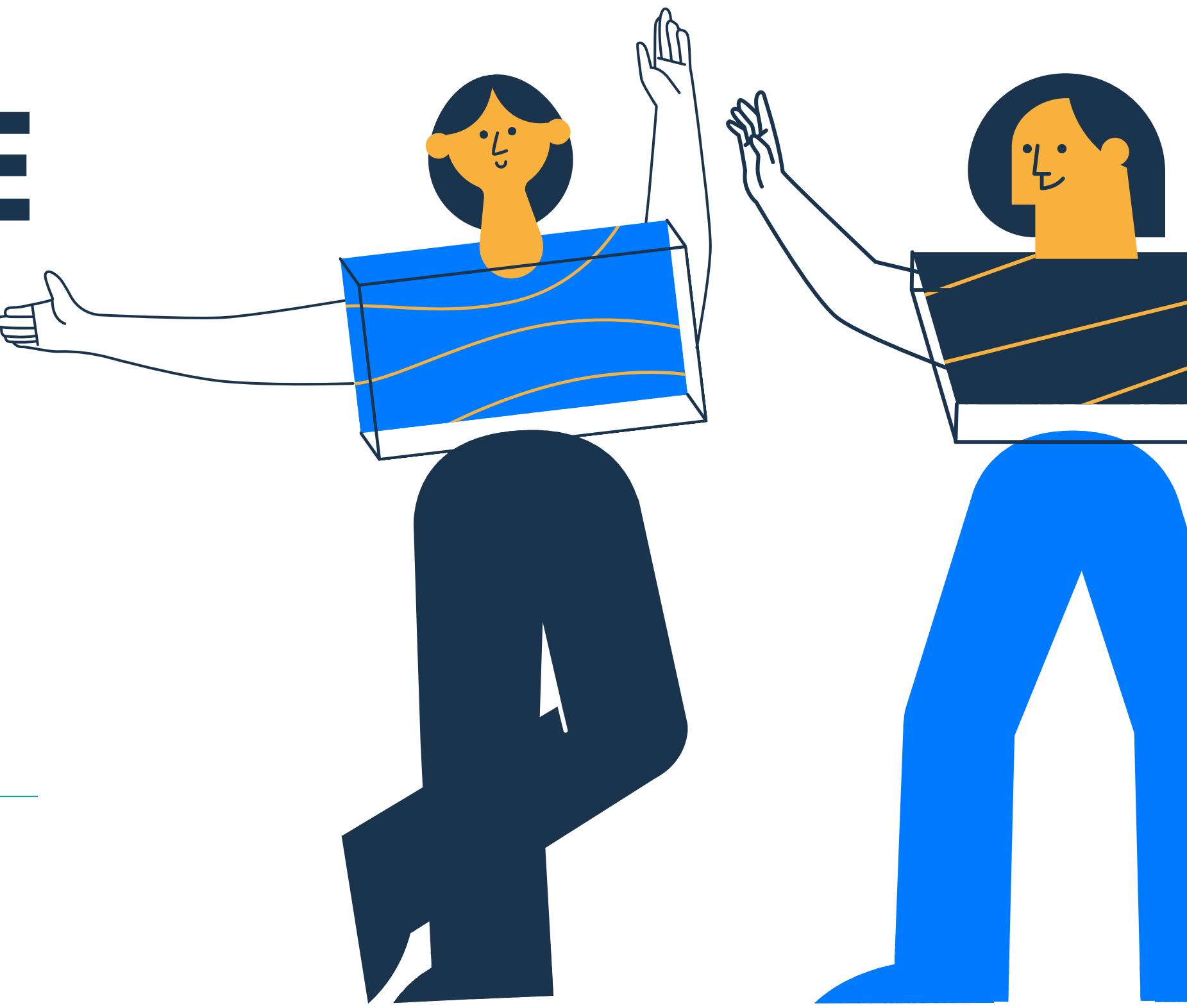


DATA SCIENCE



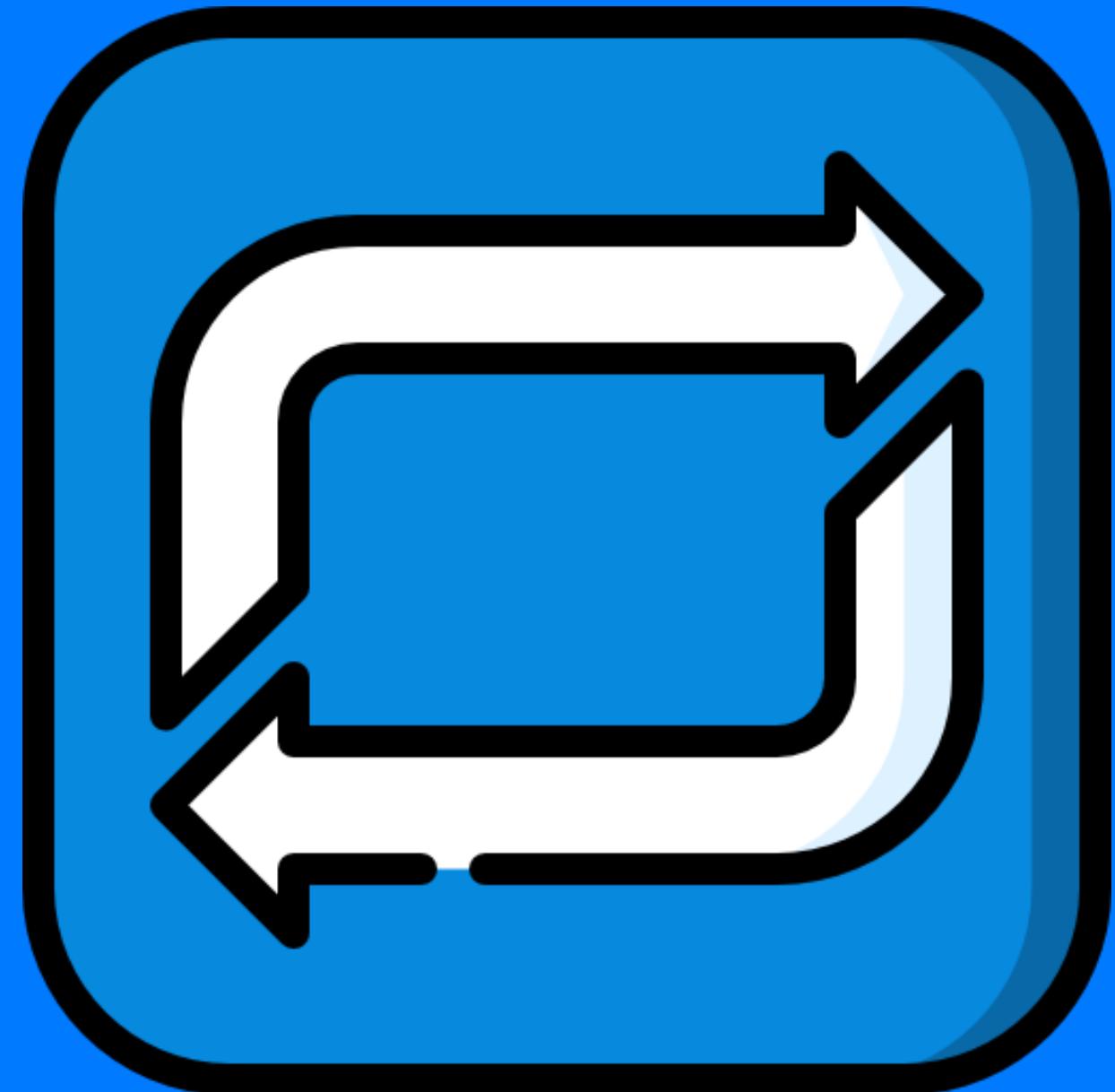
Federico Baiocco
baioccofede@gmail.com
3512075440



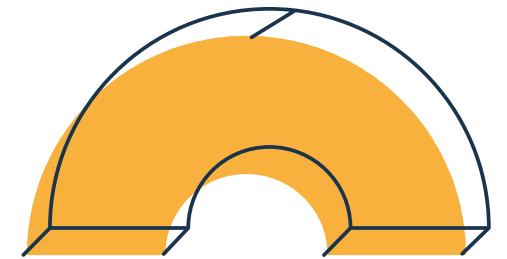
Clase 27 - Agenda

PRESENTACIONES - REPASO

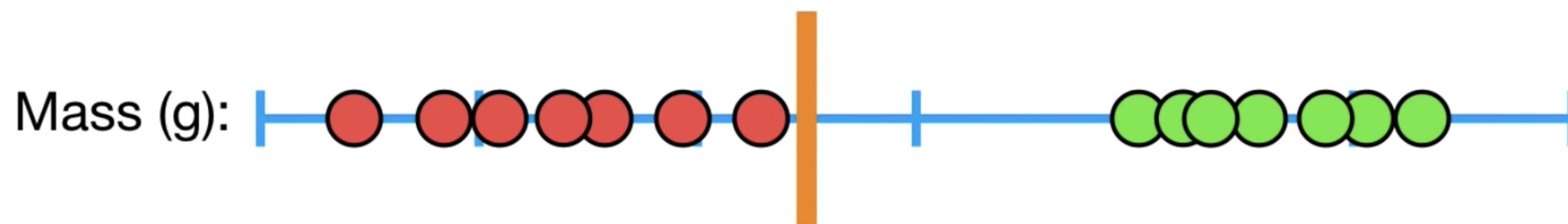
-
- SVM
 - BAGGING
 - BOOSTING
 - STACKING



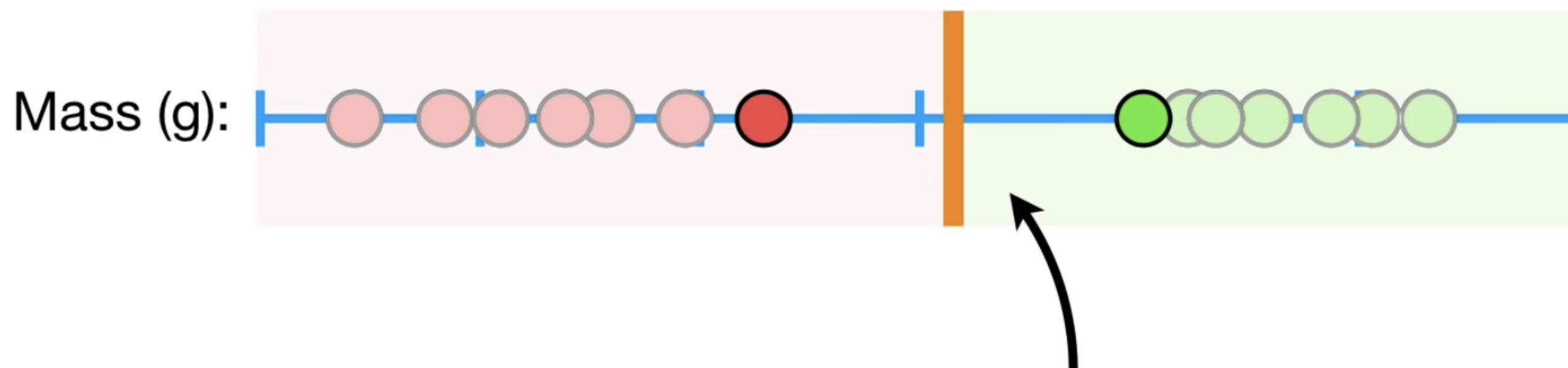
SVM



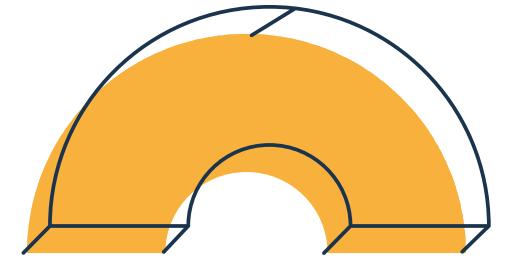
¿Cómo buscábamos el threshold?



Podíamos tomar los puntos frontera de cada grupo



SVC



Pero cuando llega un outlier ...



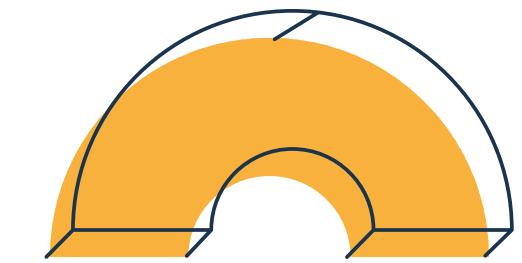
SVC



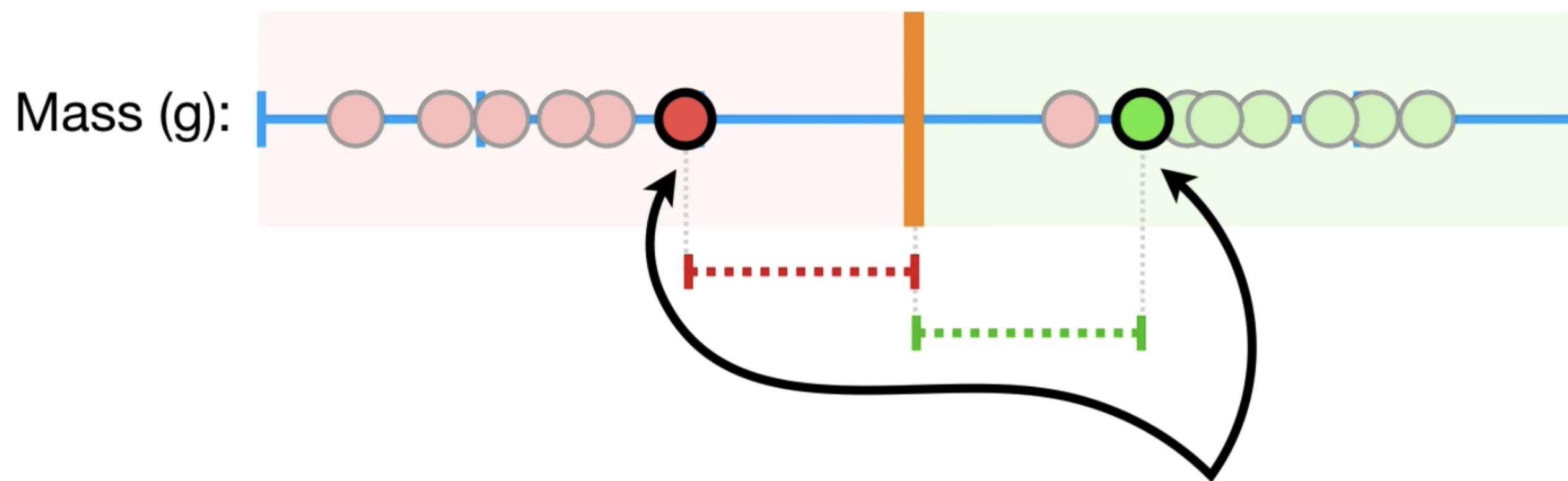
Pero cuando llega un outlier ...



SVC



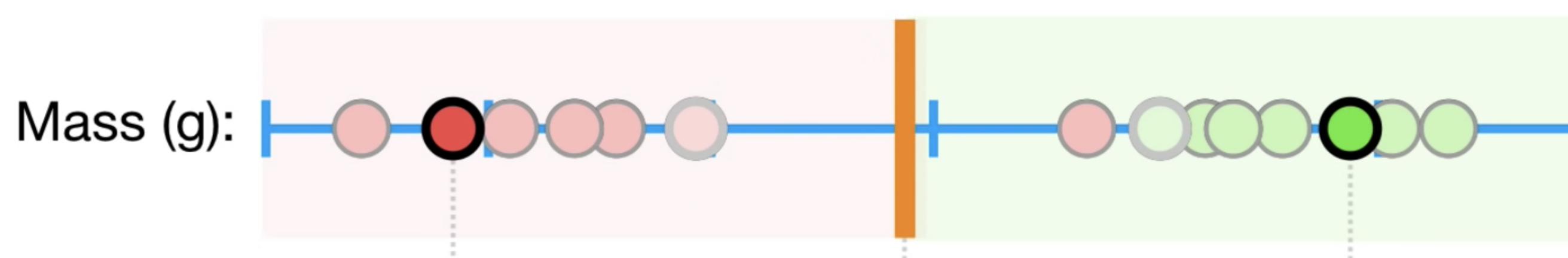
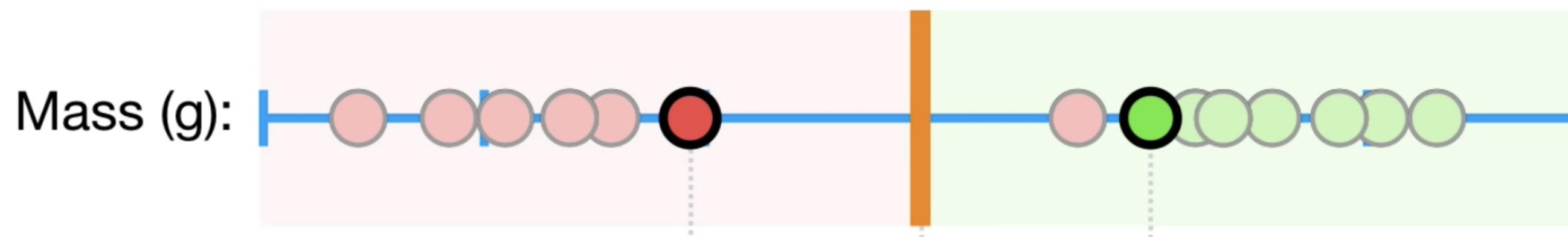
Entonces permitimos algo de error



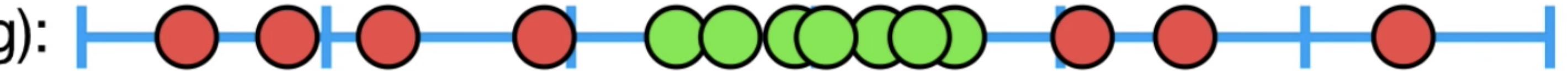
SVM

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

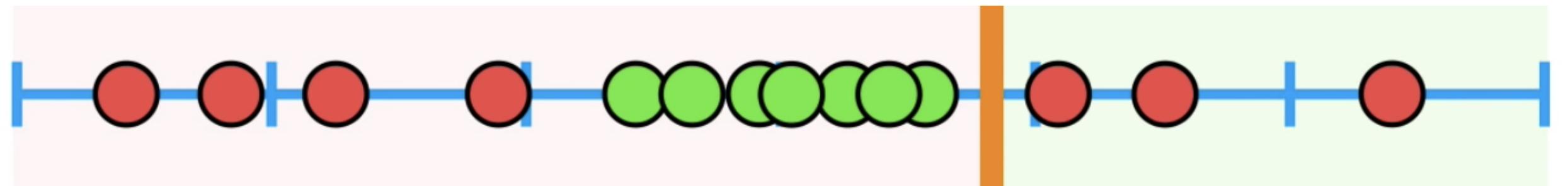
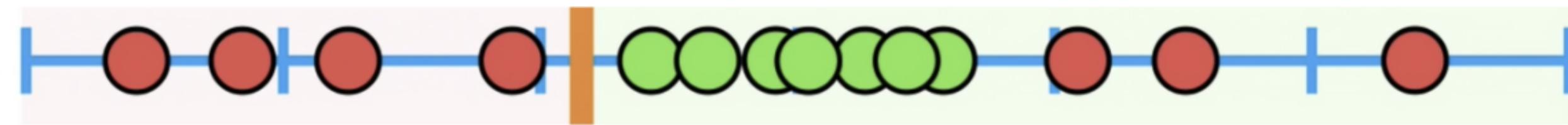
Mientras mayor sea C, menos errores estamos permitiendo sobre los datos de entrenamiento.



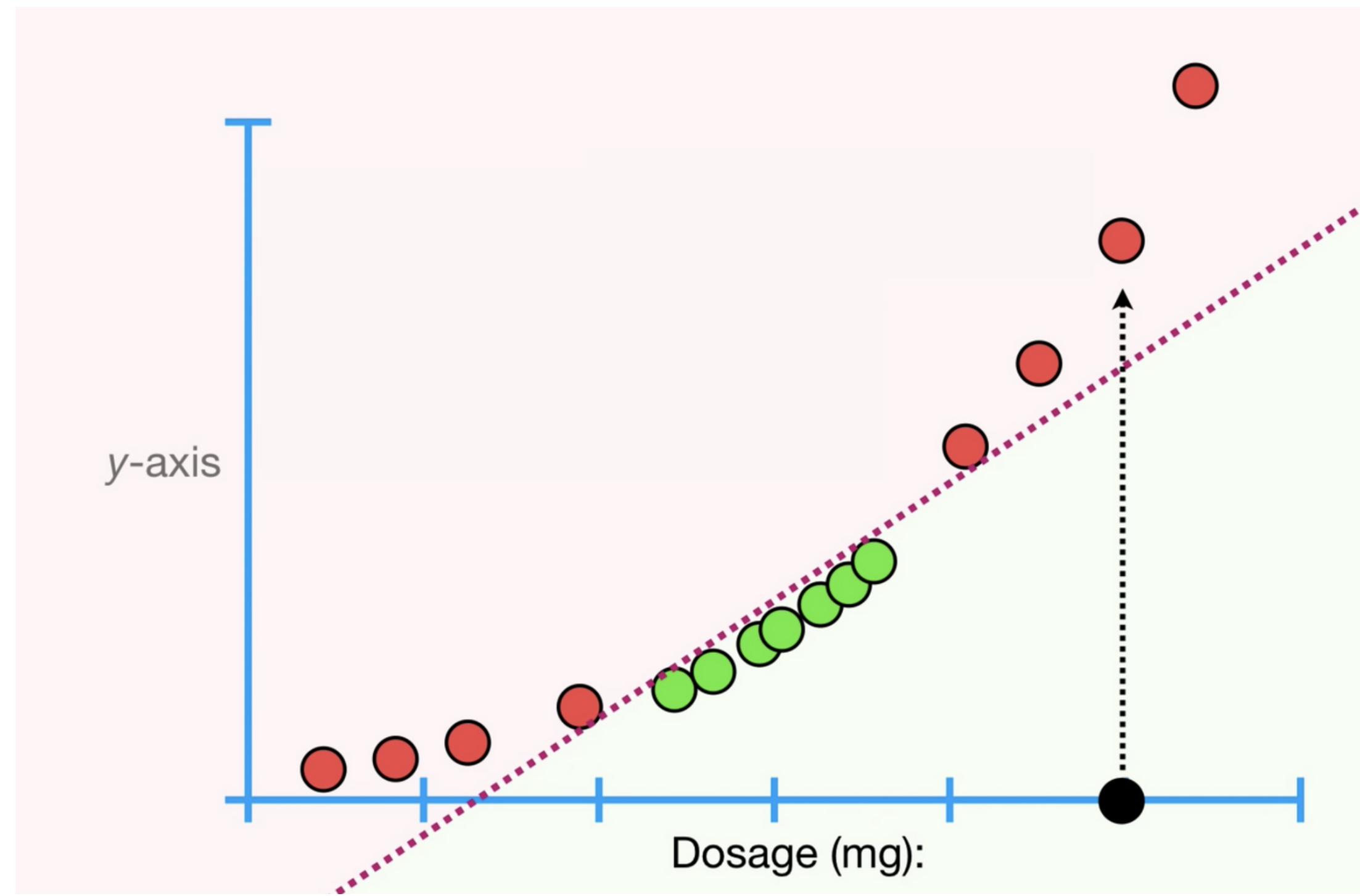
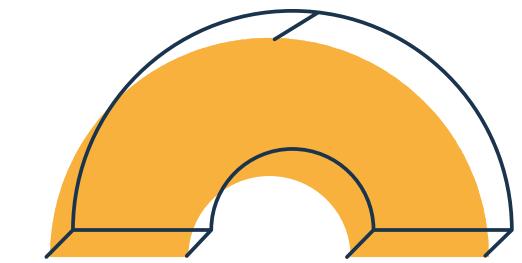
¿Y si tenemos datos que no son linealmente separables?

Dosage (mg): 

No hay un threshold que sirva



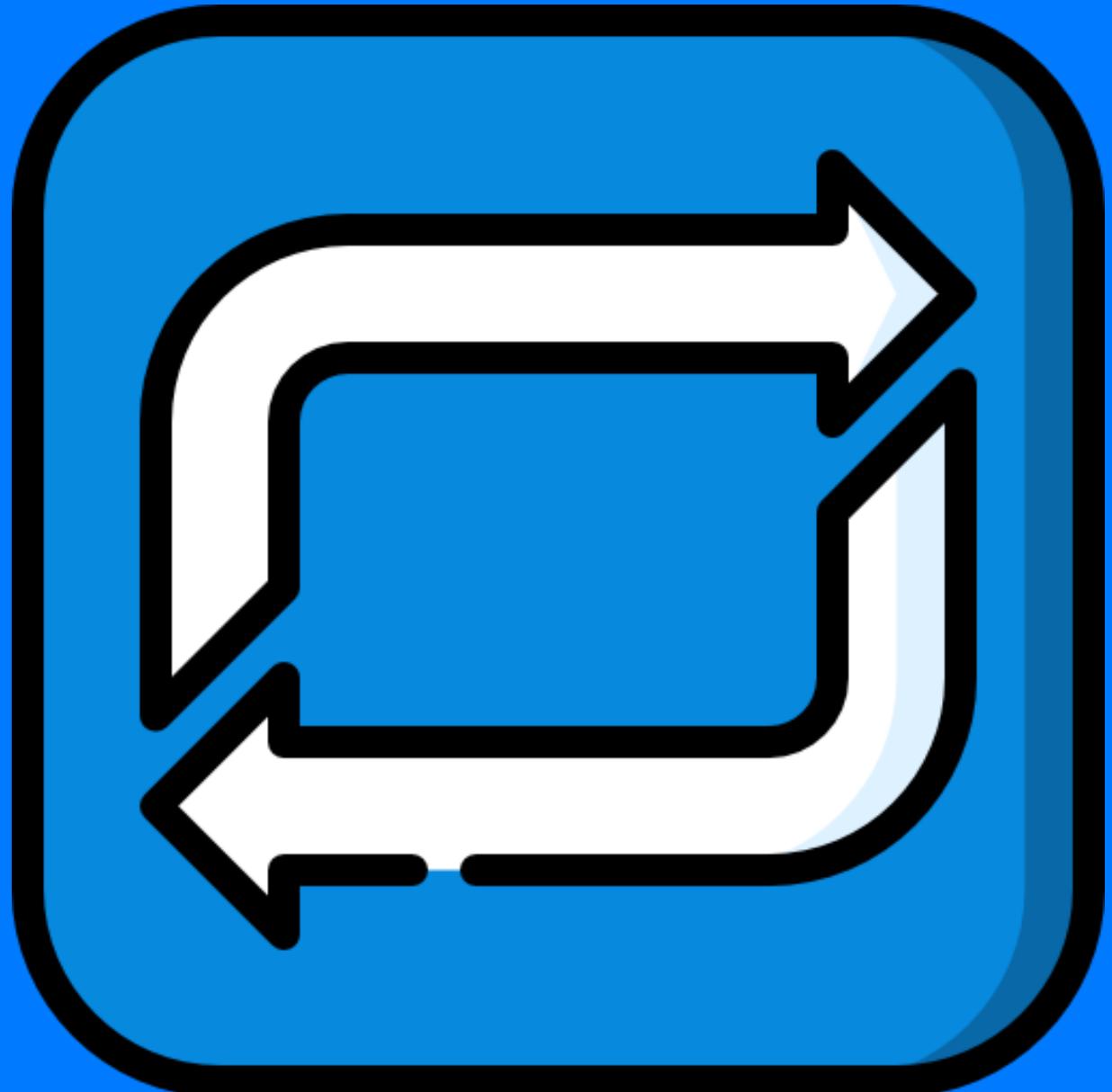
SVM



SVM

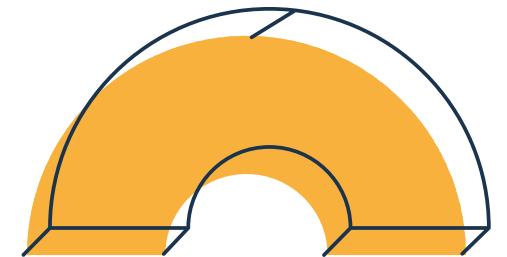
Entonces:

- 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.



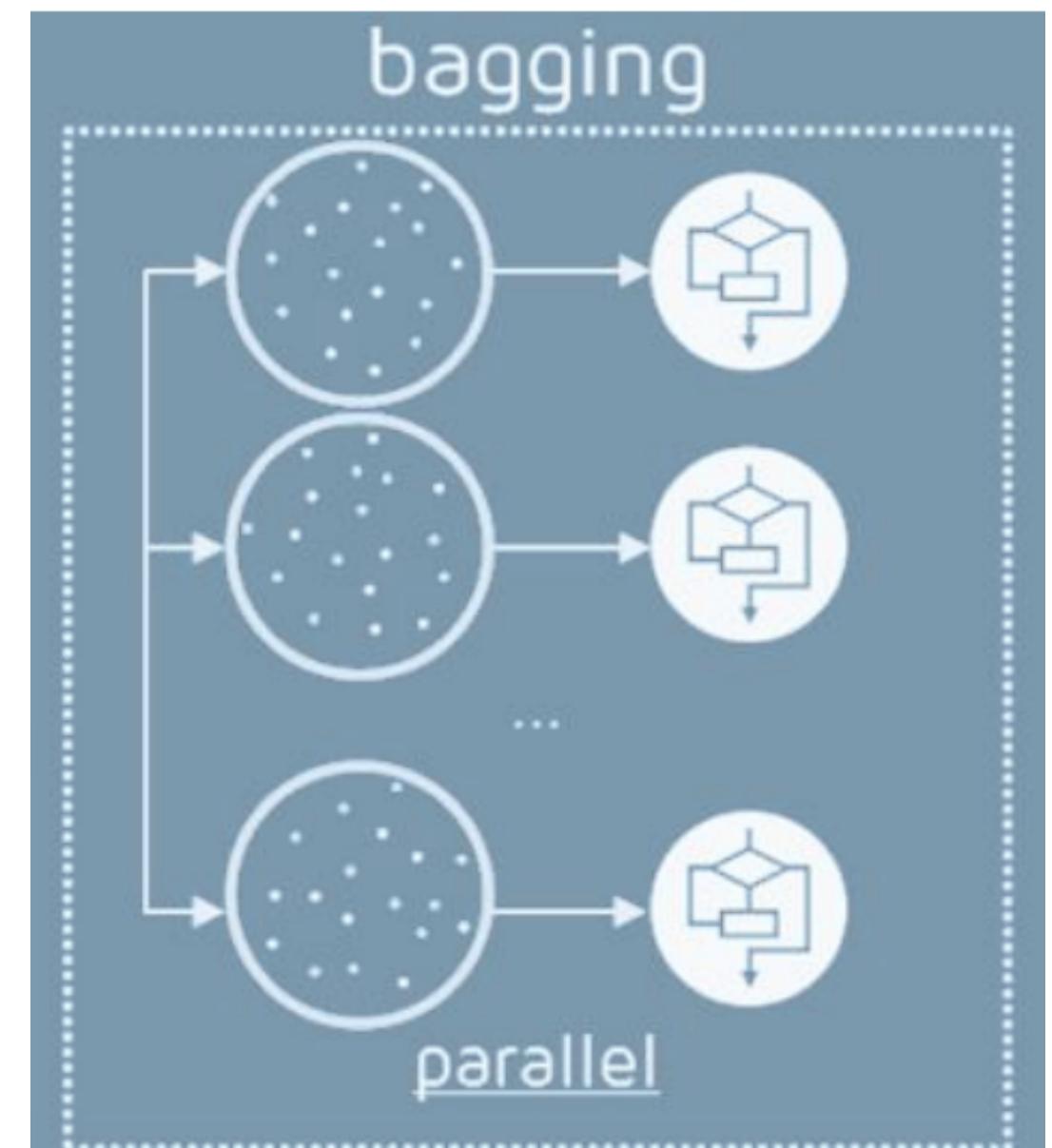
BAGGING

Bagging

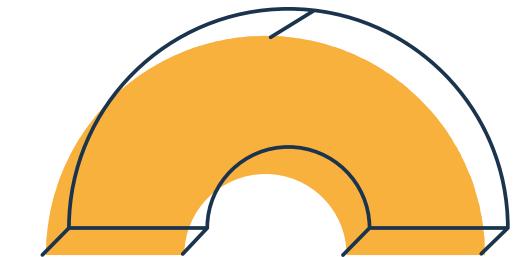


Bagging (o Bootstrap Aggregation) :

- 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"

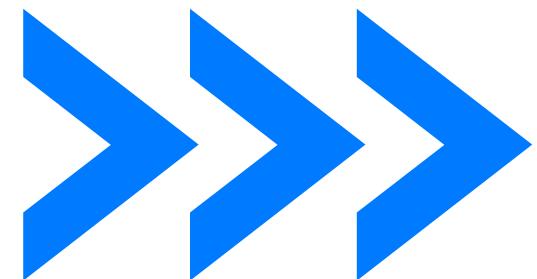


Random forest



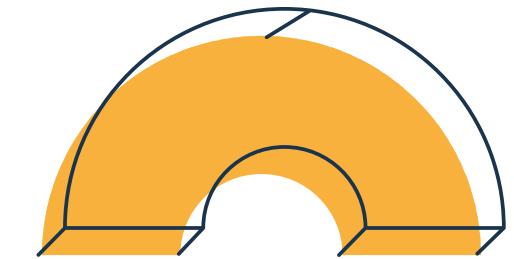
1) Creamos un dataset a partir de bootstrapping: Se seleccionan filas de manera aleatoria. Puede seleccionarse la misma fila más de una vez.

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

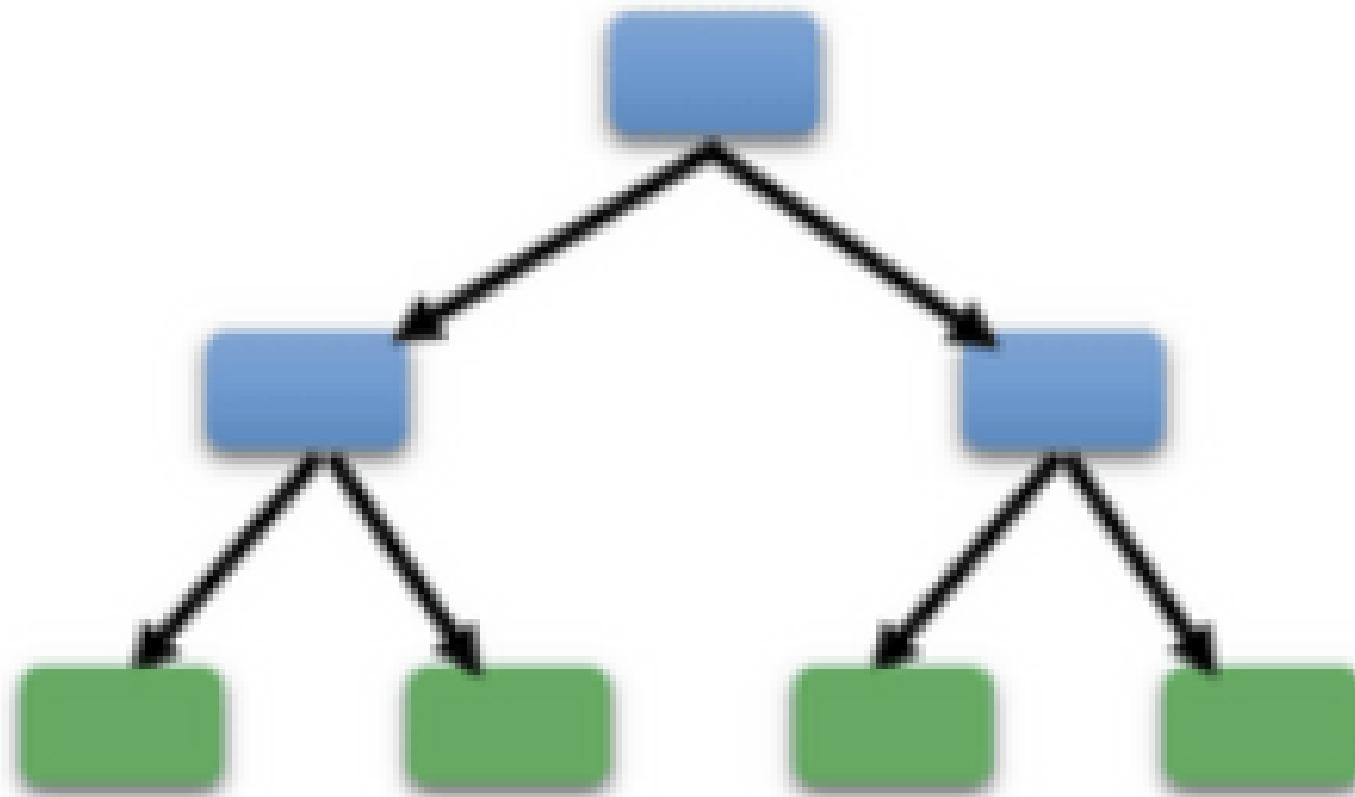


Gender	Age	EstimatedSalary	Purchased
Male	37	72000	0
Male	26	32000	0
Male	25	33000	0
Male	25	33000	0

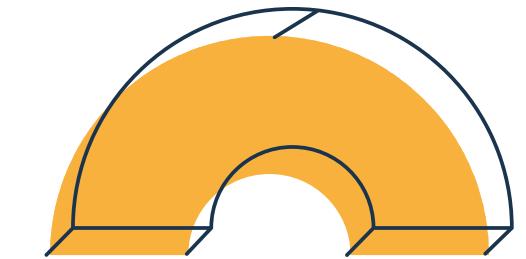
Random forest



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



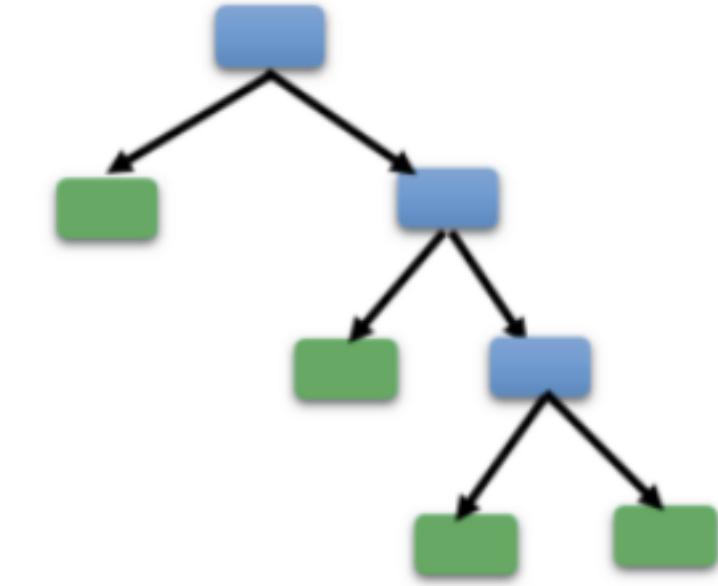
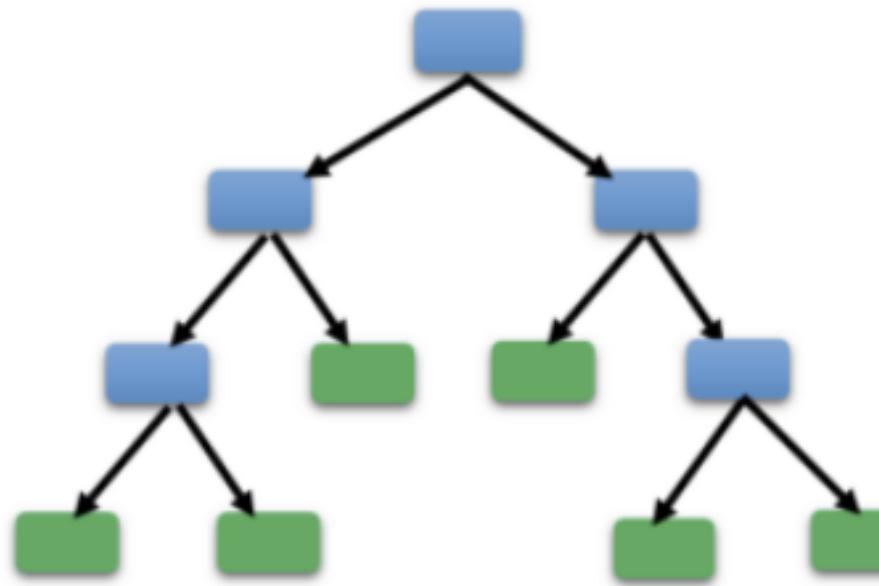
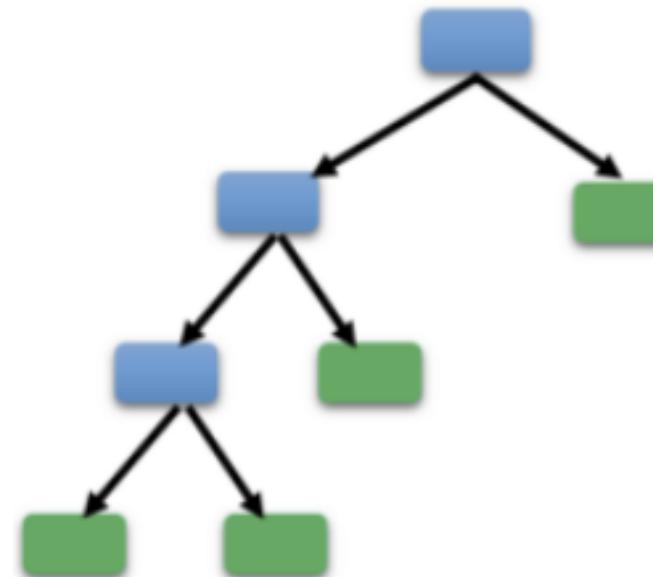
Random forest



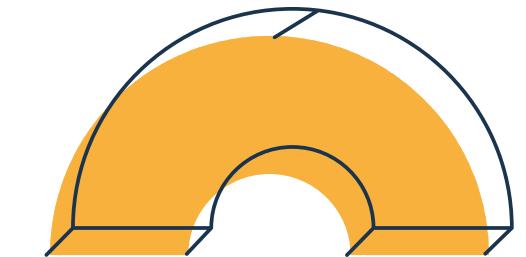
3) Vamos de nuevo al paso 1, creamos un nuevo dataset y luego un nuevo árbol (paso 2).

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

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

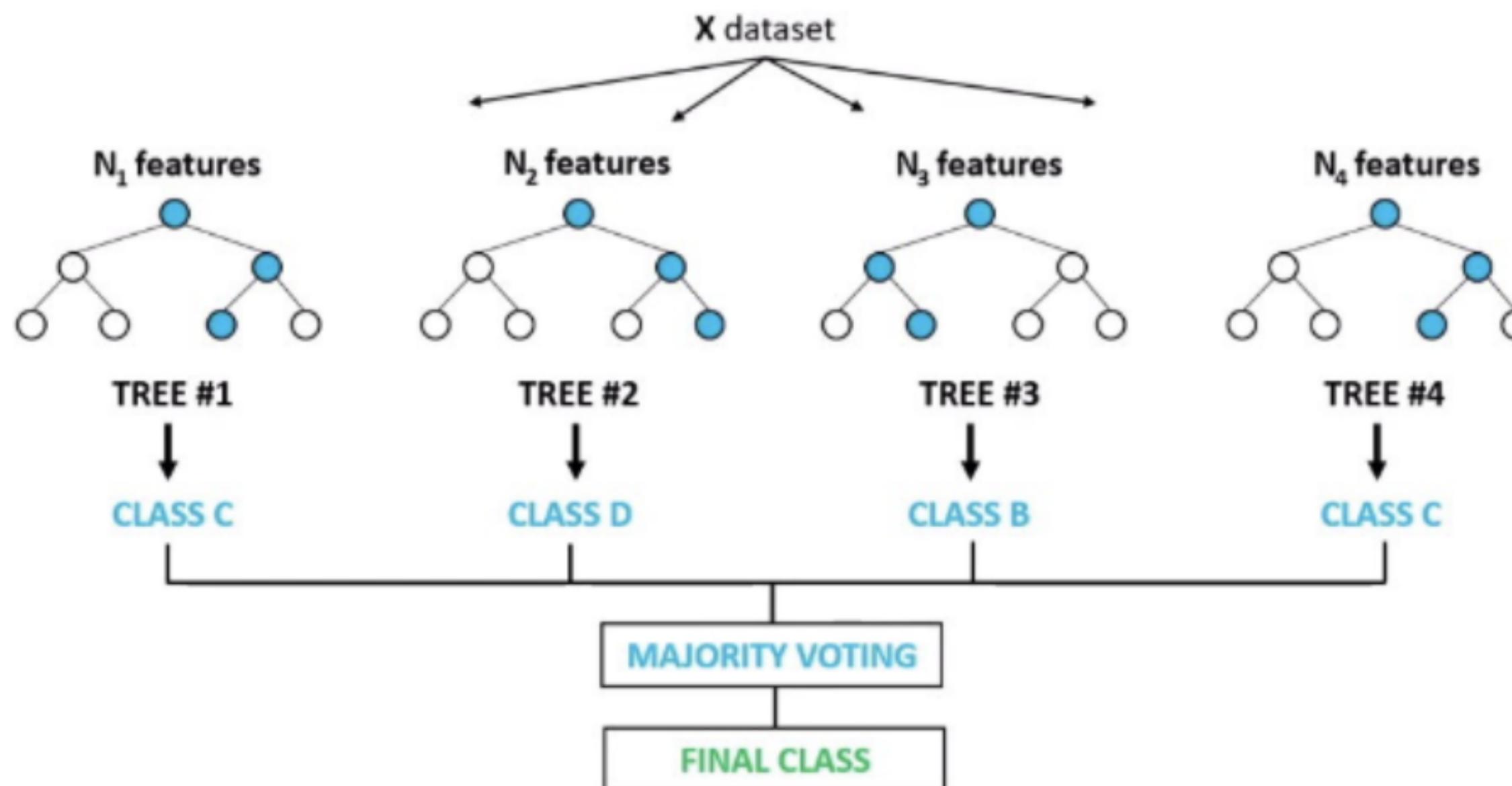


Random forest

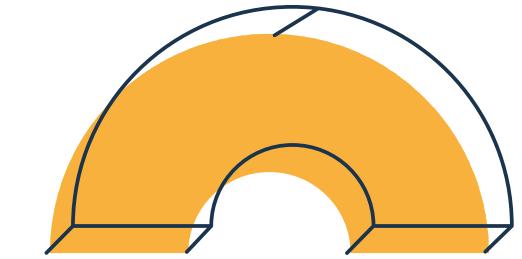


Ahora, para generar predicciones sobre nuevos datos, se utilizan todos los árboles.

Al final de todo, se hace una votación.



Random forest



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)

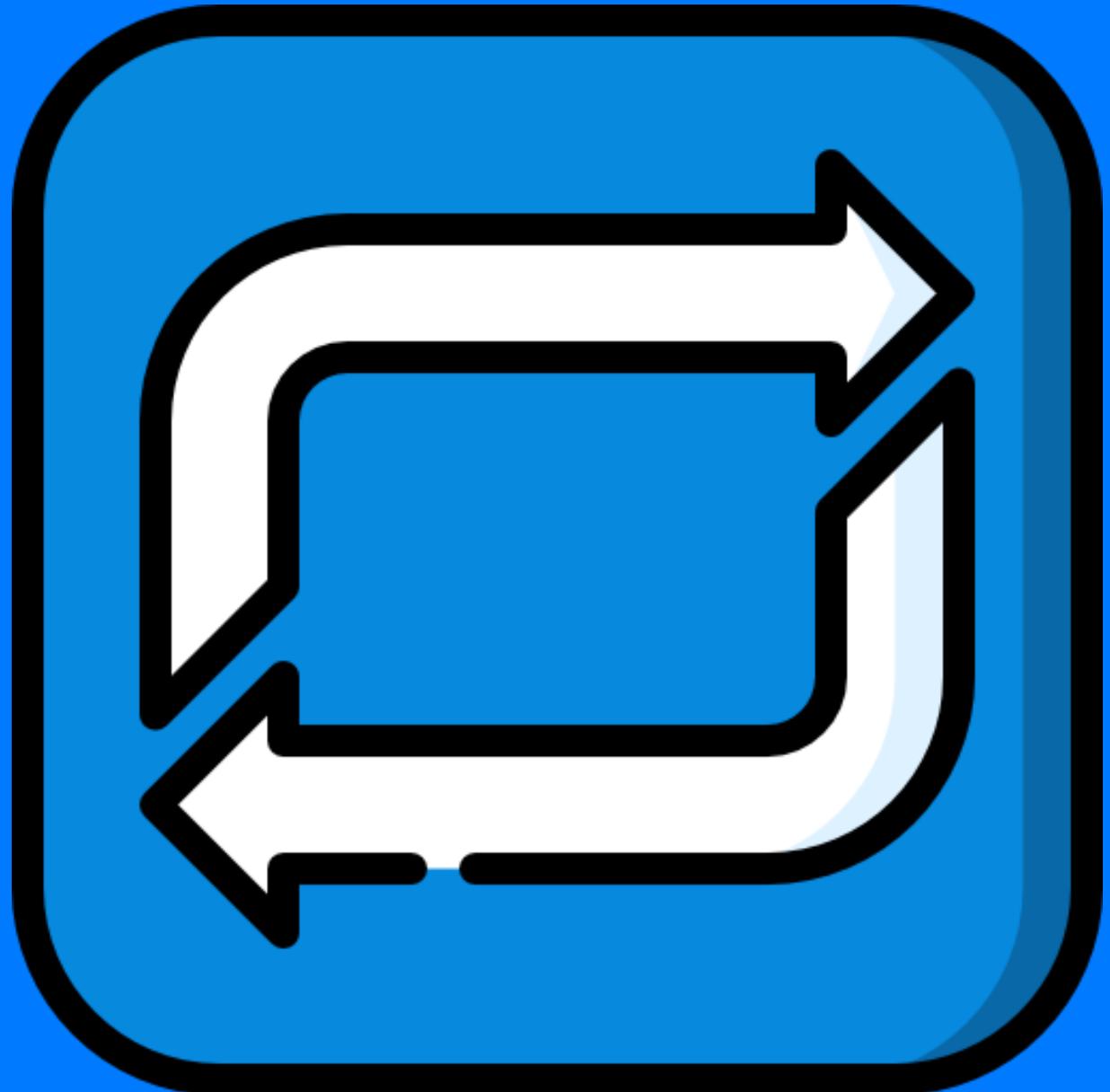
Random forest



¿ Dónde está la parte "random" de random forest ?

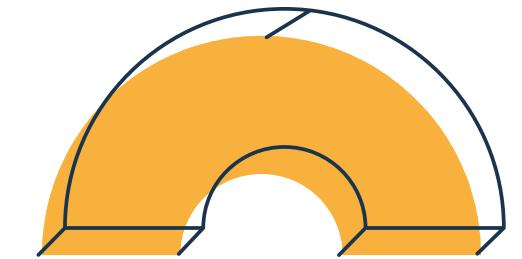
¿ Cuáles son los hiperparámetros más importantes ?

¿ Cuántas features se usan para entrenar cada árbol ?



BOOSTING

Boosting

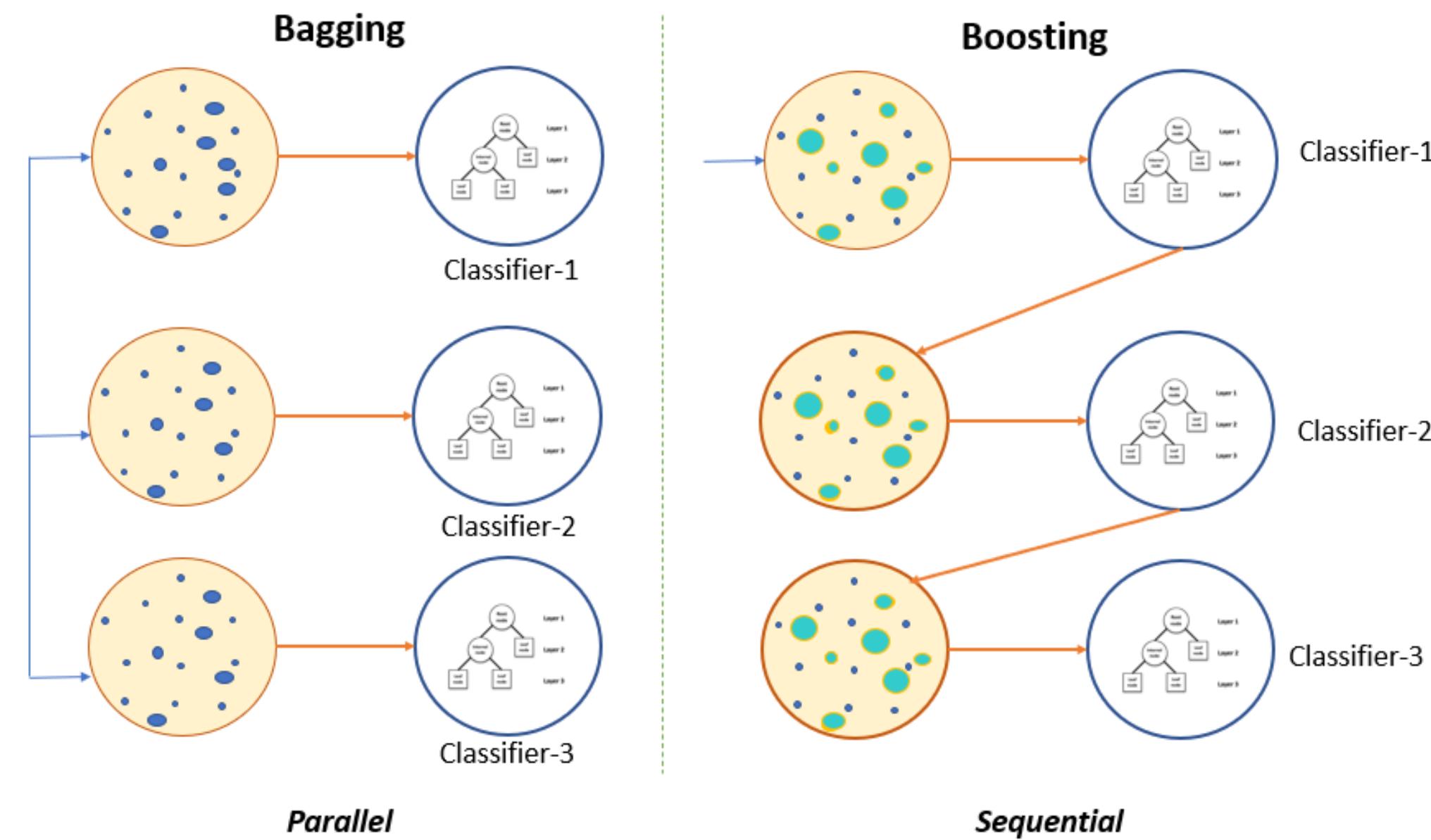
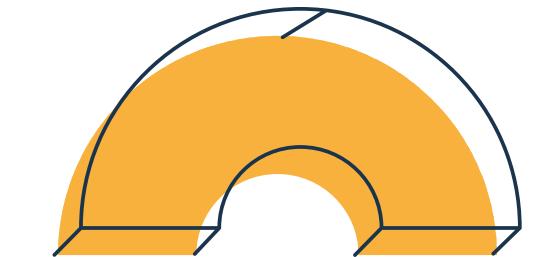


Boosting es otro método para hacer ensambles de modelos.

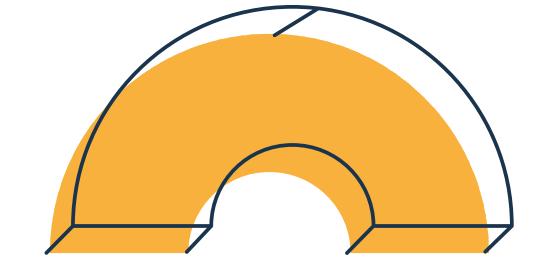
A diferencia de bagging, en este caso los modelos se ensamblan de manera secuencial

La idea principal es entrenar una secuencia de modelos en donde se le da más peso a los ejemplos que fueron clasificados erroneamente por el modelo anterior.

Boosting



Adaboost

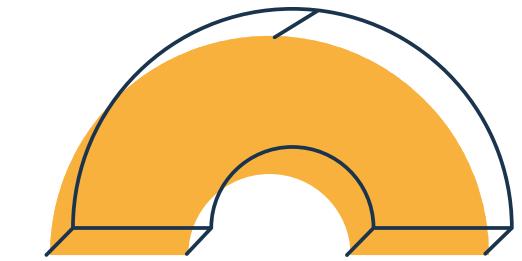


Adaboost es una de las técnicas más populares de Boosting.

Este algoritmo, mediante un entrenamiento iterativo de clasificadores débiles, le va dando mayor importancia a los datos mal clasificados anteriormente.

Adaboost -> Adaptative boosting

Bagging vs Boosting

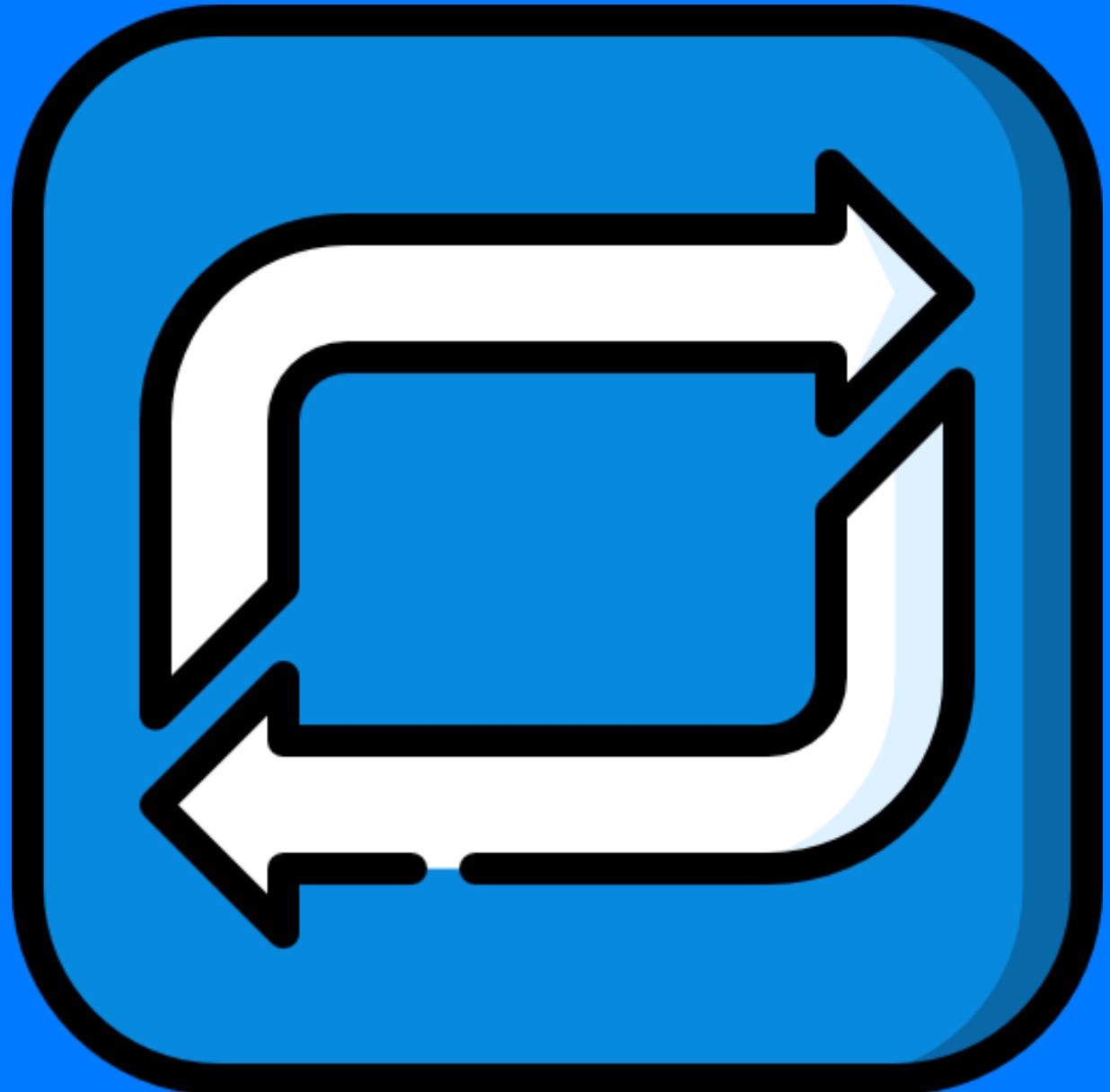


Bagging

- Se entranan modelos de manera independiente
- Fácilmente paralelizable
- Ayuda a prevenir el overfitting

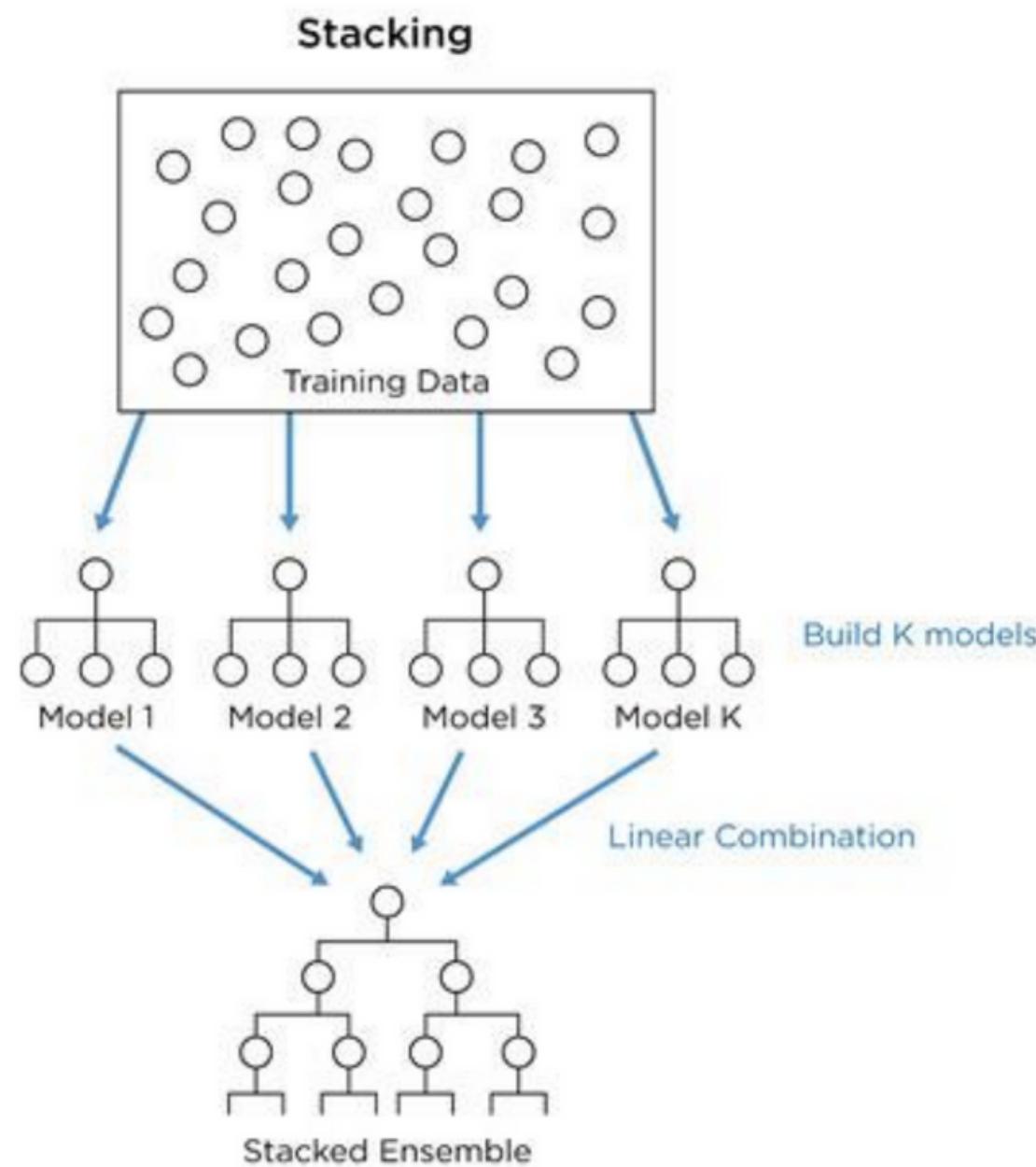
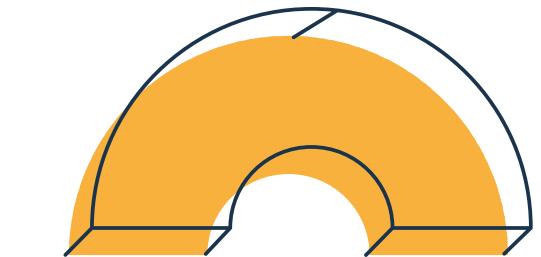
Boosting

- Modelos entrenados de manera secuencial, donde cada uno mejora las predicciones del anterior.
- Propenso a overfittear
- No se puede paralelizar fácilmente



STACKING

Stacking



Otra forma de hacer ensambles de modelos es stacking:

Cuando hacemos stacking, la idea es entrenar distintos modelos (pueden ser todos distintos). Cada uno de estos modelos genera sus propias predicciones.

Una vez que tenemos las predicciones de todos los modelos, utilizamos un "final estimator" que tomará como features (X) las predicciones de todos los modelos anteriores.