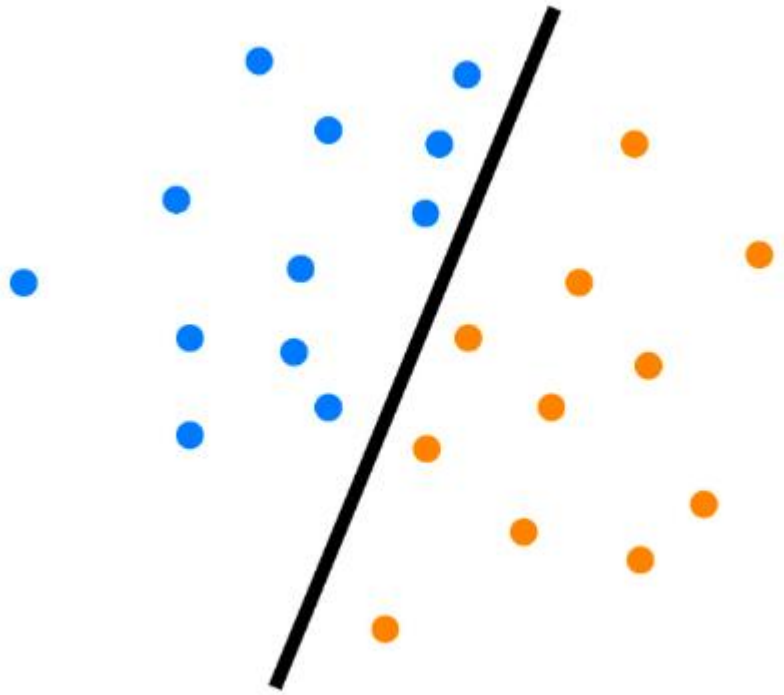
El problema de clasificación

El problema de clasificación

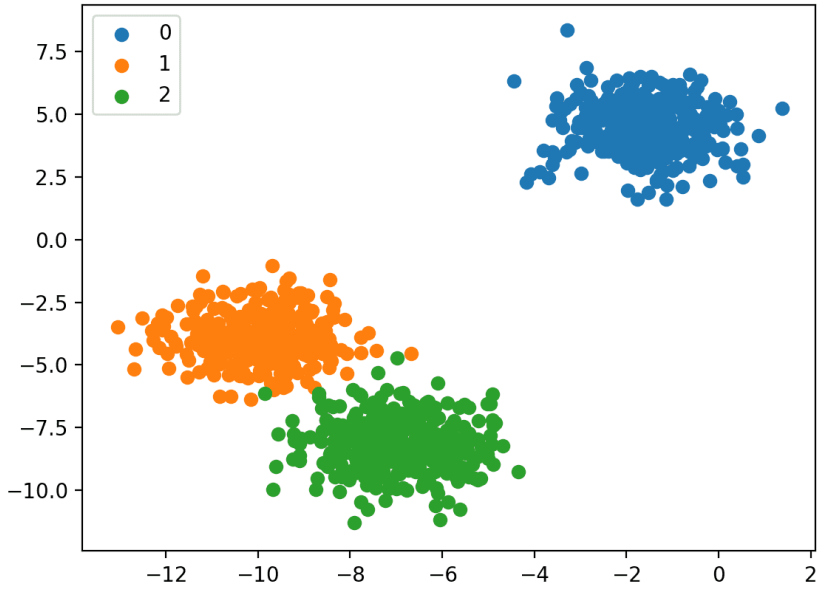
¿Aprobaré el curso o no?



El problema de clasificación

Los modelos de clasificación buscan determinar si determinados datos entrarán en alguna categoría.

Se suelen complementar con técnicas de reducción de dimensionalidad.



Dataprep for machine learning

Data desbalanceada

Si hay datos desbalanceados … , ¿qué hacemos?

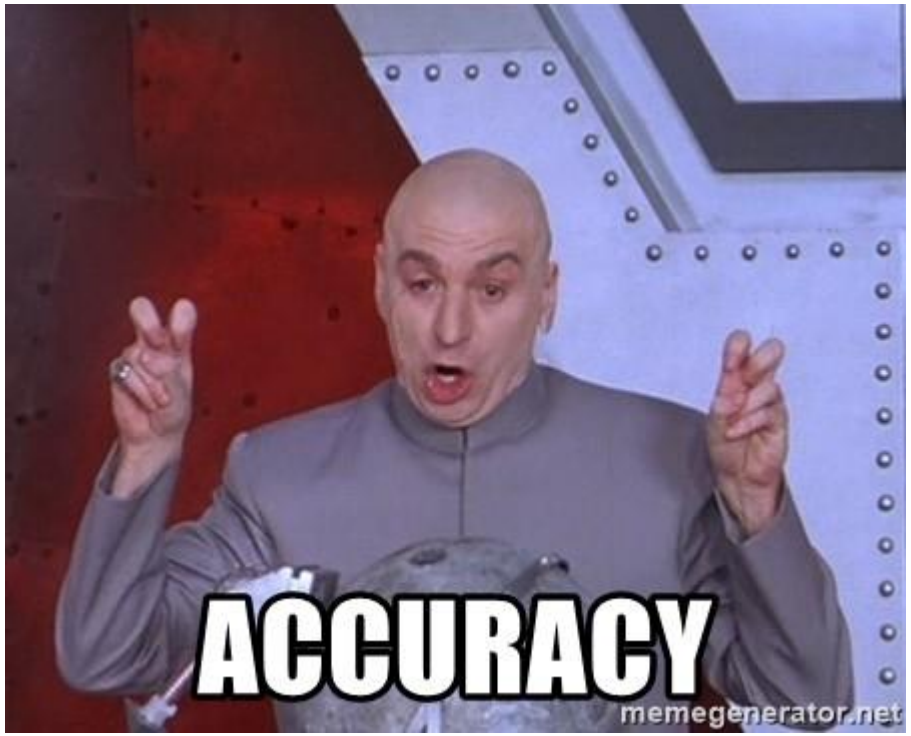
Dataprep for machine learning

Data desbalanceada

Si hay datos desbalanceados … , ¿qué hacemos?

El problema de clasificación

¿Cómo medimos su performance?



El problema de clasificación

¿Cómo medimos su performance?

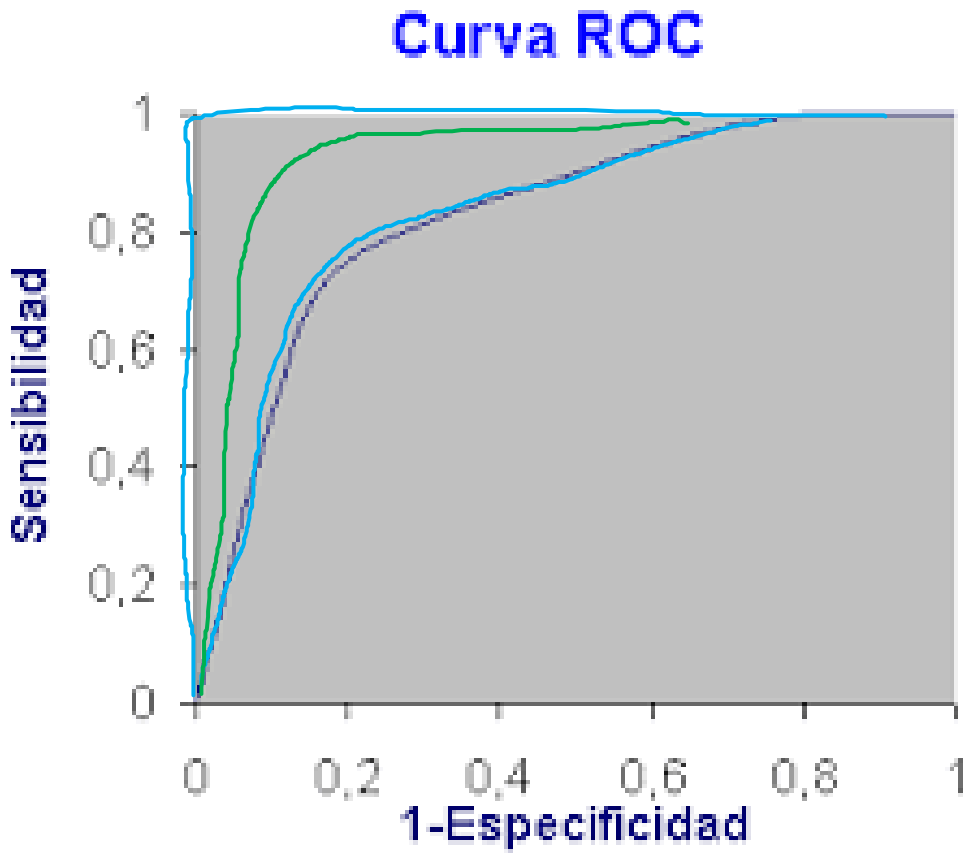
Matriz de confusión

Actual Class \ Predicted Class	Positive	Negative	
Positive	True Positive (TP)	False Negative (FN) Type II Error	**Sensitivity** $\frac{TP}{(TP+FN)}$
Negative	False Positive (FP) Type I Error	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

El problema de clasificación

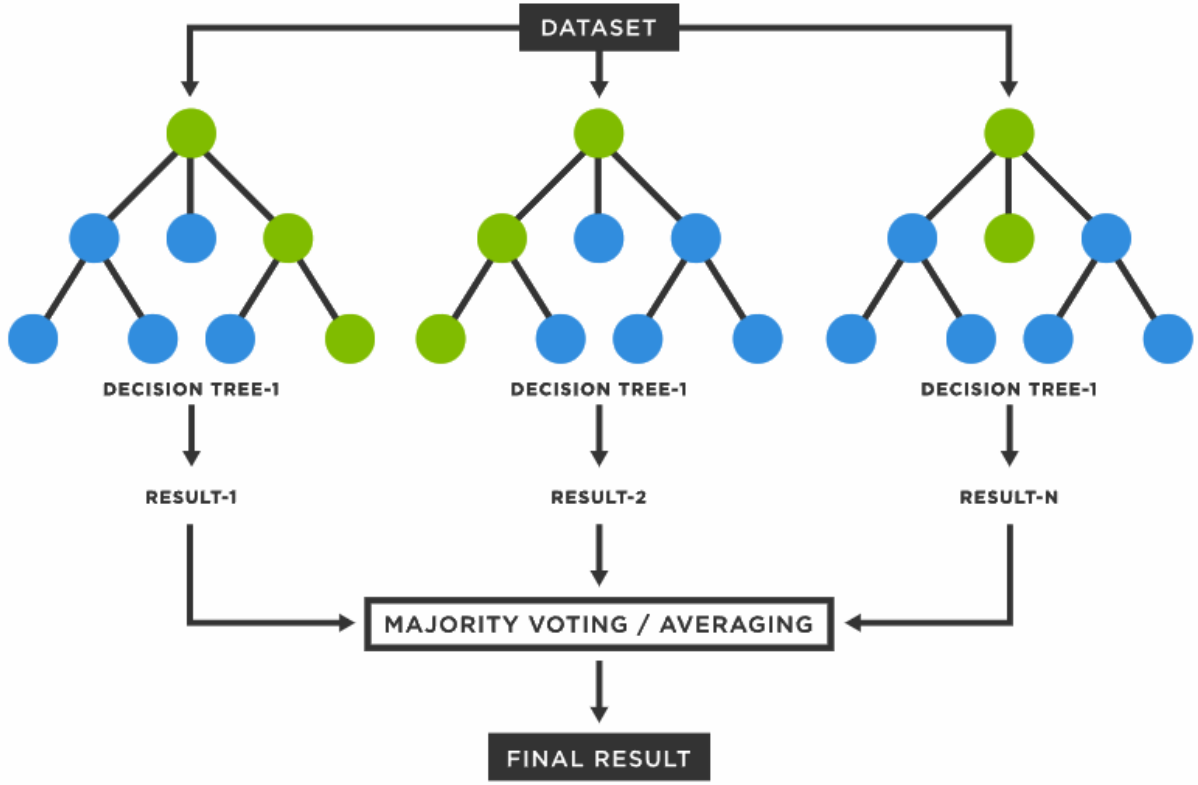
¿Cómo medimos su performance?

Curva ROC y AUC



Random Forest Classifier

Un viejo conocido



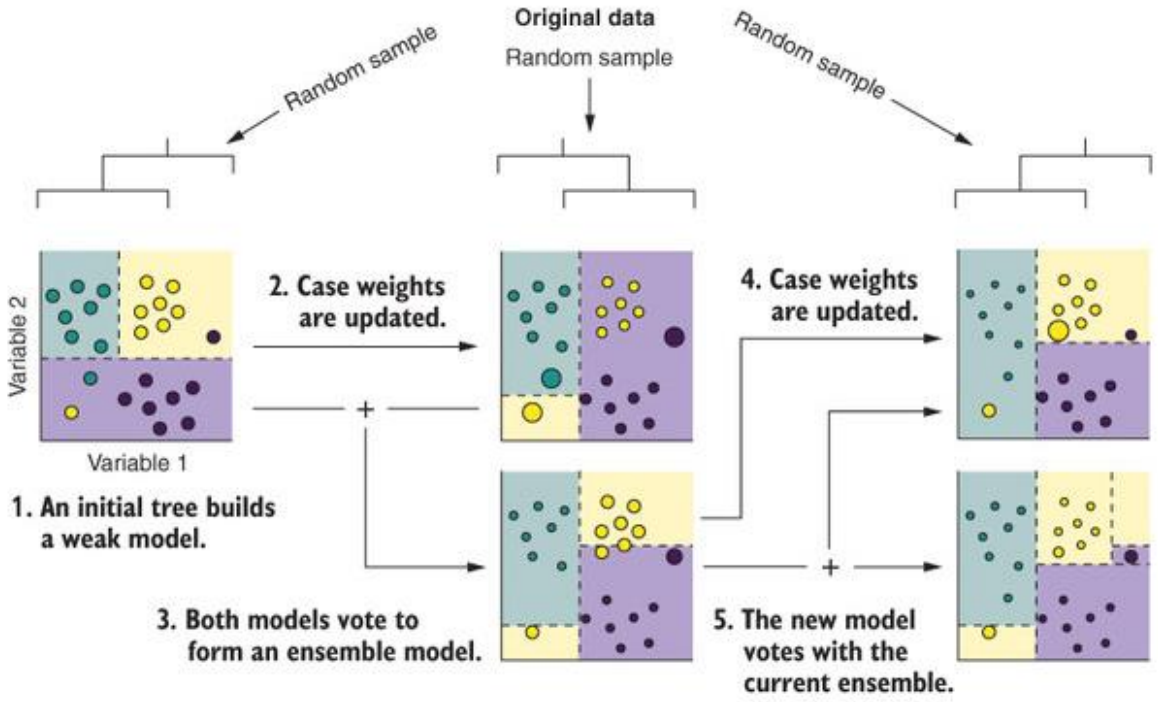
Aprender de los errores previos

Boosting

RandomForest entrena árboles en paralelo (bagging), sin embargo, ¿qué pasaría si entrenaramos los árboles de manera secuencial? A este proceso se le llama Boosting. Existen, principalmente, dos tipos de boosting: Adaptative Boosting y Gradient Boosting

Adaptative Boosting

AdaBoost entrena una muestra con bootstraping donde todos los elementos tienen el mismo peso. Se generan las predicciones y se asigna un peso mayor a aquellas que tienen error. A partir de eso, se entrena otro árbol. De esta manera, el modelo aprende de los errores



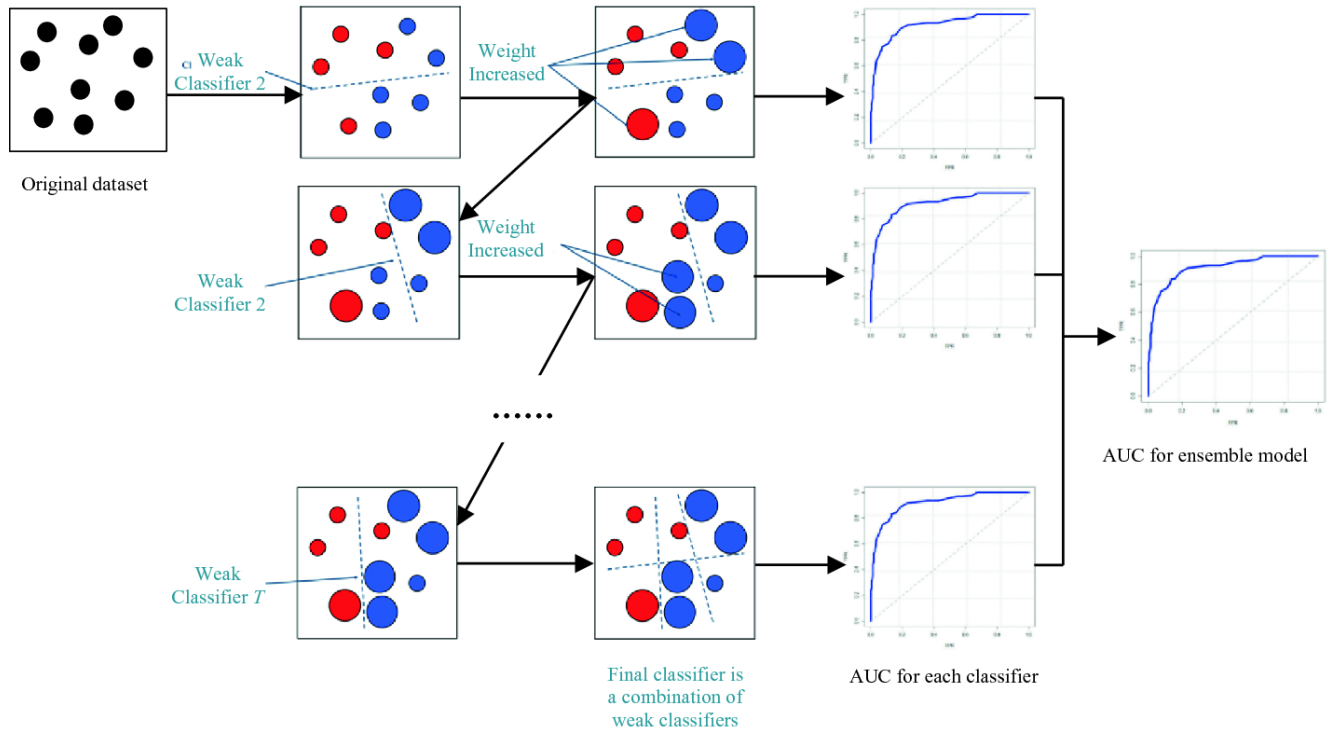
Aprender de los errores previos

Boosting

RandomForest entrena árboles en paralelo (bagging), sin embargo, ¿qué pasaría si entrenaramos los árboles de manera secuencial? A este proceso se le llama Boosting. Existen, principalmente, dos tipos de boosting: Adaptative Boosting y Gradient Boosting

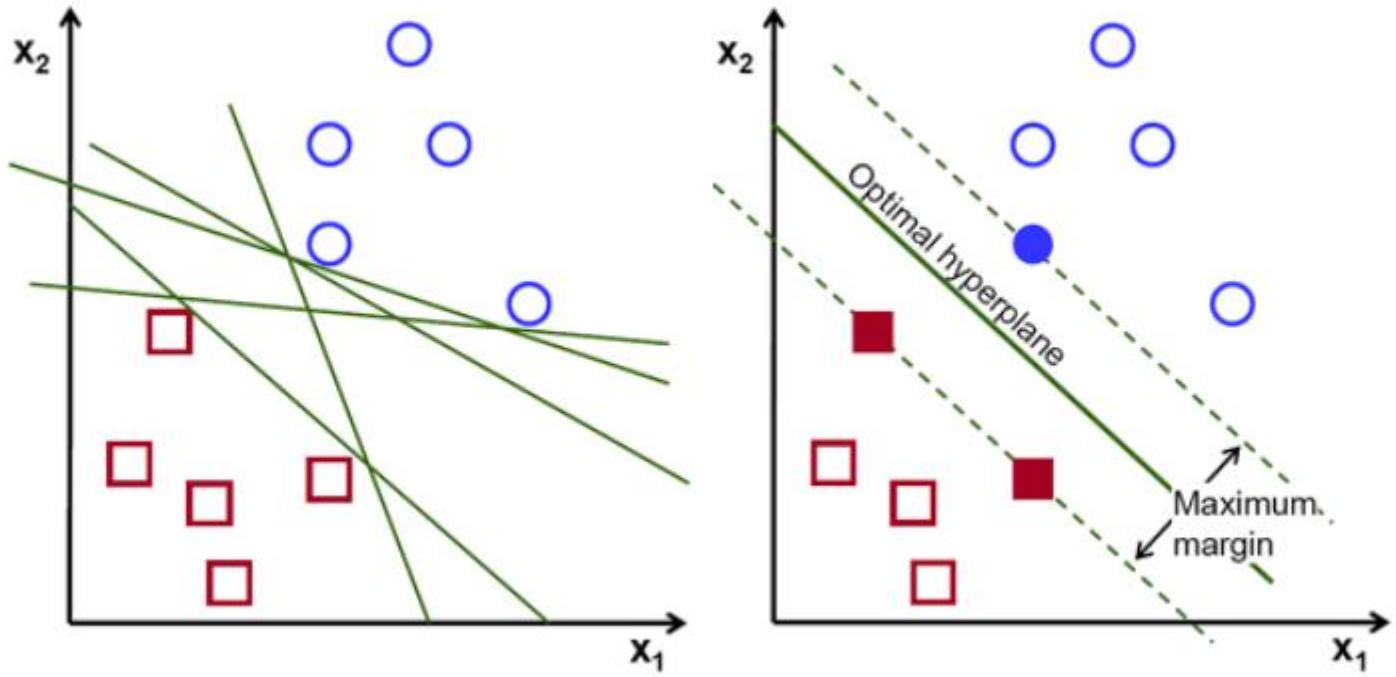
Gradient Boosting

AdaBoost entrena una muestra con bootstraping donde todos los elementos tienen el mismo peso. Se generan las predicciones. A partir de los residuos de estas, se entrena otro árbol. De esta manera, el modelo aprende de los errores



Support Vector Machine

Encontrar un hiperplano que haga que la distancia entre clases sea la máxima



Possible hyperplanes

Naive Bayes

Grupo de algoritmos basados en el teorema de Bayes.

Supuestos:
Independencia entre las características y la clase.
Cada características contribuye de igual manera a la clasificación

*Los supuestos generalmente no se cumplen en situaciones de la vidas real