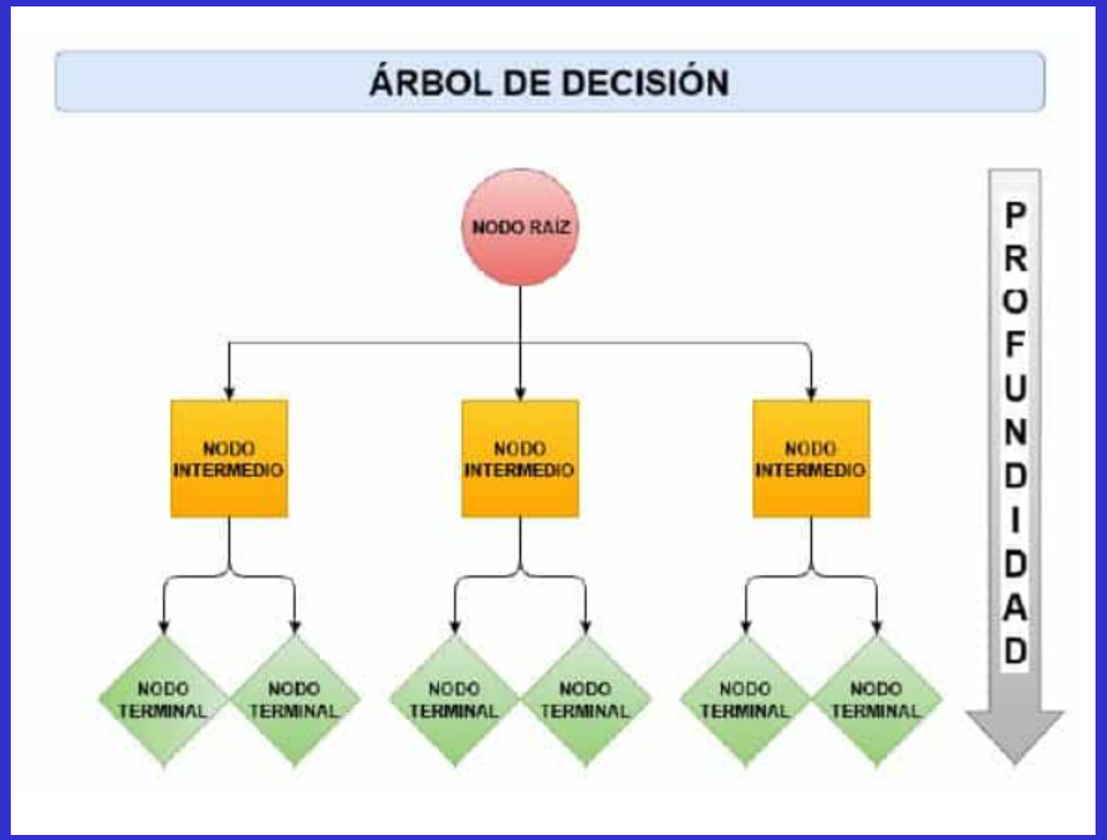


Los árboles de decisión son modelos predictivos formados por reglas binarias (si/no) con las que se consigue repartir las observaciones en función de sus atributos y predecir así el valor de la variable respuesta



Características principales

- Es un tipo de algoritmo de aprendizaje supervisado.
- Se utiliza para problemas de clasificación y regresión.
- Las variables de entrada y salida pueden ser categóricas o continuas.
- Está formado por nodos y su lectura se realiza de arriba hacia abajo.
    - Nodo raíz: Primer nodo, se produce la primera división en función de la variable más importnate.
    - Nodos internos: vuelven a dividir el conjunto de datos en función de las variables.
    - Nodo terminal u hoja: se ubican en la parte inferior, muestran la clasificación definitiva.



ARBOLES DE REGRESIÓN

- Variable dependiente es continua.
- Valores de los nodos terminales se reducen a la media de las observaciones en esa región.

ÁRBOLES DE CLASIFICACIÓN

- Variable dependiente es cualitativa.
- El valor en el nodo terminal se reduce a la moda de las observaciones del conjunto de entrenamiento que han “caído” en esa región.

Funcionamiento del algoritmo:

Algoritmo de HUNT:

1.Inducción

- Definir el objetivo: ¿Qué queremos saber a través de la clasificación ?
- Seleccionar el mejor atributo: aquiel que divide o separa mejor los datos
- Subdividir los datos dependiendo de las característicass especificadas
- Repetir el proceso hasta llegar a un nodo final u hoja (método recursivo).

2. Podar: Eliminar categorías de menor importancia o innecesarias para que el modelo sea más preciso y menos complejo. Es decir, elimina la redundancia, comprime información.

1. Índice de GINI: Mide la probabilidad de no sacar dos registros de la misma clase del nodo.
2. Entriopía: Mide la impureza en un grupo de datos

   Si Entriopía = 0 -> todos los datos pertenecen a la misma clase (puro)

   Si Entriopía = 1 -> existe la misma frecuencia para cada una de las clases de observaciones.

¿Cómo seleccionar el mejor atributo?

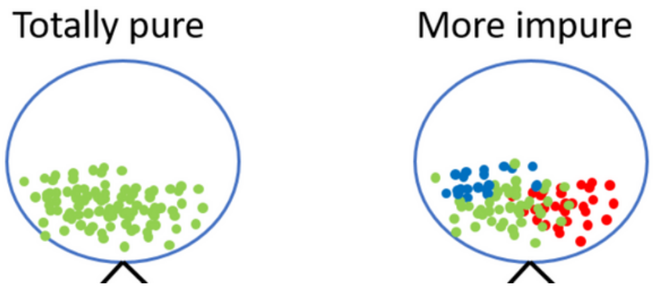
Ganancia de Información: Mide qué tan bien un atributo separa la información según su clasificación objetivo.

índice de GINI

- Mide la probabilidad de no sacar dos registros de la misma clase del nodo.

Entriopía

- Mide la impureza en un grupo de datos
  - Si Entriopía = 0 -> todos los datos pertenecen a la misma clase (puro)
  - Si Entriopía = 1 -> existe la misma frecuencia para cada una de las clases de observaciones.



Ventajas

- Son fáciles de construir, interpretar y visualizar.
- Mientras más información se tenga, mejores son los resultados
- No siempre se hace uso de todos los predictores.
- No es necesario que se cumplan los supuestos de la regresión lineal (linealidad, normalidad, homogeneidad)

Desventajas

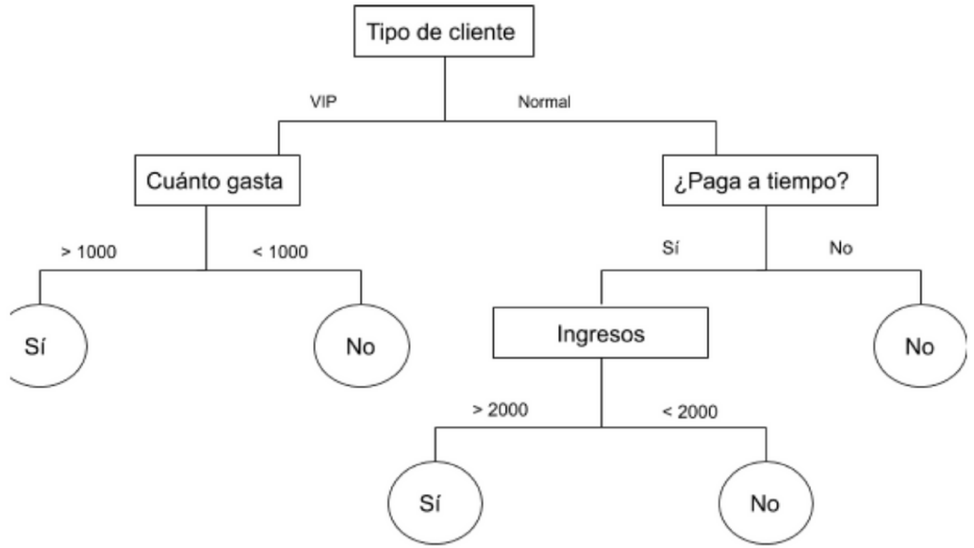
- Overfitting/Sobreajuste: La máquina se ajusta a aprender casos particulares que le enseñamos y es incapaz de reconocer nuevos datos de entrada (puntos atípicos).
- Se crean árboles sesgados si una de las clases es más numerosa que otra.
- Se pierde información cuando se utiliza para categorizar variables numéricas continuas

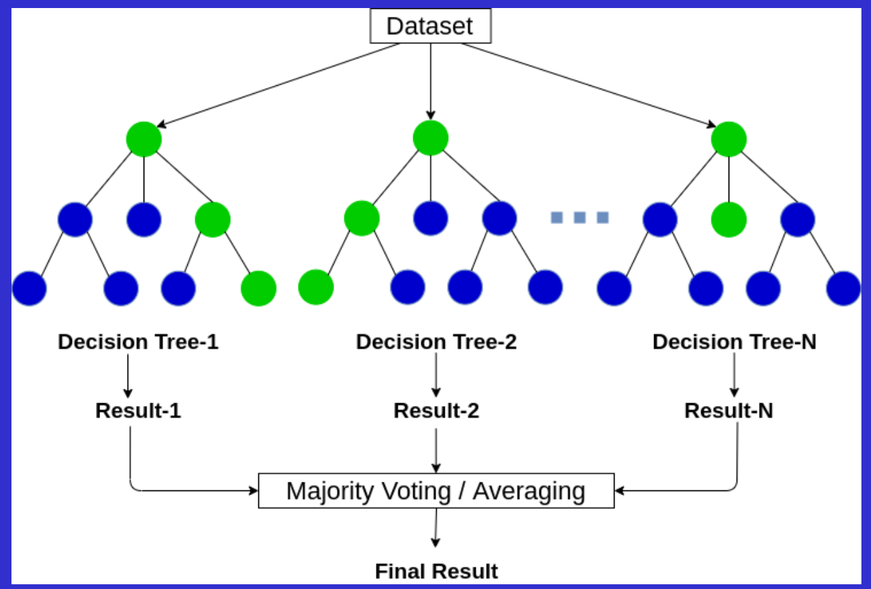
Aplicaciones y ejemplos

Un árbol de decisión sirve para abordar problemas tales como la clasificación, la predicción y la segmentación de datos con la finalidad de obtener información que pueda ser analizada para tomar decisiones futuras.

Ejemplos de preguntas a las que se pueden respondar al usar esta herramienta:

- ¿Se debe ofrecer a un determinado cliente un producto? -> Se deben tomar en cuenta las preguntas: El cliente es VIP?, Cuánto gasta el cliente en dicho producto?, Paga a tiempo?, Cuáles son sus ingresos?, etc.
- ¿Se debe desarrollar un nuevo producto o consolidar el ya existente? -> Se debe tomar en cuenta la forma en la que se realizaría y la calidad del retorno de inversión que cada una de estas ofrecería.





Bosques Aleatorios

Son un algoritmo de machine learning en la que se combinan varios árboles de decisión para tener una predicción más precisa y estable.

Funcionamiento del algoritmo:

1. Seleccionamos **k features** (columnas) de las **m totales** (siendo **k** menor a **m**) y creamos un árbol de decisión con esas **k** características.
2. Creamos **n árboles** variando siempre la cantidad de **k features** y también podríamos variar la cantidad de muestras que pasamos a esos árboles (esto es conocido como “bootstrap sample”)
3. Tomamos cada uno de los **n árboles** y le pedimos que hagan una misma clasificación. Guardamos el resultado de cada árbol obteniendo **n salidas**.
4. Calculamos los votos obtenidos para cada “clase” seleccionada y consideraremos a la más votada como la clasificación final de nuestro “bosque”.

Ventajas

- Existen muy pocas suposiciones y por lo tanto la preparación de los datos es mínima.
- Puede manejar hasta miles de variables de entrada e identificar las más significativas. Método de reducción de dimensionalidad.
- Una de las salidas del modelo es la importancia de variables.
- Incorpora métodos efectivos para estimar valores faltantes.

Desventajas

- Pérdida de interpretación
- Bueno para clasificación, no tanto para regresión. Las predicciones no son de naturaleza continua.
- En regresión, no puede predecir más allá del rango de valores del conjunto de entrenamiento.
- Poco control en lo que hace el modelo (modelo caja negra para modeladores estadísticos)

Aplicaciones

Este algoritmo es uno de los métodos más eficientes de predicción y más usados hoy día para big data, pues promedia muchos modelos con ruido e imparciales reduciendo la variabilidad final del conjunto.

Si bien las aplicaciones son las mismas, los bosques aleatorios generan predicciones mas robustas. Los grupos de árboles de clasificación se combinan y se deduce una única predicción votada en democracia por la población de árboles.

Aplicación con Python

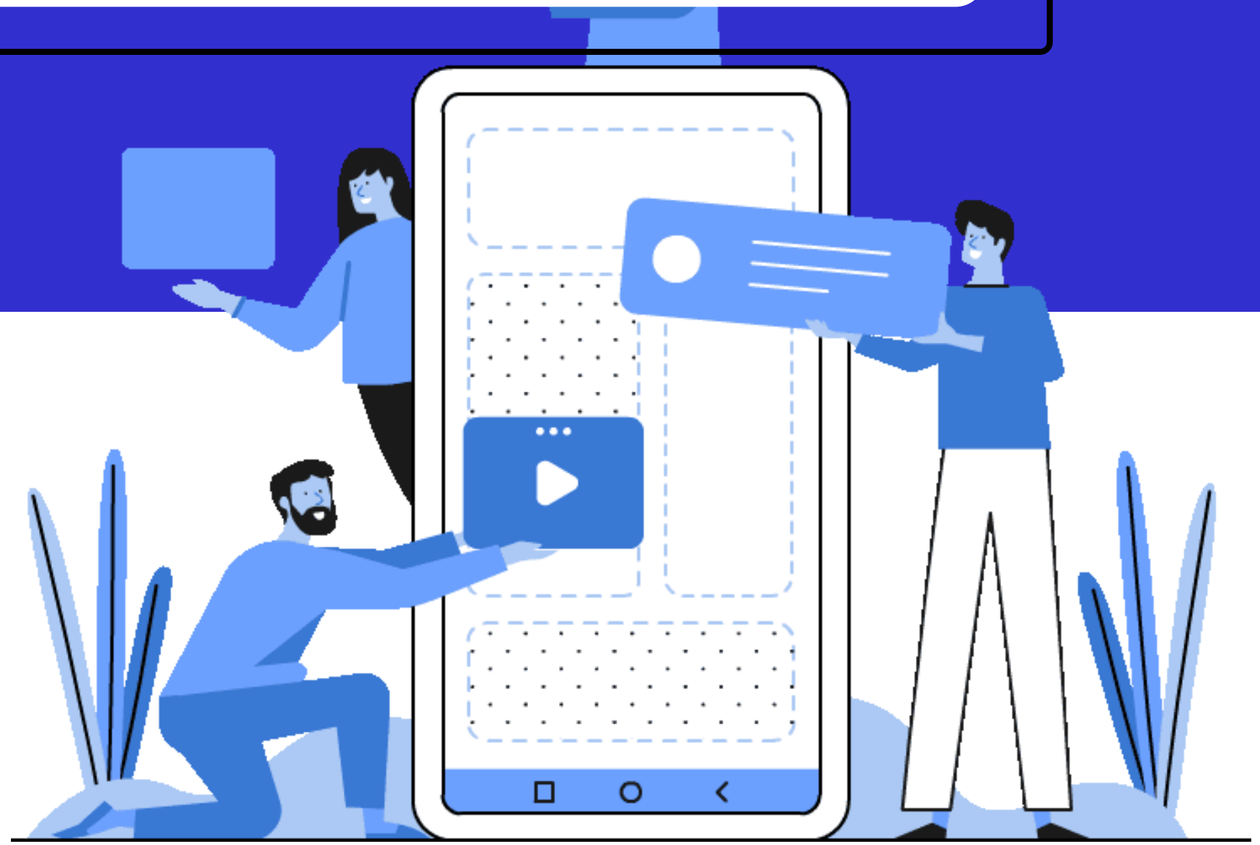
Al programar árboles de decisión y bosques aleatorios en Python, se deben decidir diferentes parámetros, dependiendo si se busca un árbol de regresión o de clasificación.
Algunos parámetros que se utilizan son:

- Criterion: Es la función utilizada para medir la calidad de la subdivisión.
- min_samples_split: Se refiere a la cantidad mínima de muestras que debe tener un nodo para poder subdividir.
- min_samples_leaf: Cantidad mínima que puede tener una hoja final.
- class_weight: Con esto se compensan los desbalances que se pueden presentar. Se asignan pesos a las etiquetas para solucionarlo.

Actividad

Base de Datos de enfermedades del corazón de Cleveland.

http://archive.ics.uci.edu/ml/datasets/Heart+Disease



Bibliografía

https://sitiobigdata.com/2019/12/14/arbol-de-decision-en-machine-learning-parte-1/#
https://www.maximaformacion.es/blog-dat/que-son-los-arboles-de-decision-y-para-que-sirven/
https://www.upgrad.com/blog/random-forest-vs-decision-tree/#:~:text=A%20decision%20tree%20combines%20some,forest%20model%20needs%20rigorous%20training.
https://bookdown.org/content/2031/ensambladores-random-forest-parte-i.html#random-forest
https://www.aprendemachinelearning.com/random-forest-el-poder-del-ensamble/