Pontificia Universidad JAVERIANA Bogotá

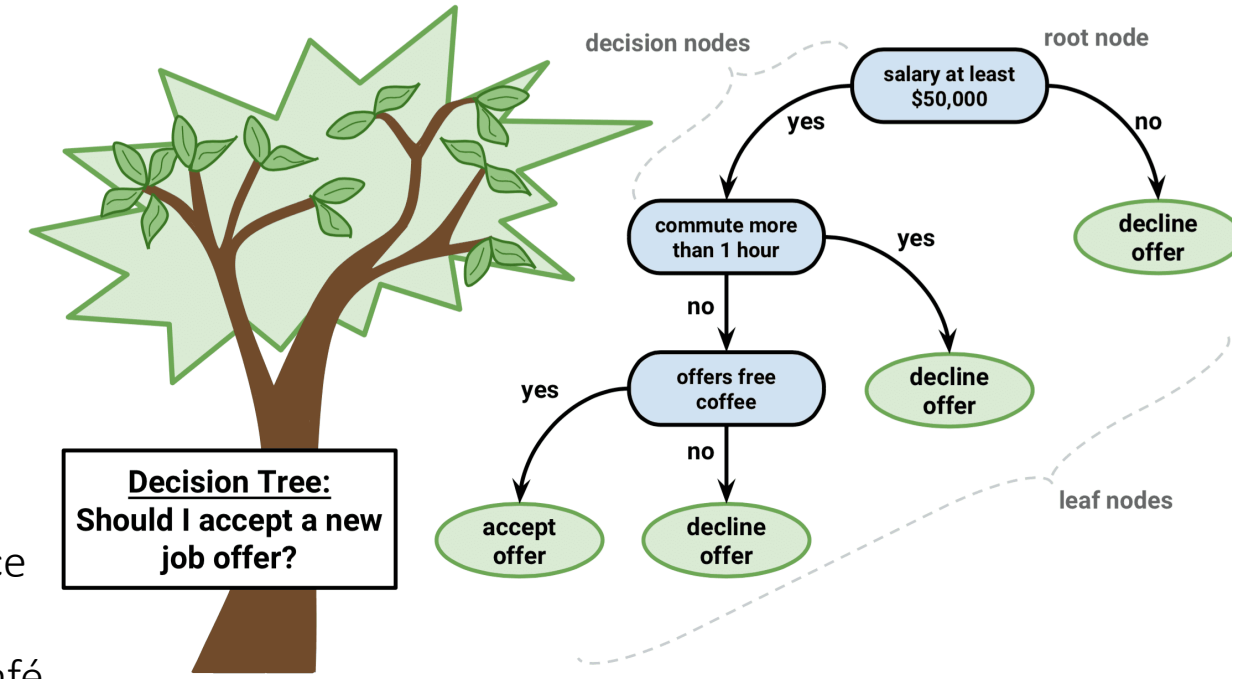
2.3 Arboles de Decisión

Educación **Continua**

Generamos experiencias educativas

Arboles de Decisión

- Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica.
- La decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta sus hojas .



De este árbol se pueden extraer las siguientes reglas (caminos)

1. Salario < $50.000 -> Rechazar.
2. Salario >= $50.000 + Transporte más de 1 hora -> Rechazar.
3. Salario >= $50.000 + Transporte menos de 1 hora + NO ofrece café gratis -> Rechazar.
4. Salario >= $50.000 + Transporte menos de 1 hora + ofrece café gratis -> Aceptar.

Usted puede construir un árbol de decisión a partir de sus propias reglas.
Sin embargo, esto puede que aplique solo para usted, y no es útil para generalizar en un grupo de interés.
En este curso nos enfocaremos en arboles de decisión que “aprende” reglas a partir de datos.

Ejemplo

Suponga que se desea construir un Árbol de Decisión para generar un sistema de recomendación de géneros de películas.

Se ha tomado una muestra de 10 personas a las que se registró:

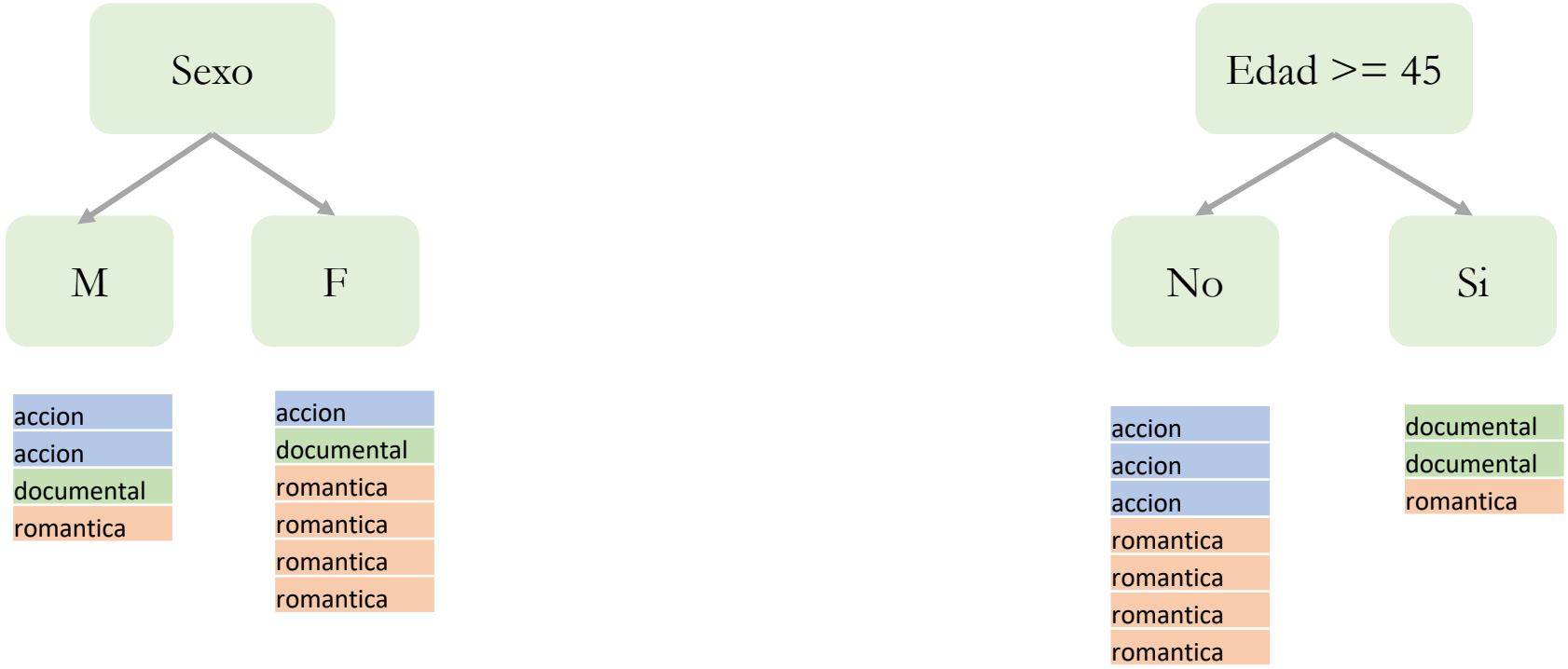
1. Sexo
2. Edad
3. Genero de película de su preferencia

Sexo	Edad	Genero de Película
F	20	Romántico
F	24	Acción
M	25	Acción
F	35	Romántico
M	35	Acción
M	40	Romántico
F	40	Romántico
F	50	Documental
F	55	Romántico
M	60	Documental

¿Cómo se Construye un Arbol de Decisión?

Pontificia Universidad JAVERIANA Bogotá

Divide y reinaras: Para todas las variables se generan divisiones de las clases y se calcula la impureza del nodo.



Al parecer la edad genera una división más “pura” (homogénea) de los géneros de películas.

Pero ¿Cómo medimos la pureza?

Indice Gini

- El índice Gini es una medida de la impureza o la heterogeneidad en un conjunto de datos utilizado comúnmente en árboles de decisión y clasificación en Machine Learning.
- Se utiliza para evaluar cuán bien un nodo de un árbol de decisión divide los datos en clases o categorías.
- Un valor de Gini más bajo indica una mayor pureza, lo que significa que los datos en el nodo están más homogéneos en términos de la variable objetivo que se está prediciendo.
- l cálculo del índice Gini se realiza mediante la siguiente fórmula:

$$GINI_{nodo} = 1 - \sum_{k=1}^{K} p_k^2$$

- Donde, k son las clases en cada nodo y p_k la proporción la clase k dentro del nodo.
- Para cada partición se saca la suma ponderada de cada uno de los nodos que genera.

Indice Gini

- Recuerde: $GINI_{nodo} = 1 - \sum_{k=1}^{K} p_k^2$
- Para la partición de la variable sexo:

$$GINI_{Sexo\_M} = 1 - \left[\left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = 0{,}625$$

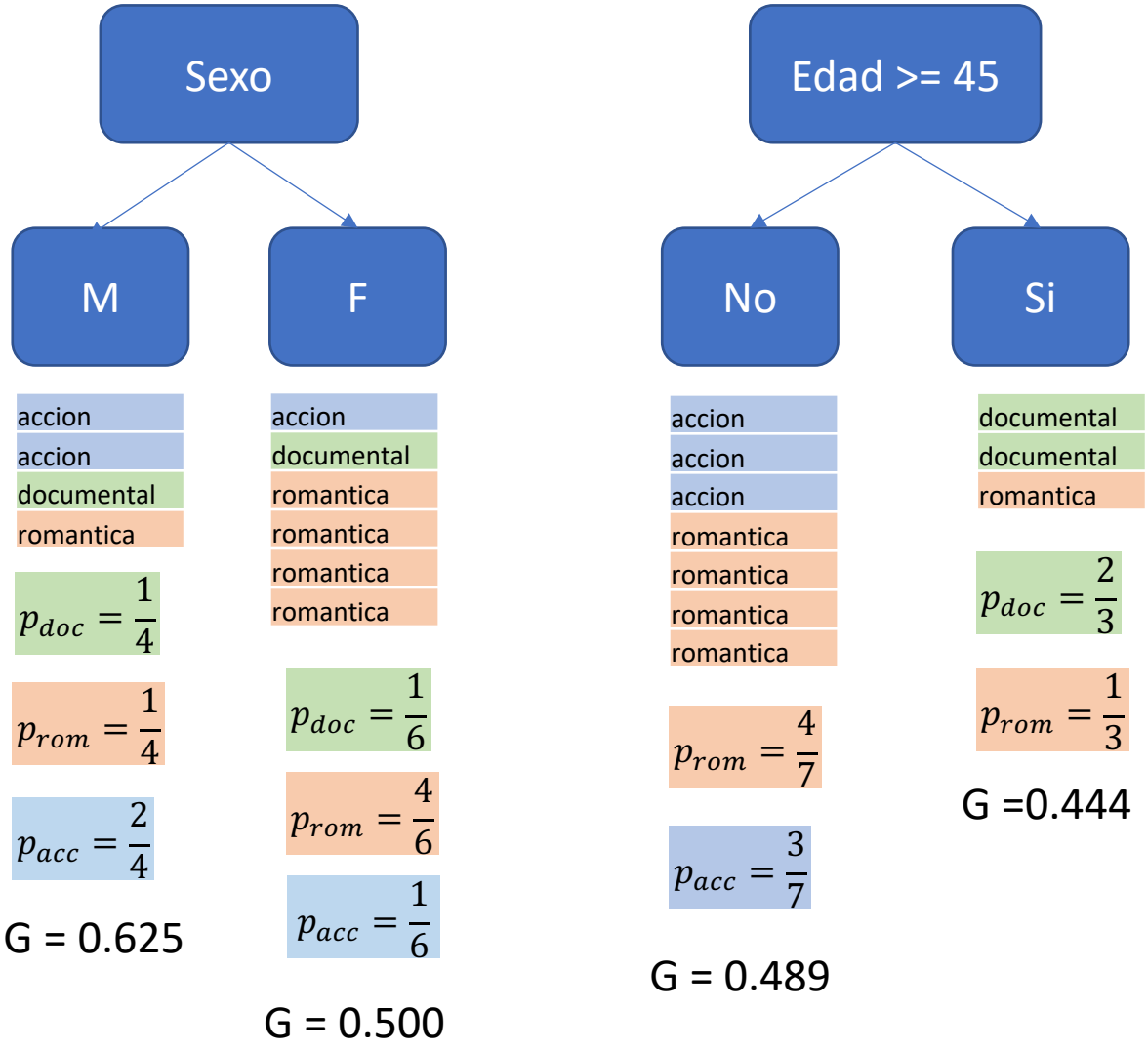
$$GINI_{Sexo\_F} = 1 - \left[\left(\frac{1}{6}\right)^2 + \left(\frac{4}{6}\right)^2 + \left(\frac{1}{6}\right)^2\right] = 0.500$$

- De esta manera:

$$GINI_{Sexo} = \left(\frac{4}{10}\right)0{,}625 + \left(\frac{6}{10}\right)0.500 = 0.550$$

$$GINI_{edad \geq 45} = \left(\frac{7}{10}\right)0{,}489 + \left(\frac{3}{10}\right)0.444 = 0.476$$

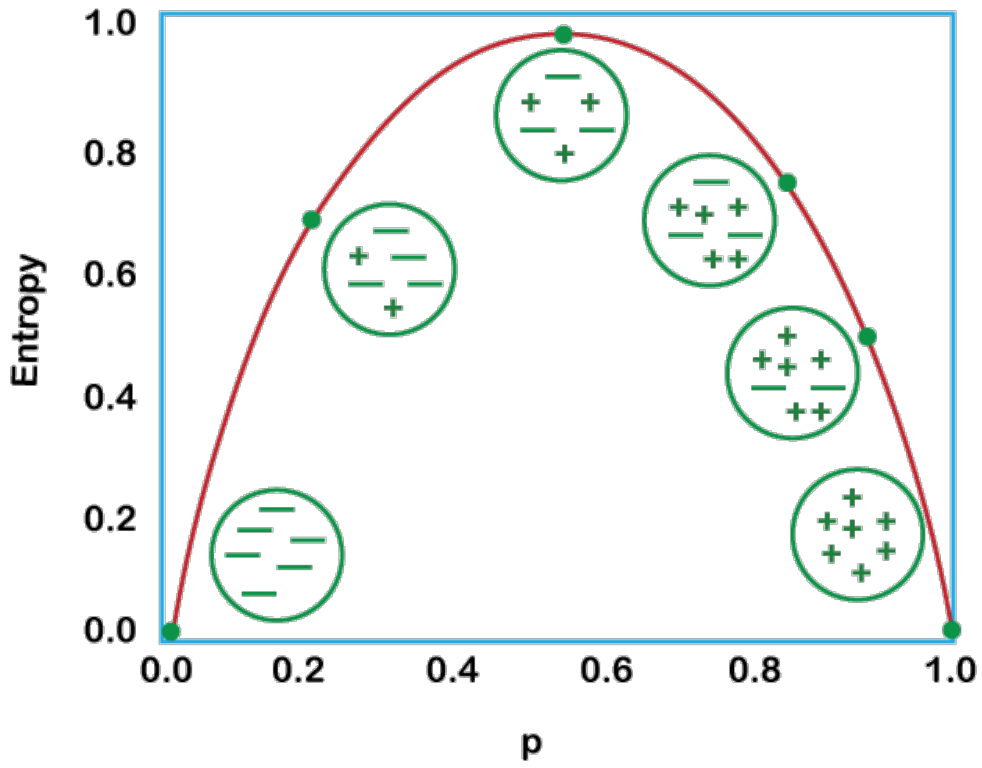
- Con este resultado la mejor partición (menor impureza) se genera con la partición edad >= 45.
- A partir de los nodos generados se pueden generar más particiones teniendo en cuenta el mismo criterio.



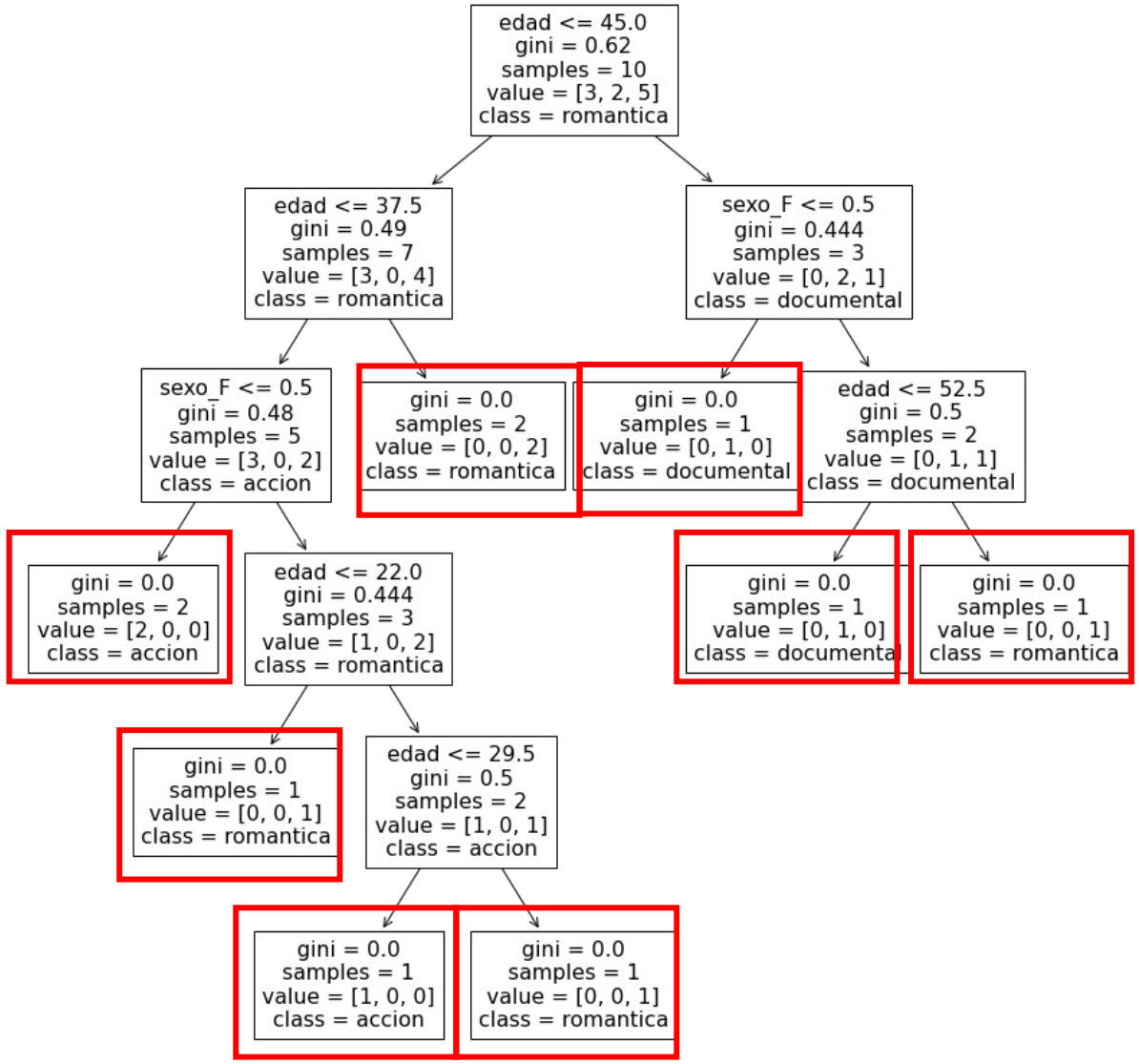
Entropía

- La entropía es otra medida de impureza o desorden en un conjunto de datos que se utiliza en árboles de decisión y clasificación en Machine Learning.
- Al igual que el índice Gini, la entropía se utiliza para evaluar cuán bien un nodo de un árbol de decisión divide los datos en clases o categorías.
- Un valor de entropía más bajo indica una mayor pureza en el nodo, lo que significa que los datos en el nodo son más homogéneos en términos de la variable objetivo que se está prediciendo.

$$ENTROPIA_{nodo} = -\sum_{k=1}^{K} p_k \log(p_k)$$



¿Cuando detener la Partición?



Observe que total hay 8 nodos terminales para 10 instancias y están compuestos por una misma clase (óseo error 0). Si bien, esto parece funcionar perfectamente, genera aun problema que se denomina sobreajuste, lo que hace que el árbol no sea efectivo para predecir nuevas observaciones.

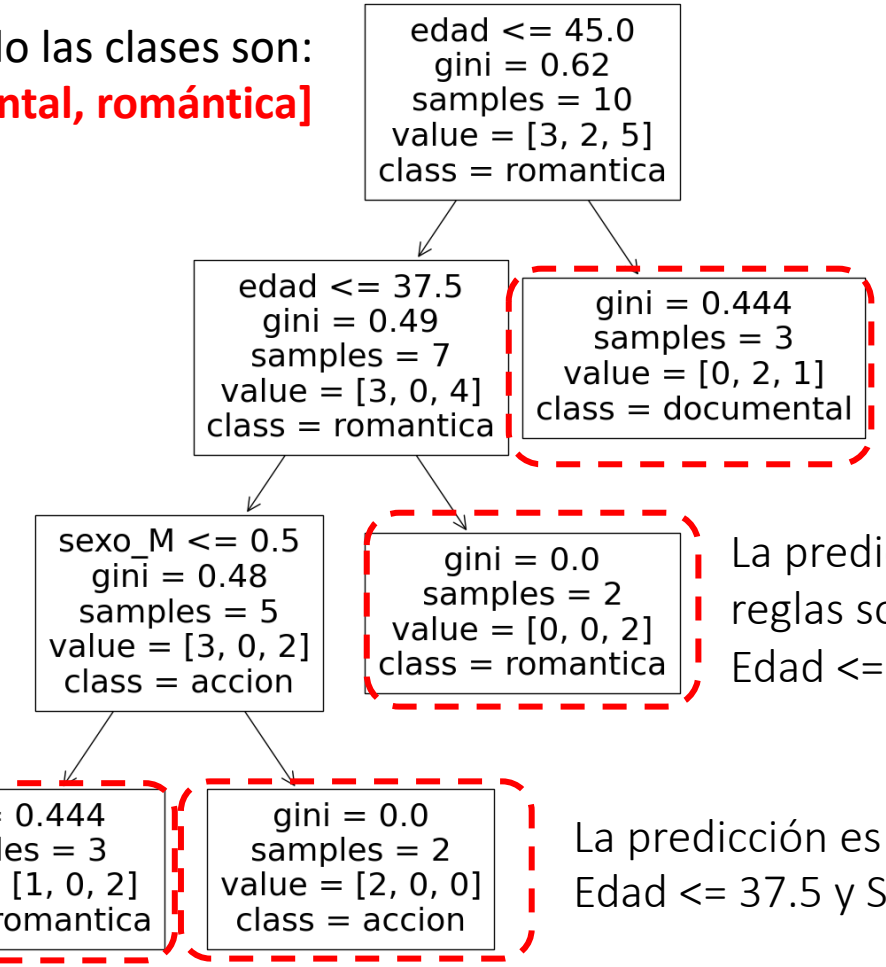
La profundidad del árbol es un parámetro que se debe calibrar para evitar el sobreajuste (overfitting).

El procedimiento para calibrar parámetros en un árbol o cualquier otro algoritmo se denomina validación cruzada. Este procedimiento lo veremos con mayor detalle más adelante.

Interpretación del árbol de decisión

Suponga que se obtiene el siguiente árbol de decisión después calibrar el criterio de parada…

En cada nodo las clases son:
[acción, documental, romántica]



El árbol según la clase mayoritaria del nodo terminal.

En este nodo las clases son:
Acción: 1 voto (instancia)
Romántica: 2 votos (instancias)

La predicción es documental y las reglas son:
Edad > 45

La predicción es romántica y las reglas son:
Edad <= 45 y Edad > 37.5

La predicción es romantica y las reglas son:
Edad <= 37.5 y Sexo Femenino

La predicción es acción y las reglas son:
Edad <= 37.5 y Sexo Masculino

Arboles de regresión

Los Árboles de Regresión son un tipo de modelo utilizado para resolver problemas de regresión.

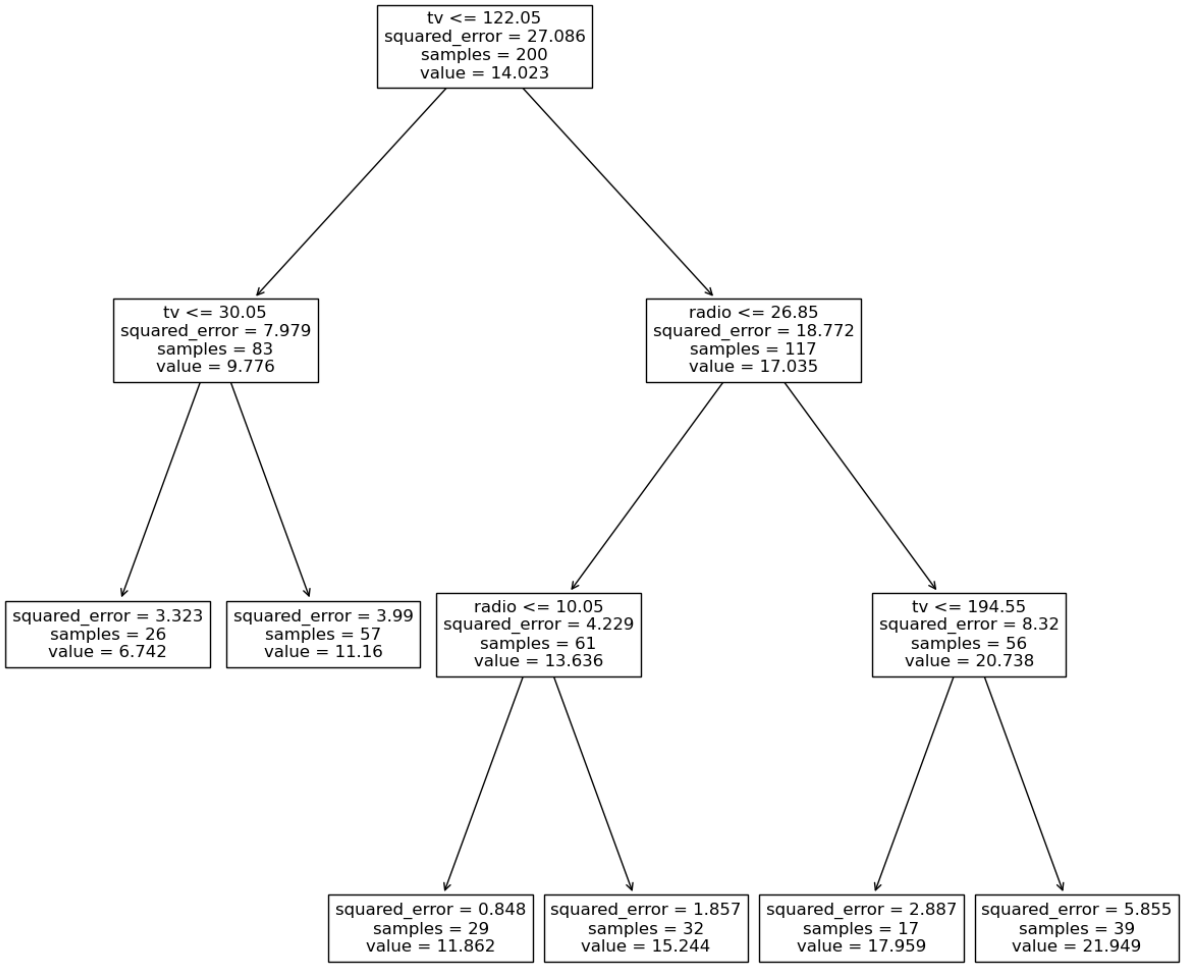
A diferencia de los Árboles de Clasificación, los Árboles de Regresión predicen valores numéricos continuos en lugar de clases discretas.

Estructura de un Árbol de Regresión

- Al igual que los Árboles de Clasificación, un Árbol de Regresión tiene una estructura jerárquica compuesta por nodos.
- Cada nodo representa una decisión basada en una característica y un valor de umbral.
- Las hojas del árbol contienen valores numéricos que representan las predicciones.
- La predicción en el nodo termina se basa en el promedio de las instancias que componen el nodo
- El criterio de división es la minimización de la suma de los errores al cuadrado de las particiones del nodo:

$$RSS_E = \sum_{j=1}^{J} \sum_{i \in R} (Y_i - \bar{Y}_{Rj})$$

Árbol de Regresión (Gastos en Publicidad)



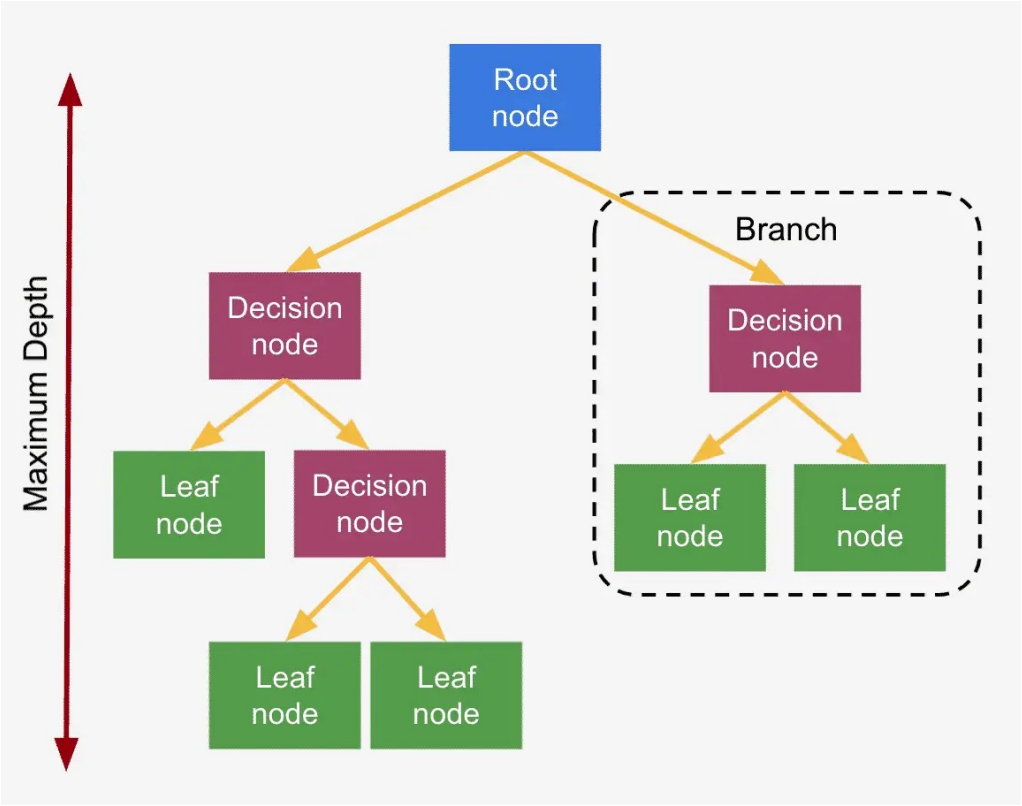
Criterios de parada

En resumen, un árbol de decisión se construye con los siguientes pasos:

1. Seleccionar una variable para dividir en el nodo actual (usando Gini, entropía u otros criterios).
2. Dividir el nodo en subnodos basados en los valores de la variable seleccionada.
3. Repetir el proceso en cada subnodo hasta que se cumpla un criterio de parada

Los criterios de parada para la construcción del árbol pueden ser:

- Profundidad máxima (**máximum depth**)
- Número mínimo de instancias por nodo (**min samples leaf**)
- El número mínimo de instancias necesarias para dividir un nodo interno (**min samples split**)
- Mínima disminución de impureza. Un nodo se dividirá si esta división induce una disminución de la impureza mayor o igual a este valor (**min impurity decrease**).



Poda del Árbol de Decisión

Busca encontrar el equilibrio óptimo entre la simplicidad del árbol y su capacidad para hacer predicciones precisas. Con esto se busca evitar el sobre ajuste.

Se basa en la idea de que un árbol más pequeño (menos profundo) con menos ramas puede ser más generalizable y menos propenso al sobreajuste.

Proceso de Poda con Construcción del árbol:

1. Inicialmente, se construye un árbol de decisión completo sin restricciones, lo que podría llevar al sobreajuste.
2. Cálculo del Costo-Complejidad (Cost Complexity Measure, CCM): Para cada nodo del árbol, se calcula un valor llamado CCM. El CCM mide la complejidad del nodo en relación con su capacidad de clasificación precisa y se calcula de la siguiente manera:

$$CCM(T) = Error(T) + CCM(T) = Error(T) + \alpha|T|$$

Donde:

CCM(T) es el valor de CCM del nodo T.

Error(T) es la tasa de error del nodo T en el conjunto de entrenamiento.

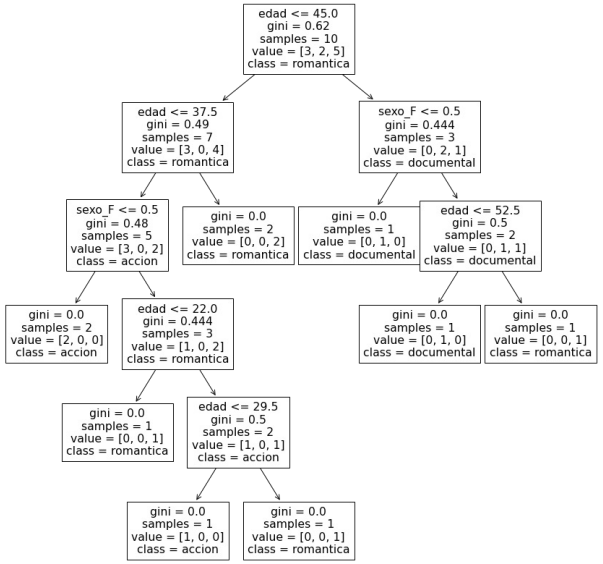
|T| es el número de nodos hoja en el subárbol TT.

α es un hiperparámetro que controla el equilibrio entre la precisión y la simplicidad del árbol.

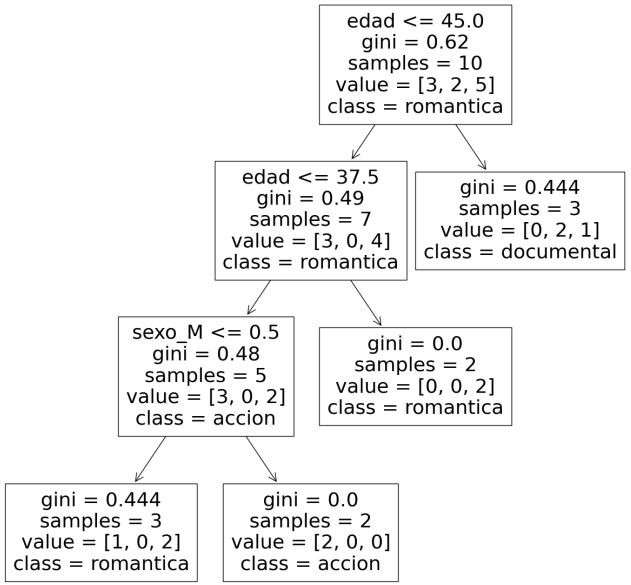
Poda del Árbol de Decisión

3. **Poda del Árbol:** A continuación, se busca el subárbol más pequeño que minimiza el valor de CCM. Esto se hace mediante la poda de las ramas con los valores más altos de CCM.
4. **Selección de Alpha:** Para determinar el valor óptimo de α, se utiliza la validación cruzada (por ejemplo, validación cruzada K-fold) para evaluar el rendimiento del árbol en diferentes valores de αα y seleccionar aquel que minimice el error de validación.
5. **Árbol Final:** El árbol resultante después de la poda con Minimal Cost-Complexity Pruning es el árbol final que equilibra la simplicidad y la capacidad de predicción.

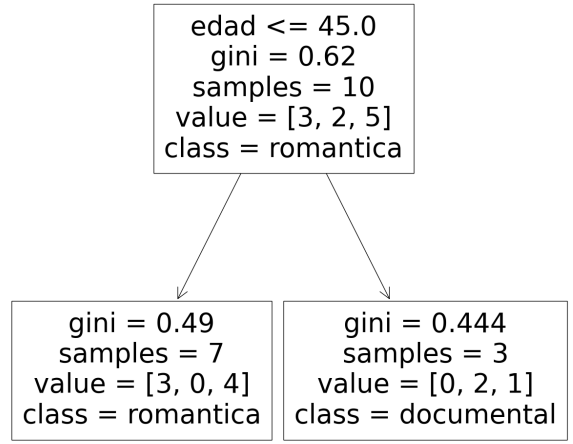
Árbol sin poda (Alpha = 0)



Árbol podado con Alpha = 0.1



Árbol podado con Alpha = 0.12



¿Qué árbol es mejor?

Ajuste de Hiperparámetros

Pontificia Universidad JAVERIANA Bogotá

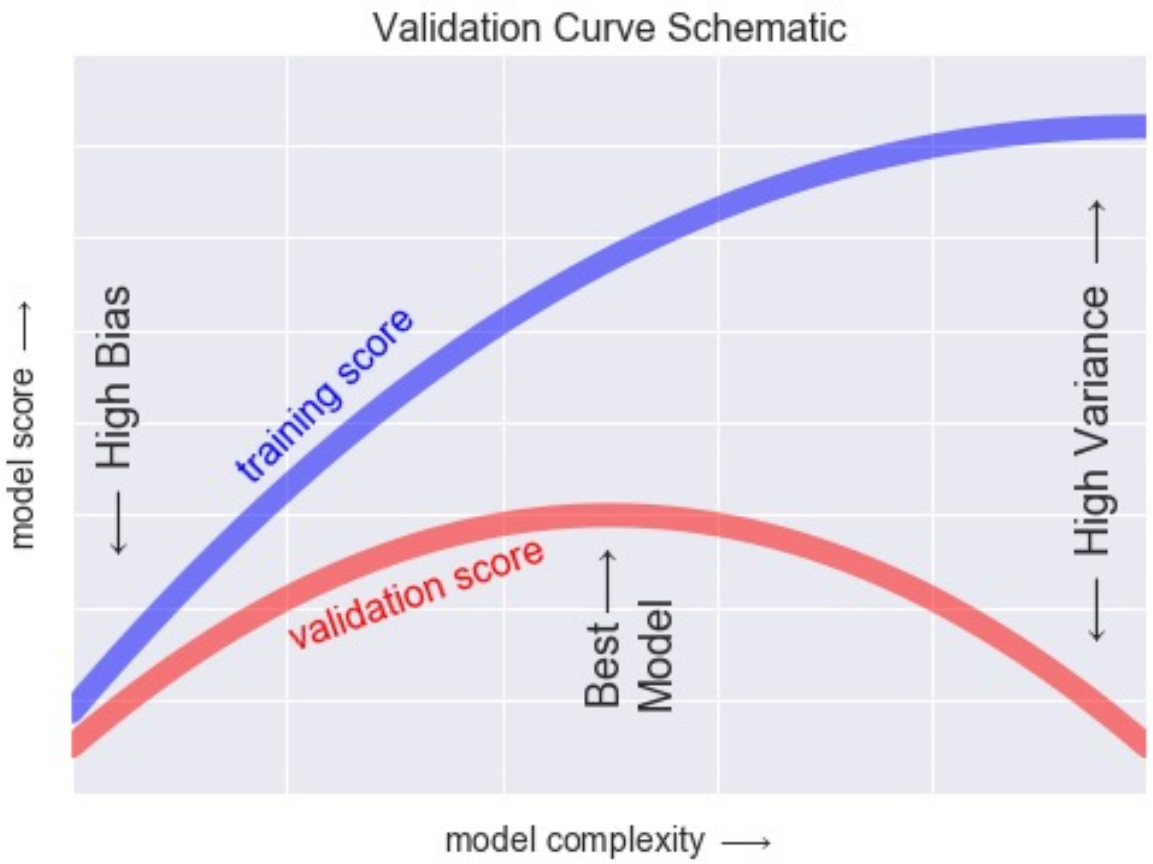
Los hiperparámetros son configuraciones ajustables en algoritmos de Machine Learning que afectan el rendimiento y el comportamiento del modelo.

El ajuste de hiperparámetros es el proceso de encontrar la mejor combinación de hiperparámetros para un modelo dado.

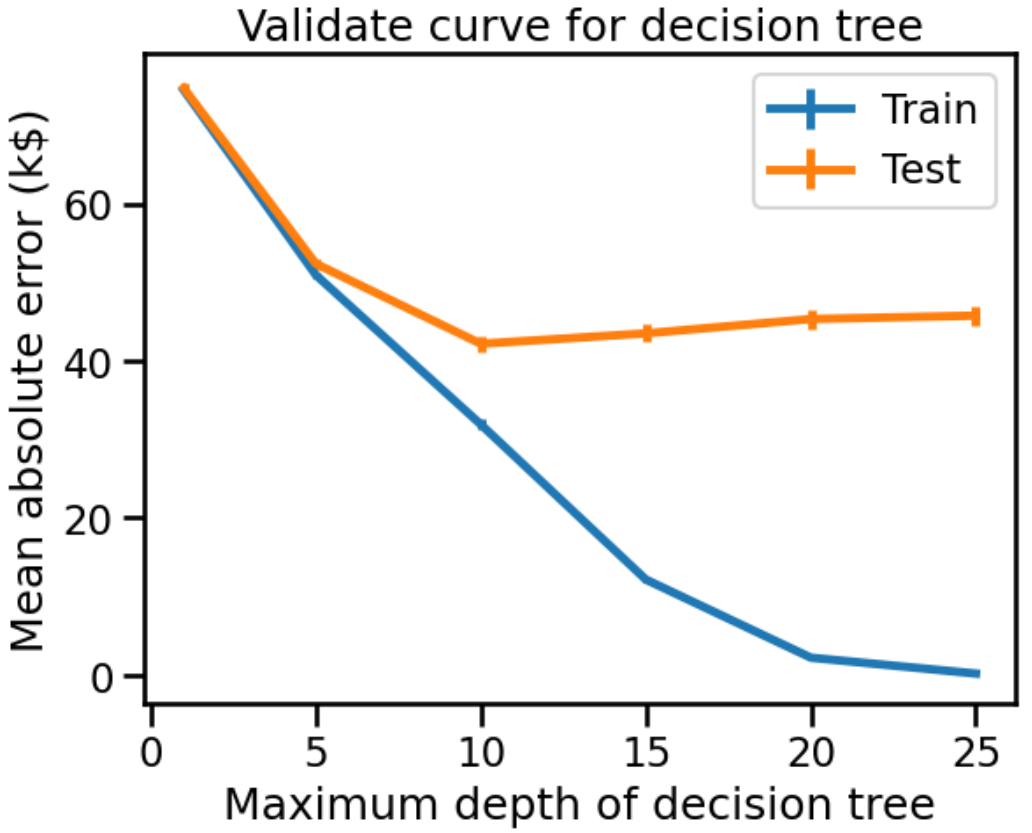
Pasos:

1. **Selección de Hiperparámetros Relevantes:** Identifica los hiperparámetros más relevantes del modelo o algoritmo.
2. **Definición de Espacio de Búsqueda:** Establece un espacio de búsqueda para cada hiperparámetro, que incluye los posibles valores que se probarán.
3. **Selección de Estrategia de Optimización:** Elige una estrategia de optimización, como búsqueda grid (todas las posibles combinaciones), búsqueda aleatoria, optimización bayesiana, etc.
4. **Validación Cruzada (Cross-Validation):** Divide los datos en conjuntos de entrenamiento y validación mediante validación cruzada para evaluar el rendimiento en diferentes combinaciones de hiperparámetros.
5. **Evaluación del Rendimiento:** Utiliza métricas de evaluación (recall, precisión, F1-score, MSE, MAPE, etc.) para medir el rendimiento del modelo en cada configuración de hiperparámetros.
6. **Selección del Mejor Conjunto de Hiperparámetros:** Identifica el conjunto de hiperparámetros que da como resultado el mejor rendimiento en los datos de validación cruzada.

Curvas de validación



El hiperparámetro óptimo se obtiene cuando comienza a haber divergencia en la curva de entrenamiento y validación.



En este ejemplo observe que al aumentar la profundidad del árbol después de 10, el MSE no disminuye. Por tanto, una profundidad de 10 puede ser un valor óptimo del hiperparámetro.

Validación Cruzada

- En general se utiliza una técnica de remuestreo para estimar el rendimiento de los modelos y ajustar hiperparámetros.
- Se utiliza un subconjunto de muestras para ajustar un modelo y las muestras restantes se utilizan para estimar la eficacia del modelo.
- Este proceso se repite varias veces y los resultados se agregan y resumen.
- En la validación cruzada las muestras se dividen aleatoriamente en k conjuntos de igual tamaño.
- Se ajusta un modelo utilizando todas las muestras excepto un subconjunto, el cual se utiliza para validación.