Apoyo Ayudantía Árboles de Decisión

Natalie Julian

Árbol de decisión

Corresponde a una técnica de Machine Learning de tipo Supervisado. Al final del entrenamiento, se obtienen secuencias de preguntas de los datos que llevan a un resultado o *predicción*. Un plus que tiene esta técnica es que es posible visualizar las secuencias de preguntas en forma de árbol, siendo fácil de explicar y comprender para la audiencia.

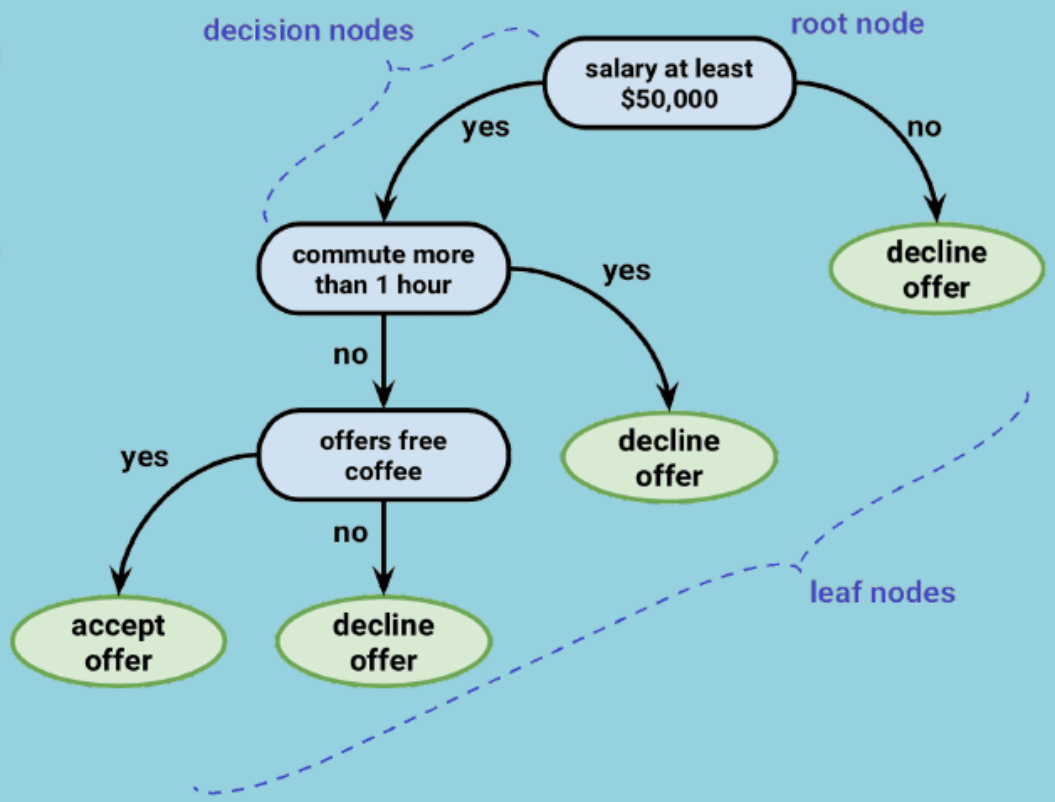
Los árboles de decisión se clasifican en:

- Árbol de Regresión: Cuando la variable Y output es numérica. Ejemplo: Determinar las secuencias de preguntas adecuadas de modo de predecir el salario de un colaborador.
- Árbol de Clasificación: Cuando la variable Y output es categórica. Ejemplo: Determinar las secuencias de preguntas adecuadas de modo de predecir el fallo o no fallo de un producto.

Ejemplo Árbol de Clasificación: ¿Debería aceptar un nuevo empleo?

Podemos observar las siguientes partes del árbol:

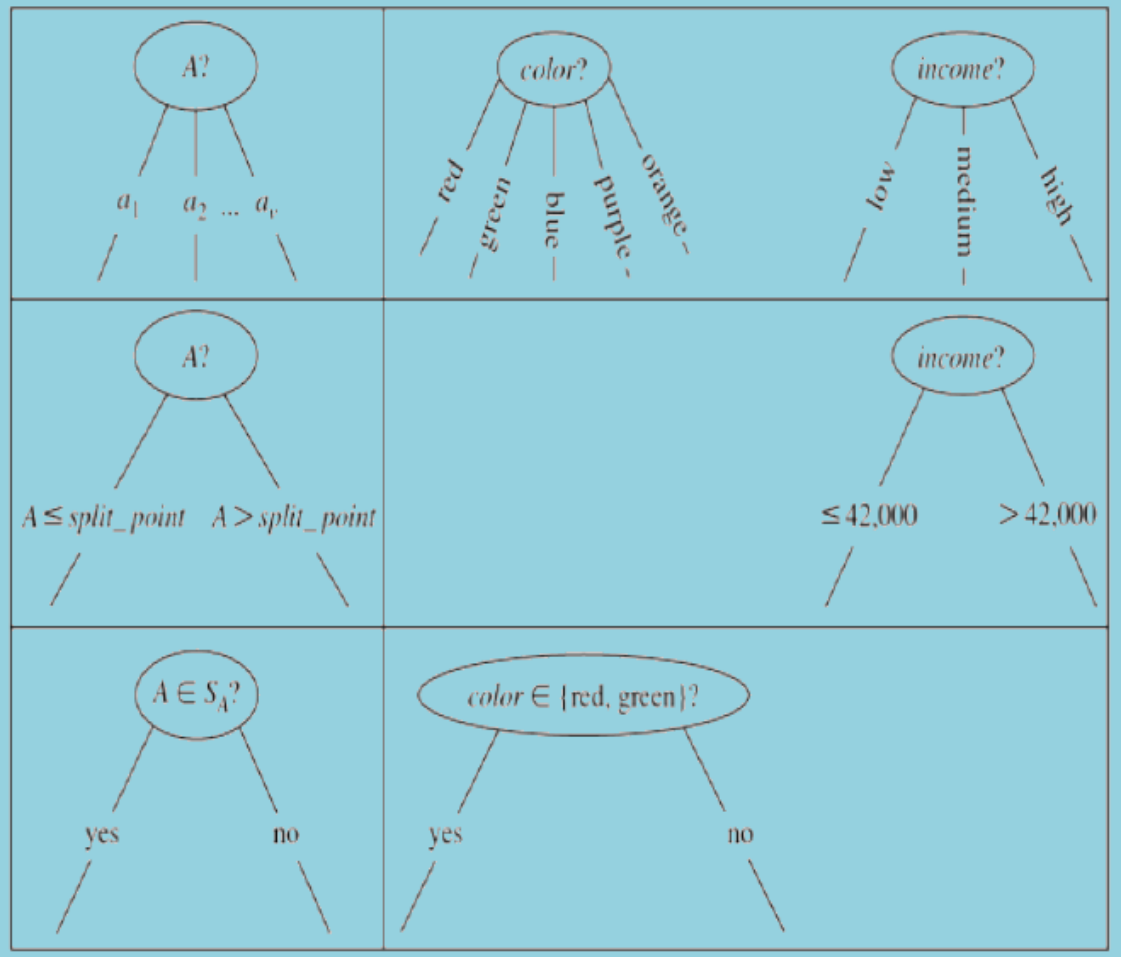
- Cada nodo de decisión (decision node) representa una pregunta sobre un atributo. El nodo raíz (root node) corresponde a la primera pregunta.
- Cada rama representa una posible respuesta a las preguntas.
- Cada hoja o nodo terminal (leaf node) representa la clase o predicción al seguir ese camino.



Árbol de clasificación: ¿Cómo se obtienen las preguntas?

1. Si todas las observaciones pertenecen a la misma clase o categoría, se crea un nodo terminal con dicha categoría.
2. Si no, se recurre a algún método para determinar la primera pregunta o criterio de separación, la idea es obtener particiones puras (que los grupos estén en la menor medida posible mezclados, es decir más homogéneos).
3. Se crea el primer nodo el criterio de separación encontrado, este proceso se itera de forma de añadir ramas y nodos de forma anidada hasta que ocurra alguno de los siguientes casos:

- Todas las tuplas pertenecen a la misma clase.
- No hay más atributos para particionar.
- Ya no hay más datos.



Entropía o Ganancia de información

La información esperada necesaria o entropía, para clasificar una tupla en alguna de las $C_1, \ldots, C_m$ clases es:

$$Info(D) = -\sum_{i=1}^{m} p_i log_2(p_i)$$

Donde p_i consiste en la probabilidad de pertenecer a la clase C_i. Para saber cuánto es el grado de entropía luego de la partición sugerida por el atributo A, calculamos:

$$Info_A(D) = \sum_{j=1}^{v} q_j \cdot Info(D_j)$$

Donde q_j corresponde a la probabilidad de pertenecer a la categoría j del atributo A.

Impureza de Gini

Los datos se dividen según el atributo A, en dos regiones, A_1 y A_2, con n_1 y n_2 elementos, respectivamente. p_{ji} corresponde a la proporción de elementos en A_j con $j = 1, 2$ que pertenecen a la clase i con $i = 1, \ldots, m$.

La impureza de Gini $I(A_j)$ se calcula:

$$= p_{j1}(1 - p_{j1}) + p_{j2}(1 - p_{j2}) + \ldots + p_{jm}(1 - p_{jm})$$

$$= \sum_{i=1}^{m} p_{ji}(1 - p_{ji}) = \sum_{i=1}^{m} (p_{ji} - p_{ji}^2) = \sum_{i=1}^{m} p_{ji} - \sum_{i=1}^{m} p_{ji}^2$$

$$= 1 - \sum_{k=1}^{m} p_{ji}^2$$

Y el índice de Gini se calcula:

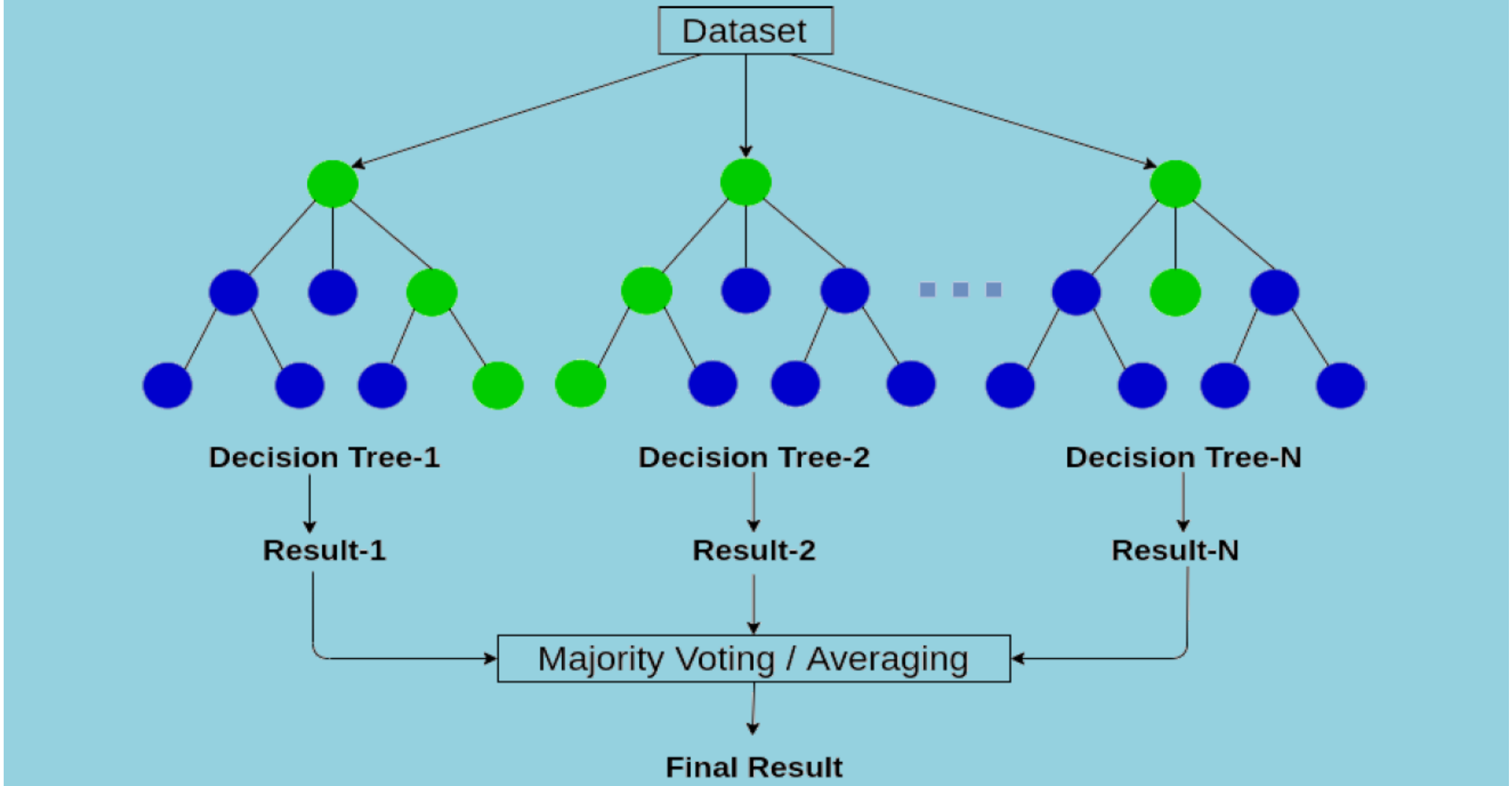
$$Gini(A) = q_1 I(A_1) + q_2 I(A_2)$$

Donde q_j corresponde a la probabilidad de pertenecer a la categoría j del atributo A.

Bosque de árboles

Corresponde a una técnica de Machine Learning que se basa en un conjunto de árboles de decisiones seleccionando aleatoriamente submuestras (con reemplazamiento) para elaborar cada árbol.

El punto clave del Bosque de árboles es utilizar una serie de árboles de decisión (diferentes individuos y diferentes variables), con el fin de mejorar la tasa de clasificación correcta. La diferencia con el bagging es que en el bosque de árboles también se toma una muestra de los features, es decir, no se utilizan todas las variables como en el bagging.


Dataset
Decision Tree-1
Decision Tree-2
Decision Tree-N
Result-1
Result-2
Result-N
Majority Voting / Averaging
Final Result

Matriz de confusión

La matriz de confusión (tambien conocidas como tablas de clasificación) (Matriz de Confusión) entregan información para evaluar la capacidad predictiva del modelo, hay que tener en consideración lo siguiente:

1. Un modelo puede ser correcto y tener malas propiedades de clasificación
2. En general, modelos con probabilidades estimadas cercanas a 0.5 tendrán bajo poder de clasificación.
3. Capacidad predictiva depende del punto de corte.

En las matrices de confusión se cruzan las predicciones con las tablas reales en una tabla de 2×2. En lo siguiente positivo se refiere a los éxitos (1) y negativos se refiere a los fracasos (0):

		Valores Reales	
		Positivos	**Negativos**
Valores Predichos	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (VP)	Verdaderos Negativos (VN)

Desde una matriz de confusión se pueden extraer diversos indicadores que nos permiten evaluar el modelo:

- **Sensibilidad:** = $\frac{VP}{VP+FN}$
- **Especificidad:** = $\frac{VN}{VN+FP}$
- **Precisión** = $\frac{VP}{VP+FP}$
- **Exactitud** = $\frac{VP+VN}{VP+VN+FP+FN}$

La **sensibilidad** representa la proporción de positivos capturados correctamente por el modelo, sobre el total de positivos reales. Representa que tan bien el modelo califica los casos "positivos" de nuestros datos.

La **especificidad** representa la proporción de casos negativos capturados correctamente por el modelo, sobre el total de negativos reales. Representa que tan bien el modelo califica los casos "negativos" de nuestros datos.

La **exactitud** mide la proporción de casos clasificados correctamente, independiente de si es positivo o negativo. Esta medida no es recomendada cuando la base de datos es desbalanceada.

La **presición** es el porcentaje de casos positivos clasificados correctamente.