MINISTERIO DEL TRABAJO

Ciencia de Datos (Arboles)

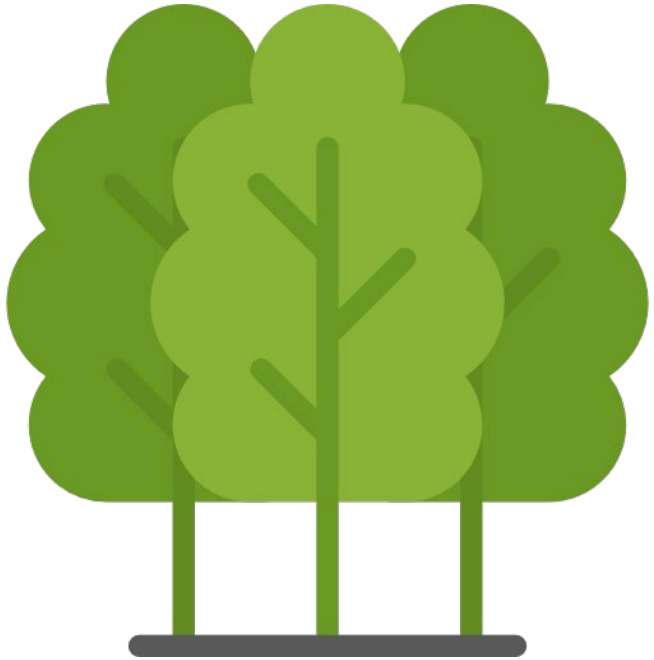
Centro de Servicios y Gestión Empresarial
SENA Regional Antioquia

Conceptos

Arboles

En machine learning, los árboles son una categoría de algoritmos ampliamente utilizados para la clasificación y la regresión.

Hay varios tipos de árboles y técnicas relacionadas. Los tipos más comunes de árboles en machine learning son:



Tipos de arboles

Árboles de Decisión (Decision Trees):

- Los árboles de decisión son estructuras jerárquicas en forma de árbol que se utilizan para tomar decisiones basadas en condiciones lógicas.
- Cada nodo en el árbol representa una pregunta o una prueba sobre una característica específica del conjunto de datos.
- Las ramas que salen de un nodo representan las posibles respuestas a esa pregunta.
- Las hojas del árbol representan las decisiones finales o predicciones.

Tipos de arboles

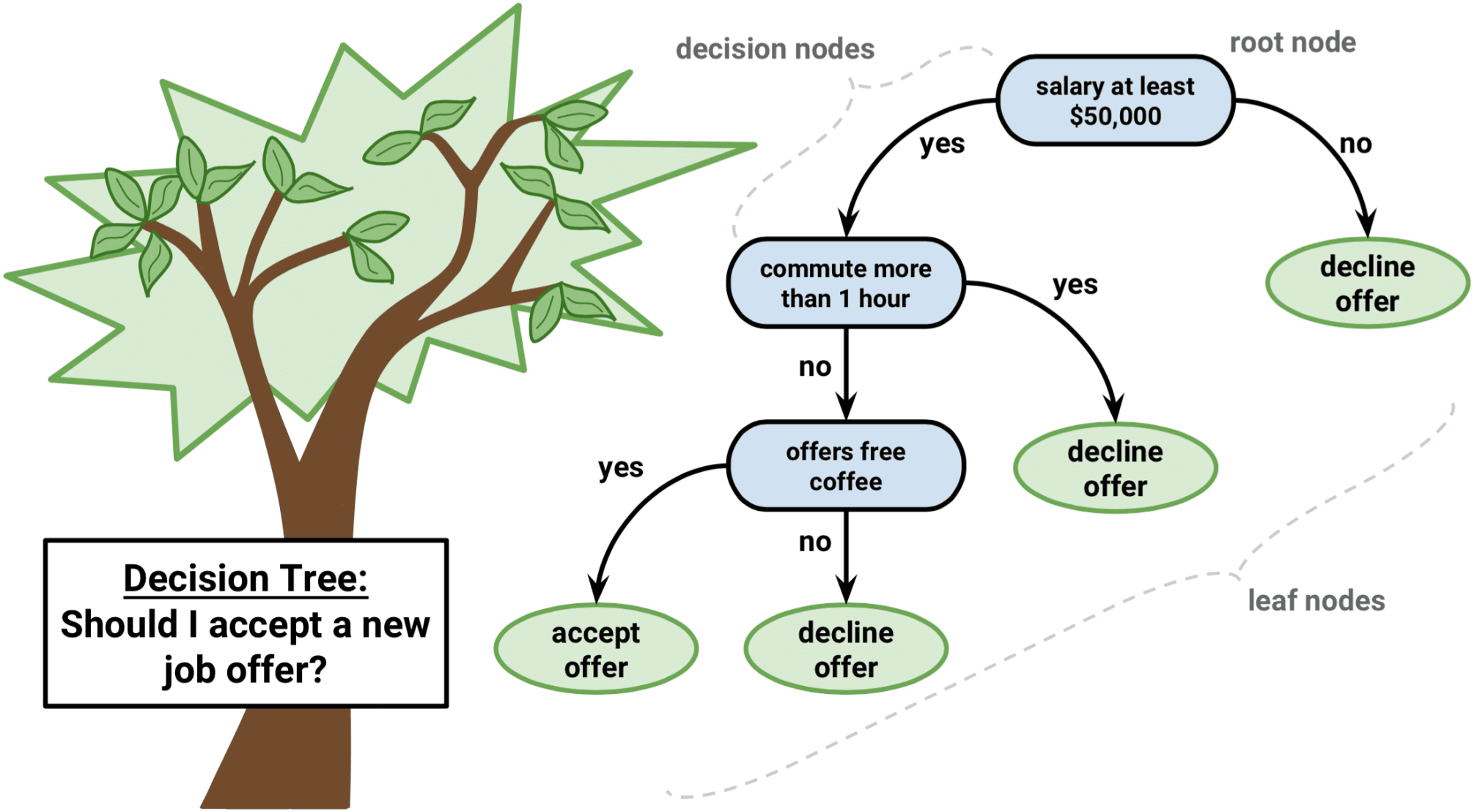
Árboles de Clasificación (Classification Trees):

- Los árboles de clasificación se utilizan para problemas de clasificación, donde el objetivo es predecir una etiqueta de clase o categoría para un dato dado.
- Cada hoja del árbol corresponde a una clase específica.

Árboles de Regresión (Regression Trees):

- Los árboles de regresión se utilizan para problemas de regresión, donde el objetivo es predecir un valor numérico (por ejemplo, una cantidad) en lugar de una etiqueta de clase.
- Las hojas del árbol contienen valores numéricos que representan la predicción

Tipos de arboles



Elementos

Nodo Raíz (Root Node):

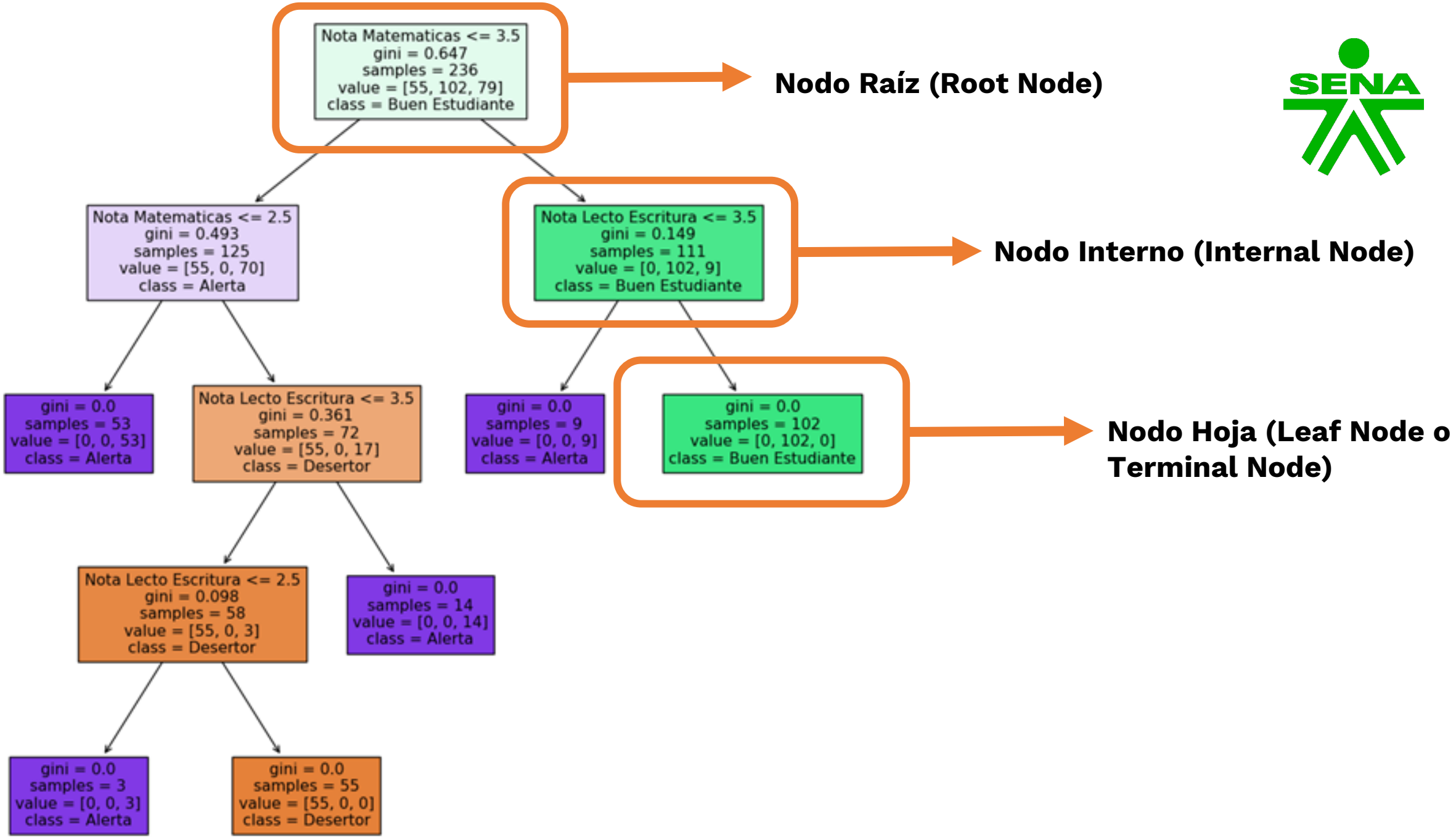
- Es el nodo superior del árbol y el punto de partida para tomar decisiones.
- Representa la característica más importante que divide el conjunto de datos en dos o más subconjuntos.

Nodo Interno (Internal Node):

- Son los nodos que no son hojas ni la raíz.
- Representan pruebas o decisiones basadas en características específicas.
- Dividen el conjunto de datos en subconjuntos más pequeños.

Nodo Hoja (Leaf Node o Terminal Node):

- Son los nodos finales del árbol.
- Representan las predicciones o decisiones finales.
- No se dividen más y contienen el resultado final del proceso de clasificación o regresión.


Nota Matematicas <= 3.5
gini = 0.647
samples = 236
value = [55, 102, 79]
class = Buen Estudiante
Nota Matematicas <= 2.5
gini = 0.493
samples = 125
value = [55, 0, 70]
class = Alerta
Nota Lecto Escritura <= 3.5
gini = 0.149
samples = 111
value = [0, 102, 9]
class = Buen Estudiante
gini = 0.0
samples = 53
value = [0, 0, 53]
class = Alerta
Nota Lecto Escritura <= 3.5
gini = 0.361
samples = 72
value = [55, 0, 17]
class = Desertor
gini = 0.0
samples = 9
value = [0, 0, 9]
class = Alerta
gini = 0.0
samples = 102
value = [0, 102, 0]
class = Buen Estudiante
Nota Lecto Escritura <= 2.5
gini = 0.098
samples = 58
value = [55, 0, 3]
class = Desertor
gini = 0.0
samples = 14
value = [0, 0, 14]
class = Alerta
gini = 0.0
samples = 3
value = [0, 0, 3]
class = Alerta
gini = 0.0
samples = 55
value = [55, 0, 0]
class = Desertor
Nodo Raíz (Root Node)
Nodo Interno (Internal Node)
Nodo Hoja (Leaf Node o Terminal Node)

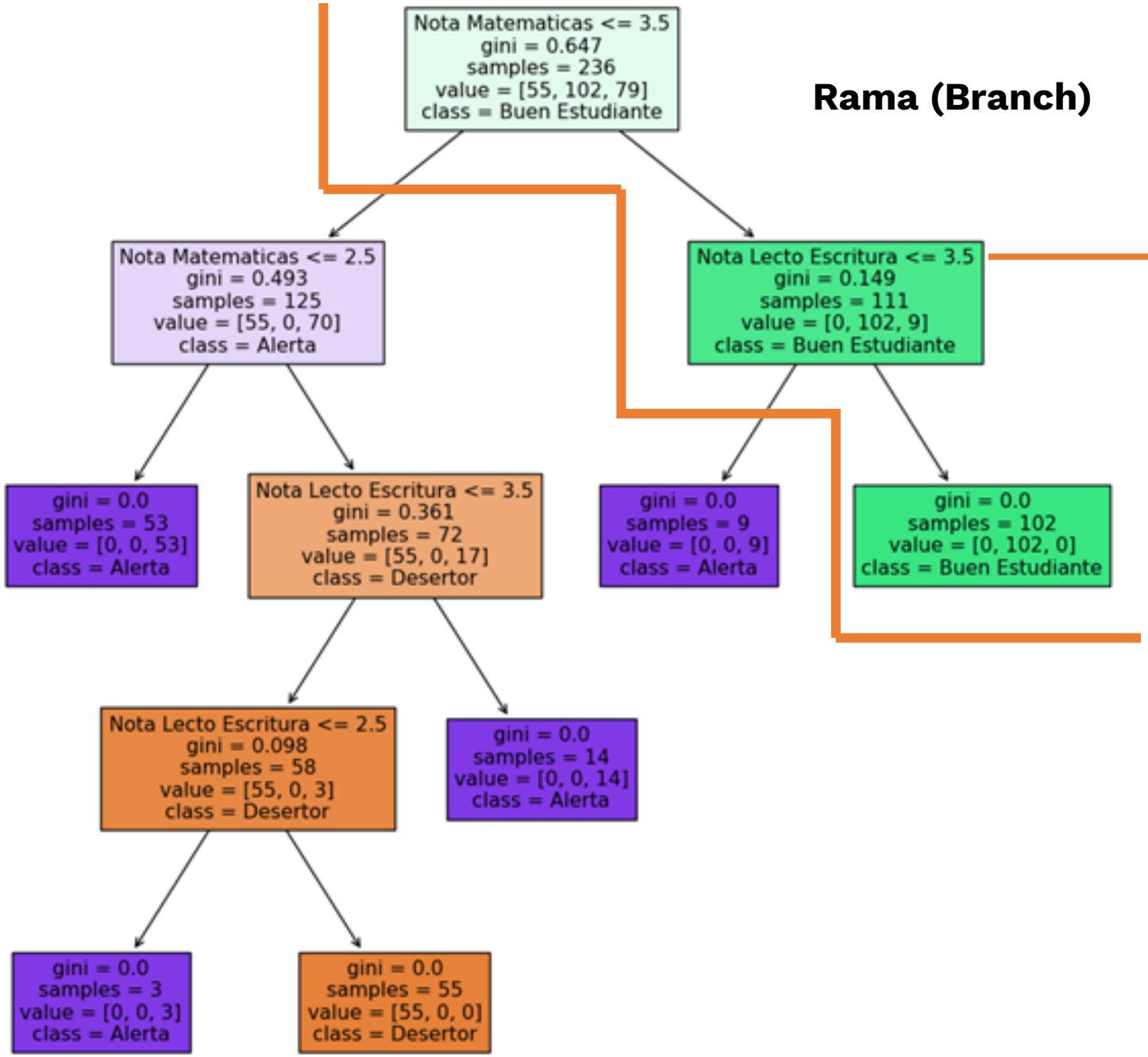
Elementos

Rama (Branch):

- Las conexiones entre nodos representan el flujo de decisión.
- Cada rama sale de un nodo y se dirige hacia otro nodo (ya sea interno o una hoja) basado en el resultado de la prueba realizada en ese nodo.

Valor de División (Split Value):

- Es el valor de una característica que se utiliza para dividir el conjunto de datos en un nodo interno.
- Por ejemplo, si la característica es "Edad" y el valor de división es 30, entonces los datos se dividen en dos subconjuntos: aquellos con edades menores o iguales a 30 y aquellos con edades mayores a 30.



Rama (Branch)

Valor de División (Split Value)

Elementos

Criterio de División (Split Criterion):

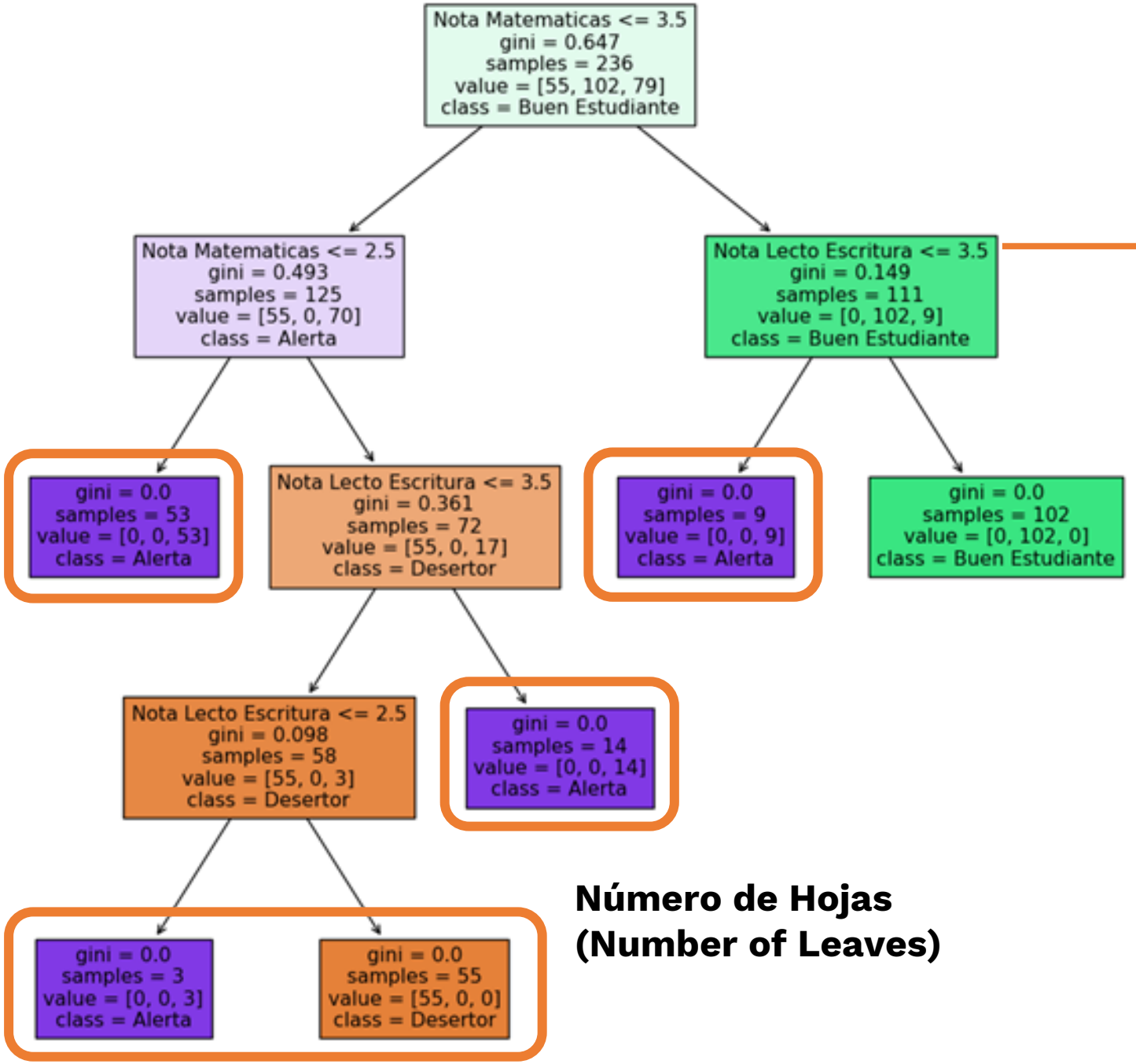
- Representa la medida utilizada para decidir cómo dividir los datos en un nodo.
- Ejemplos comunes de criterios son la ganancia de información (Information Gain) y la impureza de Gini (Gini Impurity) para árboles de clasificación, y el error cuadrático medio (Mean Squared Error) para árboles de regresión.

Profundidad del Árbol (Tree Depth):

- Es la longitud máxima del camino desde la raíz hasta cualquier hoja en el árbol.
- La profundidad del árbol puede controlarse como un hiperparámetro y afecta la complejidad y el sobreajuste del modelo.

Número de Hojas (Number of Leaves):

- Representa la cantidad total de nodos hoja en el árbol.
- Puede variar según la complejidad del árbol y la cantidad de divisiones realizadas.



Criterio de División (Split Criterion)

Profundidad del Árbol (Tree Depth)

Número de Hojas (Number of Leaves)

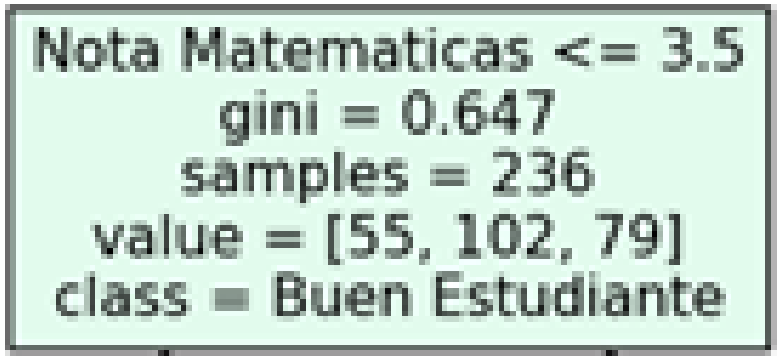
Elementos

Gini (Gini impurity):

- El índice de Gini es una medida de impureza en un nodo hoja.
- Mide cuán mezcladas o impuras son las muestras de diferentes clases en el nodo hoja.
- Un valor de Gini de 0 significa que todas las muestras en el nodo pertenecen a una sola clase, es decir, es un nodo puro.
- Un valor de Gini más alto indica una mayor impureza en el nodo, con muestras distribuidas entre varias clases.

Samples:

- Indica el número total de muestras que llegan a ese nodo hoja.
- Representa cuántas instancias de datos se encuentran en ese nodo específico.



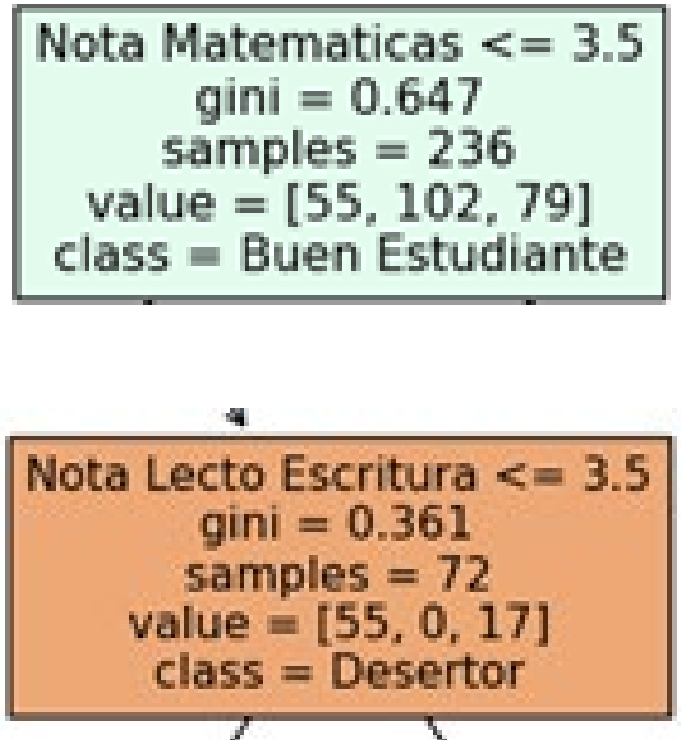
Elementos

Value:

- Es una lista que muestra la distribución de clases en ese nodo hoja.
- Cada elemento en la lista representa el recuento de muestras pertenecientes a una clase específica en ese nodo.
- Por ejemplo, si tienes tres clases (Clase A, Clase B, Clase C), el valor podría verse como [10, 5, 3], lo que significa que hay 10 muestras de la Clase A, 5 de la Clase B y 3 de la Clase C en ese nodo hoja.

Class:

- Indica la clase de destino que se asigna a ese nodo hoja.
- Es la clase que se predice para las muestras que llegan a ese nodo en particular.



G R A C I A S

Presentó: Alvaro Pérez Niño
Instructor Técnico
Correo: aperezn@misena.edu.co
http://centrodeserviciosygestionempresarial.blogspot.com/