Subject: Probability and Statistics

Class XVIII: Decision trees, Random Forest

UdeA

Bioengineering

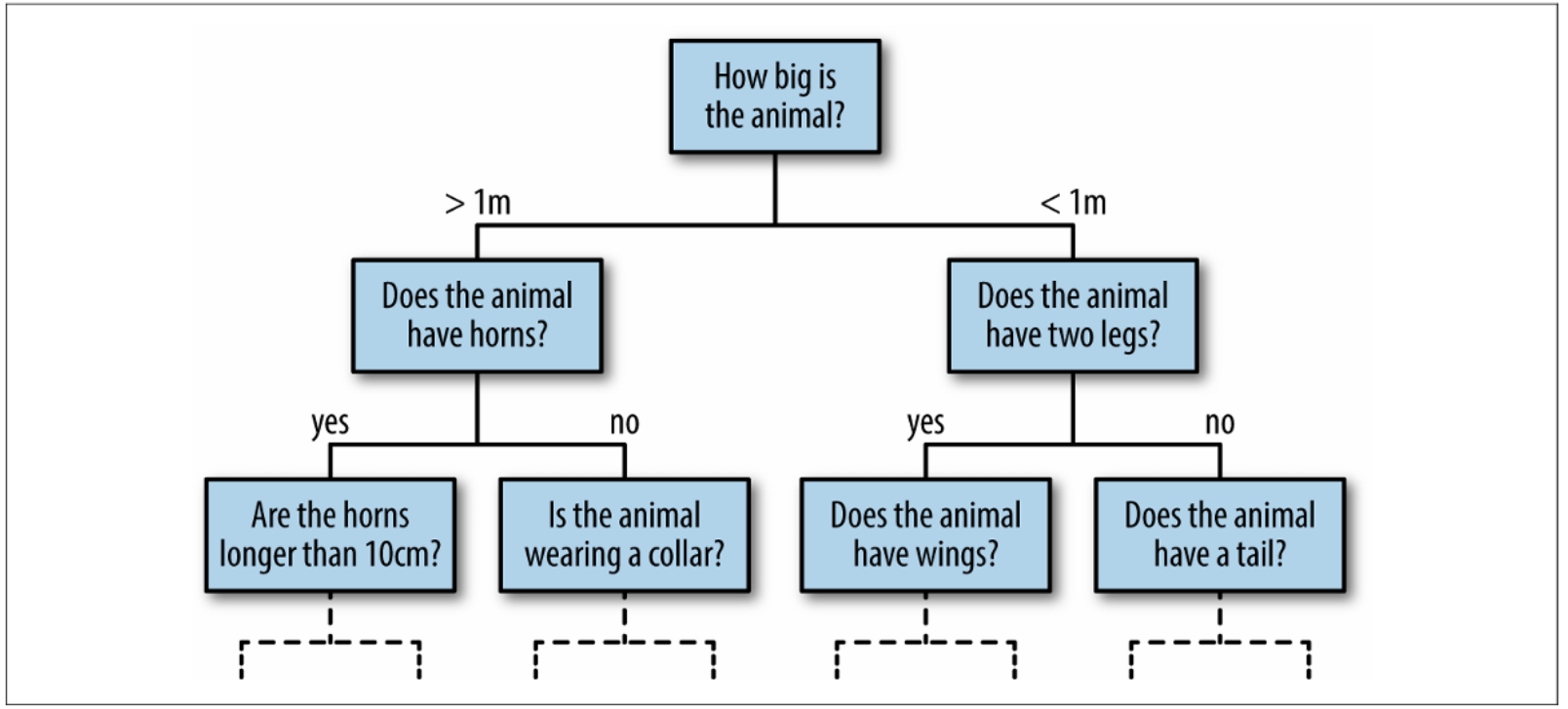
Francisco José Campuzano Cardona

Bioengineerer, MSc in Engineering

Decision Trees

UdeA

Los árboles de decisión son una forma extremadamente intuitiva de clasificar cosas. Básicamente la idea es hacer ciertas preguntas para ir acotando la información y llegar a una clasificación. Supongamos que queremos clasificar un animar, podemos usar el siguiente árbol binario:

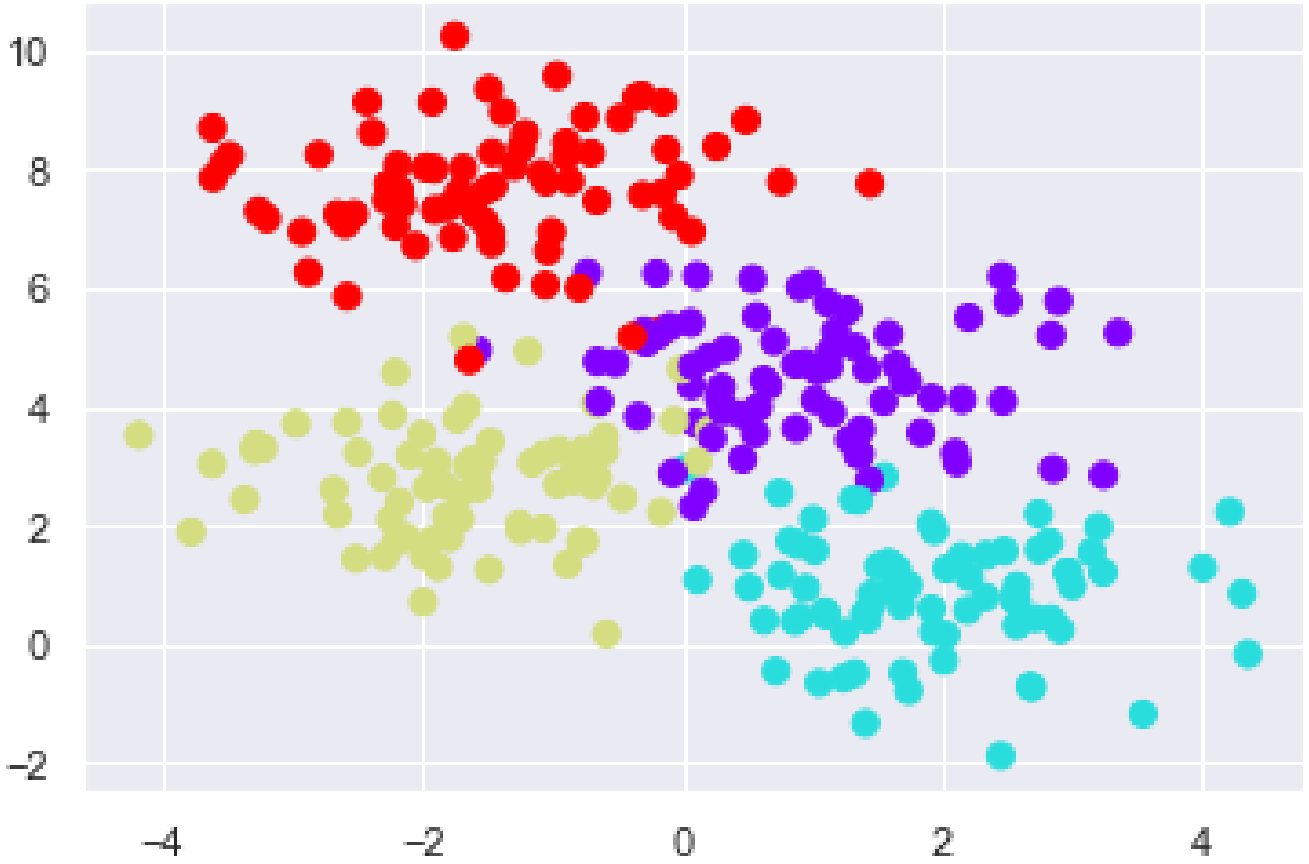


Entonces un árbol de decisión es una forma muy intuitiva y eficiente de clasificar, pero el truco está en la preguntas… ¿Qué preguntas debo hacer para ir separando apropiadamente los datos en grupos que me permitan llegar a las clases?

Entonces en aprendizaje automático esas “preguntas” son mas bien líneas rectas que actúan como umbrales, y entonces la preguntas es si los datos están por encima o por debajo de este umbral, de modo que estas líneas parten los datos en dos. Entonces un nodo de un árbol de decisión, corresponde a una partición a partir de un umbral puesto sobre una característica.

UdeA

Supongamos que tenemos estos datos clasificados en 4 clases



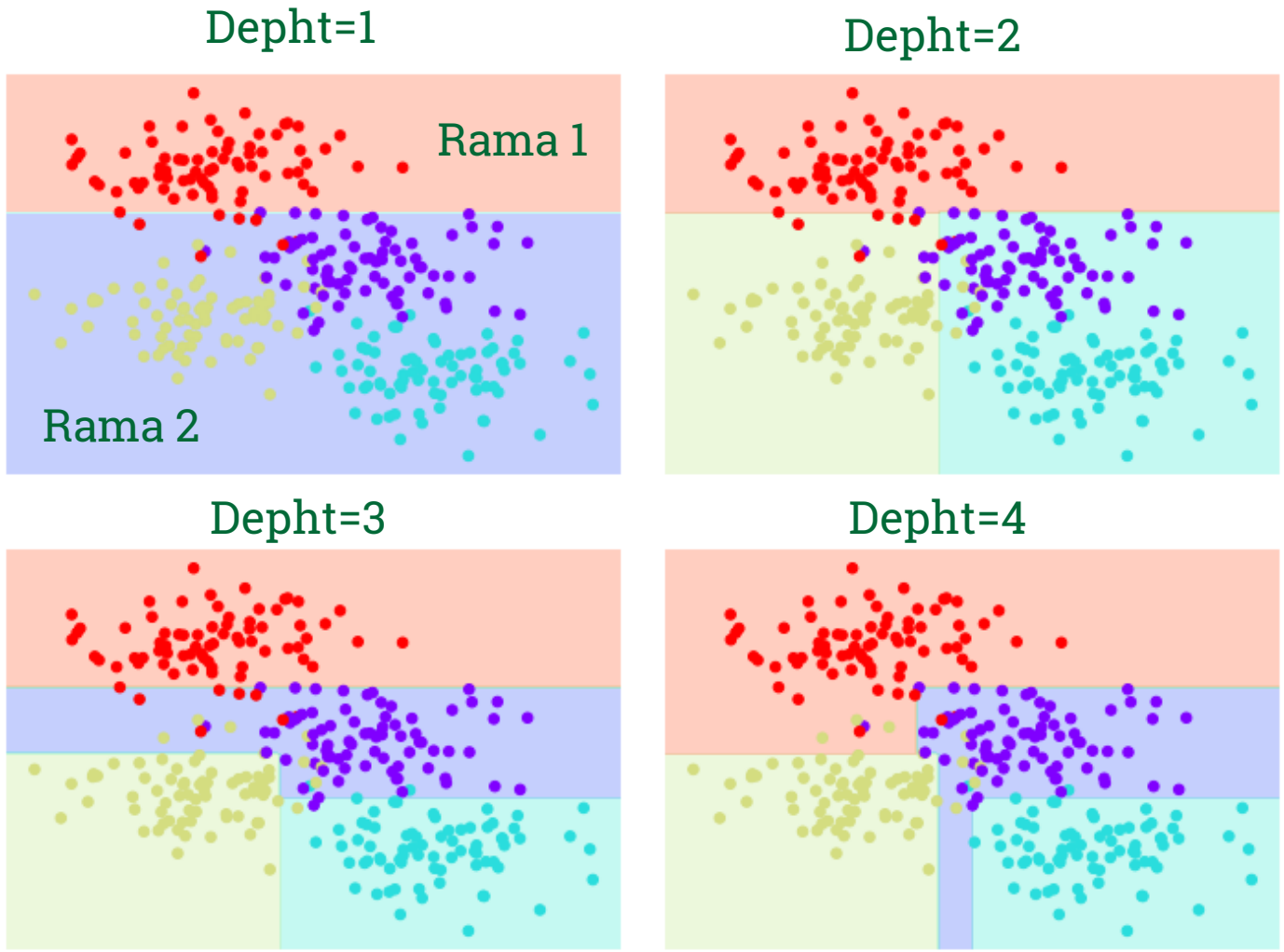
Decision Trees

UdeA

Entonces cada “pregunta” será una línea recta que parta los grupos en dos. Esto en cada nodo. La profundidad del árbol son las veces que se hace estas preguntas.

Las ramas se subdividen el función de la necesidad.

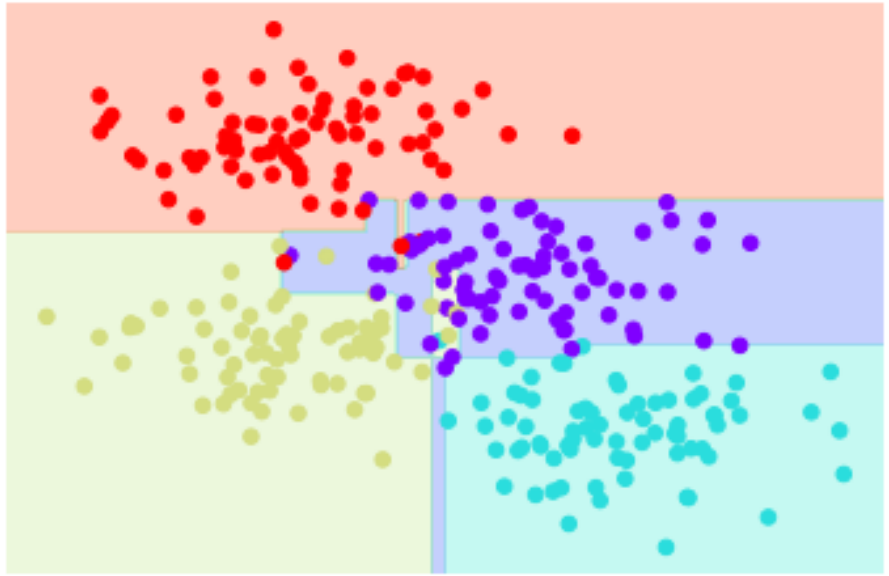
.



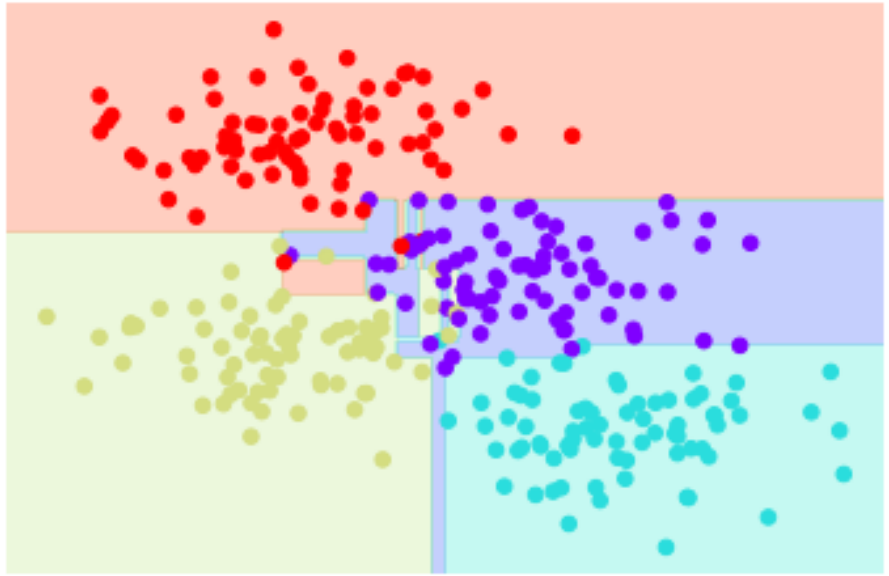
UdeA

Pero a medida que aumenta la profundidad, se generan zonas de clasificación muy particulares, que evidentemente no responden a la distribución de los datos, y esto es un claro sobreajuste.

Depht=6

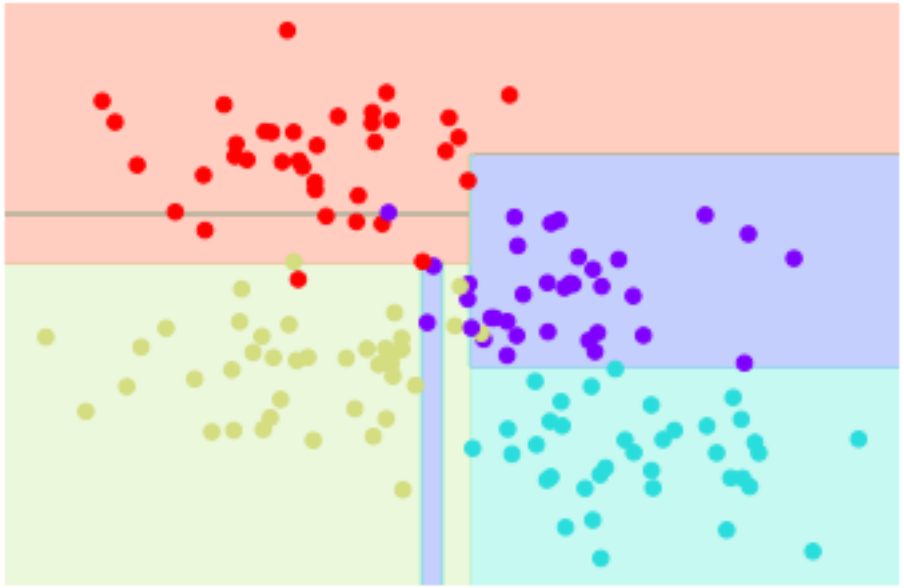


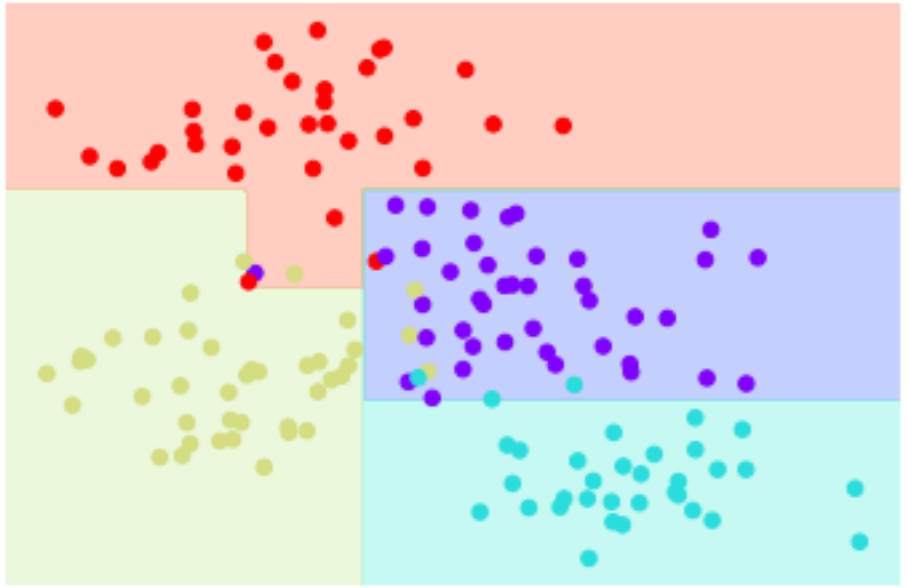
Depht=10



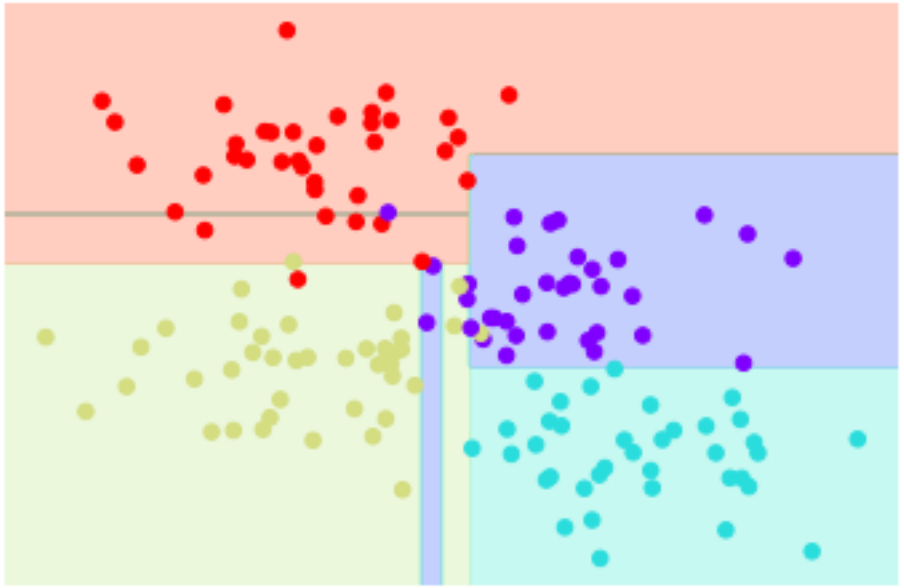
UdeA

De hecho, es muy fácil subreajustar un modelo de arboles de decisión, con lo que se termina ajustando particularidades de la muestra y no de la distribución original de la que provienen.
Otra forma de evidenciar el sobre ajuste es por ejemplo, separar los datos en dos subgrupos aleatorios y ver el ajuste del modelo en cada uno.



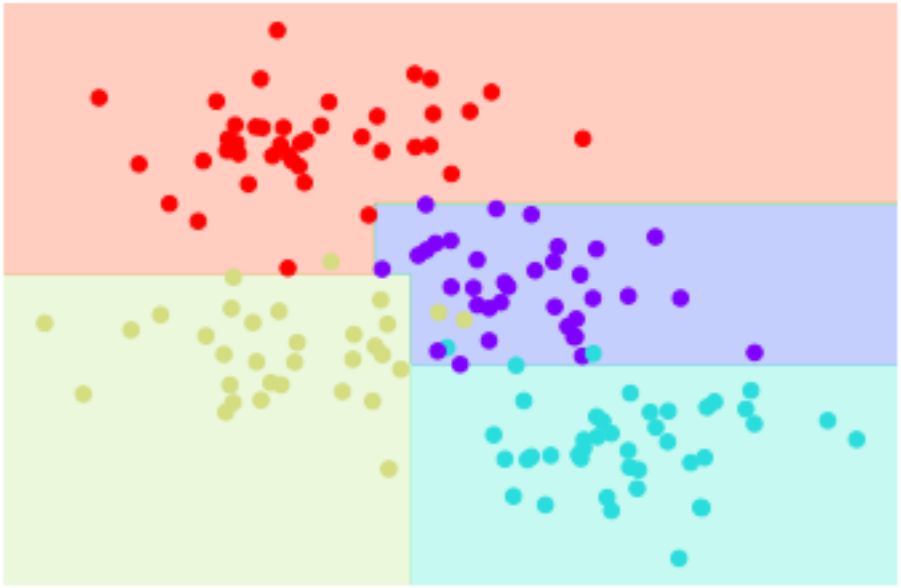


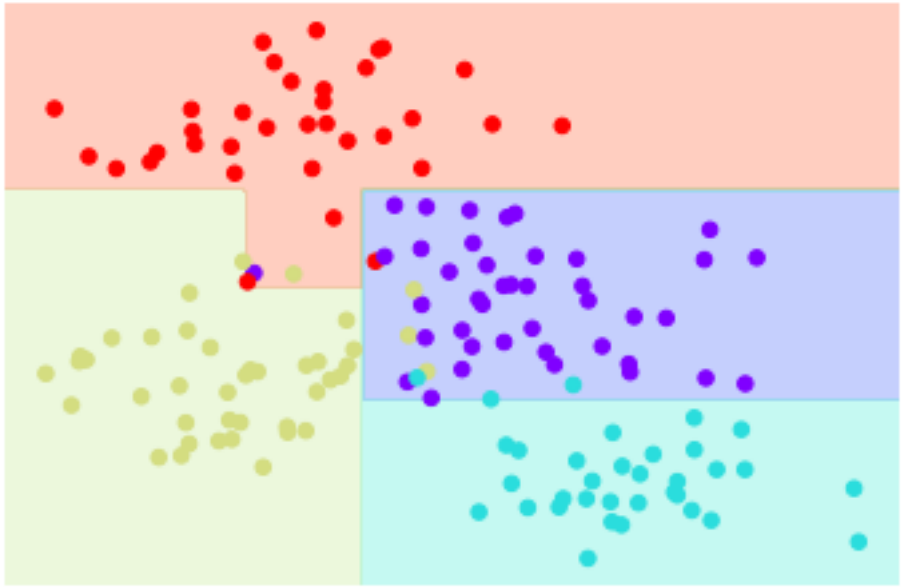
Ahora si se particionan los datos más de 2 veces, podríamos encontrar un modelo promedio mejorado.



UdeA

Entonces el efecto del sobreajuste se puede mejorar combinando varios estimadores sobreajustados.





Random Forest

UdeA

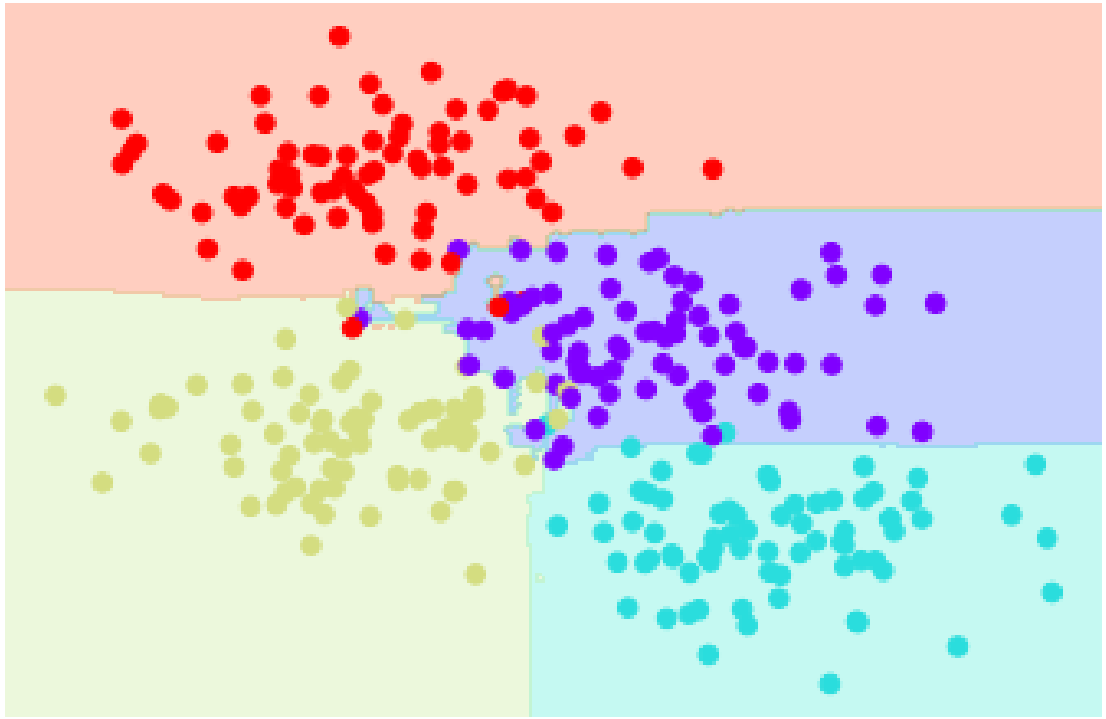
Entonces si generamos muchos arboles para combinarlos, ya tendríamos un **bosque.** Y estos arboles se generan con particiones aleatorias de los datos, entonces de ahí el nombre del método **Random Forest.**

Entonces se usa una técnica de muestreo llamada *bagging,* que consiste en tomar muestras aleatorias de cierto tamaño, con reemplazo, es decir que en cualquier muestreo, todos los datos tiene posibilidad de ser elegidos, no se retiran del sorteo los que previamente se sacaron.

Adicionalmente, el algoritmo *Random Forest*, submuestrea las variables de forma aleatoria

UdeA

Entonces al usar el algoritmo *Random Forest*, se hace un *bagging (Bootstrap)* del set de datos de entrenamiento, con lo que se generan múltiples arboles, que se promedian para obtener un único árbol.



Muestras del 80%

Random Forest Regression

UdeA

En lo visto hemos considerado al algoritmo *Random Forest*, como un método de clasificación, cuando las etiquetas son una variable categórica. Pero también puede ser usado para el caso donde las etiquetas sean una variable continua, en cuyo caso sería un método de regresión.

La sintaxis usada en Sklearn es muy simular, pero usamos la clase

RandomForestRegressor()

Veamos

Random Forest

UdeA

En los modelos *Random Forest,* se puede determinar la importancia que tiene cada *feature*, lo que es útil para eliminar variables del modelo.

Este es un atributo del modelo ajustado y se puede saber su importancia a través de:

Modelo.feature_importances_ Lo cual arroja valores cuya suma es 1, es decir, indican algo así como un fracción de importancia.

Los nombre de las features están en:
Modelo.feature_names_in_

Subject: Probability and Statistics

UdeA

Bioengineering

¡Thanks!

Francisco José Campuzano Cardona

Bioengineering. MSc in Engineering