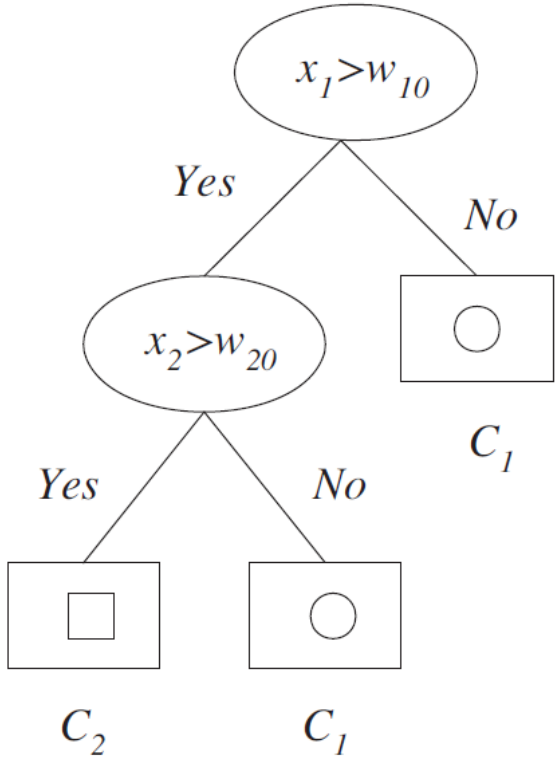
Aprendizaje Supervisado: Árboles de Decisión

CURSO GRUPO BANCOLOMBIA

Universidad Nacional de Colombia

Árboles de Decisión

- Un árbol de decisión es una estructura jerárquica que implementa la estrategia: **“divide y vencerás”**.
- Un modelo de árbol de decisión divide los datos, con una secuencia de preguntas.
- Las preguntas se conocen como nodos y cada nodo implementa una función simple para separar los datos.
- Un dato de entrada es evaluado en el nodo principal y después se evalúa en los nodos correspondientes
- El nodo final de la rama determina la clase a la que pertenecen los datos.



Tomado de: Introduction to Machine Learnin, Ethem Alpaydin, 2010

Árboles de Decisión

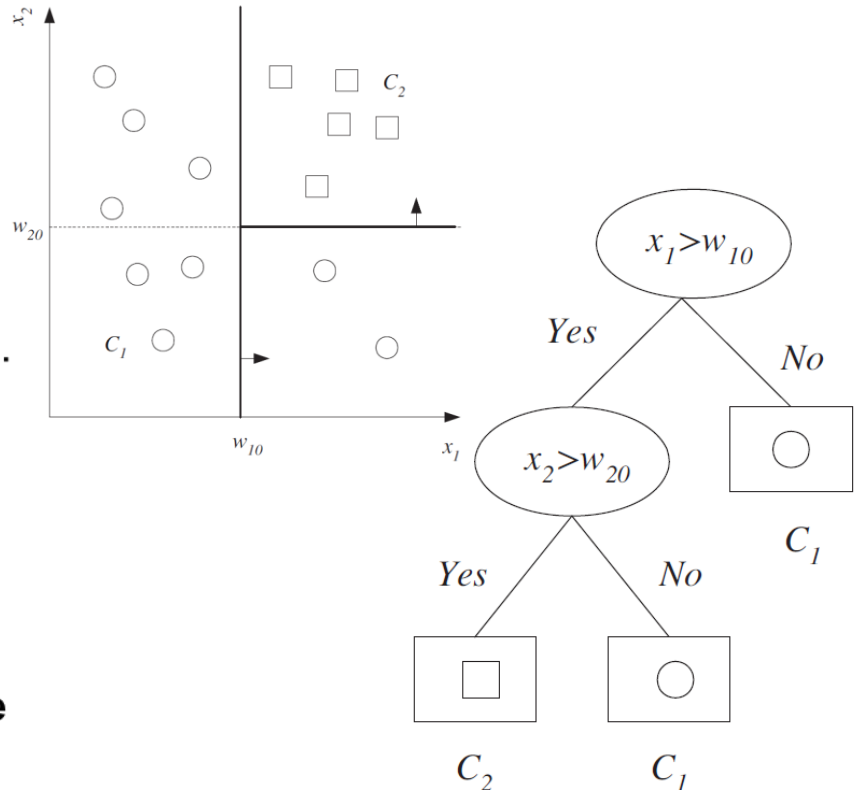
El árbol de decisión **clasifica** los datos dependiendo de su posición con respecto al valor de w_{10} y w_{20}.

- Clases: círculos C_1 y cuadrados C_2.
- Características: los datos están ubicados en el plano (x_1, x_2) .
- Cada nodo evalúa sólo una característica (x_1 ó x_2).
- La división se hace con una comparación:

$$f_m(\boldsymbol{x}) : x_j > w_{m0}$$

- w_{m0} es un umbral adecuadamente elegido.
- Para cada nodo, se evalúan todas las posibles opciones de w_{m0}

¿Cómo se determina el mejor w_{m0}? Con la **medida de impureza**.



Tomado de: Introduction to Machine Learnin, Ethem Alpaydin, 2010

Medida de impureza

Una división es pura si todos los datos de las ramas resultantes pertenecen a la misma clase.

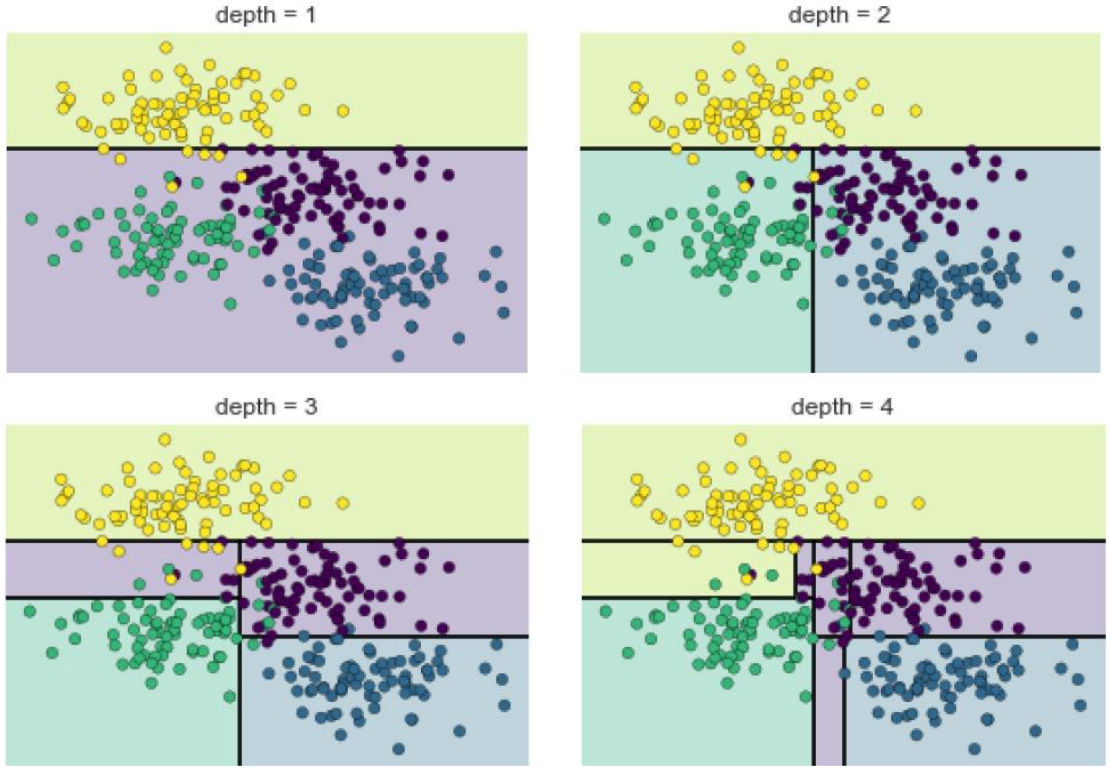
- Para un nodo m, N_m es el numero de datos de entrenamiento que alcanzan a llegar hasta el nodo m.
- Para el nodo principal, el número de datos de entrenamiento es N.
- N_m^i de N_m pertenecen a la clase C_i con
- La estimación de la probabilidad de la clase C_i es $\sum_i N_m^i = N_m$
- El nodo m es puro si p_m^i para todo i es 0 o 1 $\hat{P}(C_i|\boldsymbol{x}, m) \equiv p_m^i = \frac{N_m^i}{N_m}$
- Es 0 cuando ninguna de las instancias son de clase C_i.
- Es 1 si todas esas instancias son de C_i.
- Si la división es pura, no se divide más y se agrega un nodo final etiquetado con la clase para la cual p_m^i es 1.

Tomado de: Introduction to Machine Learnin, Ethem Alpaydin, 2010

Profundidad del Árbol de Decisión

- En los algoritmos de Árboles de Decisión, la profundidad es un parámetro configurable.
- Cada profundidad divide el subconjunto de datos así:

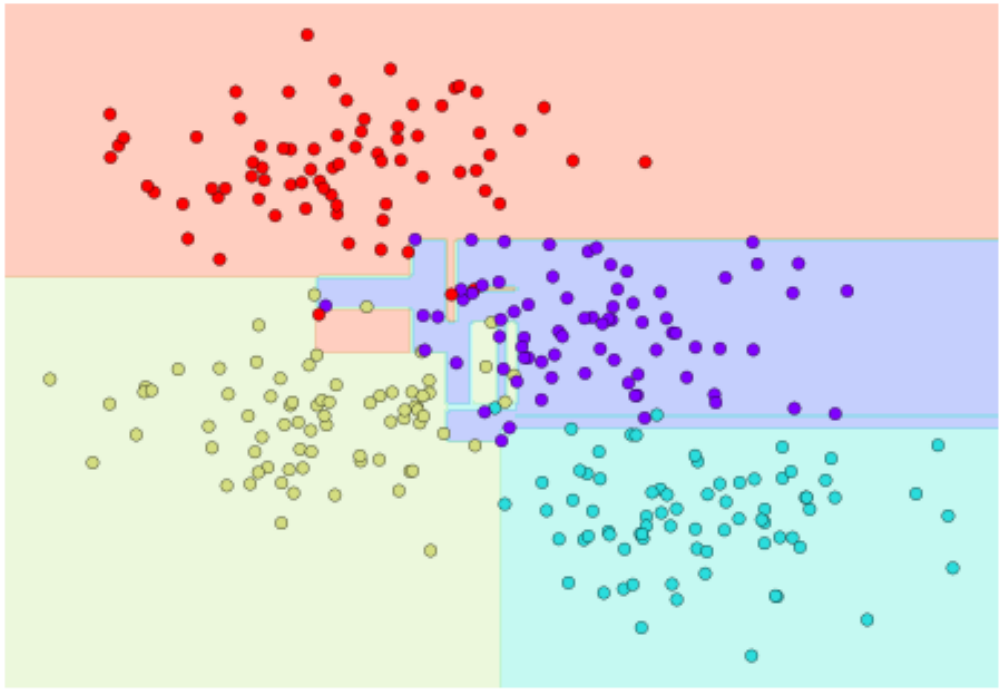
Considerando los siguientes datos bidimensionales con cuatro clases.



Tomado de: Python Data Science Handbook, Jake VanderPlas

Sobreajuste del clasificador

A medida que aumenta la profundidad de árbol, se obtienen regiones de clasificación con formas muy extrañas.
El Modelo de Árbol de Decisión puede llegar a **sobre ajustarse** a los datos.



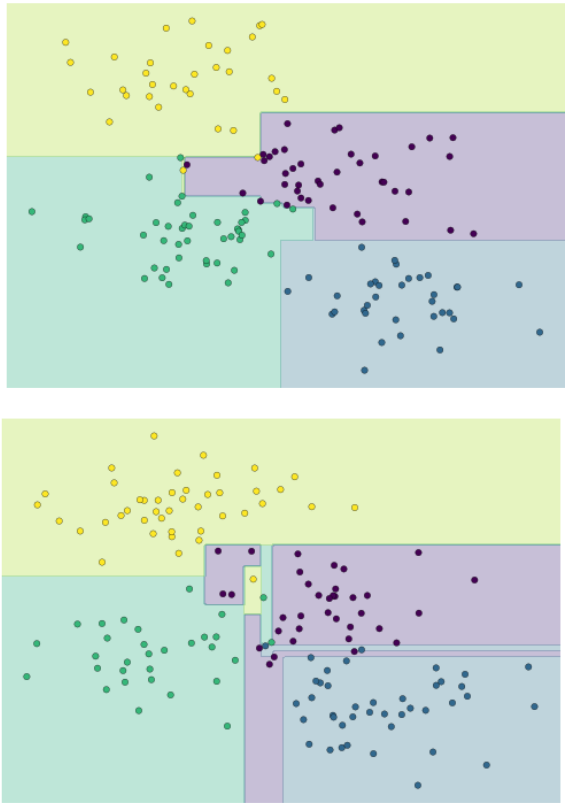
Tomado de: Python Data Science Handbook, Jake VanderPlas

Random Forests

Dos Árboles de Decisión con la Mitad de los Datos

- En algunos lugares, los dos árboles producen resultados consistentes.
- Mientras que en otros lugares, los dos árboles dan clasificaciones muy diferentes.
- Las inconsistencias tienden a ocurrir cuando la clasificación es menos segura y, por lo tanto, al usar la información de **ambos árboles**, se podría obtener un mejor resultado.
- Se podría esperar que el uso de la información de **muchos árboles** mejoraría aún más los resultados.

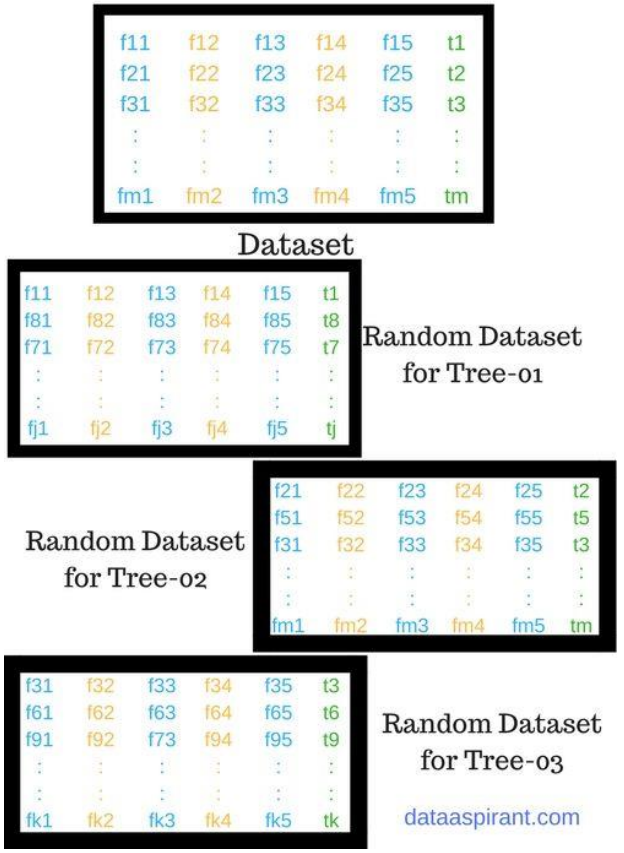
Un conjunto de árboles de decisión se conoce como un Random Forests.



Tomado de: Python Data Science Handbook, Jake VanderPlas

Algoritmo de Random Forests

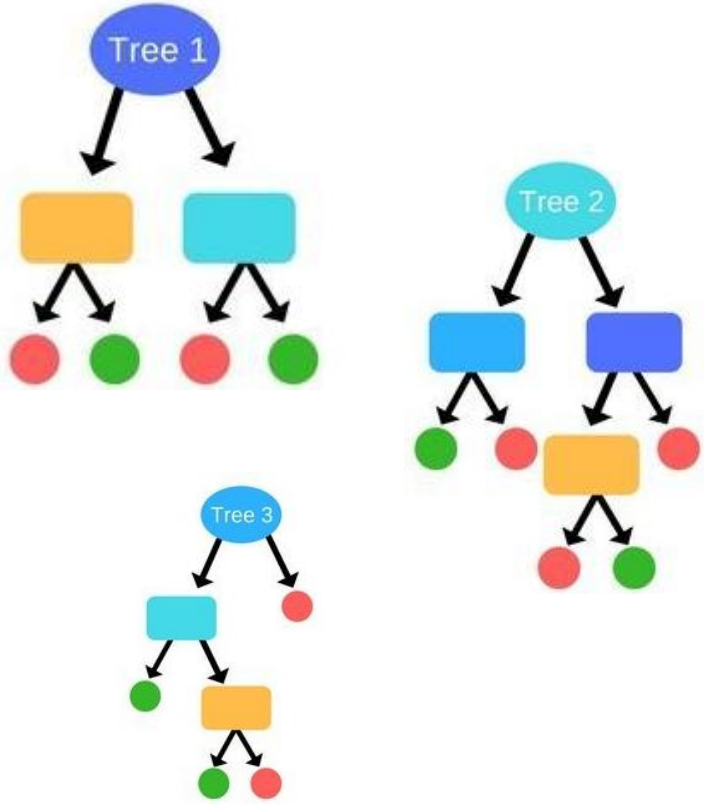
1. Se separa una porción de la base de datos original. La separación es aleatoria.
2. Se crea un árbol de decisión con la parte separada de la base de datos.
3. Se realiza otra separación aleatoria de la base de datos original.
4. Nuevamente se construye otro árbol de decisión para la nueva porción de los datos.
5. Esto se repite hasta que termine de calcular el número de estimadores (árboles de decisión) seleccionados por el usuario.



Tomado de: dataaspirant.com

Predicción con Random Forests

- Se toman las características del nuevo dato y se evalúan en cada árbol de decisión creado.
- Si se crearon 100 árboles de decisión, las características serán evaluadas en los 100 árboles.
- Se suman los votos de la predicción de cada árbol de decisión.
- La clase que tengas más votos será la clase ganadora.



Tomado de: dataaspirant.com

Ventajas de Random Forests

- Tanto el entrenamiento como la predicción son muy rápidos, debido a la simplicidad de los árboles de decisión subyacentes. Además, ambas tareas se pueden paralelizar directamente, porque los árboles individuales son entidades completamente independientes.
- Los múltiples árboles permiten una clasificación probabilística: un voto mayoritario entre los estimadores da una estimación de la probabilidad.
- El modelo no paramétrico es extremadamente flexible y, por lo tanto, puede funcionar bien en tareas que otros estimadores no ajustan.

Una desventaja principal de los bosques aleatorios es que los resultados no son fácilmente interpretables: es decir, si desea sacar conclusiones sobre el significado del modelo de clasificación, los bosques aleatorios pueden no ser la mejor opción.

Para Evaluar un Clasificador

- La exactitud es la relación entre el número de las predicciones correctas y número de predicciones totales. O sea, es la proporción de predicciones correctas.

$$Accurracy = \frac{\#\ predicciones\ correctas}{\#\ predicciones\ totales}$$

- El mejor valor posible es 1.

¡A codificar!