Árbol de Regresión

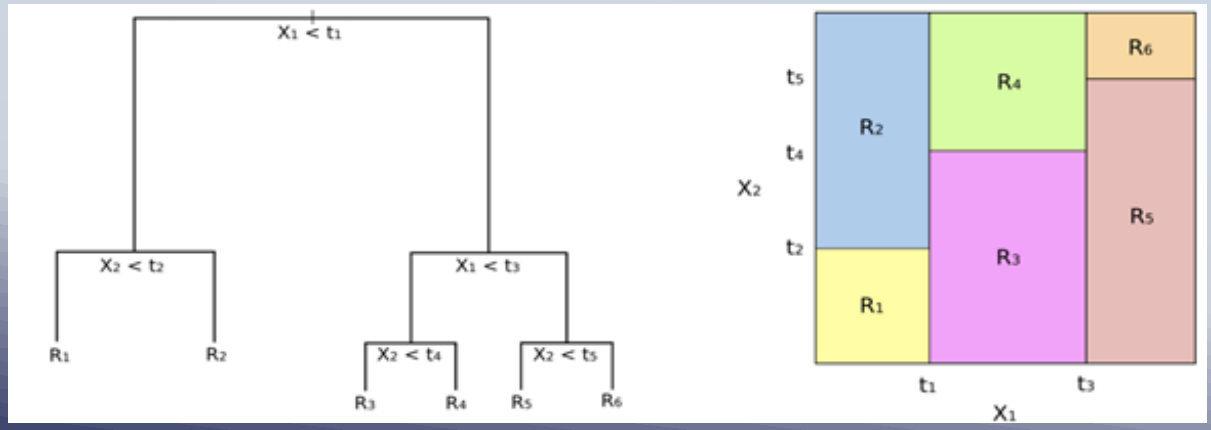
Los árboles de decisión son algunos de los algoritmos más frecuentes para la toma de decisiones en Machine Learning .

Aunque su capacidad predictiva es superada por otros algoritmos, son de uso frecuente por su sencilla implementación y fácil interpretación.

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.

Para dividir el espacio muestral en sub-regiones es preciso aplicar una serie de reglas o decisiones, para que cada sub-región contenga la mayor proporción posible de individuos de una de las poblaciones.

Si una sub-región contiene datos de diferentes clases, se subdivide en regiones más pequeñas hasta fragmentar el espacio en sub-regiones menores que integran datos de la misma clase.



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

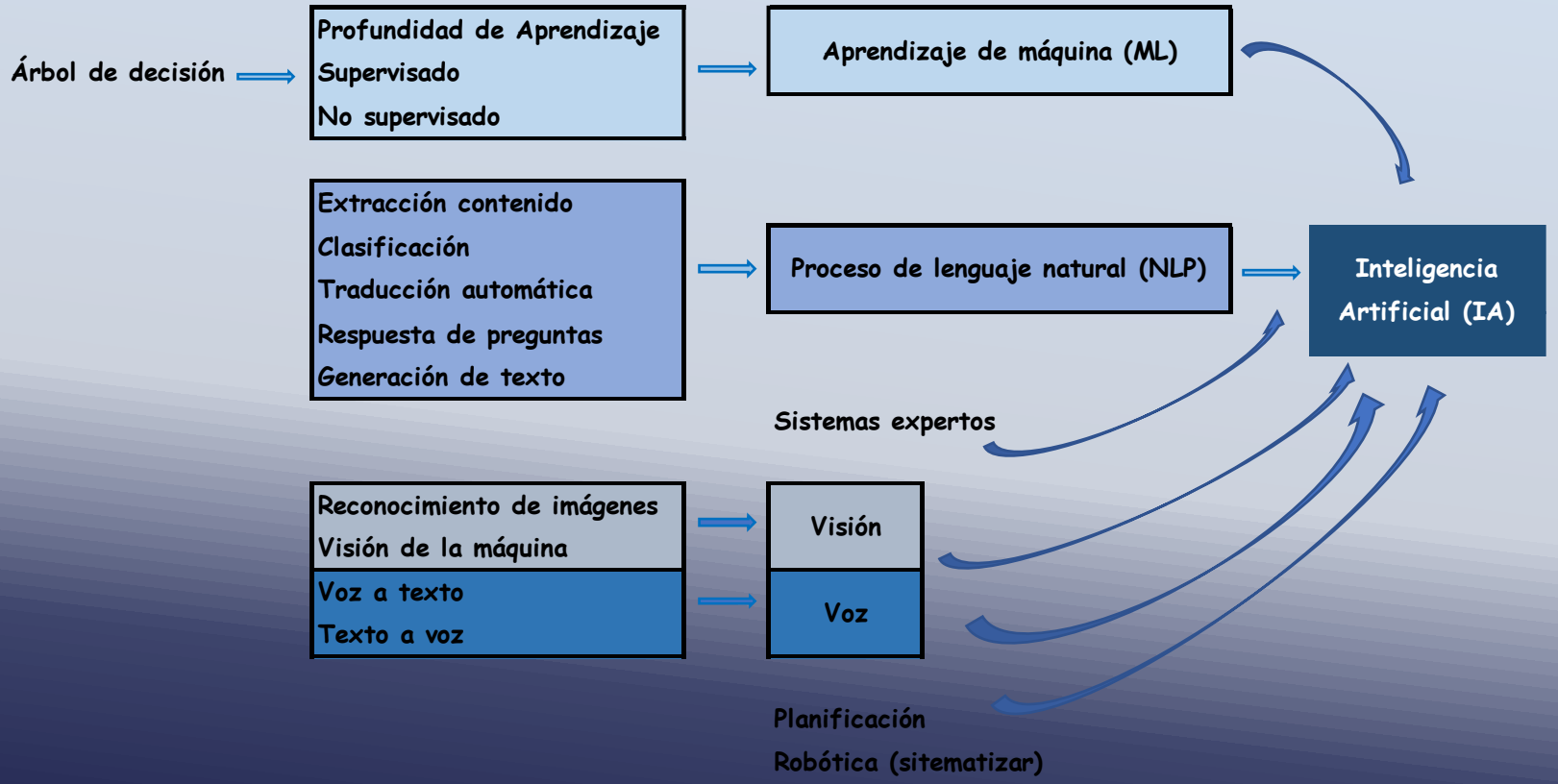
- Variable dependiente: estaríamos ante un problema de regresión.
- Variable categórica: nos enfrentaríamos a un problema de clasificación.

En 1984, los pioneros de la metodología del árbol de clasificación con aplicación al aprendizaje automático, también llamada metodología CART, (Classification and Regression Trees), fueron Leo Breiman, Jerome Friedman, Richard Olshen y Charles Stone.

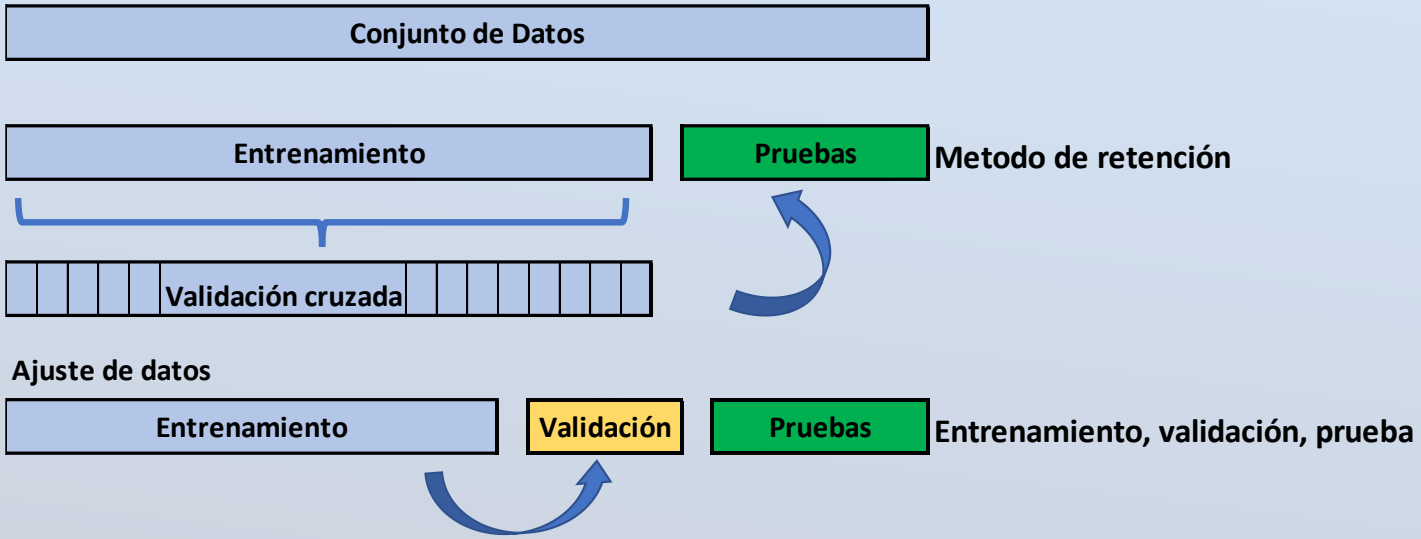
Los algoritmos del árbol de decisión son de aprendizaje automático y se clasifican en dos tipos:

- Supervisados.
- No supervisados.

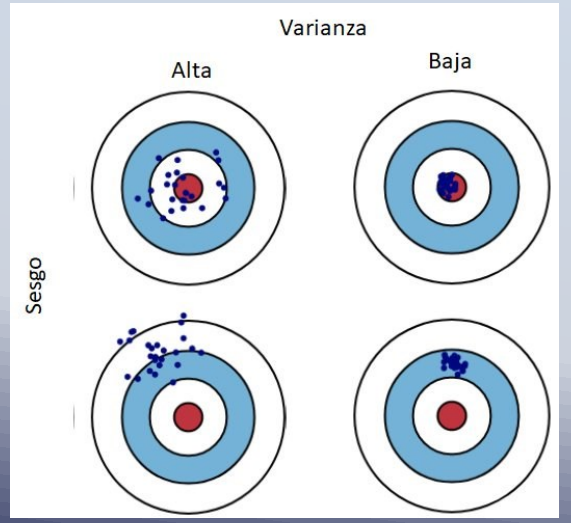
Un árbol de decisión se considera un algoritmo supervisado de aprendizaje automático, para que aprenda del modelo se necesita una variable dependiente en el entrenamiento.



Conjunto de entrenamiento y conjunto de prueba

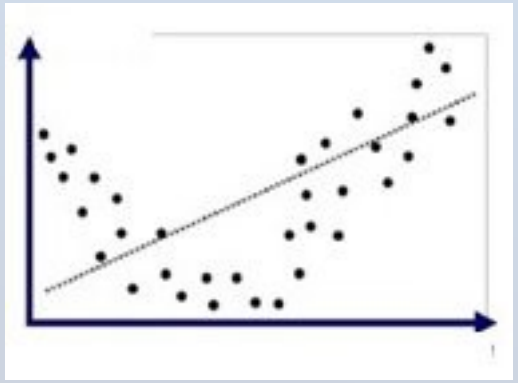


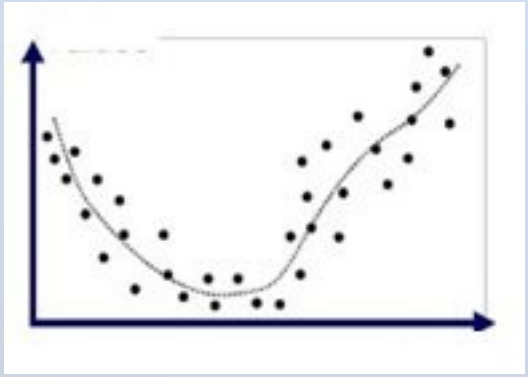
Sesgo y varianza

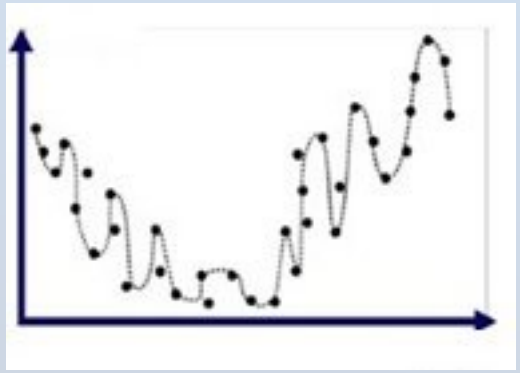


Sobreajuste (overfitting)

Explicar el conjunto de datos de entrenamiento, en lugar de encontrar patrones que generalizan. El modelo se ajusta bien al conjunto de datos de entrenamiento pero falla en el conjunto de datos de test.



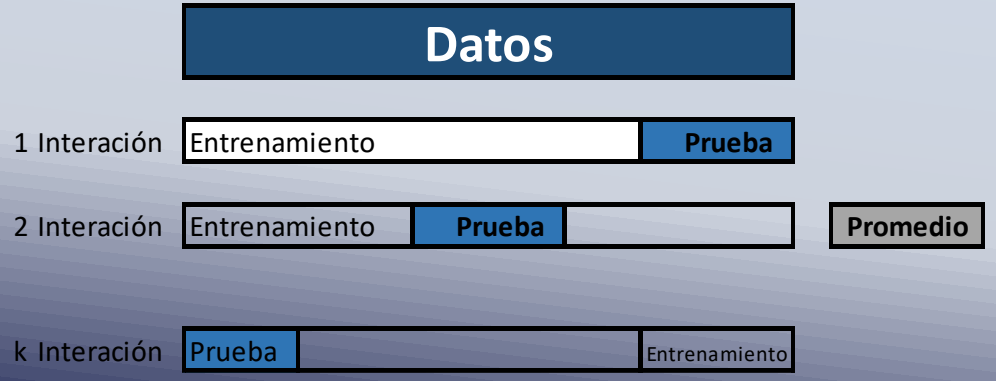




Validación Cruzada.

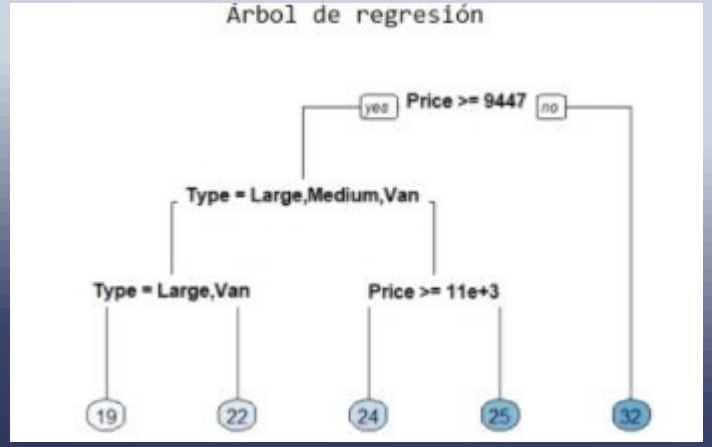
Técnicas para validar modelos o clasificaciones

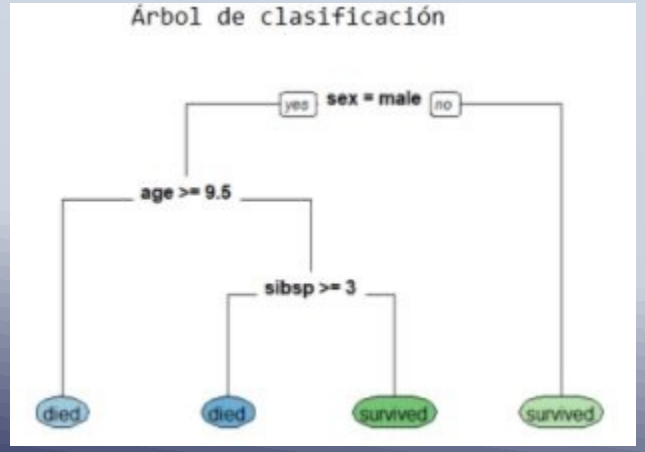
Se estima la precisión con la que el modelo queda ajustado a la predicción o ajuste



Árboles de Regresión vs Árboles de Clasificación

Regresión	Clasificación
Variable dependiente es continua	Variable dependiente es categórica
Valores de los nodos terminales se reducen a la media de las observaciones en esa región.	El valor en el nodo terminal se reduce a la moda de las observaciones del conjunto de entrenamiento que han “caído” en esa región.



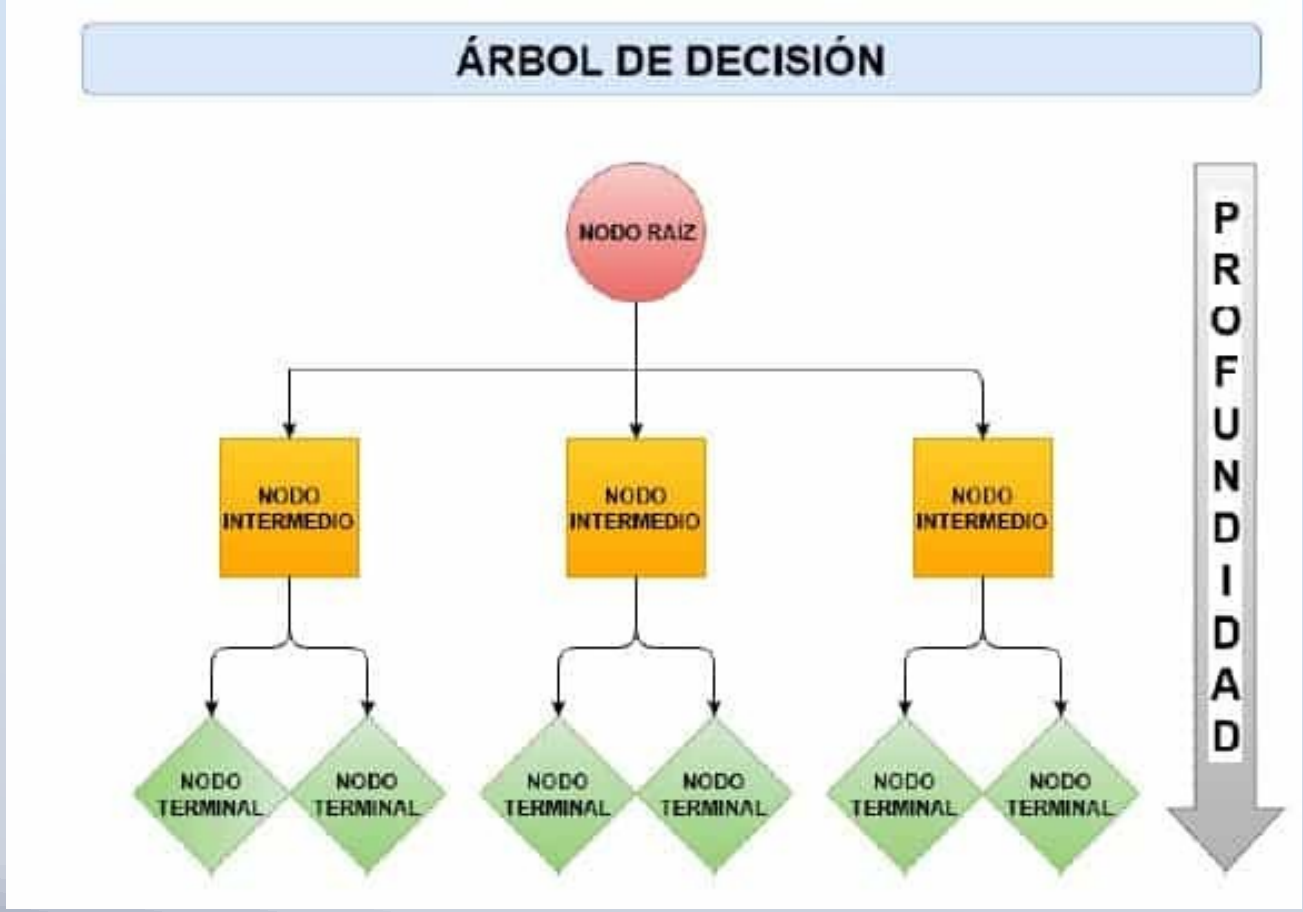


Estructura de un árbol de decisión

Los árboles de decisión están formados por nodos y se ve de arriba hacia abajo.

Dentro de un árbol de decisión existen diferentes tipos de nodos:

- Primer nodo o nodo raíz: es la primer división en asociado a la variable más importante.
- Nodos internos o intermedios: después de la primer división se encuentran los nodos, que vuelven a dividir la muestra de datos en función de las variables.
- Nodos terminales u hojas: se ubican en la parte inferior del gráfico, indicando la clasificación final



Y se tiene la profundidad de un árbol, que se determina por el número máximo de nodos de una rama.

Ventajas

- Son fáciles de construir, interpretar y visualizar.
- Útil en exploración de datos: identificar importancia de variables a partir de cientos de variables.
- Al tener ausencia de información no se llega hasta el nodo final, pero sí se puede hacer predicciones promediando las hojas del sub-árbol que se tengan.
- No es preciso que se cumplan los supuestos como en la regresión lineal (linealidad, normalidad de los residuos, homogeneidad de la varianza, etc.).
- 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.
- Permiten relaciones no lineales entre las variables explicativas y la variable dependiente.
- Se pueden clasificar variables numéricas.
- Es un método no paramétrico (i.e., no hay suposición acerca del espacio de distribución y la estructura del clasificador)

Desventajas

- 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.
- Se ven influenciadas por los *outliers*, creando árboles con ramas muy profundas que no predicen bien para nuevos casos. Se deben eliminar dichos *outliers*.
- No suelen ser muy eficientes con modelos de regresión.
- Crear árboles demasiado complejos puede conllevar que no se adapten bien a los nuevos datos. La complejidad resta capacidad de interpretación.
- Se pueden crear árboles sesgados si una de las clases es más numerosa que otra.
- Se pierde información cuando se utilizan para categorizar una variable numérica continua.
- Inestabilidad: un pequeño cambio en los datos puede modificar ampliamente la estructura del árbol. Por lo tanto, la interpretación no es tan directa como parece.

Árbol de Regresión

Un modelo de árbol de regresión es una descripción condicional de Y dado X. Los dos componentes fundamentales del modelo son: un árbol binario b, nodos terminales y el vector de parámetros $\theta = (\theta_1, \theta_2, \dots, \theta_b)$ Donde el parámetro θi esta asociado al nodo terminal Ni.

Un árbol de regresión crea un modelo explicativo y predictivo para una variable cuantitativa dependiente basada en variable explicativas cuantitativas y cualitativas.

Regla de división

Para dividir un nodo de un árbol binario se consideran 2 pasos:

1) Se selecciona una variable Xi
2) Si la variable seleccionada es cuantitativa, se selecciona aleatoriamente un valor r y se asigna al nodo hijo de las observaciones que cumplan la condición Xi≤r, las restantes se asignan al otro nodo hijo. Si la variable selecciona resulta cualitativa, se selecciona un subconjunto A de las categorías Xi y las observaciones con valores pertenecientes al conjunto A se asignan al nodo hijo, las restantes al otro nodo

Ramificación de un árbol

- La decisión de hacer divisiones estratégicas afecta altamente la precisión del árbol.
- Los criterios de decisión son diferentes para árboles de clasificación y regresión.
- Existen varios algoritmos para decidir la ramificación.
- La creación de subnodos incrementa la homogeneidad de los subnodos resultantes. Es decir, la pureza del nodo se incrementa respecto a la variable objetivo.
- Se prueba la división con todas las variables y se escoge la que produce subnodos más homogéneos.

Algunos algoritmos más comunes para la selección:

1. Índice Gini,
2. Chi Cuadrado,
3. Ganancia de la información y
4. Reducción en la varianza

1) Índice Gini

Seleccionamos aletoriamente dos items de una población, entonces estos deben ser de la misma clase y la probabilidad de esto es 1 si la población es “pura”.

$$GINI(t) = 1 - \sum_{i=1}^{n} (p_i)^2$$

1. Variable objetivo categórica: “*Éxito*” o “*Fracaso*”
2. Solo divisiones binarias.
3. A mayor valor de índice Gini, mayor la homogeneidad
4. CART (Classification and Regression Tree) usa el método de Gini para la división binaria

2) Chi Cuadrado

Un algoritmo para encontrar la significancia estadística de las diferencias entre sub_nodos y un nodo padre. Se mide a partir de la suma de los cuadrados de las diferencia estandarizadas entre las frecuencias observadas y esperadas de la variable objetivo.

$$T^2 = \sum_{i=1}^{c} \sum_{J=1}^{K} \frac{(Oij - Eij)^2}{Eij}$$

1. Variable objetivo categórica “Éxitos” o “Fracasos”.
2. Dos o más divisiones
3. **A más alto valor de Chi-Cuadrado, más alta la significancia estadística de las diferencias entre cada nodo y el nodo padre.**

3) Entropía de Shannon

El término de entropía (del griego tropos= cambio, transformación) lo utiliza Clausius en 1851, es uno de los que trabajaron en la segunda ley de la termodinámica. La entropía debe ser definida tomando en cuenta consideraciones estadísticas y probabilísticas. Una definición es: “una medida de incertidumbre promedio, la cual se calcula a partir de la probabilidad de ocurrencia de cada uno de los eventos”, otra “Medida de desorden de un sistema”

Para realizar le entropía es necesario el uso de la Teoría de Probabilidad, cuya función es manejar una gran cantidad de acontecimientos o eventos que individualmente son casuales, pero en conjunto predecibles en términos probabilísticos.

$$H = -\sum_{i=1}^{n} P_i * log_2 P_i$$

Donde:

H Información contenida en una muestra

Pi Proporción de individuos de una misma especie, respecto al total de individuos. Pi=ni/N

Ganancia de información

- Un nodo menos impuro requiere menos información para ser descrito mientras un nodo más impuro necesita más información.
- La teoría de la información es una medida para definir este grado de desorganización en un sistema denominado como Entropía.
- Muestra completamente homogénea = entropía 0.
- Muestra igualmente dividida (50% – 50%) = entropía 1.

4) Reducción en la varianza (regresión)

Los algoritmos anteriores se aplicaban para problemas de clasificación con variables objetivo categóricas. La reducción en la varianza es un algoritmo usado para variables objetivo continuas (problemas de regresión). Este algoritmo usa la fórmula estándar de la varianza para escoger el criterio de división. La división con la varianza más baja se escoge para dividir la población:

Parámetros del modelo y como evitar sobreajuste en árboles de decisión

El sobreajuste es uno de los desafíos más importantes en el proceso de modelación de árboles de decisión. Si no se definen límites, el árbol tendrá un 100% de precisión en el conjunto de datos de entrenamiento. En el peor caso tendrá una hoja por cada observación.

Dos formas de evitar el sobreajuste:

- Definir restricciones sobre el tamaño del árbol y
- Podar el árbol.

Definir restricciones sobre el tamaño del árbol (prepruning)

Uso de parámetros para definir un árbol. Los parámetros son independientes de la herramienta de programación (R & Python)

1. **Mínimo de observaciones para dividir un nodo**
   - Mínimo número de muestras u observaciones que se requieren en un nodo para ser considerado para ramificación.
   - Valores más altos previenen que el modelo aprenda relaciones muy específicas.
   - Valores demasiado altos pueden causar un pobre ajuste del modelo. El parámetro debe ajustarse usando validación cruzada.

2. Mínimo número de observaciones para un nodo terminal

- Valores más bajos son necesarios para problemas de clases no balanceadas.

3. Máxima profundidad del árbol (vertical)

- Una mayor profundidad permite aprender relaciones más específicas.
- Debe ajustarse con validación cruzada.

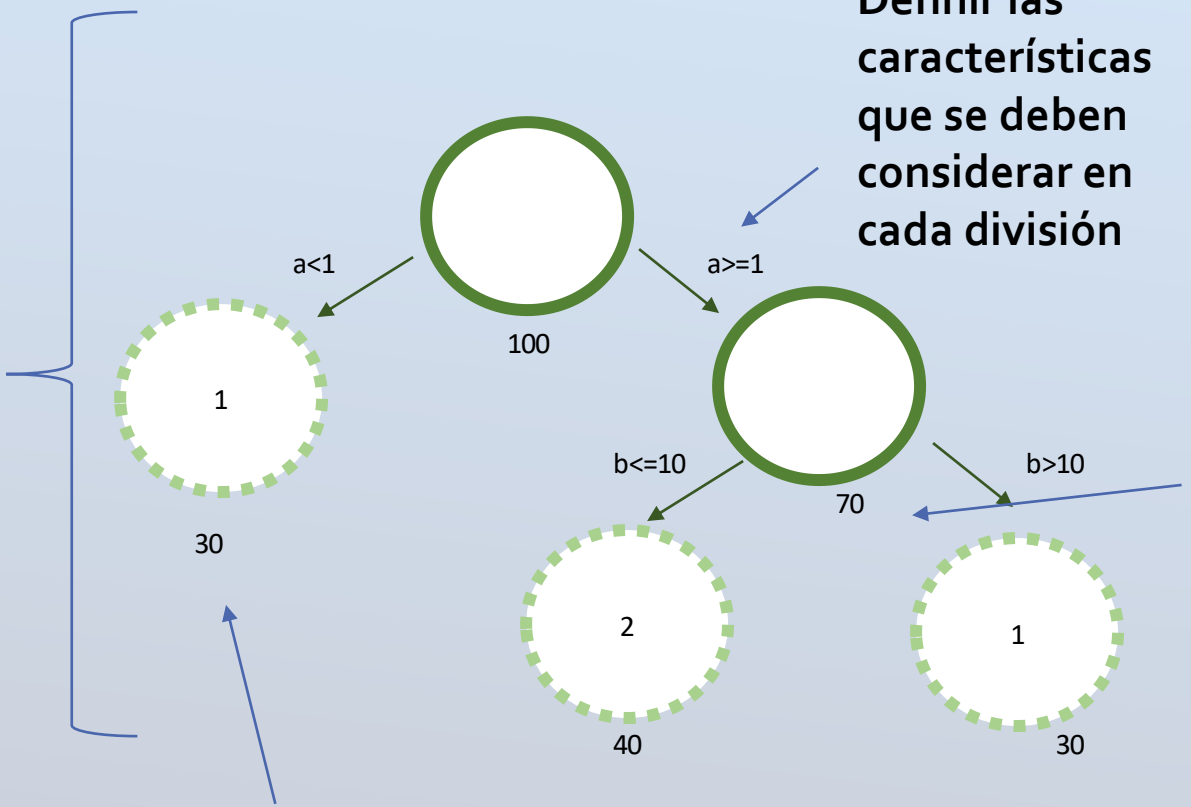
4. Máximo número de nodos hoja

- Se puede definir en lugar de máxima profundidad. Profundidad n = máximo 2^n hojas

5. Máximo número de atributos a considerar para la ramificación

- Seleccionados aleatoriamente.
- Como regla general, la raíz cuadrada del número total de atributos funciona apropiadamente. Sin embargo, se debe probar hasta un 30%-40% del número total de atributos.

Ajustar la profundidad aseguraría que no se dividirán más incluso si las hojas tienen la muestra mínima requerida



Definir las características que se deben considerar en cada división

No permitiría dividir ningún nodo con menos de 70 muestras

No permitir que ningún nodo hoja tenga menos de 30 muestras

Poda del árbol (postpruning)

Un árbol de decisión con restricciones no verá en una carretera un camión “descompuesto” más adelante y adoptará un “greddy approach” al tomar la izquierda (para esquivar el obstaculo). Al contrario, si usamos el proceso de poda, en efecto se mirará algunos pasos adelante para tomar una decisión antes de llegar a donde esta el camión descompuesto.

Por lo tanto sabemos que hacer la poda es mejor. Su implementación es sencilla.

- Construir el árbol a un profundidad extensa.
- Remover las hojas que den un valor negativo comparado desde la raíz. Existen varios criterios que pueden ser utilizados, algunos basados en heurísticas y otros en parámetros de regularización (penalización).

Ejemplo: Supongamos que un nodo de decisión da una ganancia de -15 (pérdida de 15) y la siguiente ramificación da una ganancia de 25. Un árbol de decisión simple pararía en el paso uno, sin embargo, el proceso de poda considerará la ganancia general de +10 y mantendrá ambas hojas.

Revisar las librerías como **:** python Scikit, o xgboost en Python y en R rpart para ver este tipo de opciones

Arboles de decisión y modelos lineales

- Si la relación entre la variable dependiente y la(s) independiente se aproxima a un modelo lineal, la regresión lineal dará mejores resultados que un modelo de árbol de decisión.
- Si la relación es compleja y altamente no lineal, entonces el árbol de decisión tendrá mejores resultados de que un método clásico de regresión.
- Si se quiere construir un modelo que sea fácil de explicar, entonces un modelo de árbol de decisión será mejor que un modelo lineal.

Ejemplo

Sexo

Estatura

Clase A / B