

DATA SCIENCE



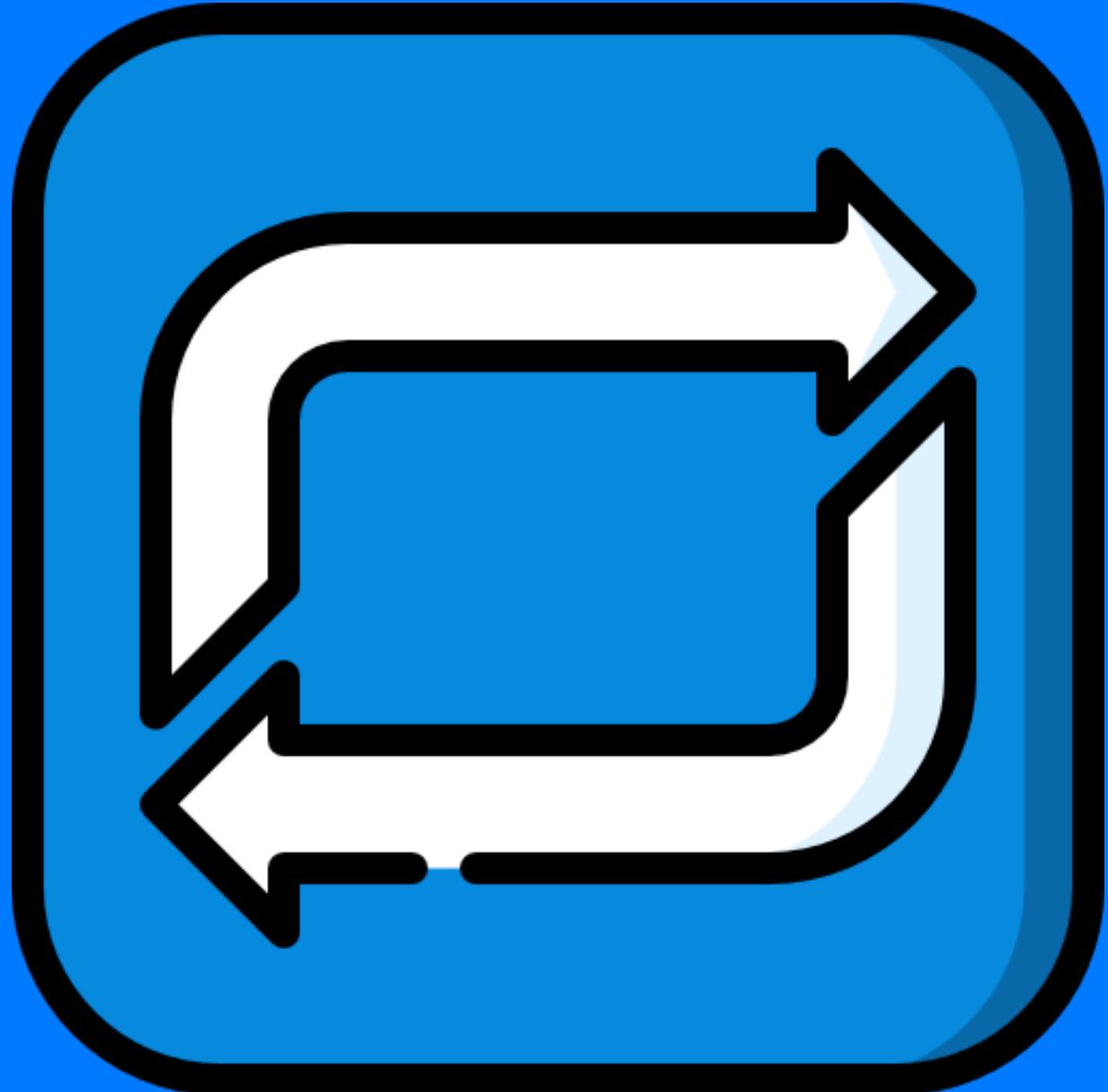
Federico Baiocco
baioccofede@gmail.com
3512075440



Clase 25 - Agenda

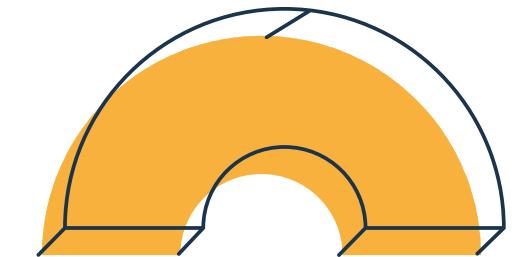
ENSAMBLES - BAGGING

¿Cómo vienen con el curso?



Repaso SVM

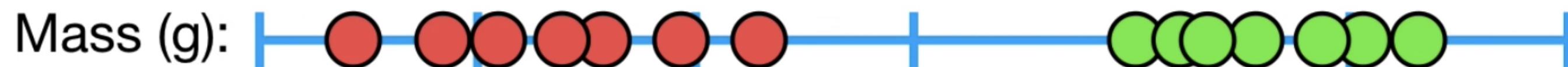
SVM



Queremos clasificar ratones como obeso/no obeso de acuerdo a su masa.

Los puntos verdes representan ratones obesos.

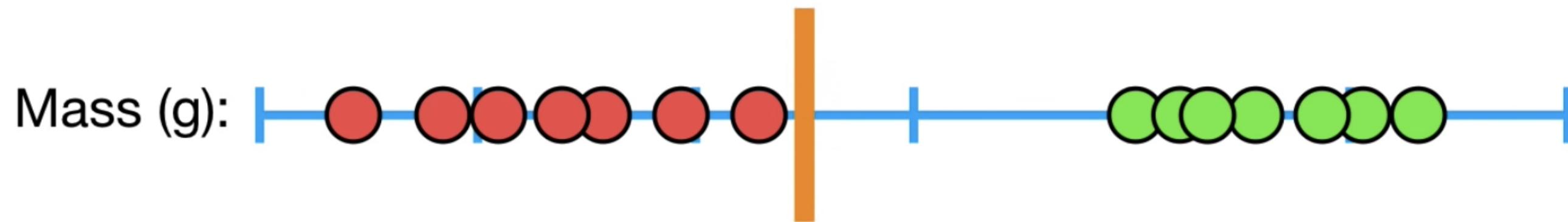
Los puntos rojos representan ratones que no son obesos.



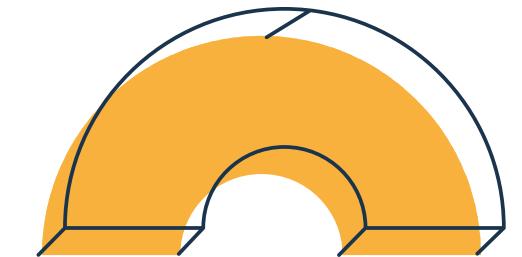
SVM



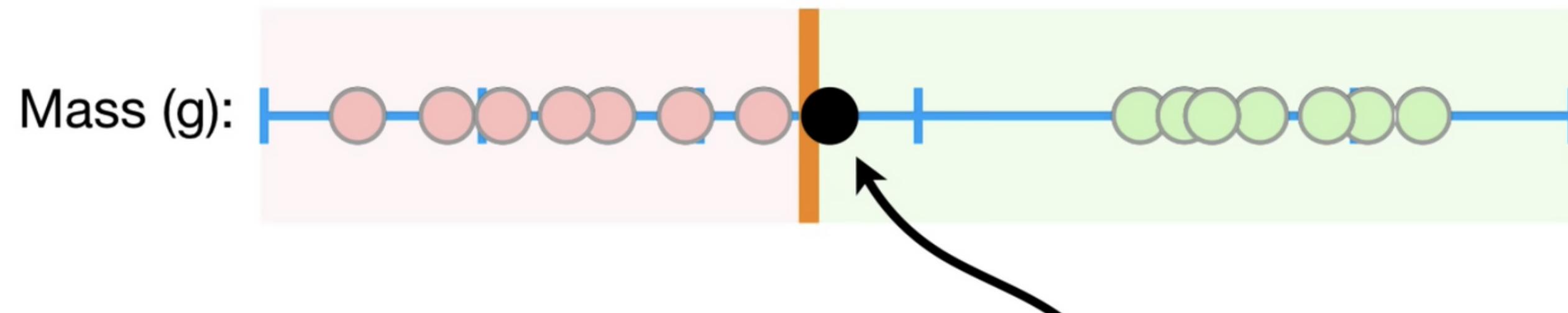
¿Cómo buscábamos el threshold?



SVM



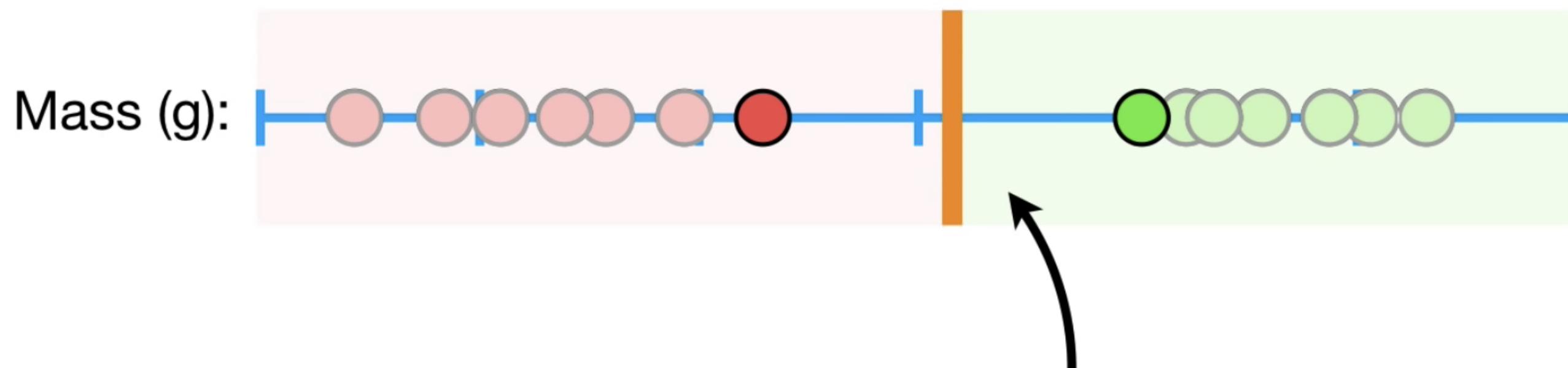
Si nos entraba una observación como esta, la clasificamos como verde cuando en realidad no tenía mucho sentido



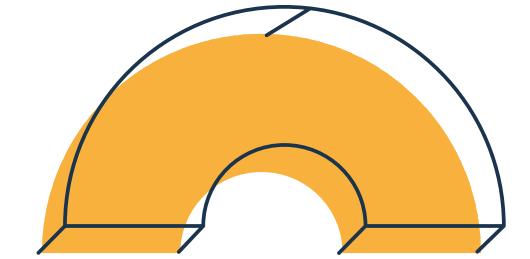
SVM



Entonces tomamos los puntos frontera de cada grupo



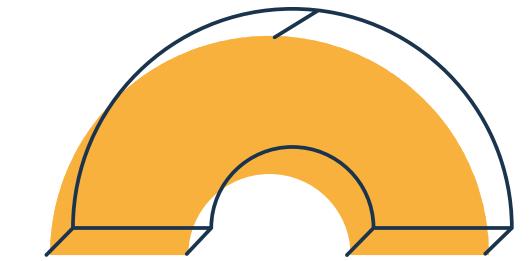
SVM



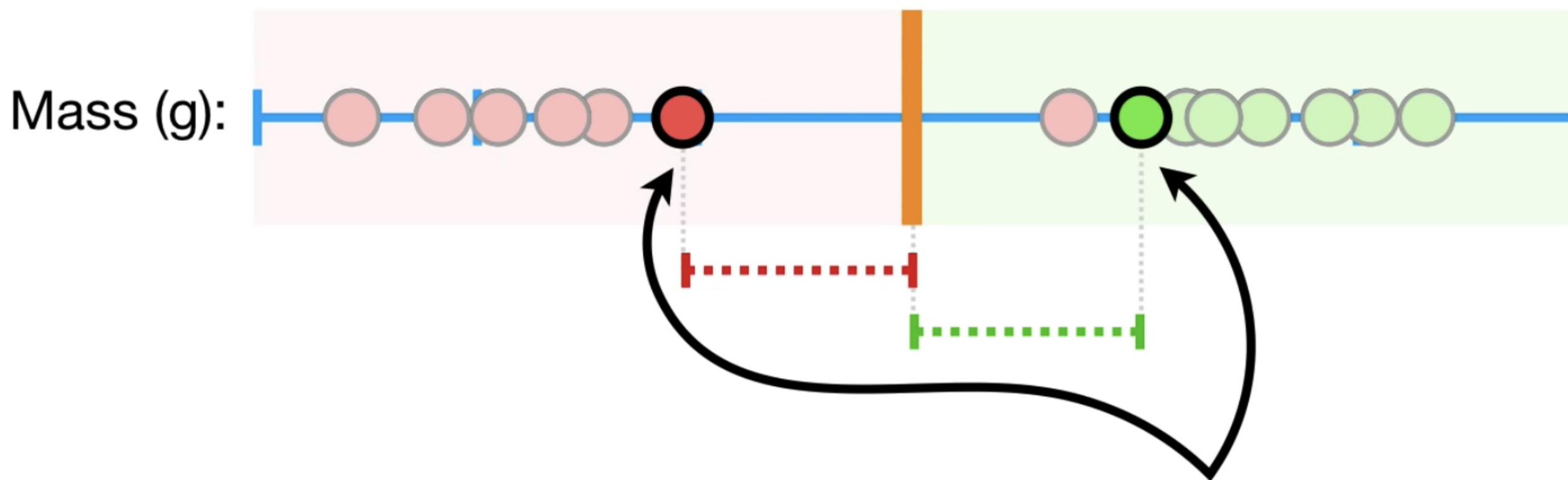
Pero cuando llega un outlier ...



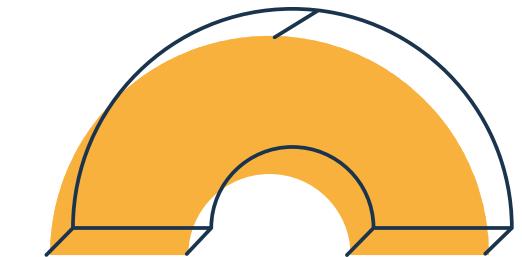
SVM



Entonces permitimos algo de error

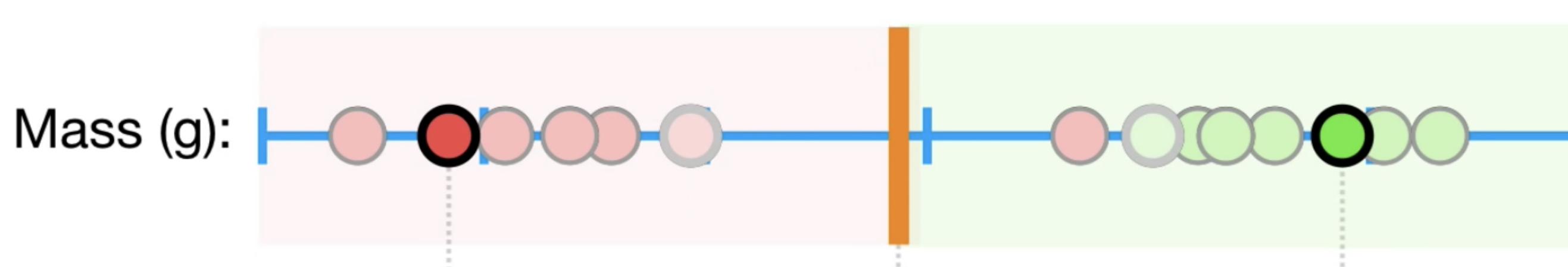


SVM

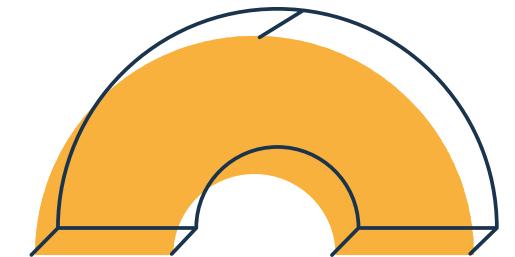


Para decidir que tanto error permitimos, utilizamos el hiperparámetro C.

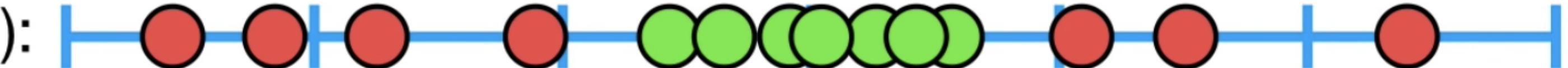
Mientras mayor sea C, menos errores estamos permitiendo sobre los datos de entrenamiento.



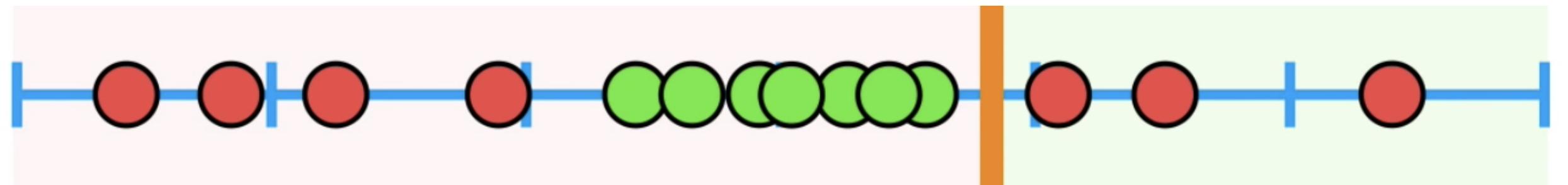
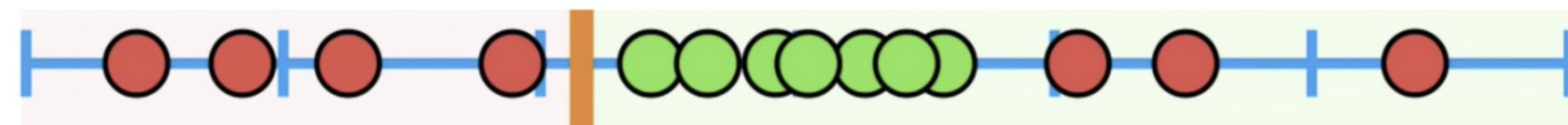
SVM



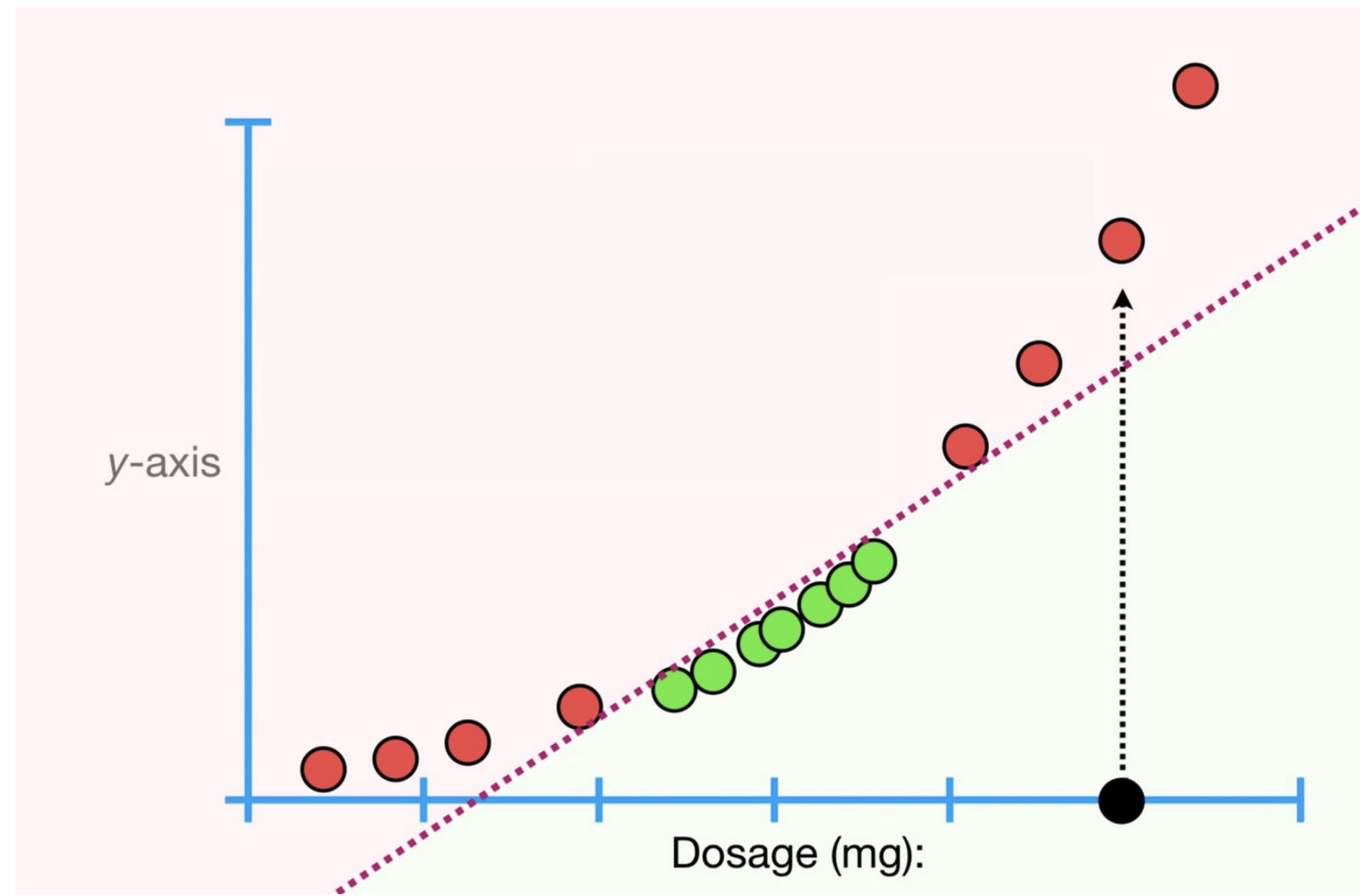
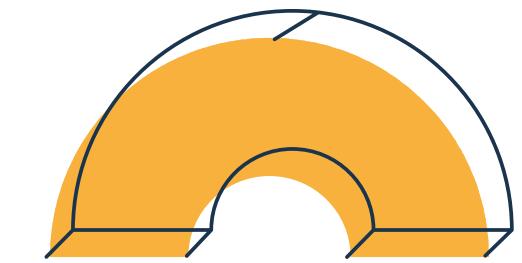
¿Y si tenemos datos que no son linealmente separables?

Dosage (mg): 

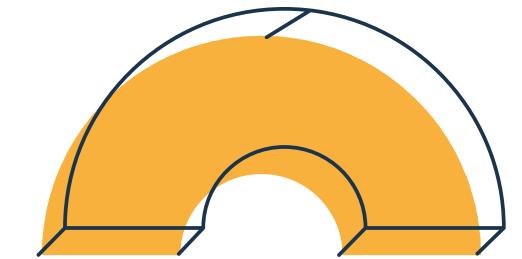
No hay un threshold que sirva



SVM

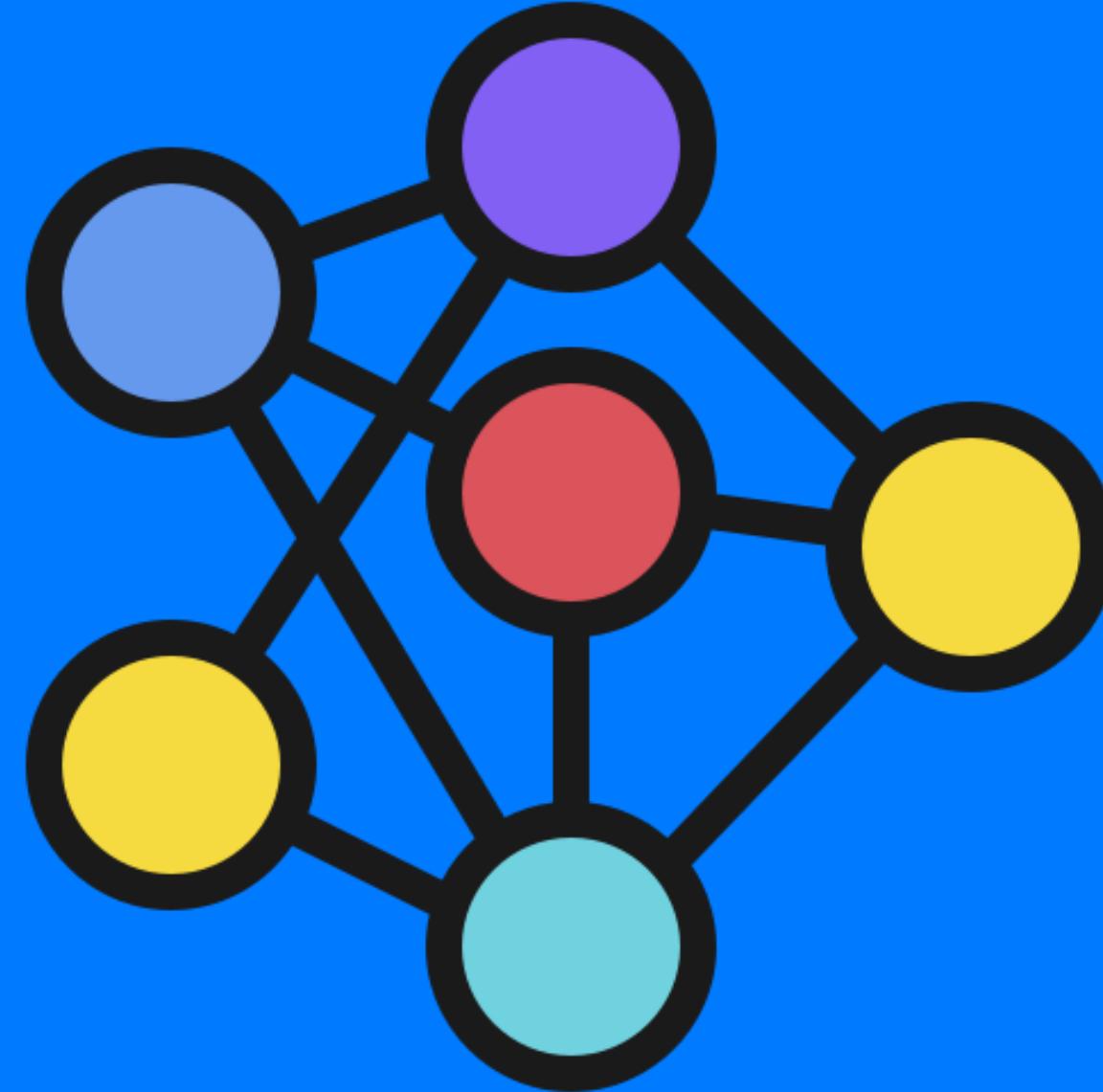


SVM



Entonces:

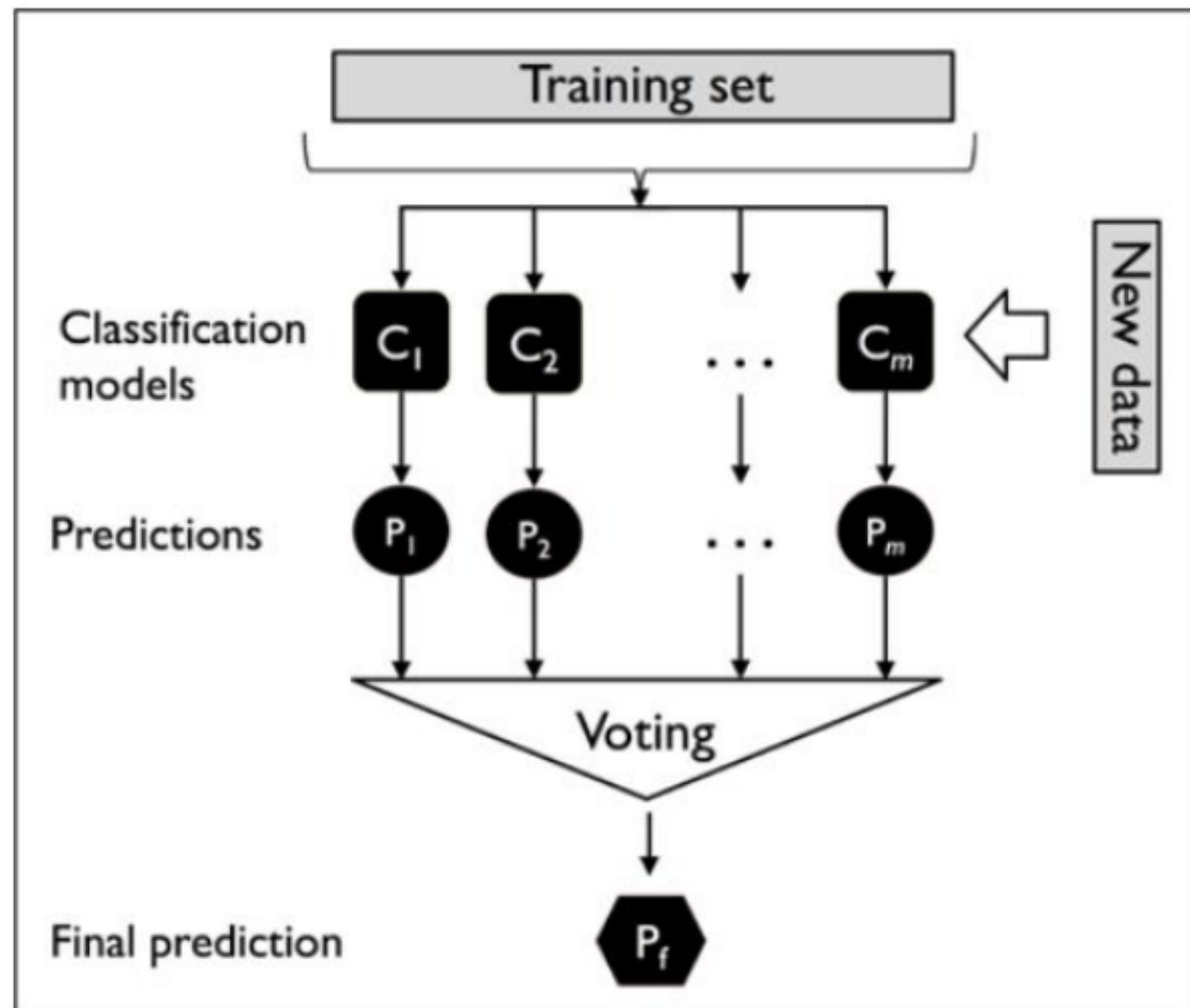
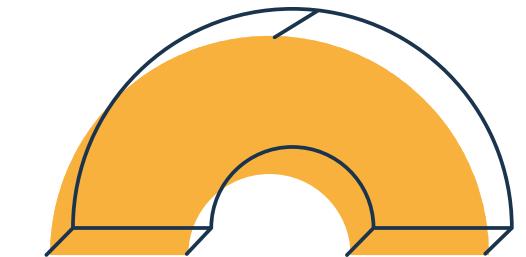
- Para datos linealmente separables: Las support vector machines funcionan muy bien
- Para datos casi linealmente separables: Las support vector machines funcionan bien eligiendo un buen valor de C.
- Para datos que no son linealmente separables: Podemos proyectar los datos a un espacio en el que sean perfectamente/casi linealmente separables.



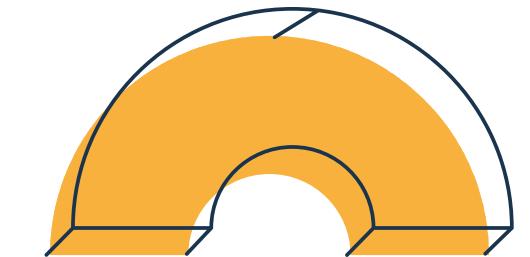
Ensambles

La idea detrás de los ensambles es entrenar muchos modelos distintos y hacerlos "votar"

Ensambles



Ensamblles



- Si todos los modelos son parecidos entre sí, esto no va a tener mucho sentido.
- Necesitamos modelos diferentes entre sí
- Para que los modelos sean diferentes podemos:
 - Entrenarlos con distintos conjuntos de datos
 - Usar técnicas de modelado diferentes

Ensamblles

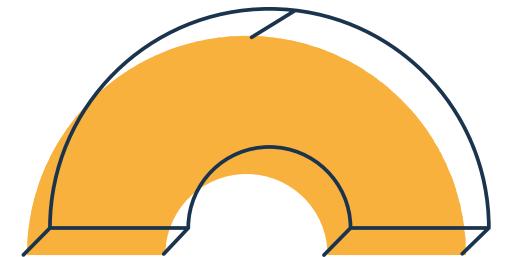


Existen distintas técnicas de ensamble:

- BAGGING
- BOOSTING
- STACKING

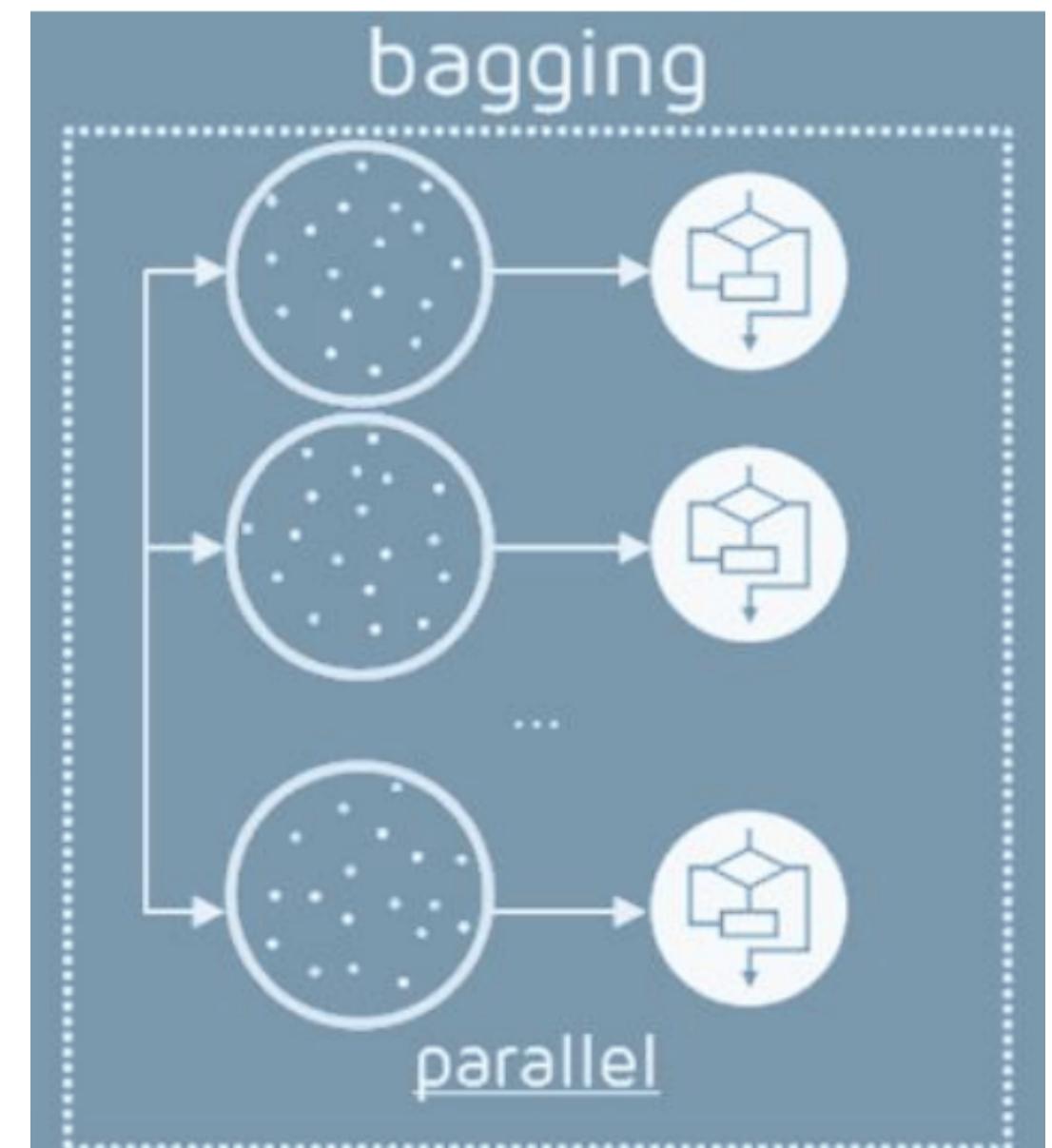
En esta clase vamos a ver BAGGING.

Bagging



Bagging (o Bootstrap Aggregation) :

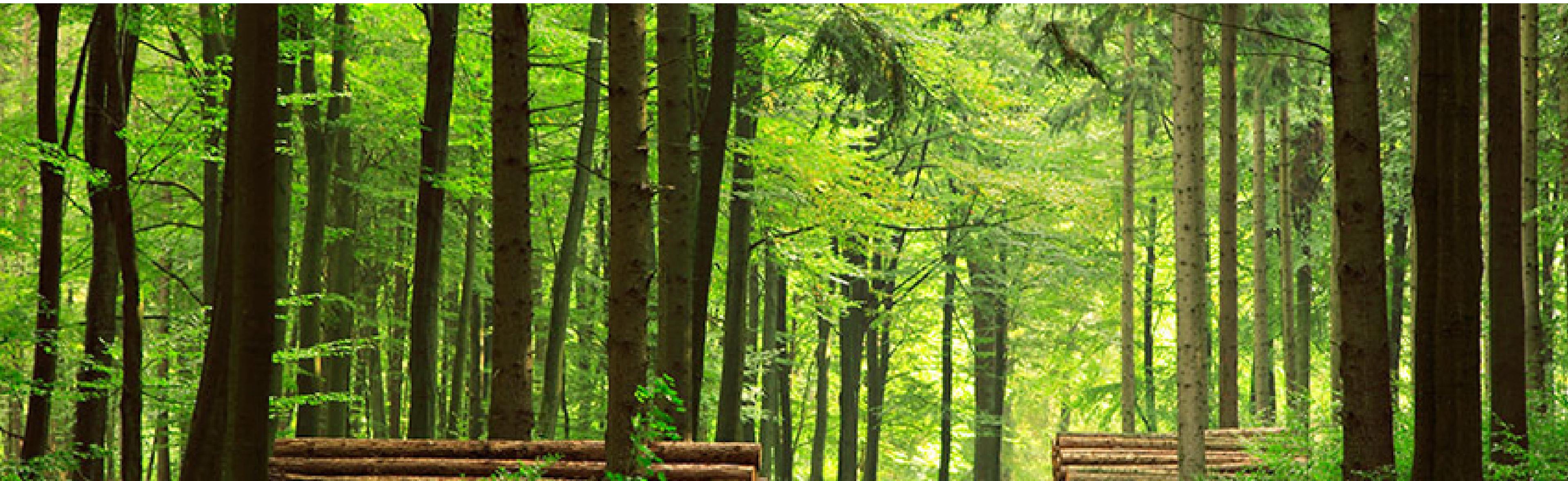
- Dada una muestra de datos, se extraen varias muestras
- Esta selección se realiza de manera aleatoria
- Una vez que forman las muestras, se entranan modelos de manera separada.
- La predicción final se hace por "votación"

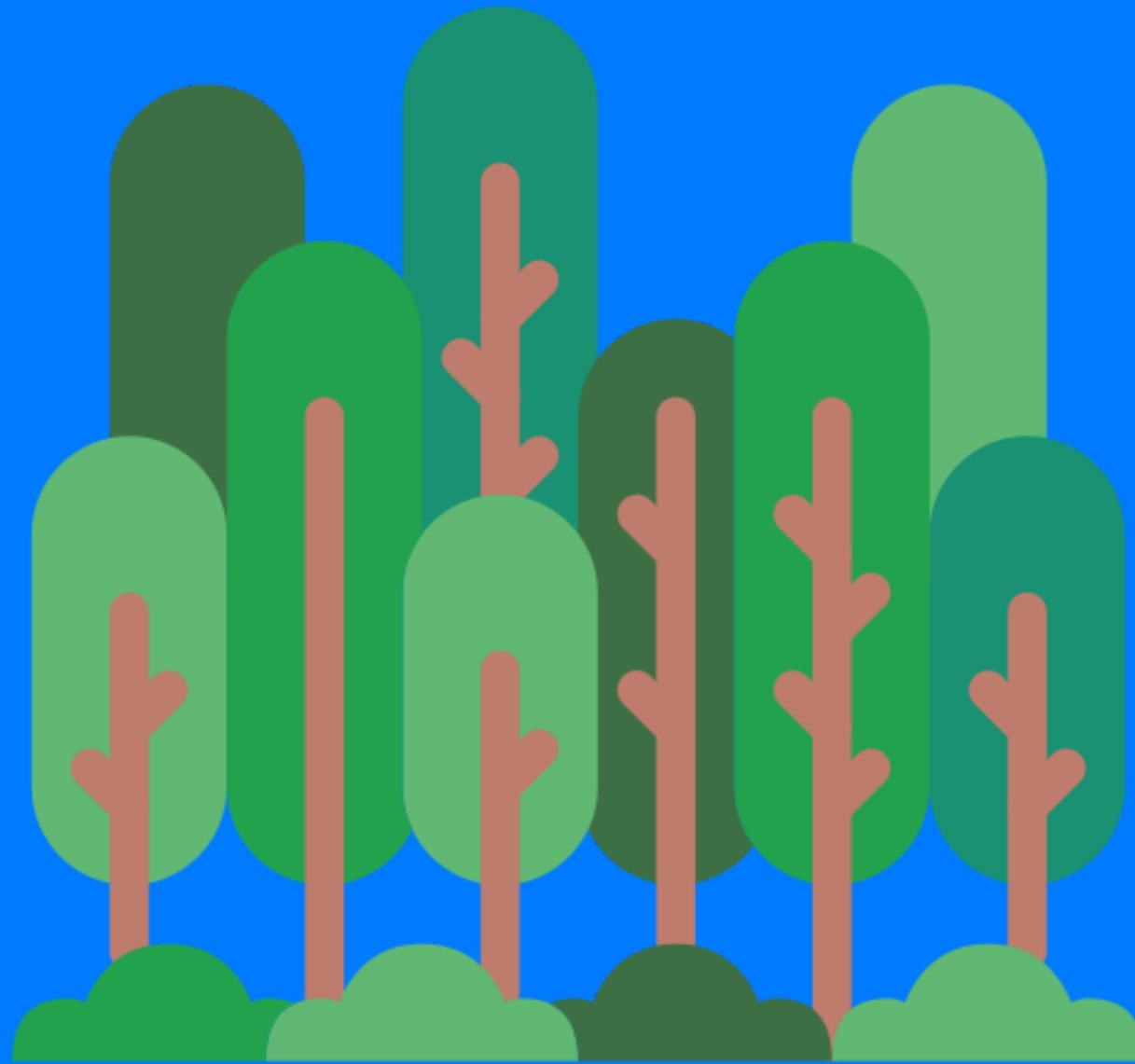


Bagging



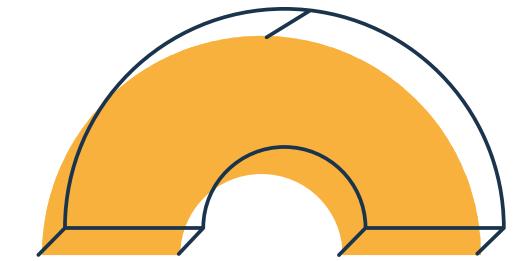
Esta técnica se puede aplicar uniendo cualquier modelo de los que vimos, pero se suele utilizar en árboles para armar [random forests](#)





Random forest

Random forest



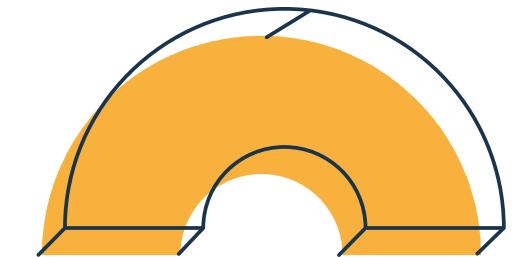
Cuando utilizábamos árboles de decisión, veíamos que a medida que la profundidad del árbol aumentaba, tendíamos a overfittear el modelo.

Los árboles tienden a "memorizar" los datos de entrenamiento.

Acá es donde surge **random forest** como una mejora.

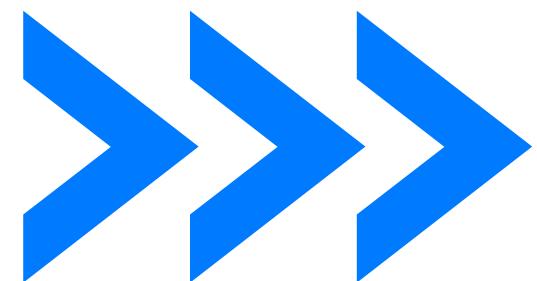
Random forest combina la simplicidad de los árboles, con flexibilidad. Esto resulta en un modelo que funciona mucho mejor a la hora de predecir sobre datos nuevos.

Random forest



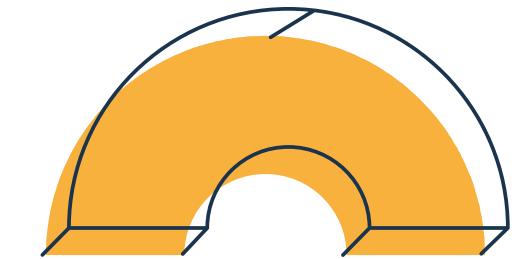
1) Creamos un dataset a partir de bootstrapping: Se seleccionan filas de manera aleatoria. Puede seleccionarse la misma fila más de una vez.

Gender	Age	EstimatedSalary	Purchased
Male	26	32000	0
Male	37	72000	0
Female	49	36000	1
Male	25	33000	0

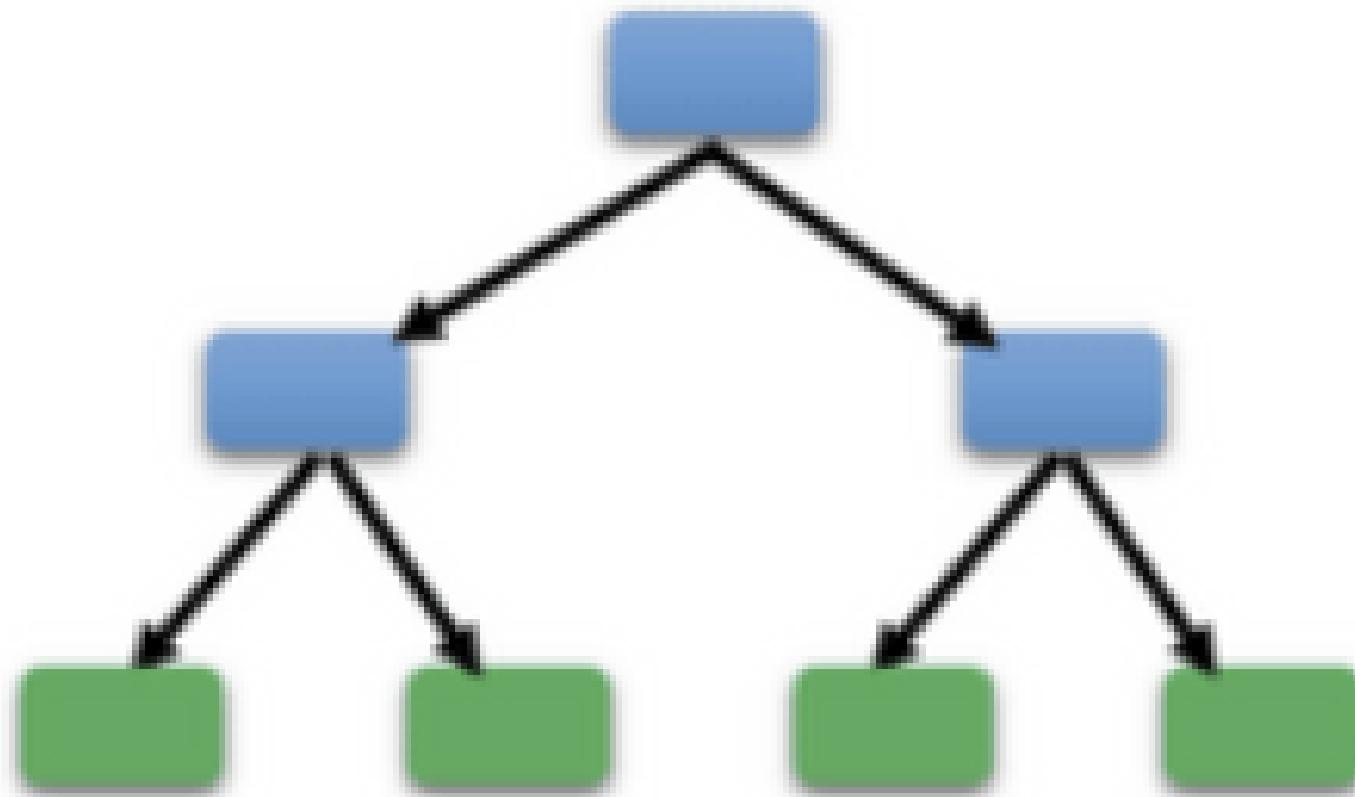


Gender	Age	EstimatedSalary	Purchased
Male	37	72000	0
Male	26	32000	0
Male	25	33000	0
Male	25	33000	0

Random forest



2) Crear un decision tree utilizando el dataset creado mediante bootstrapping, pero utilizando únicamente un subset de features



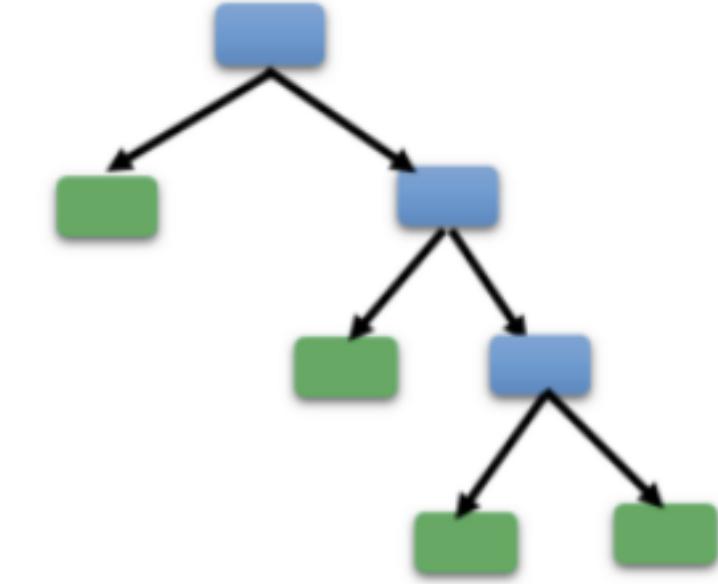
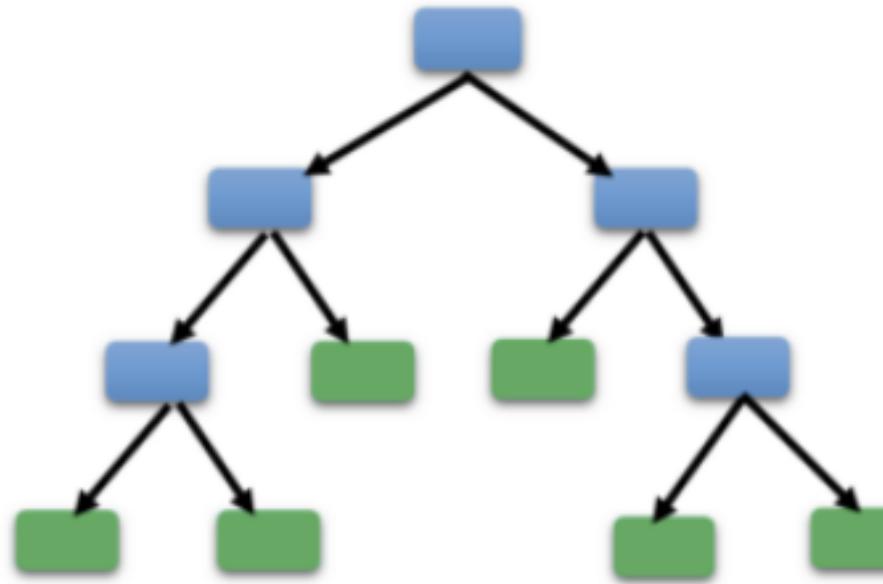
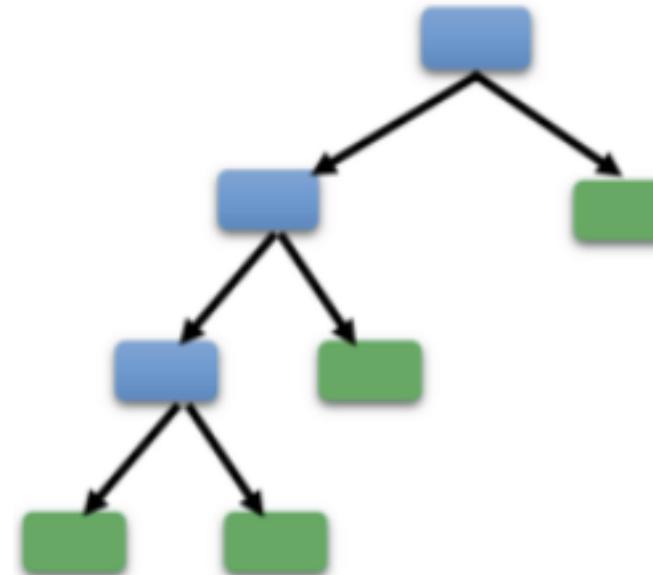
Random forest



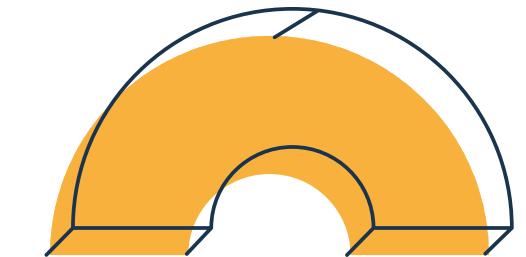
3) Vamos de nuevo al paso 1, creamos un nuevo dataset y luego un nuevo árbol (paso 2).

La idea es que entrenemos muchos (por ej: cientos) de árboles distintos.

¿ En qué van a ser distintos cada uno de los árboles?

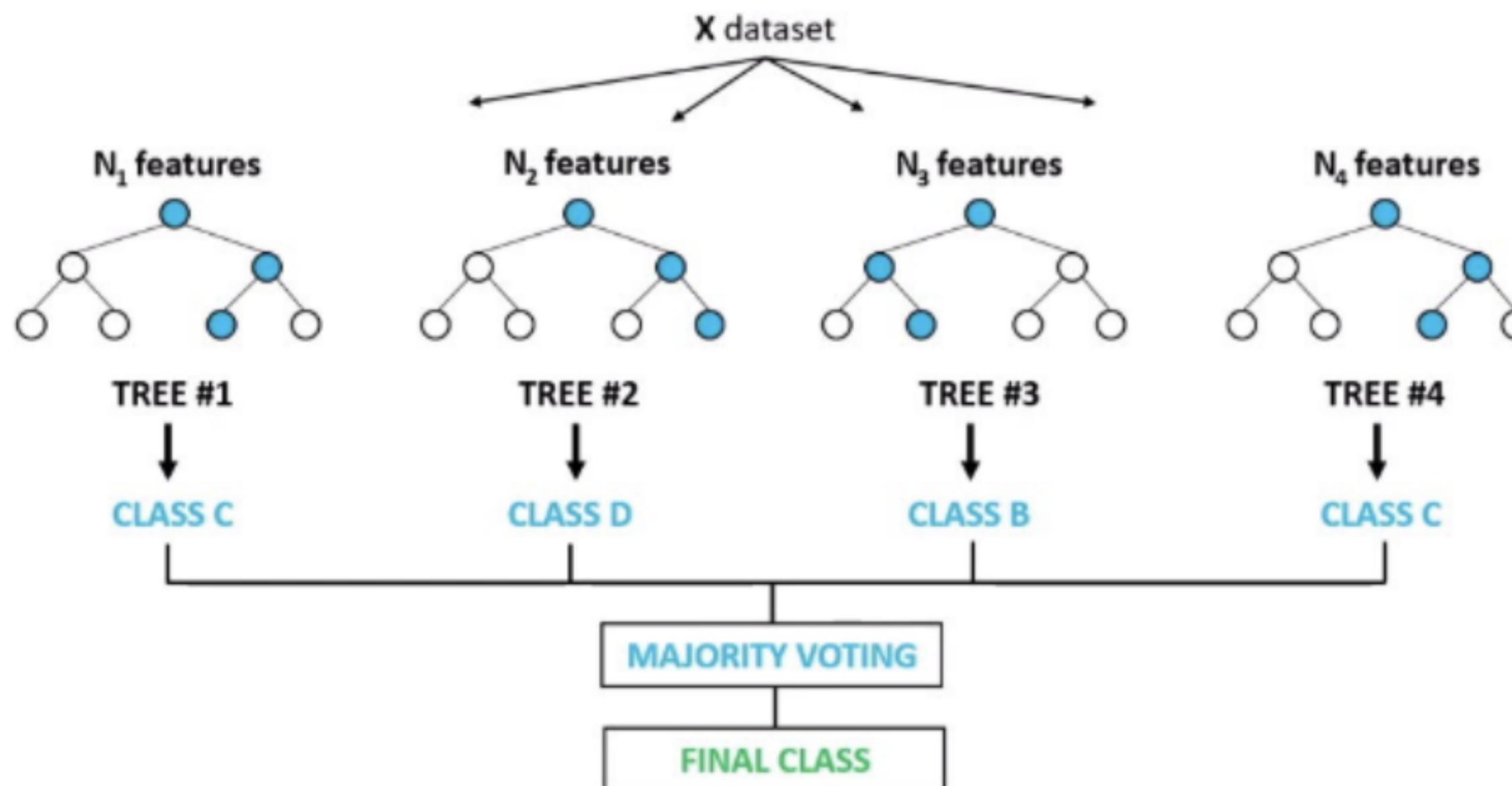


Random forest

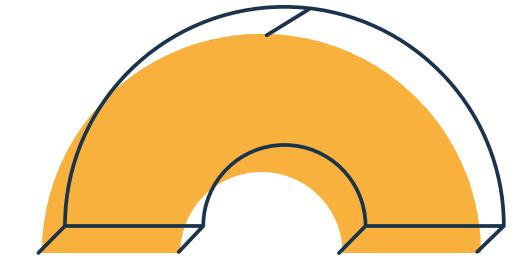


Ahora, para generar predicciones sobre nuevos datos, se utilizan todos los árboles.

Al final de todo, se hace una votación.

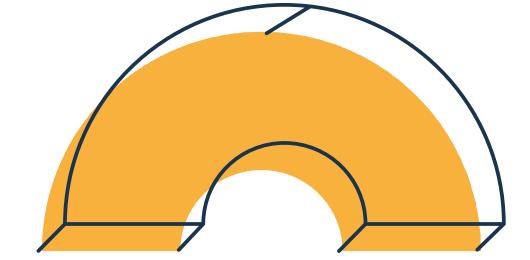


Random forest



BAGGING = Bootstrapping + Aggregate

Random forest



Ventajas:

- Es un modelo robusto. (outliers)
- Luego de entrenar un random forest podemos evaluar la importancia de las distintas features
- Al entrenarse muchos árboles por separado, se puede paralelizar el proceso para que sea más rápido.
- Interpretabilidad (al estar formados por árboles de decisión)



BREAK!

PRÁCTICA