Algoritmos basados en Árboles de decisión

TAE IA
Formación de Talento Altamente Especializado

Árboles de decisión

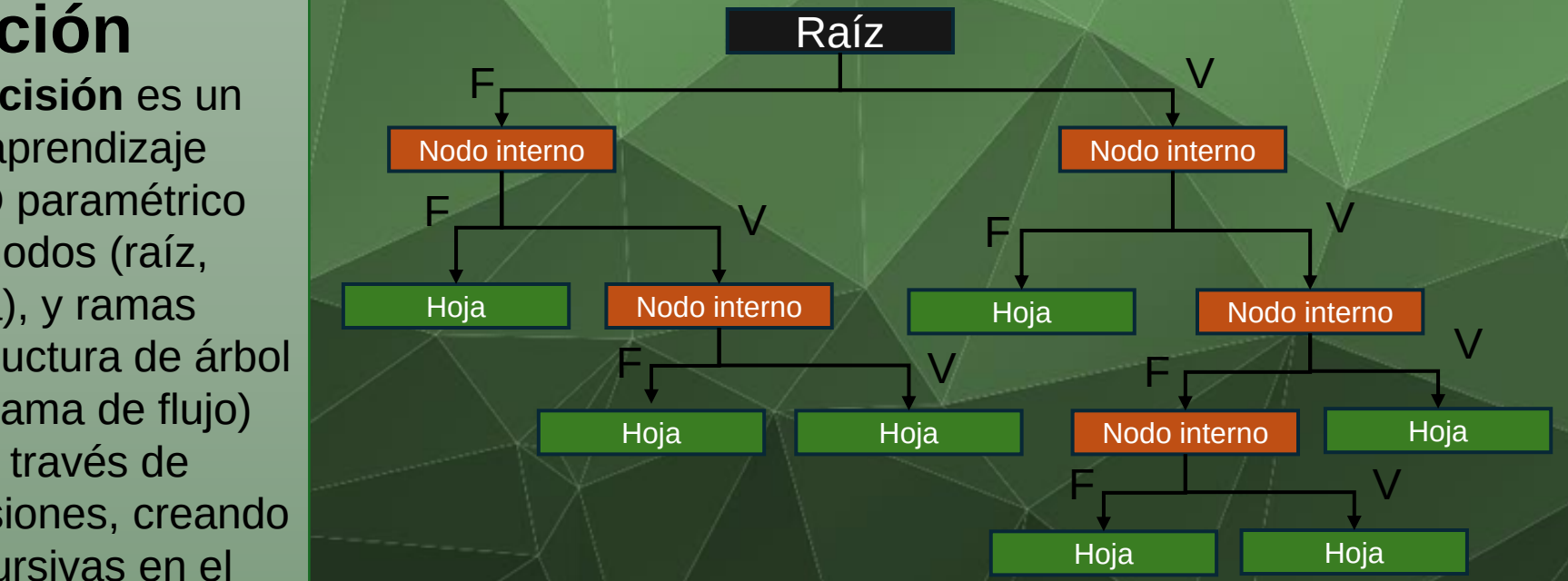
Definición

Un árbol de decisión es un algoritmo de aprendizaje supervisado NO paramétrico conteniendo nodos (raíz, internos, hoja), y ramas formando una estructura de árbol jerárquica (diagrama de flujo) avanzando a través de sentencias y decisiones, creando particiones recursivas en el espacio dimensional de los datos.

Árboles de decisión

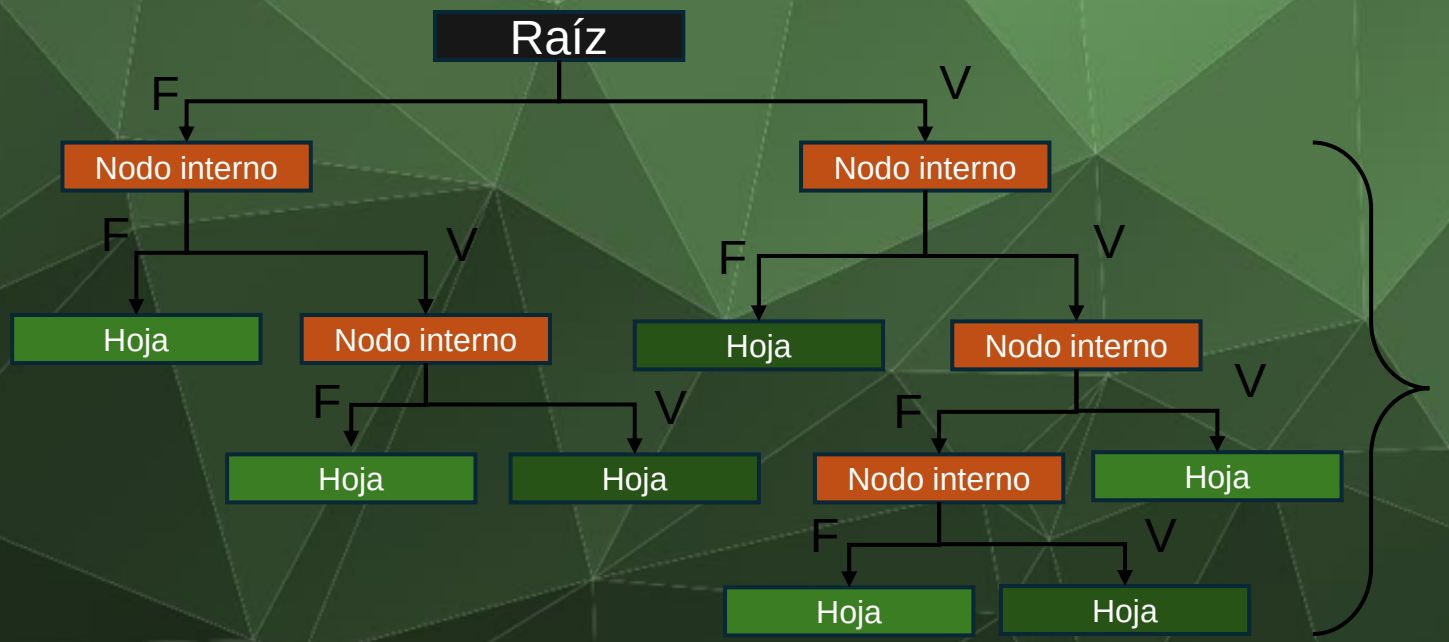
Definición

Un árbol de decisión es un algoritmo de aprendizaje supervisado NO paramétrico conteniendo nodos (raíz, internos, hoja), y ramas formando una estructura de árbol jerárquica (diagrama de flujo) avanzando a través de sentencias y decisiones, creando particiones recursivas en el espacio dimensional de los datos.



TAE IA
Formación de Talento Altamente Especializado

Árboles de decisión



Sentencia Inicial

Sentencias adicionales

Resultado (clase o valor continuo)

TAE IA
Formación de Talento Altamente Especializado

¿Cómo construir un árbol?

CART

(Classification And Regression Trees)

Pureza de un nodo:

Medición de la homogeneidad de las clases dentro de cada nodo. Un nodo que contiene solamente una clase se dice que es puro.

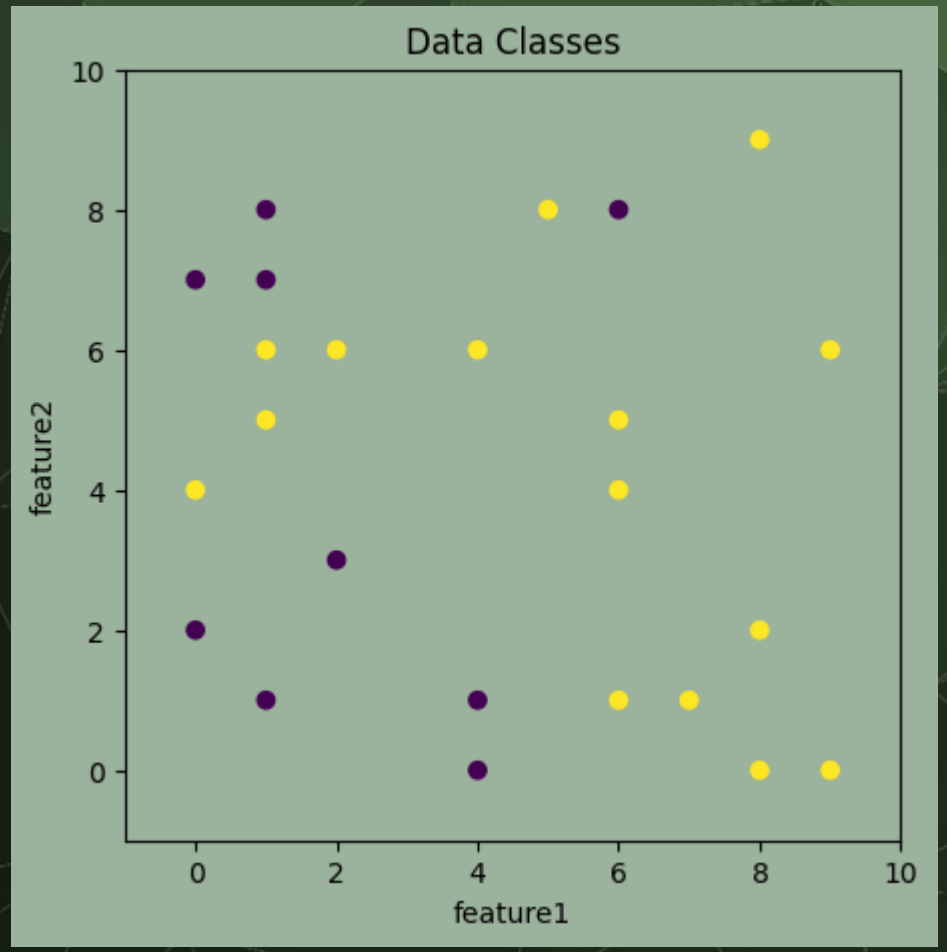
Índice Gini:

Mide Pureza de un nodo: A mayor Índice Gini, menor pureza.
1-prob(categoria1)2 – prob(categoria2)**2**

Ganancia de información:

Índice Gini antes – promedio de índice Gini actual.

¿Cómo construir un árbol?



Índice Gini:

Mide Pureza de un nodo: A mayor Índice Gini, menor pureza.

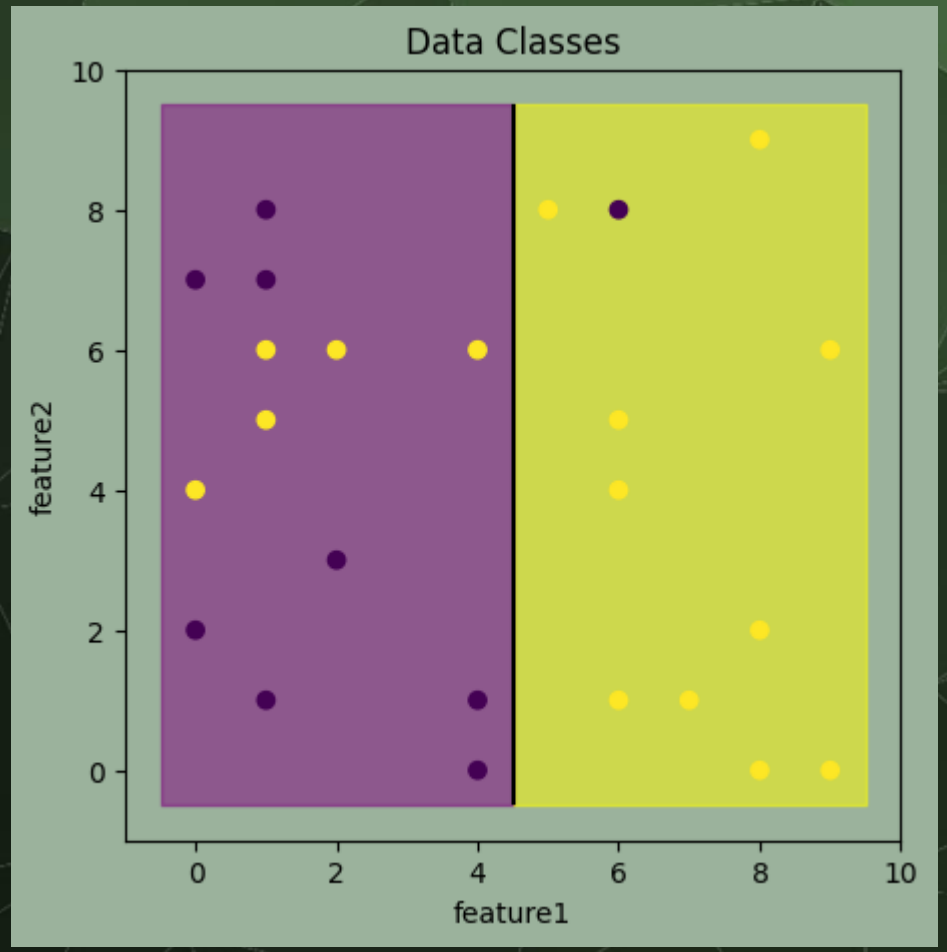
1-prob(categoria1)2 – prob(categoria2)**2**

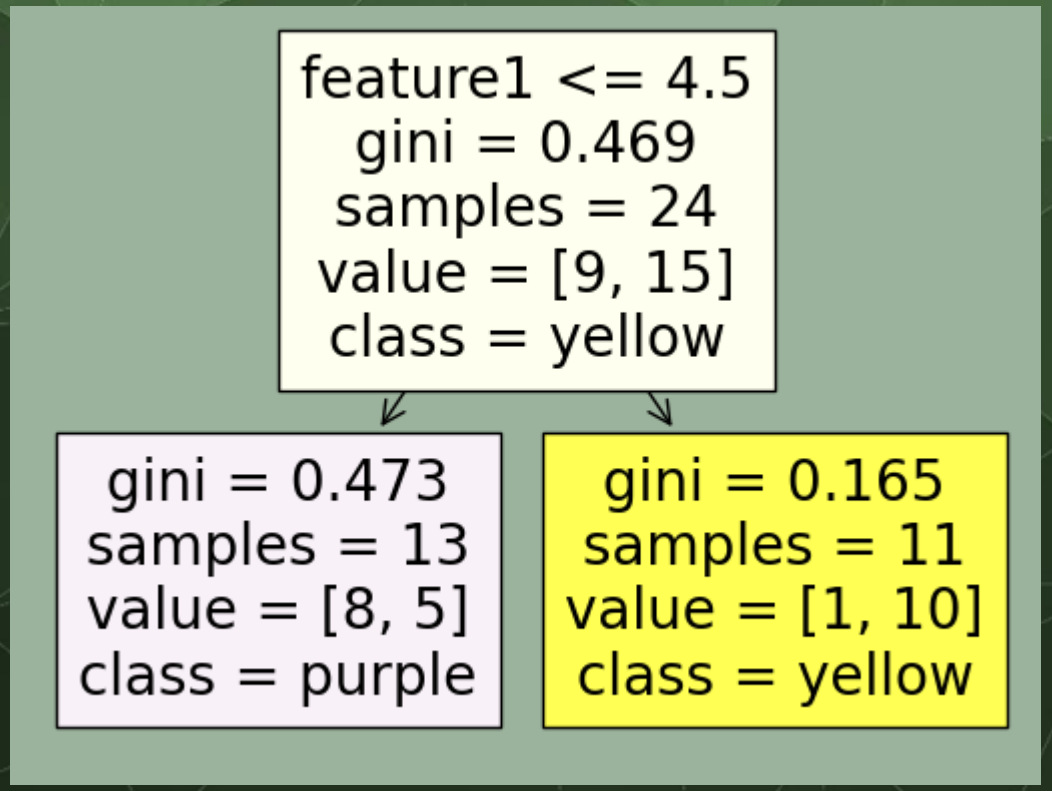
gini = 1 - prob(categoria1)2 - prob(categoria2)**2**

gini = 1 – (9/24)2 – (15/24)**2 = 0.469**

TAE IA
Formación de Talento Altamente Especializado

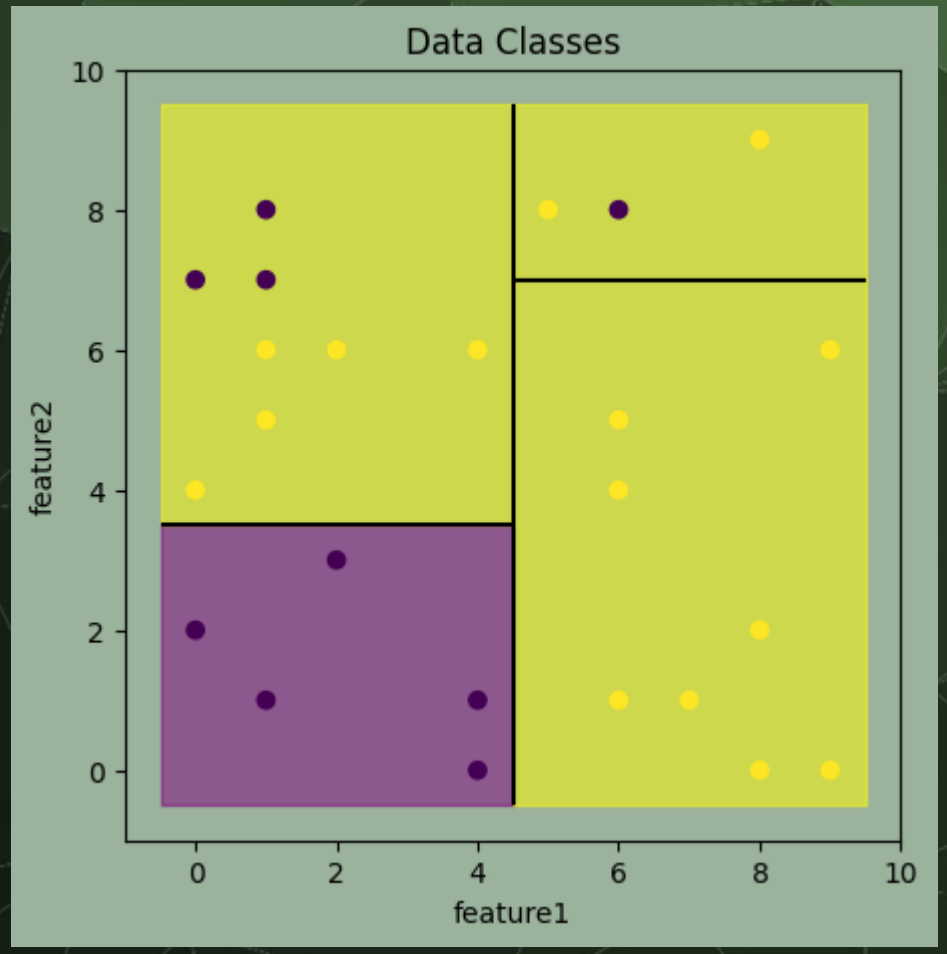
¿Cómo construir un árbol?

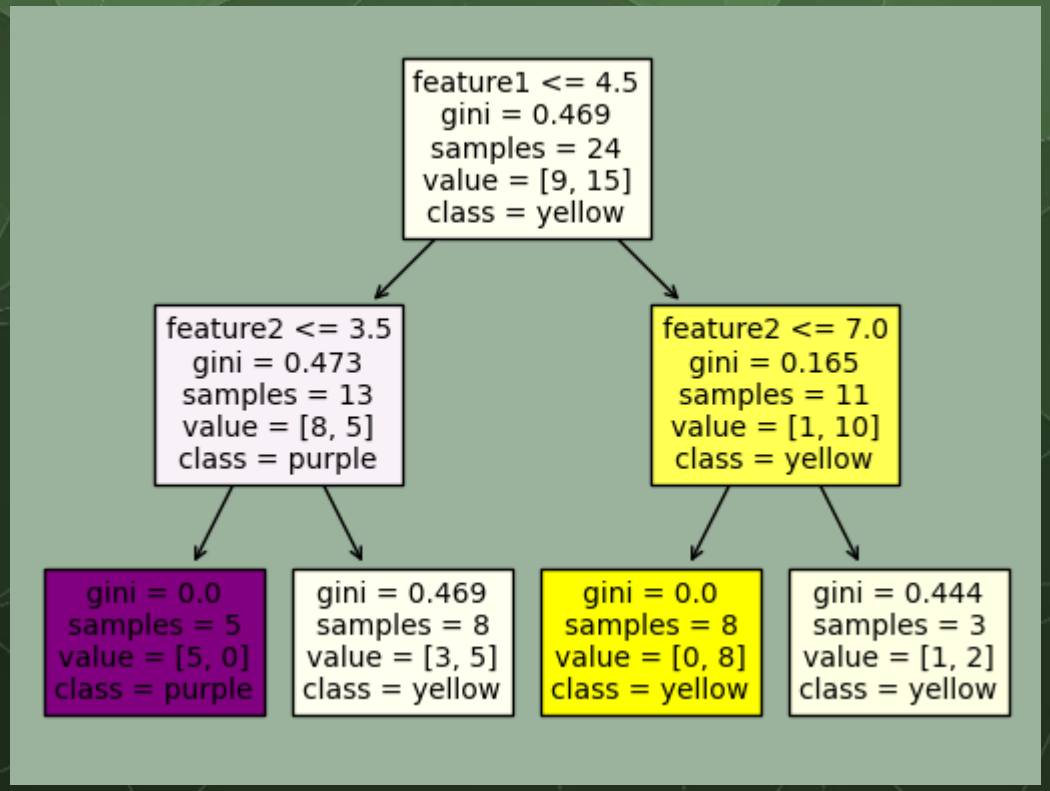




TAE IA
Formación de Talento Altamente Especializado

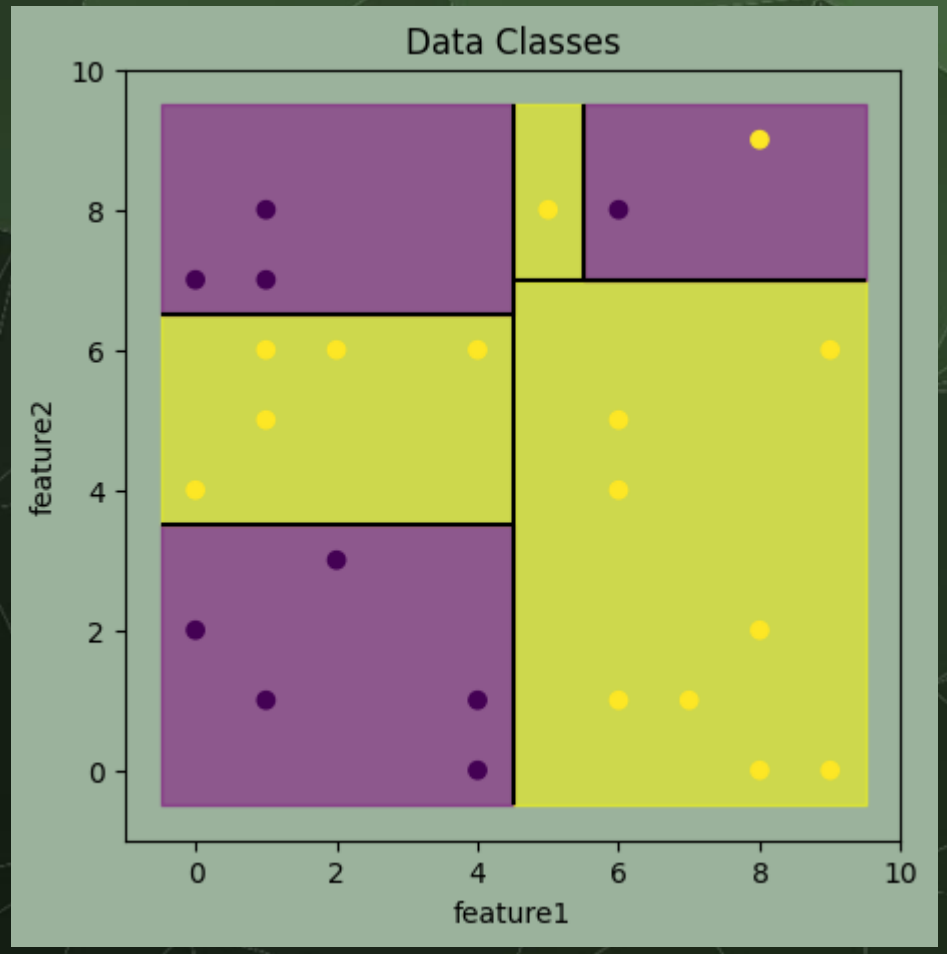
¿Cómo construir un árbol?

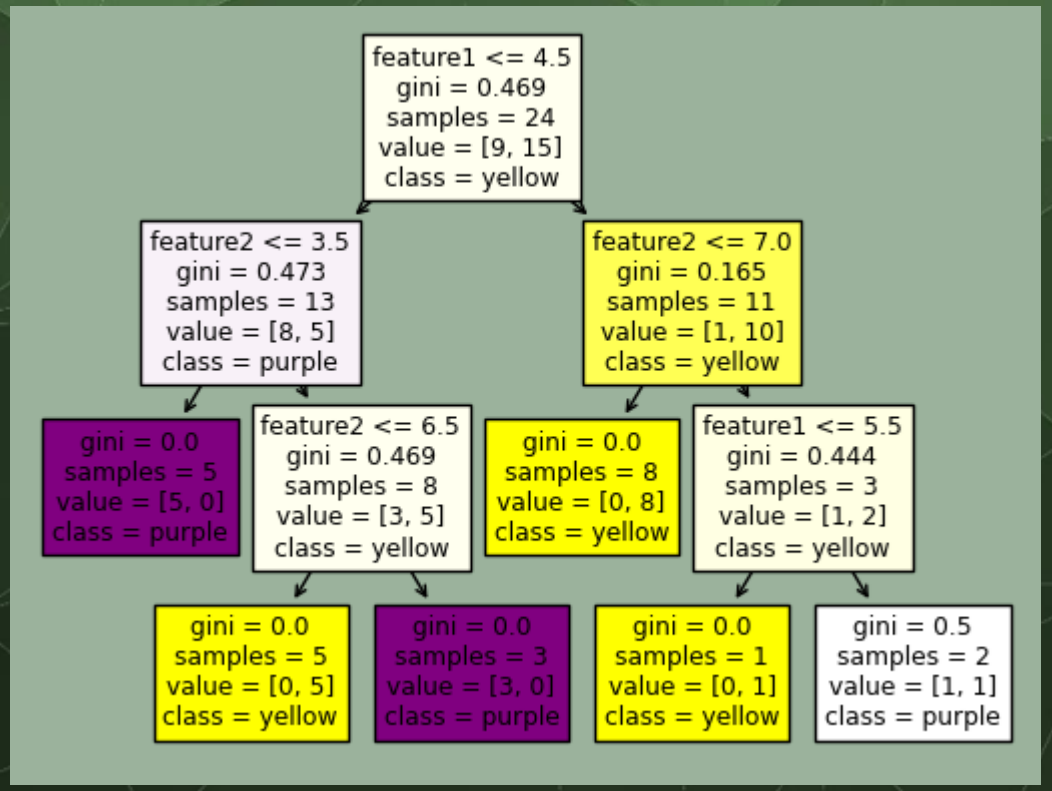




TAE IA
Formación de Talento Altamente Especializado

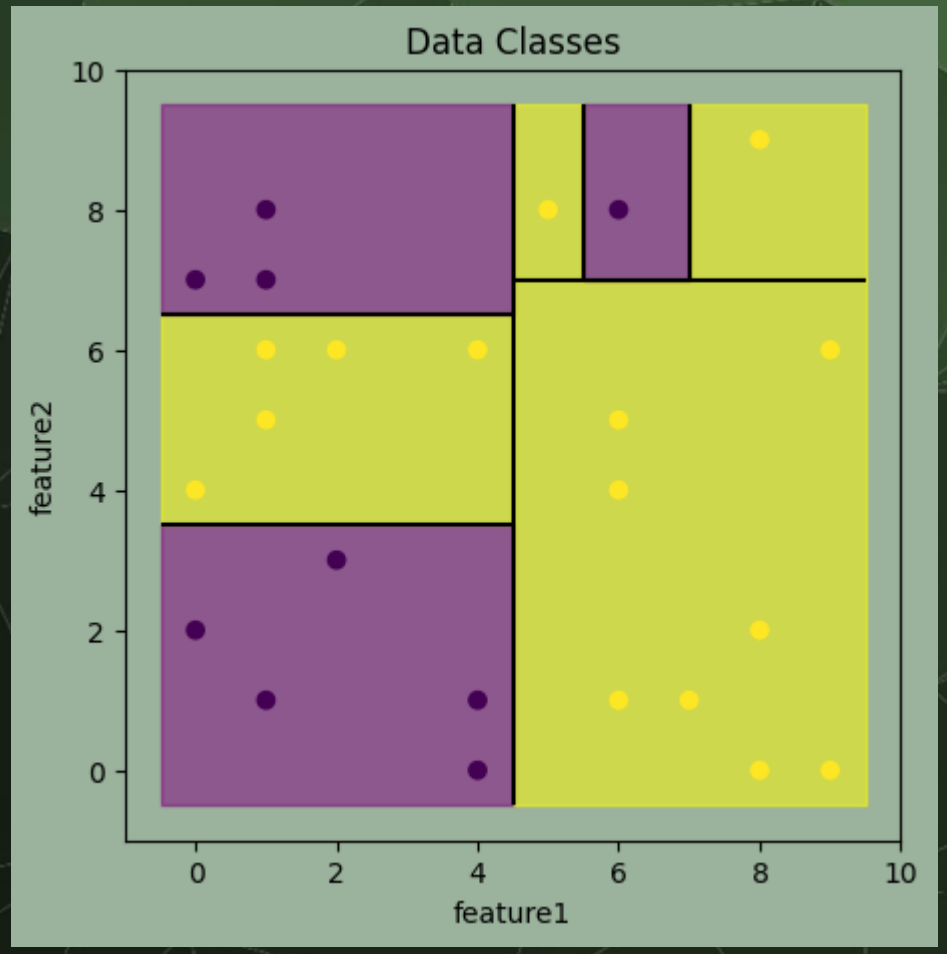
¿Cómo construir un árbol?

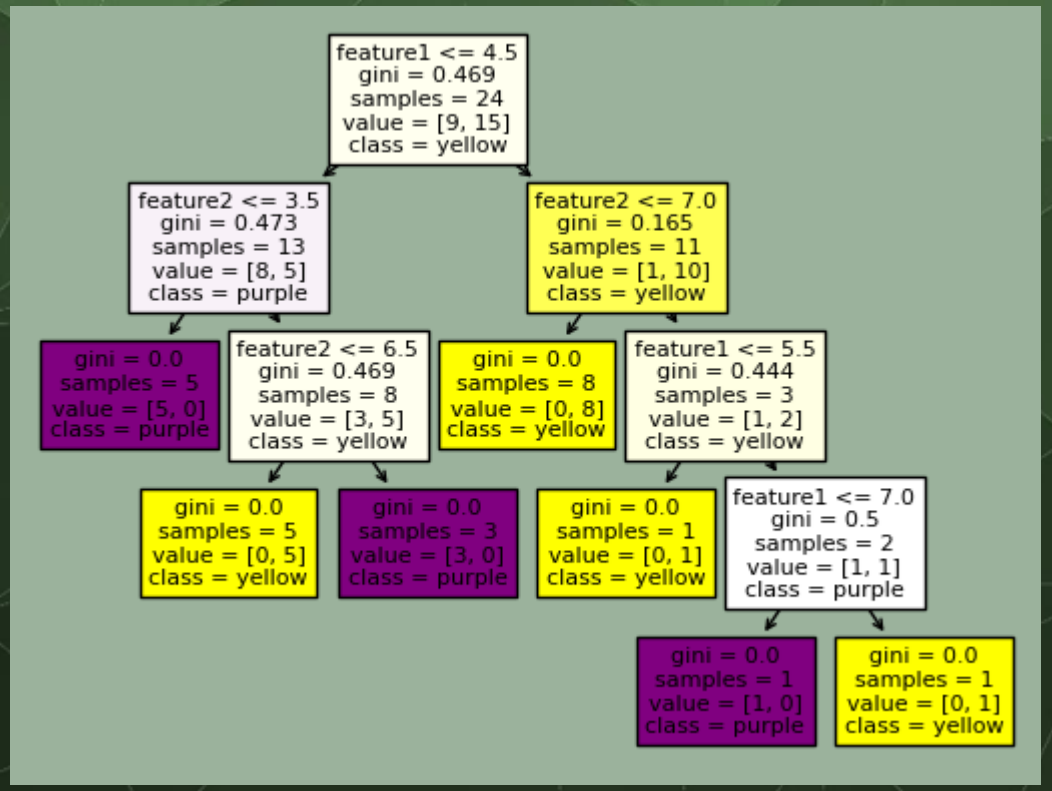




TAE IA
Formación de Talento Altamente Especializado

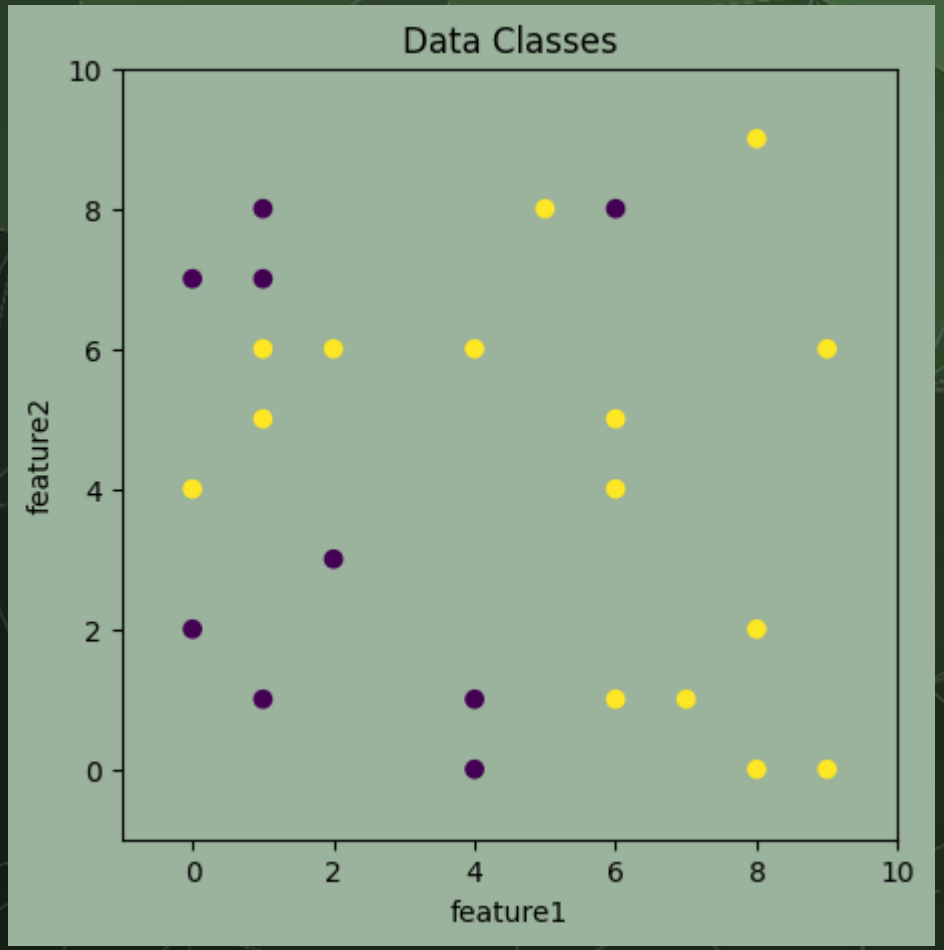
¿Cómo construir un árbol?





TAE IA
Formación de Talento Altamente Especializado

¿Cómo construir un árbol?



Para todas las variables disponibles, tenemos que elegir solo un umbral: Para elegir el umbral:

1) **Reducimos el espacio de búsqueda con los datos disponibles: obtenemos el punto medio de los valores ordenados para cada feature.**

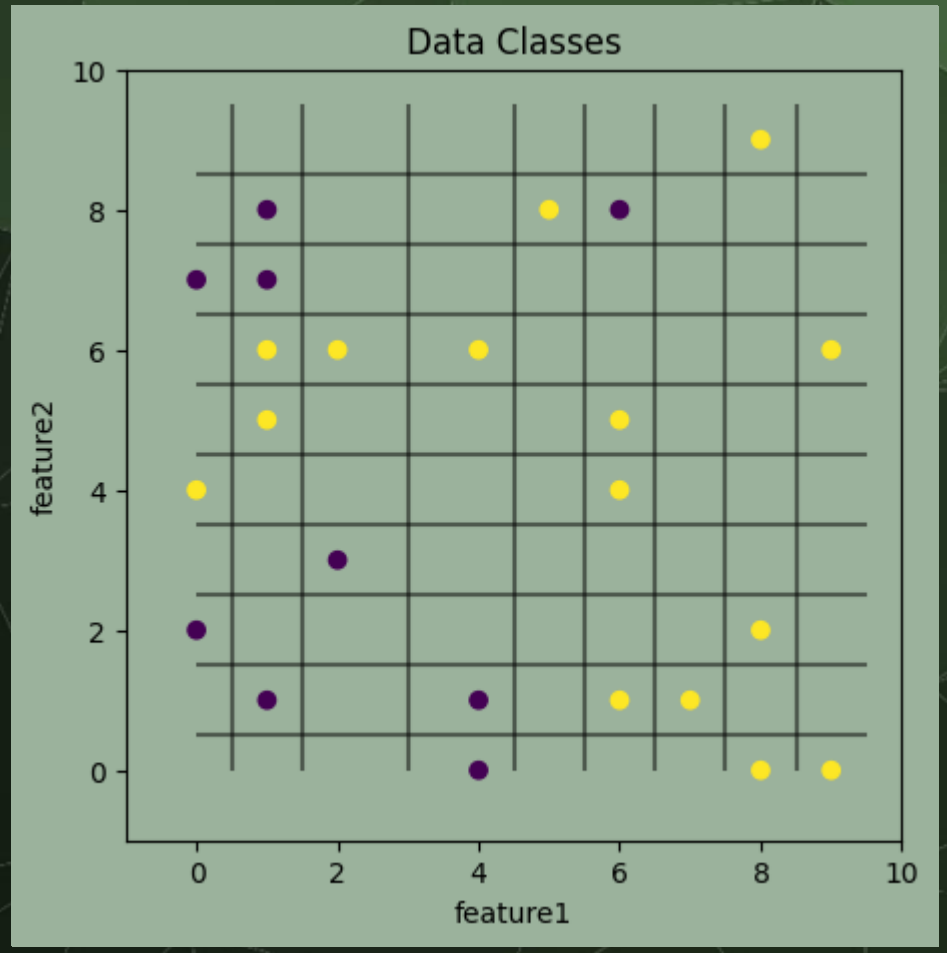
2) **Encontramos la ganancia de información para cada umbral. Es decir, el promedio ponderado de cada índice de Gini obtenido de dividir el espacio en dos subconjuntos.**

3) **Elegimos aquel que tengan la mayor ganancia de información (equivalente a obtener el valor menor de índice de Gini calculado en el paso anterior) de todos los umbrales disponibles.**

4) **Repetimos en cada sub-espacio disponible.**

TAE IA
Formación de Talento Altamente Especializado

¿Cómo construir un árbol?



Para todas las variables disponibles, tenemos que elegir solo un umbral: Para elegir el umbral:

1) **Reducimos el espacio de búsqueda con los datos disponibles: obtenemos el punto medio de los valores ordenados para cada feature.**

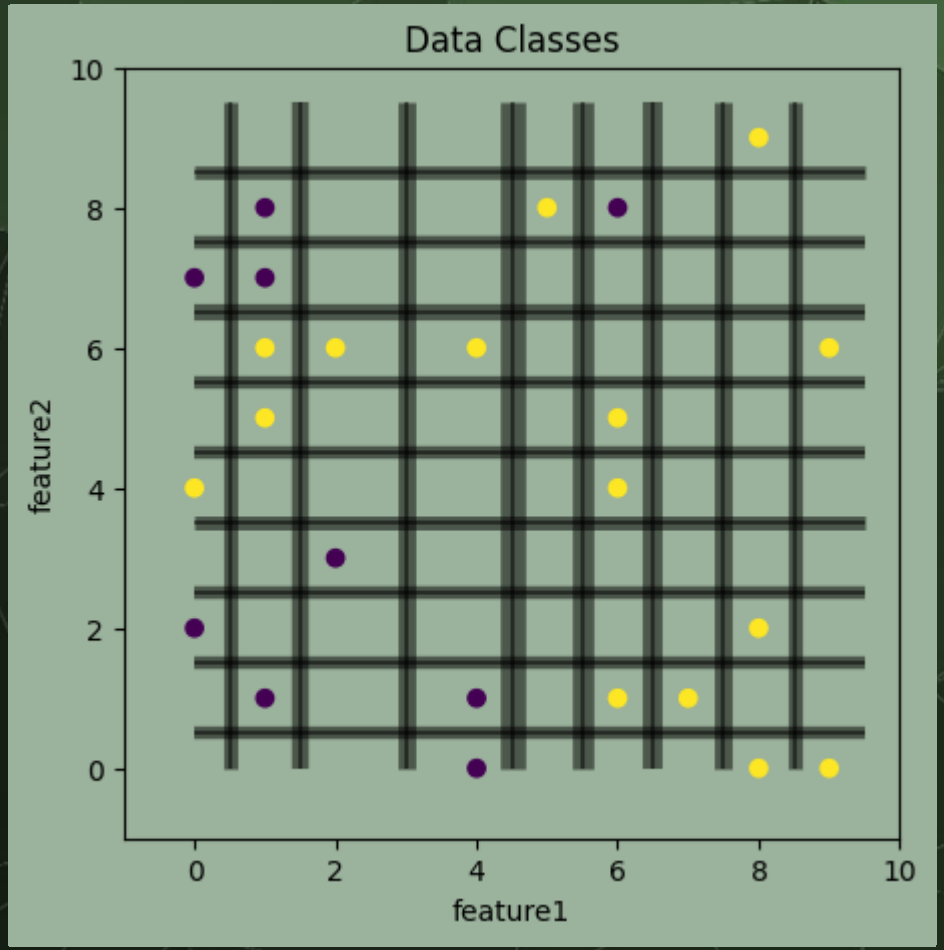
2) **Encontramos la ganancia de información para cada umbral. Es decir, el promedio ponderado de cada índice de Gini obtenido de dividir el espacio en dos subconjuntos.**

3) **Elegimos aquel que tengan la mayor ganancia de información (equivalente a obtener el valor menor de índice de Gini calculado en el paso anterior) de todos los umbrales disponibles.**

4) **Repetimos en cada sub-espacio disponible.**

TAE IA
Formación de Talento Altamente Especializado

¿Cómo construir un árbol?



Para todas las variables disponibles, tenemos que elegir solo un umbral: Para elegir el umbral:

1) Reducimos el espacio de búsqueda con los datos disponibles: obtenemos el punto medio de los valores ordenados para cada feature.

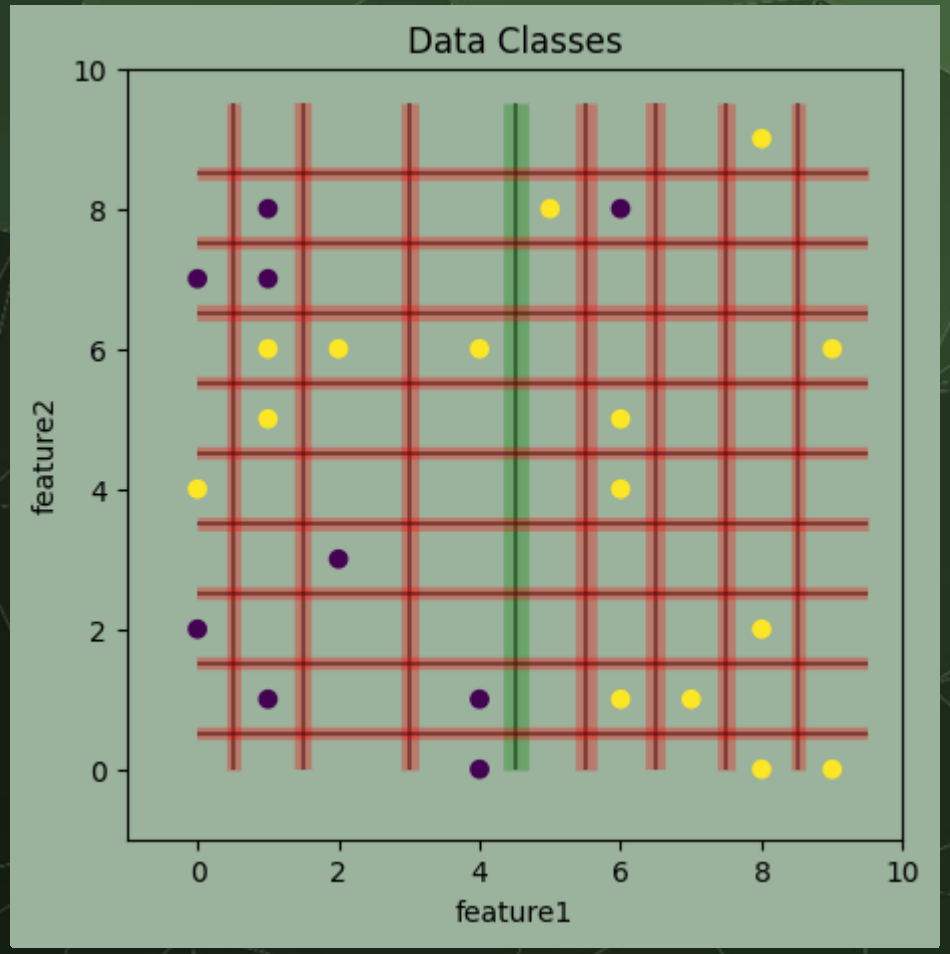
2) Encontramos la ganancia de información para cada umbral. Es decir, el promedio ponderado de cada índice de Gini obtenido de dividir el espacio en dos subconjuntos.

3) Elegimos aquel que tengan la mayor ganancia de información (equivalente a obtener el valor menor de índice de Gini calculado en el paso anterior) de todos los umbrales disponibles.

4) Repetimos en cada sub-espacio disponible.

TAE IA
Formación de Talento Altamente Especializado

¿Cómo construir un árbol?



Para todas las variables disponibles, tenemos que elegir solo un umbral: Para elegir el umbral:

1) **Reducimos el espacio de búsqueda con los datos disponibles: obtenemos el punto medio de los valores ordenados para cada feature.**

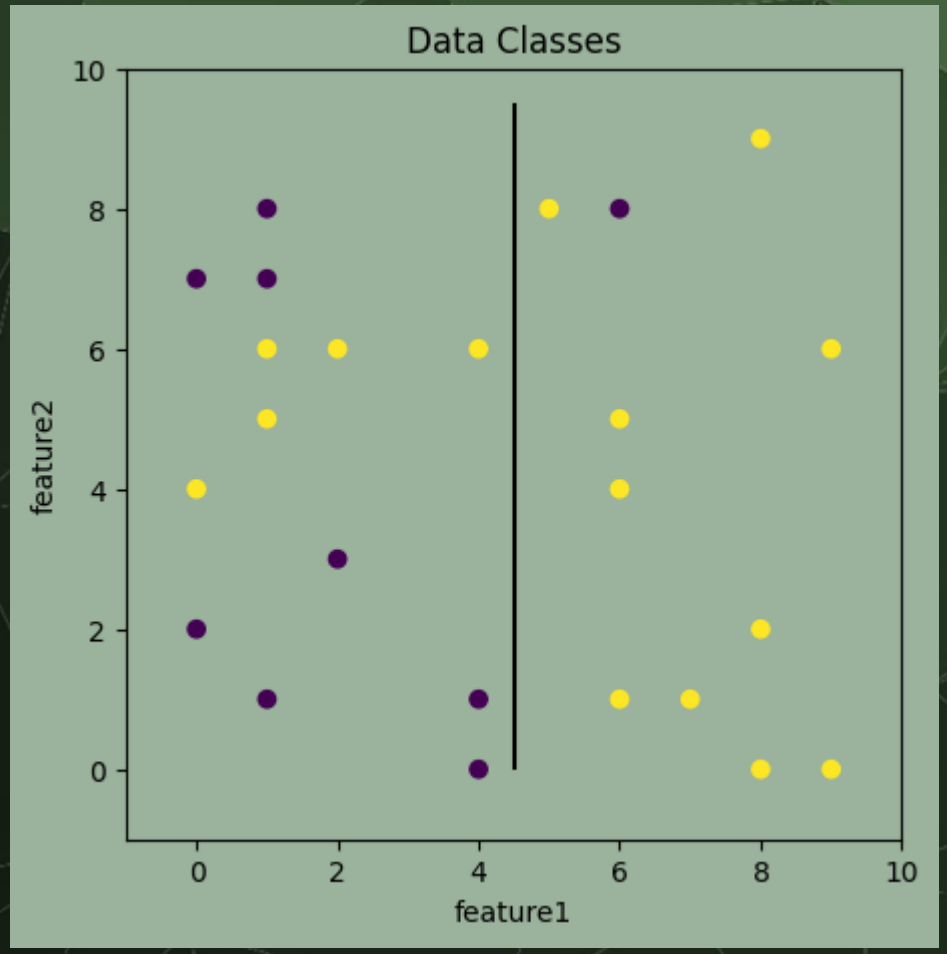
2) **Encontramos la ganancia de información para cada umbral. Es decir, el promedio ponderado de cada índice de Gini obtenido de dividir el espacio en dos subconjuntos.**

3) **Elegimos aquel que tengan la mayor ganancia de información (equivalente a obtener el valor menor de índice de Gini calculado en el paso anterior) de todos los umbrales disponibles.**

4) **Repetimos en cada sub-espacio disponible.**

TAE IA
Formación de Talento Altamente Especializado

¿Cómo construir un árbol?



Para todas las variables disponibles, tenemos que elegir solo un umbral: Para elegir el umbral:

1) **Reducimos el espacio de búsqueda con los datos disponibles: obtenemos el punto medio de los valores ordenados para cada feature.**

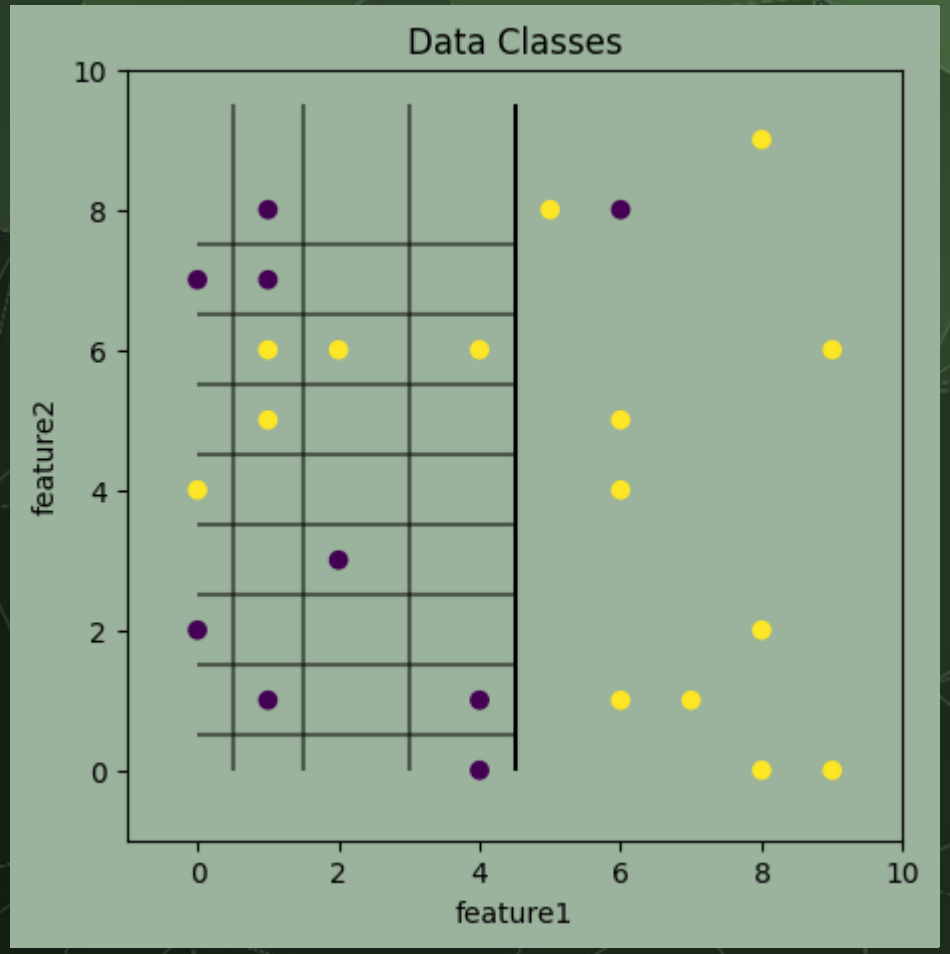
2) **Encontramos la ganancia de información para cada umbral. Es decir, el promedio ponderado de cada índice de Gini obtenido de dividir el espacio en dos subconjuntos.**

3) **Elegimos aquel que tengan la mayor ganancia de información (equivalente a obtener el valor menor de índice de Gini calculado en el paso anterior) de todos los umbrales disponibles.**

4) **Repetimos en cada sub-espacio disponible.**

TAE IA
Formación de Talento Altamente Especializado

¿Cómo construir un árbol?



Para todas las variables disponibles, tenemos que elegir solo un umbral: Para elegir el umbral:

1) **Reducimos el espacio de búsqueda con los datos disponibles: obtenemos el punto medio de los valores ordenados para cada feature.**

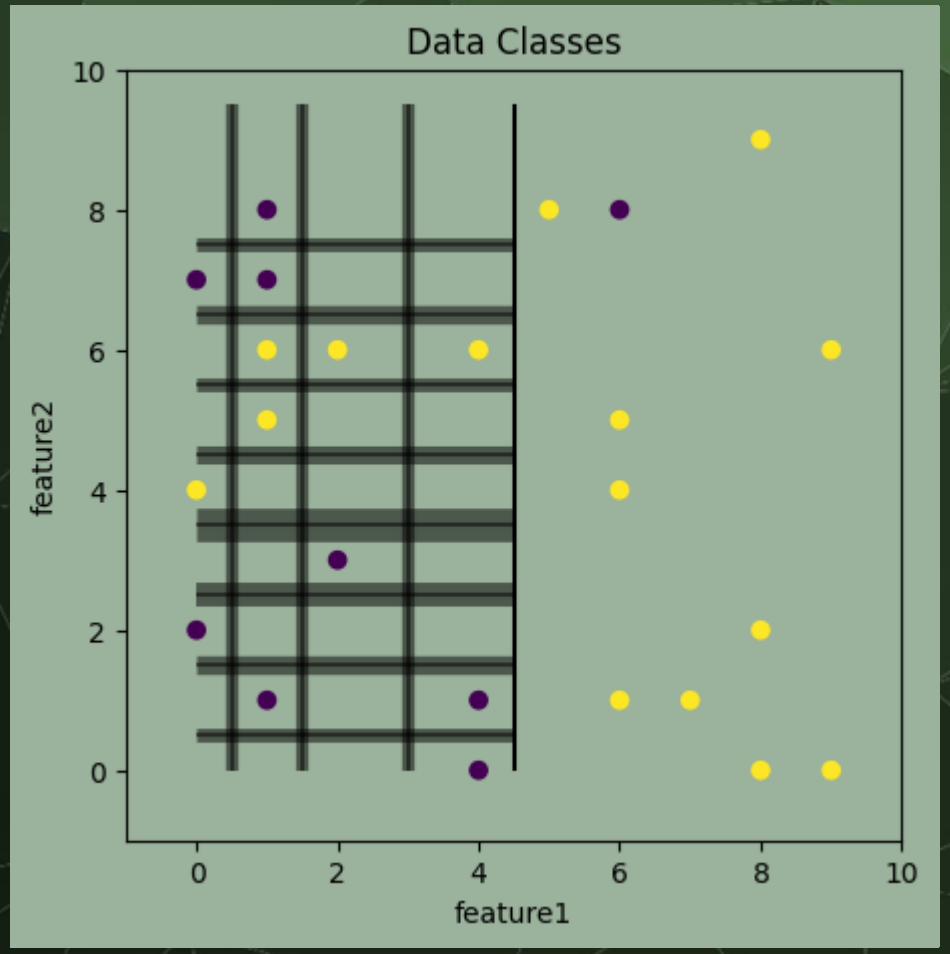
2) **Encontramos la ganancia de información para cada umbral. Es decir, el promedio ponderado de cada índice de Gini obtenido de dividir el espacio en dos subconjuntos.**

3) **Elegimos aquel que tengan la mayor ganancia de información (equivalente a obtener el valor menor de índice de Gini calculado en el paso anterior) de todos los umbrales disponibles.**

4) **Repetimos en cada sub-espacio disponible.**

TAE IA
Formación de Talento Altamente Especializado

¿Cómo construir un árbol?



Para todas las variables disponibles, tenemos que elegir solo un umbral: Para elegir el umbral:

1) **Reducimos el espacio de búsqueda con los datos disponibles: obtenemos el punto medio de los valores ordenados para cada feature.**

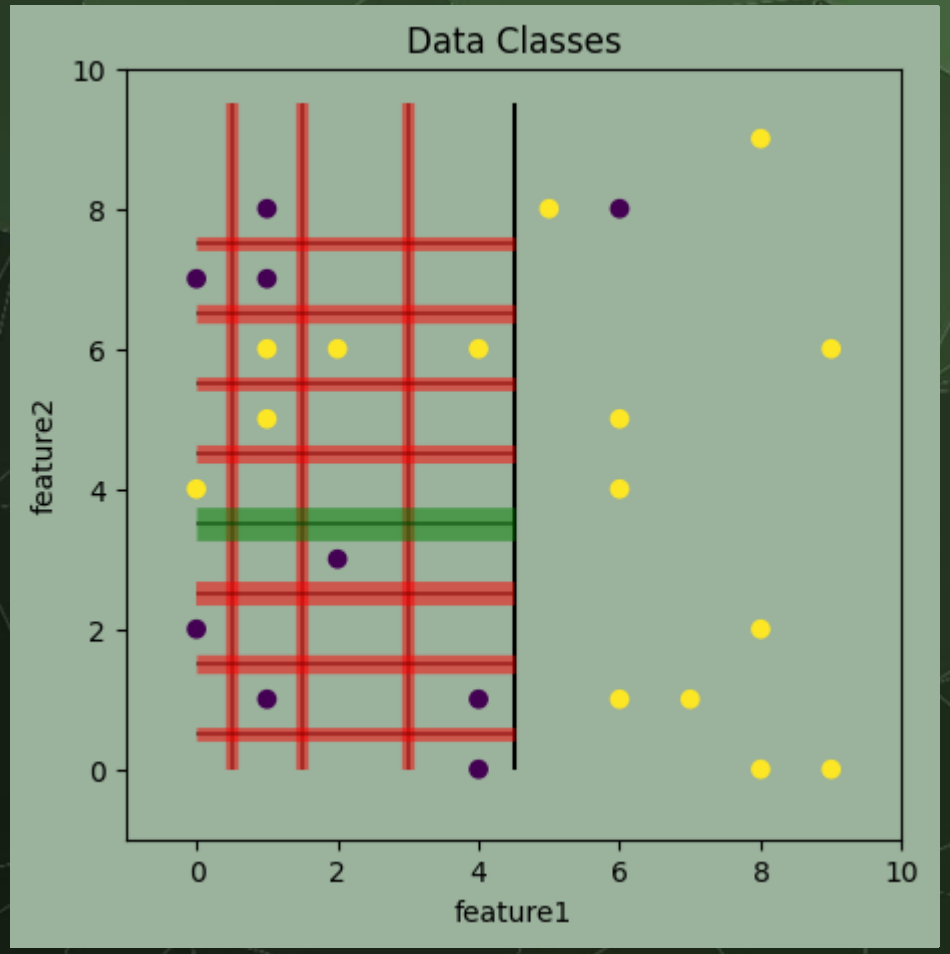
2) **Encontramos la ganancia de información para cada umbral. Es decir, el promedio ponderado de cada índice de Gini obtenido de dividir el espacio en dos subconjuntos.**

3) **Elegimos aquel que tengan la mayor ganancia de información (equivalente a obtener el valor menor de índice de Gini calculado en el paso anterior) de todos los umbrales disponibles.**

4) **Repetimos en cada sub-espacio disponible.**

TAE IA
Formación de Talento Altamente Especializado

¿Cómo construir un árbol?



Para todas las variables disponibles, tenemos que elegir solo un umbral: Para elegir el umbral:

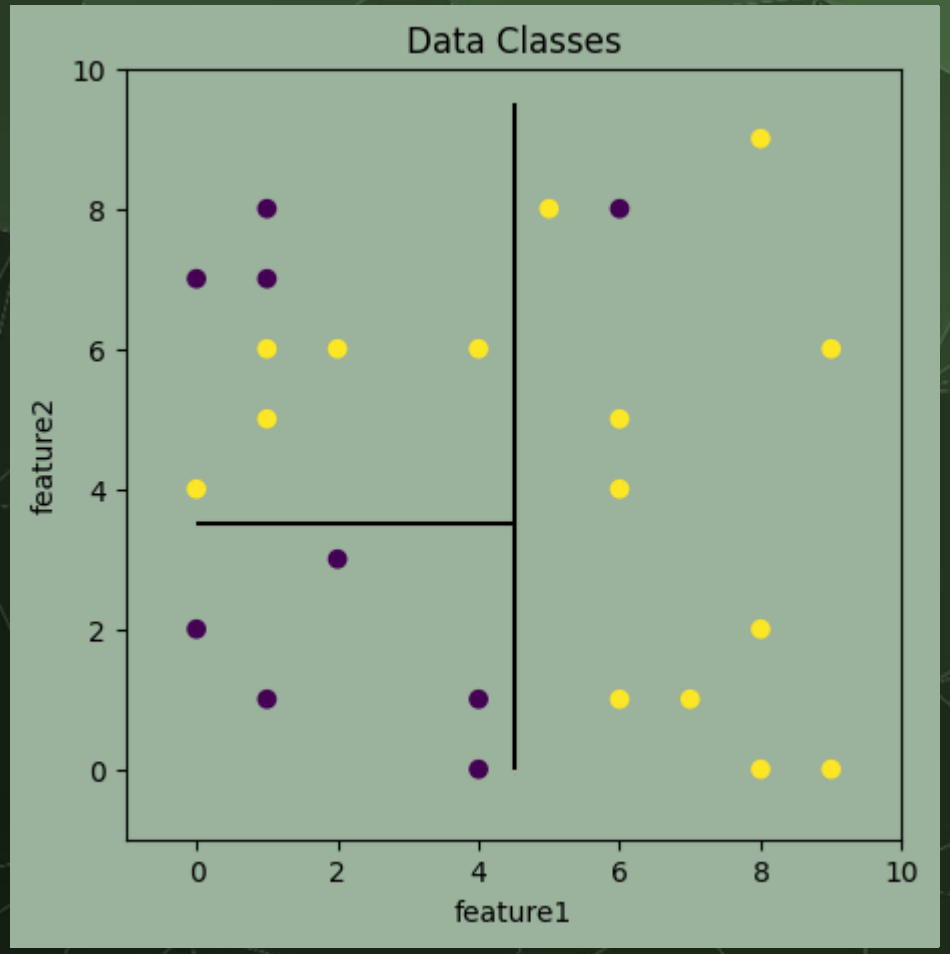
1) Reducimos el espacio de búsqueda con los datos disponibles: obtenemos el punto medio de los valores ordenados para cada feature.

2) Encontramos la ganancia de información para cada umbral. Es decir, el promedio ponderado de cada índice de Gini obtenido de dividir el espacio en dos subconjuntos.

3) Elegimos aquel que tengan la mayor ganancia de información (equivalente a obtener el valor menor de índice de Gini calculado en el paso anterior) de todos los umbrales disponibles.

4) Repetimos en cada sub-espacio disponible.

¿Cómo construir un árbol?



Para todas las variables disponibles, tenemos que elegir solo un umbral: Para elegir el umbral:

1) **Reducimos el espacio de búsqueda con los datos disponibles: obtenemos el punto medio de los valores ordenados para cada feature.**

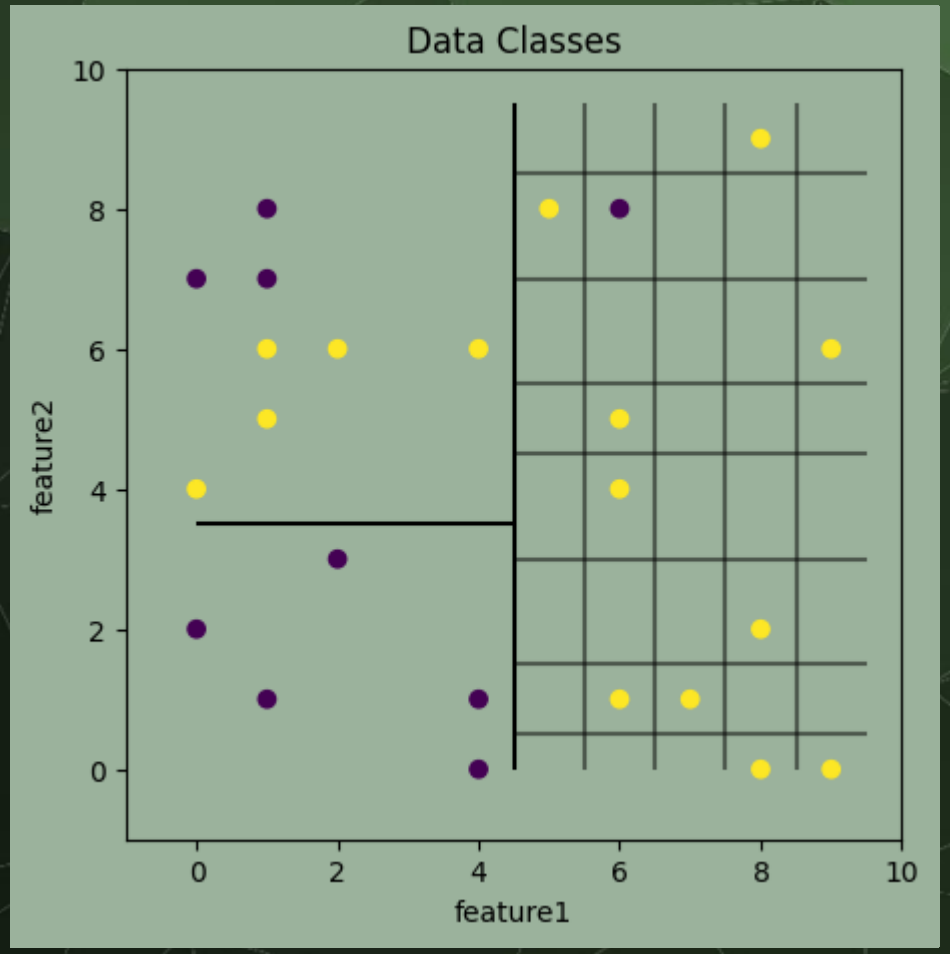
2) **Encontramos la ganancia de información para cada umbral. Es decir, el promedio ponderado de cada índice de Gini obtenido de dividir el espacio en dos subconjuntos.**

3) **Elegimos aquel que tengan la mayor ganancia de información (equivalente a obtener el valor menor de índice de Gini calculado en el paso anterior) de todos los umbrales disponibles.**

4) **Repetimos en cada sub-espacio disponible.**

TAE IA
Formación de Talento Altamente Especializado

¿Cómo construir un árbol?



Para todas las variables disponibles, tenemos que elegir solo un umbral: Para elegir el umbral:

1) **Reducimos el espacio de búsqueda con los datos disponibles: obtenemos el punto medio de los valores ordenados para cada feature.**

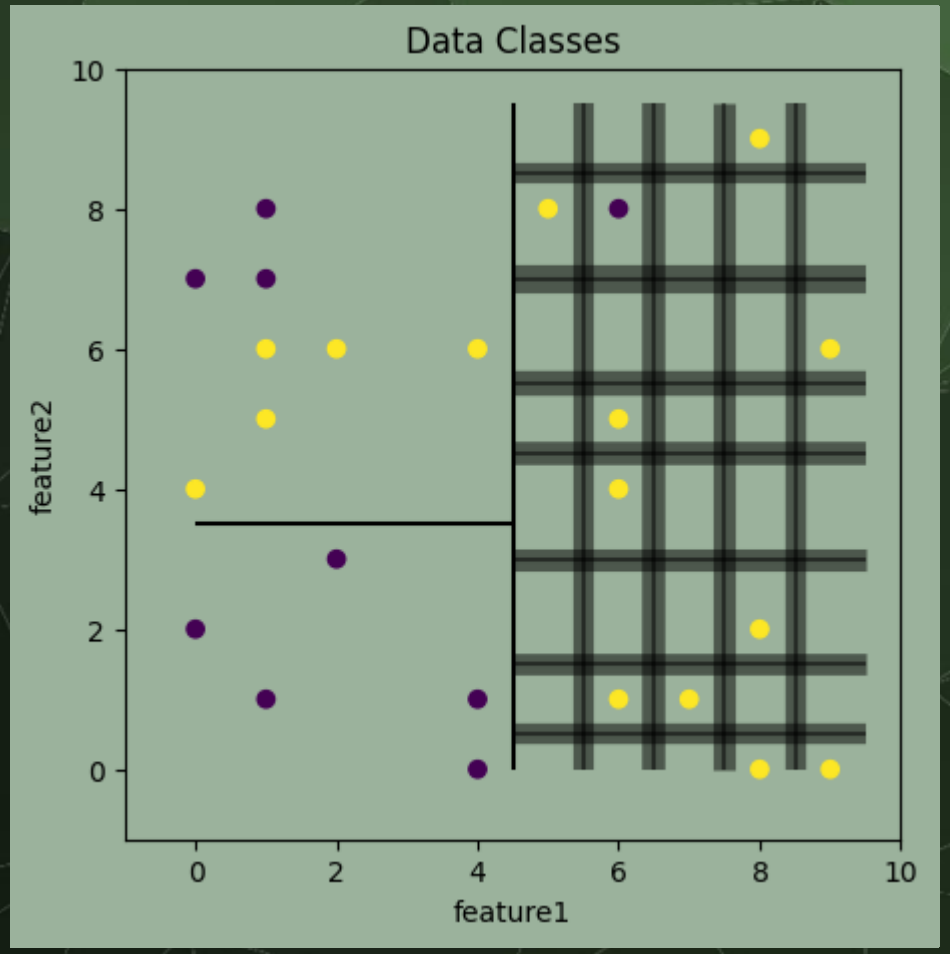
2) **Encontramos la ganancia de información para cada umbral. Es decir, el promedio ponderado de cada índice de Gini obtenido de dividir el espacio en dos subconjuntos.**

3) **Elegimos aquel que tengan la mayor ganancia de información (equivalente a obtener el valor menor de índice de Gini calculado en el paso anterior) de todos los umbrales disponibles.**

4) **Repetimos en cada sub-espacio disponible.**

TAE IA
Formación de Talento Altamente Especializado

¿Cómo construir un árbol?



Para todas las variables disponibles, tenemos que elegir solo un umbral: Para elegir el umbral:

1) **Reducimos el espacio de búsqueda con los datos disponibles: obtenemos el punto medio de los valores ordenados para cada feature.**

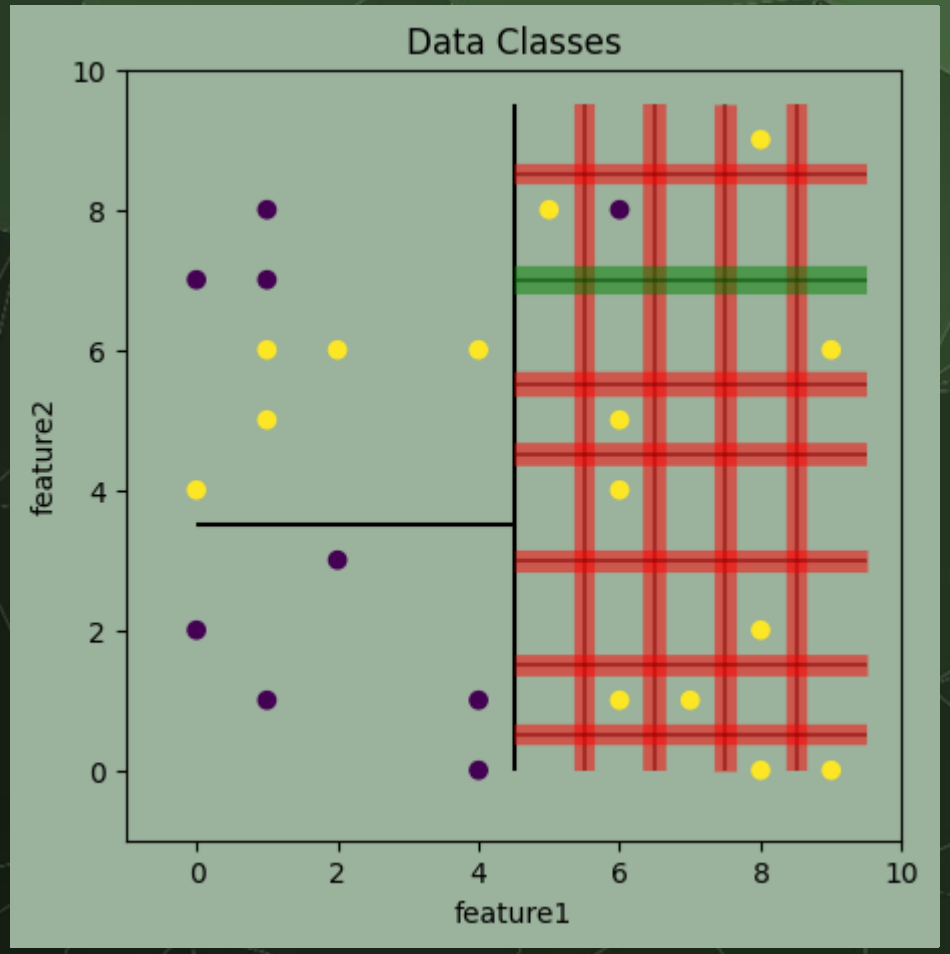
2) **Encontramos la ganancia de información para cada umbral. Es decir, el promedio ponderado de cada índice de Gini obtenido de dividir el espacio en dos subconjuntos.**

3) **Elegimos aquel que tengan la mayor ganancia de información (equivalente a obtener el valor menor de índice de Gini calculado en el paso anterior) de todos los umbrales disponibles.**

4) **Repetimos en cada sub-espacio disponible.**

TAE IA
Formación de Talento Altamente Especializado

¿Cómo construir un árbol?



Para todas las variables disponibles, tenemos que elegir solo un umbral: Para elegir el umbral:

1) **Reducimos el espacio de búsqueda con los datos disponibles: obtenemos el punto medio de los valores ordenados para cada feature.**

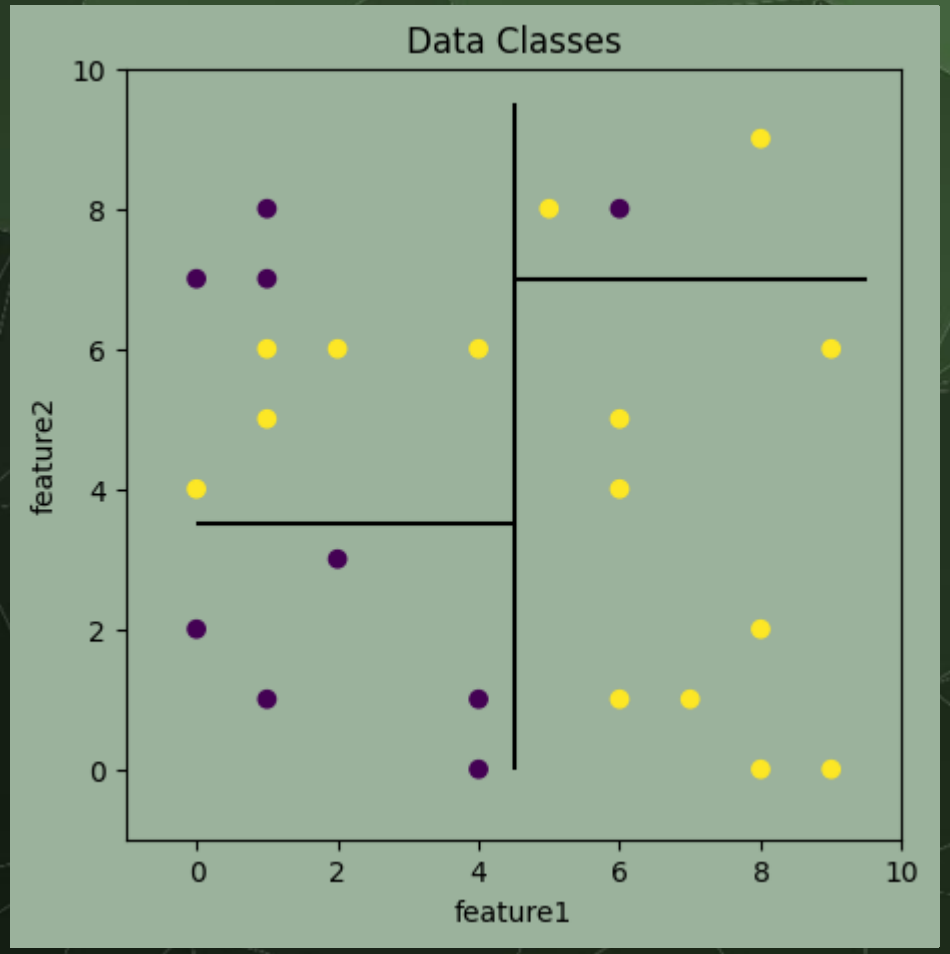
2) **Encontramos la ganancia de información para cada umbral. Es decir, el promedio ponderado de cada índice de Gini obtenido de dividir el espacio en dos subconjuntos.**

3) **Elegimos aquel que tengan la mayor ganancia de información (equivalente a obtener el valor menor de índice de Gini calculado en el paso anterior) de todos los umbrales disponibles.**

4) **Repetimos en cada sub-espacio disponible.**

TAE IA
Formación de Talento Altamente Especializado

¿Cómo construir un árbol?



Para todas las variables disponibles, tenemos que elegir solo un umbral: Para elegir el umbral:

1) **Reducimos el espacio de búsqueda con los datos disponibles: obtenemos el punto medio de los valores ordenados para cada feature.**

2) **Encontramos la ganancia de información para cada umbral. Es decir, el promedio ponderado de cada índice de Gini obtenido de dividir el espacio en dos subconjuntos.**

3) **Elegimos aquel que tengan la mayor ganancia de información (equivalente a obtener el valor menor de índice de Gini calculado en el paso anterior) de todos los umbrales disponibles.**

4) **Repetimos en cada sub-espacio disponible.**

TAE IA
Formación de Talento Altamente Especializado

¿Cómo construir un árbol?

En regresión se realiza lo mismo… pero cambia un poco:

1) **Reducimos el espacio de búsqueda con los datos disponibles: obtenemos el punto medio de los valores ordenados para cada feature.**

2) **Encontramos la suma total del error (MSE) entre el valor verdadero (y) y el promedio de "y" de cada espacio obtenido al seprararlo en dos subconjuntos.**

3) **Elegimos aquel tengan el menor error de todos los umbrales disponibles.**

4) **Repetimos en cada sub-espacio disponible. Pero debemos definir un valor mínimo de datos en cada hoja/nodo, si no, vamos a realizar muchísimas particiones.**

Ventajas y Desventajas

Ventajas y Desventajas

- Abordan relaciones complejas (no lineales)
- Pueden trabajar bien con datos con alta dimensionalidad.
- Requieren un mínimo preprocesamiento de datos.
- Son robustos ante los datos aberrantes.
- No tienen sensibilidad ante transformaciones monótonas.
- El mismo algoritmo te puede proporcionar la importancia de las variables.
- Es fácil de interpretar: La explicabilidad es relativamente sencilla.

Ventajas y Desventajas

- Su desempeño usualmente es pobre.
- Tienen un poder predictivo limitado.
- No son buenos extrapolando información.
- Los árboles son modificados por pequeñas perturbaciones en los datos.
- Tienden a sobreajustar.
- Pueden tornarse excesivamente complejos.

Ventajas y Desventajas

No todo está perdido

Existe una manera de perder las desventajas y mantener la mayoría de las ventajas

Ventajas y Desventajas

No todo está perdido

Existe una manera de perder las desventajas y mantener la mayoría de las ventajas

Métodos de Ensamble

Combinan las predicciones de estimadores (modelos) base con un algoritmo para mejorar la generalización y robustez.

Ventajas y Desventajas

No todo está perdido

Existe una manera de perder las desventajas y mantener la mayoría de las ventajas

Métodos de Ensamble

- **Bagging**
- **Boosting**

Bagging

Bagging

Definición

Algoritmo de ensamble donde se busca seleccionar una muestra de datos aleatoria con reemplazo (cada elemento se puede elegir más de una vez), y con estos se entrenan diversos modelos simples de forma independiente. Dependiendo de la tarea se realiza el promedio del valor obtenido o se obtiene la mayoría de votos.

Bagging

Información adicional:

- También es conocido como (B)oostrap(Agg)regation.
- El algoritmo “Random Forest” se basa en aplicar esta técnica, con árboles de decisiones.

Random Forest

- Se entrenan muchos árboles de decisión, limitando las características que pueden ver.
- Al entrenar tantos árboles, se desacopla el error (se logra generalizar).
- No tiende a sobreajustarse debido a que contempla muchos subconjuntos de datos para realizar la respuesta.

Random Forest

Random Forest

TAE IA
Formación de Talento Altamente Especializado

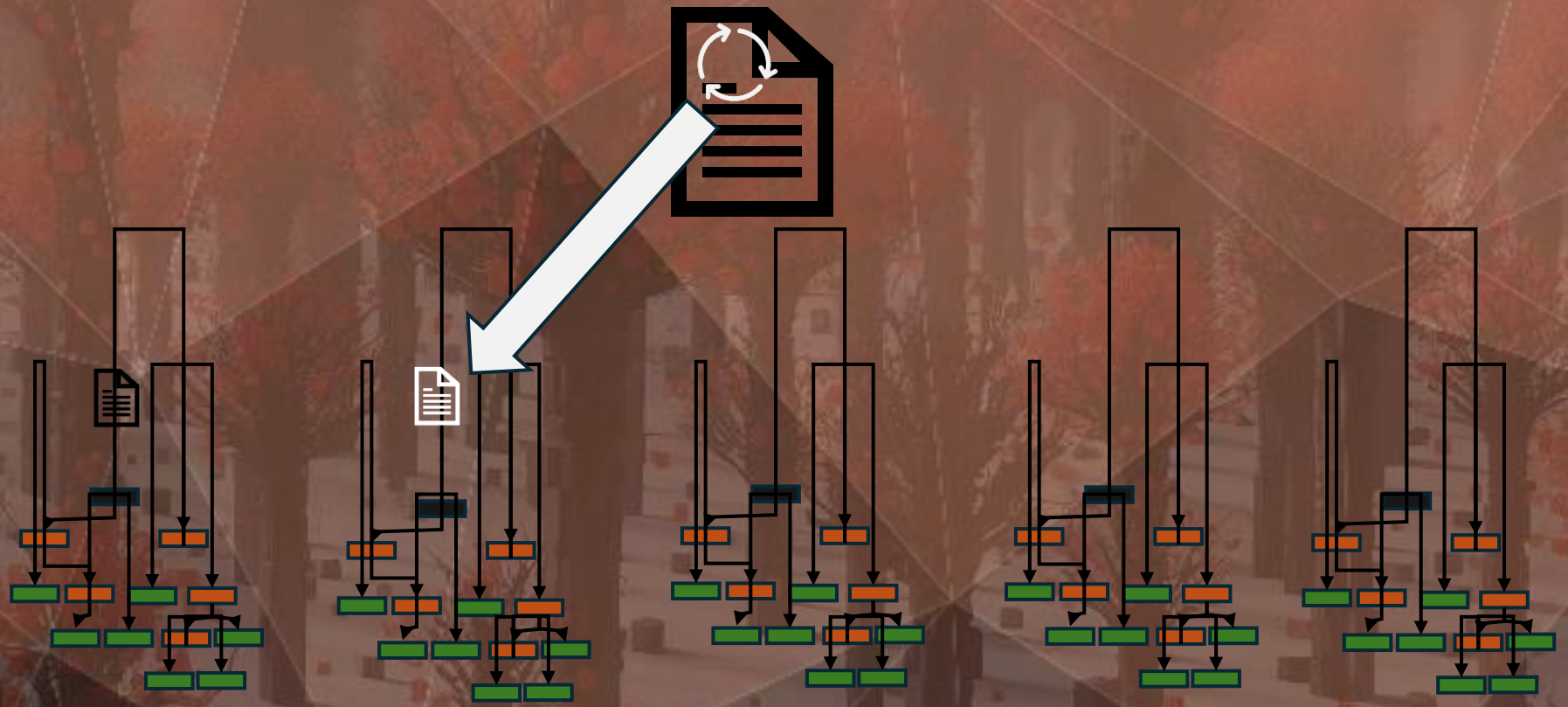
Random Forest

TAE IA
Formación de Talento Altamente Especializado

Random Forest

Random Forest

Random Forest



TAE IA
Formación de Talento Altamente Especializado

Random Forest

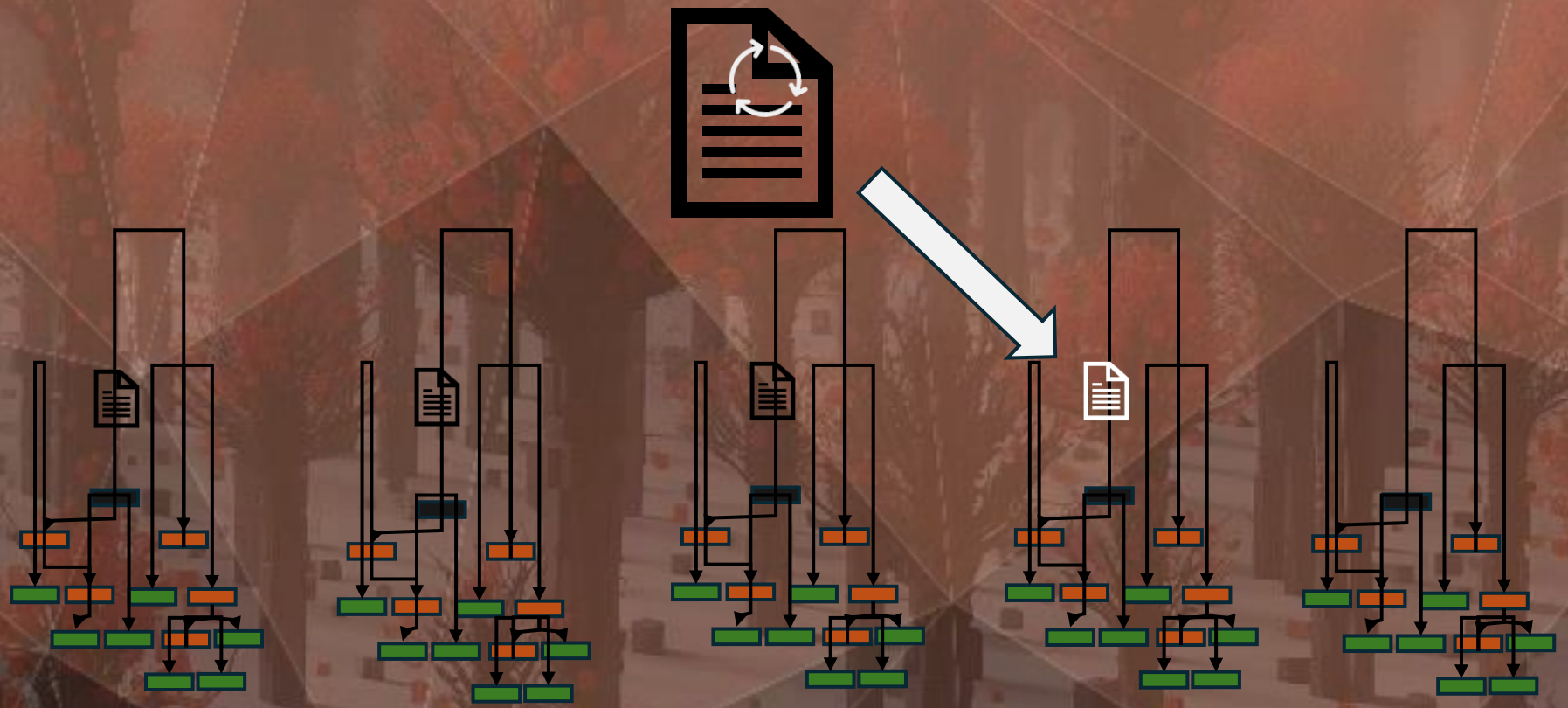
Random Forest

Random Forest

Random Forest

Random Forest

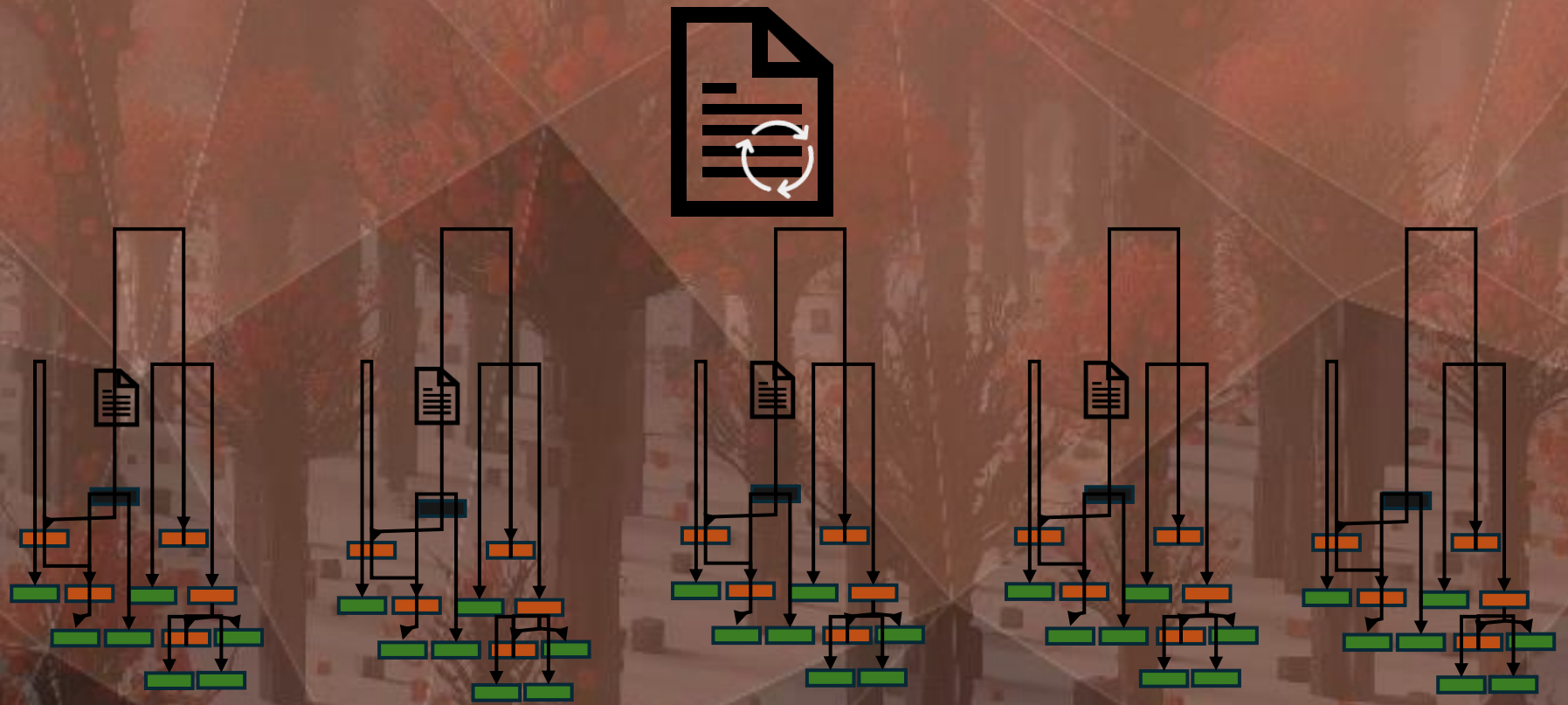
Random Forest



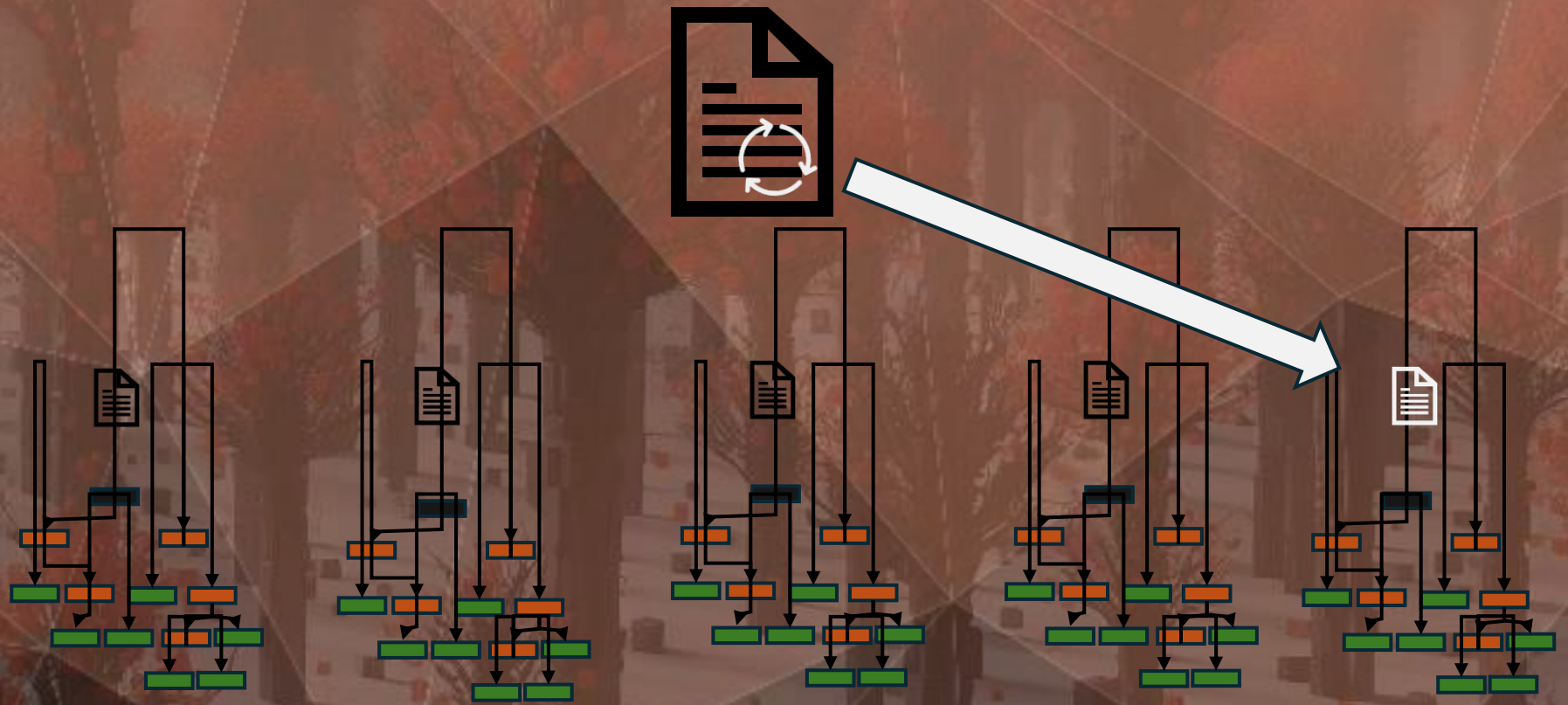
TAE IA
Formación de Talento Altamente Especializado

Random Forest

Random Forest



Random Forest

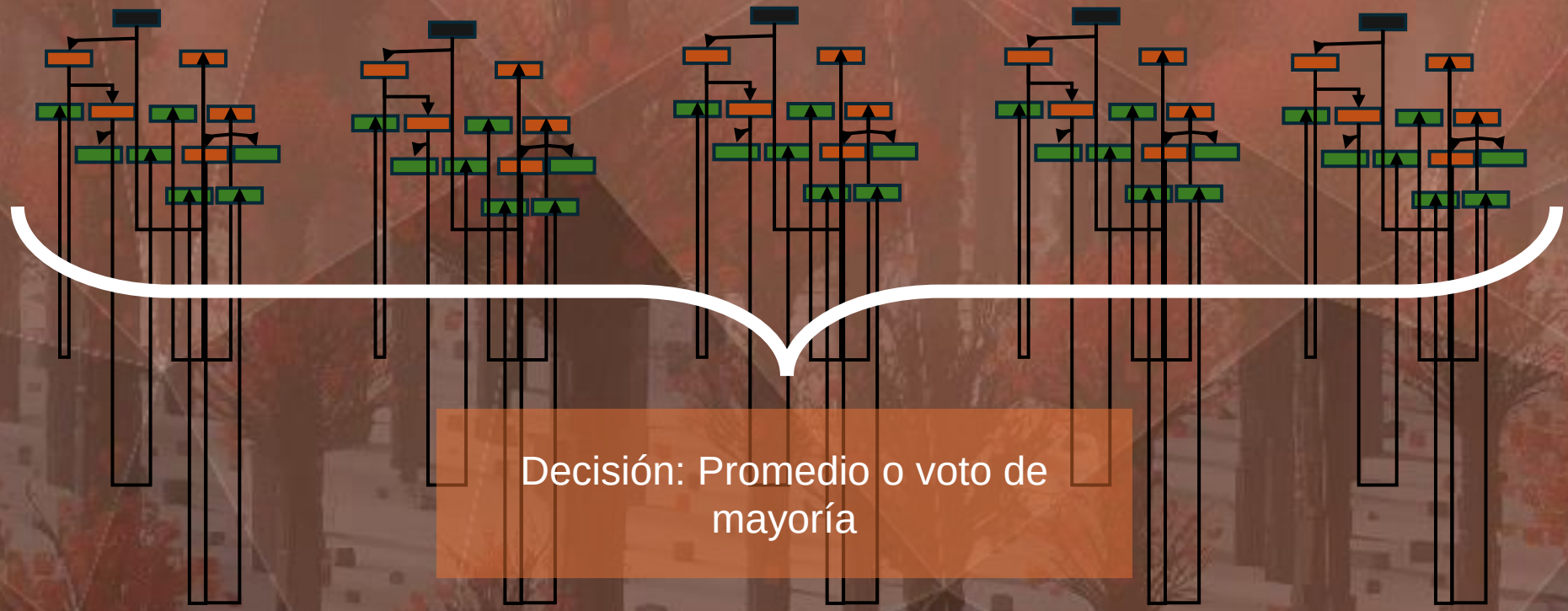


TAE IA
Formación de Talento Altamente Especializado

Random Forest

Random Forest

Random Forest



Decisión: Promedio o voto de mayoría

Boosting

Boosting

Definición

Algoritmo de ensamble donde se busca reducir los errores de manera secuencial, es decir: Cada modelo nuevo se entrena dándole importancia a observaciones donde el modelo anterior se equivocó. La contribución de cada nuevo modelo se ve modificada por un factor llamado taza de aprendizaje (eta).

Boosting

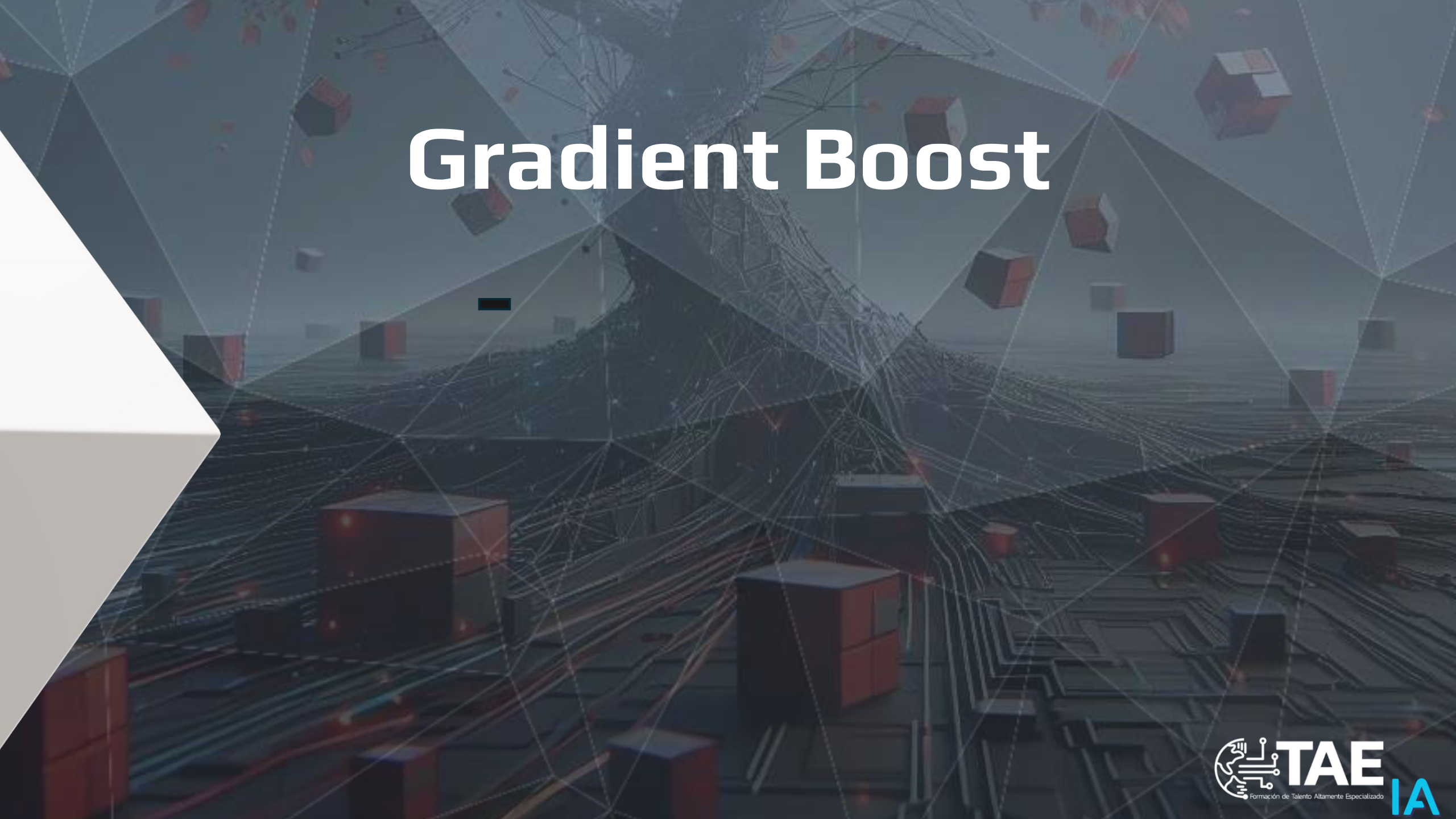
Información adicional

Boosting es una técnica muy popular y se utiliza para muchos proyectos.

“Pequeños pasos en la dirección correcta te llevan a un buen lugar.”

Algunos de los algoritmos más usados que utilizan esta técnica son: XGBoost, CatBoost, LightGBM.

Gradient Boost



Gradient Boost

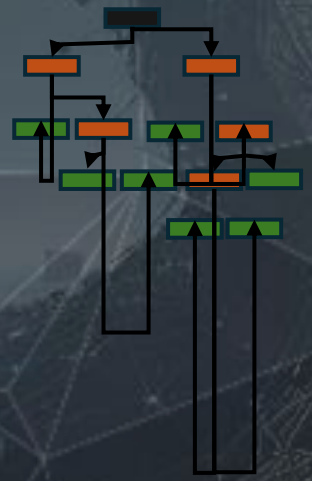
- Predicción inicial: Promedio

Gradient Boost

- Predicción inicial: Promedio

Encontramos el error de
nuestra predicción inicial
con la realidad: Creamos un árbol de
decisión para predecir este error, y le
asigamos un peso.

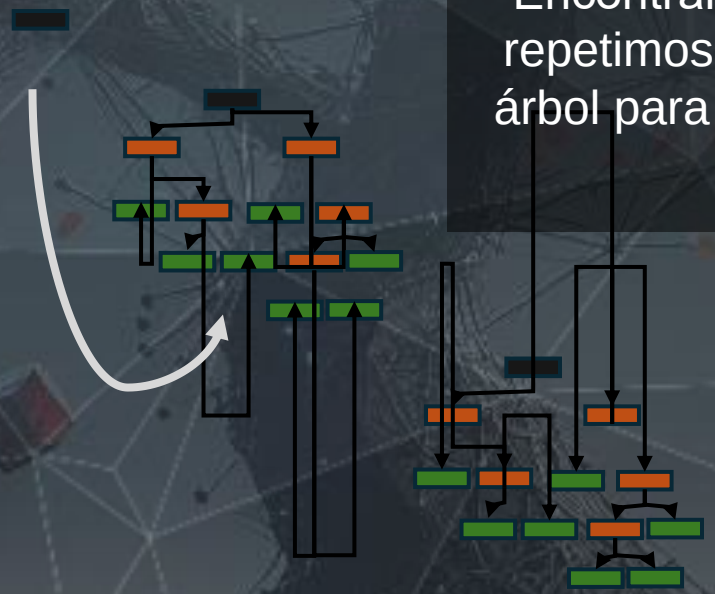
Gradient Boost



Gradient Boost

A la Predicción inicial le sumamos este resultado que es la predicción del error. La nueva predicción va acercarse un poco más al valor real.

Gradient Boost

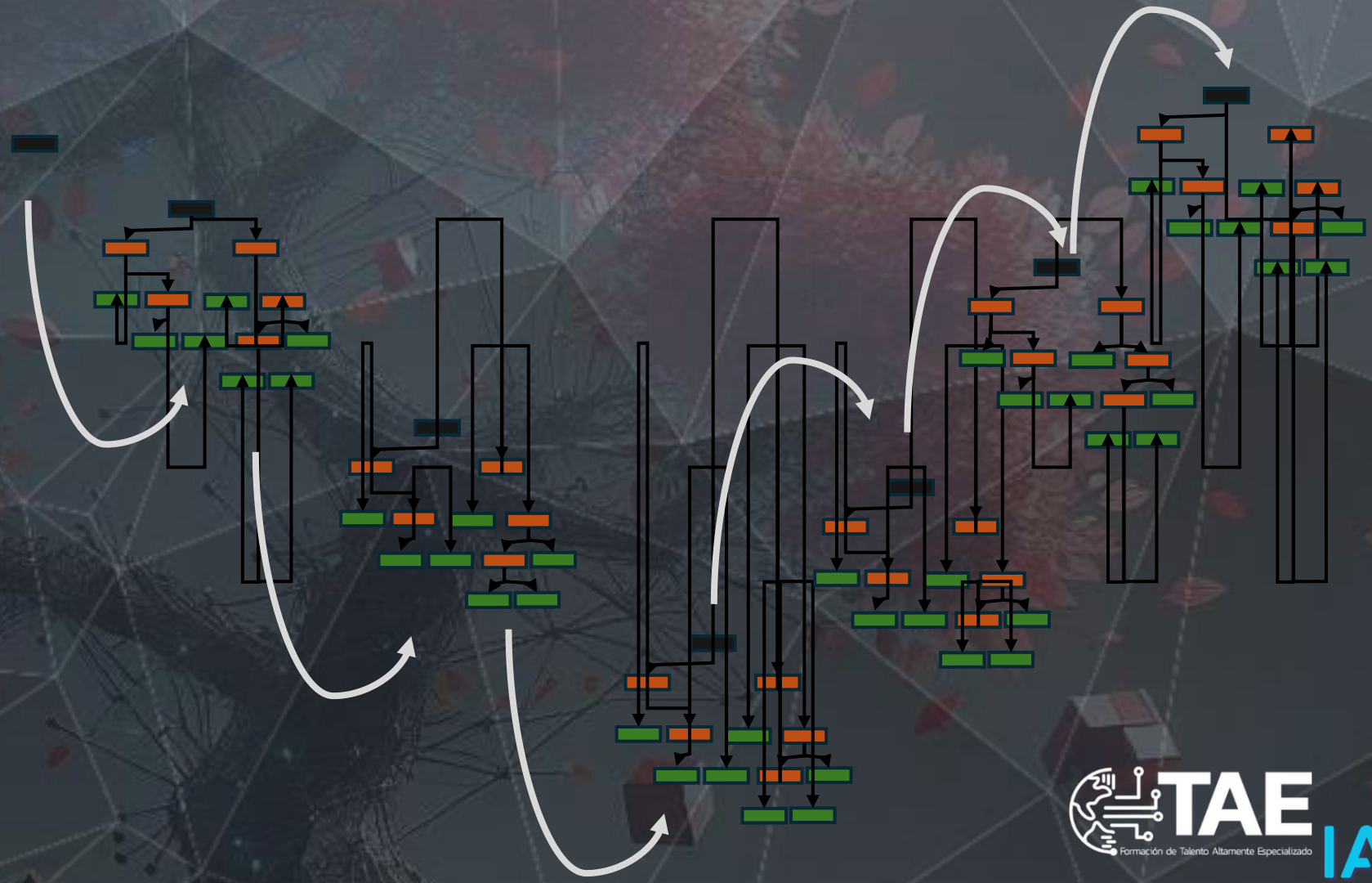


Encontramos el “nuevo error” y repetimos con la creación de un árbol para predecir estos nuevos errores.

Gradient Boost

Iteramos… hasta terminar el número de árboles deseado (mala práctica), o hasta que los nuevos árboles no contribuyan a la mejora del modelo.

Gradient Boost



TAE IA
Formación de Talento Altamente Especializado

Ventajas y Desventajas de los métodos de ensamble

Ventajas y Desventajas de los métodos de ensamble

- Abordan relaciones complejas (no lineales)
- Pueden trabajar bien con datos con alta dimensionalidad.
- Requieren un mínimo preprocesamiento de datos.
- Son robustos ante los datos aberrantes.
- No tienen sensibilidad ante transformaciones monótonas.
- El mismo algoritmo te puede proporcionar la importancia de las variables.
- Desempeño usualmente bueno.

Ventajas y Desventajas de los métodos de ensamble

- El resultado es excesivamente complejo, es decir la explicabilidad se vuelve complicada.

- La extrapolación les cuesta trabajo

Consejos:

- Algoritmos funcionan muy bien con datos tabulares: Columnas tienen significados distintos. Incluso superan redes neuronales y deep learning.
- Utilizar estos algorimos como un acercamiento inicial es relativamente sencillo por el poco preprocesamiento.
- Es tentador involucrar todas las variables posibles cuando existe tánta facilidad de tener un modelo, esto no es una buena práctica, se deben de seleccionar las variables adecuadas.
- Al usar un árbol de decision se debe hacer solo si la representación final debe ser sencilla. No servirá para modelar cosas complejas.

Referencias

- Scikit Learn (2024) Decision Trees . Recuperado el 29 de Julio del 2024 de: https://scikit-learn.org/stable/modules/tree.html
- IBM (2024). ¿Qué es un árbol de decision? Recuperado el 29 de Julio del 2024 de: https://www.ibm.com/mx-es/topics/decision-trees
- L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth, Belmont, CA, 1984.
- Hastie, T.; Tibshirani, R. and Friedman, J. (2008) The elements of statistical learning Data Mining, Inference, and Prediction. Springer. Standford California.
- XGBOOST (2024). XGBoost Documentation. Recuperado el 29 de julio del 2024 de: https://xgboost.readthedocs.io/en/stable/
- Friedman, J. (1999) Greedy Function Approximation: A Gradient Boosting Machine. Recuperado el 08 de Agosto del 2024 de: https://jerryfriedman.su.domains/ftp/trebst.pdf
- Armon, A., and Shwartz-Ziv, R. (2021). Tabular Data: Deep Learning is Not All You Need. Recuperado el 09 de Agosto del 2024 de: https://arxiv.org/abs/2106.03253

TAE IA
Formación de Talento Altamente Especializado

Gracias por su atención
¿Dudas?