KIBERNUM
IT Academy

Módulo 5 – Aprendizaje de Máquina Supervisado

Validación Cruzada

Especialización en Ciencia de Datos

Contenido

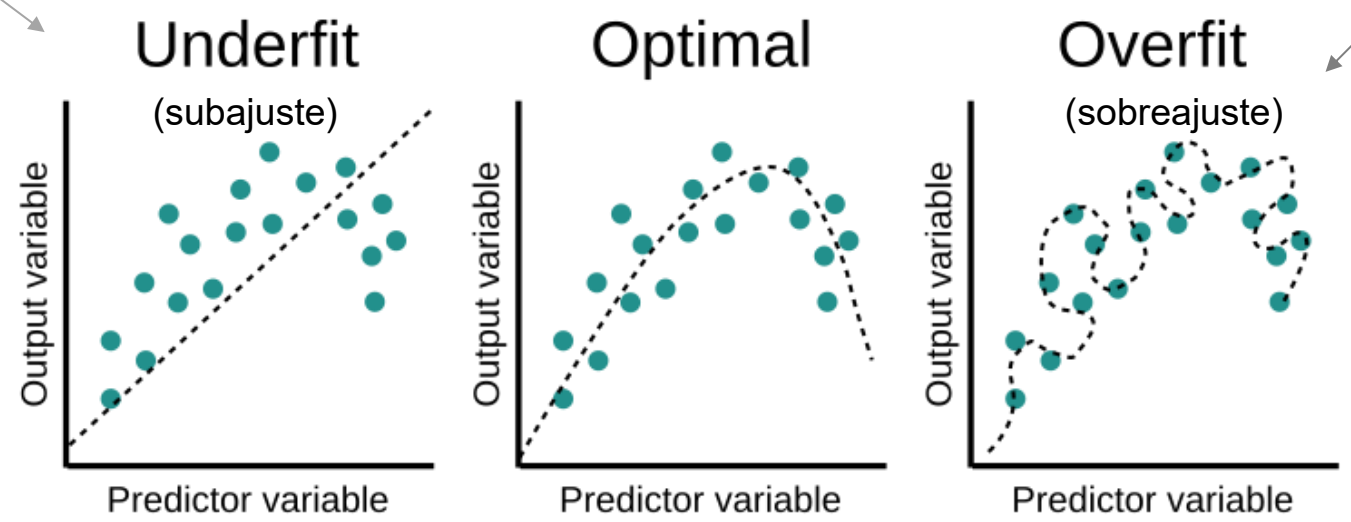
- Nivel de ajuste del modelo.
- Validación cruzada.

Nivel de Ajuste de un Modelo

Nivel de Ajuste del Modelo

El modelo no ha capturado la lógica subyacente de los datos.

El modelo se concentró demasiado en los datos de entrenamiento pero "perdió el foco".



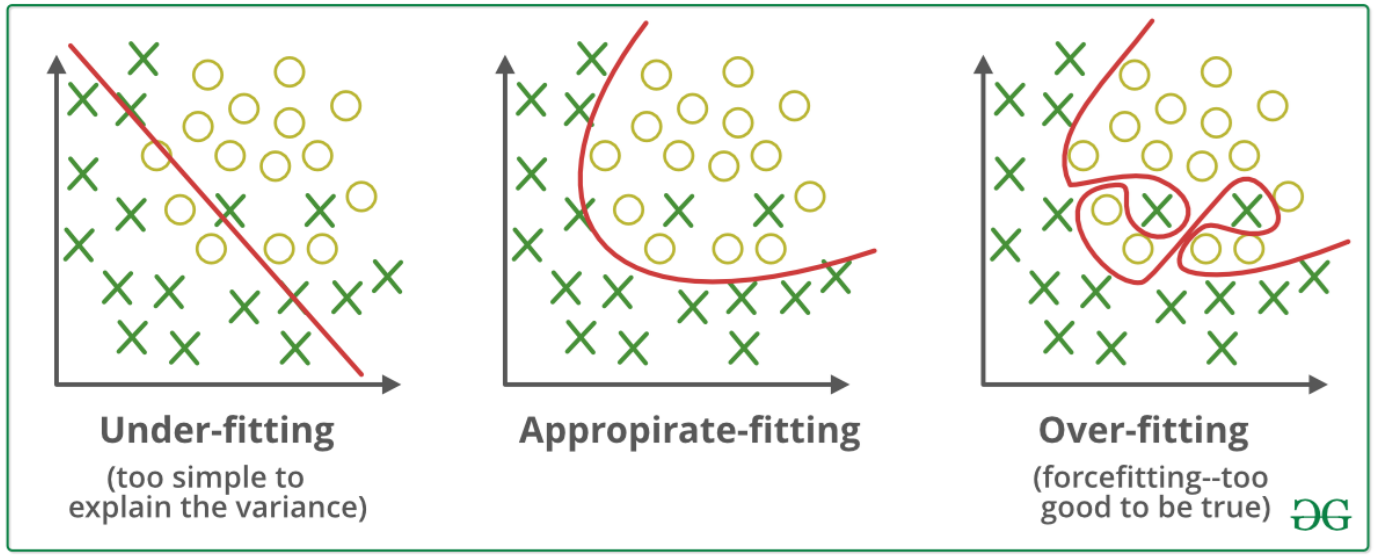
Más que un modelo que tenga ajuste perfecto, lo que se busca es un modelo que tenga buena **capacidad de generalización**.

Nivel de Ajuste del Modelo

A continuación, se presentan 3 escenarios de ajuste de un modelo de clasificación: **Subajustado, óptimo y sobreajustado.**

Subajustado
(muy simple para explicar la varianza de los datos)

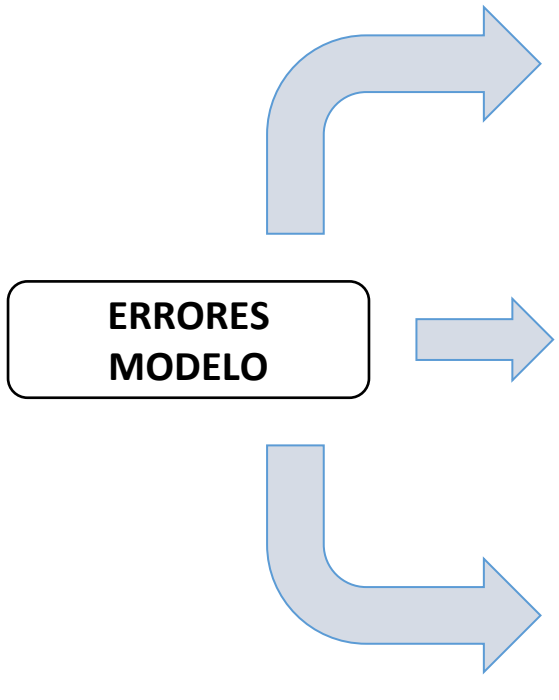
Sobreajustado
(ajuste forzado, muy bueno para ser verdadero)



Más que un modelo que tenga ajuste perfecto, lo que se busca es un modelo que tenga buena **capacidad de generalización**.

Desviación v/s Varianza

Cuando ajustamos un modelo de machine learning tratamos que sea lo más preciso, sin embargo, nunca estará libre de errores.



ERROR DE DESVIACIÓN

Es la diferencia entre la predicción esperada del modelo y el valor real.

ERROR DE VARIANZA

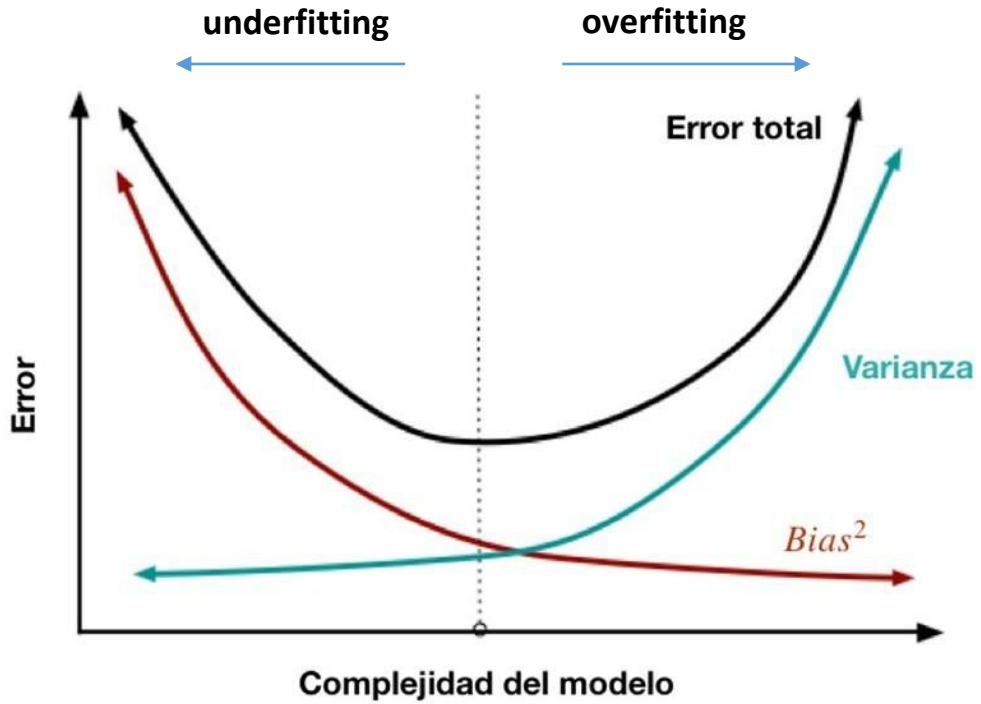
Cantidad que la estimación de la función objetivo cambiara si se utiliza diferentes datos de entrenamiento.

ERROR IRREDUCIBLE

No se puede reducir más, ruido de los datos, variables desconocidas.

Desviación v/s Varianza

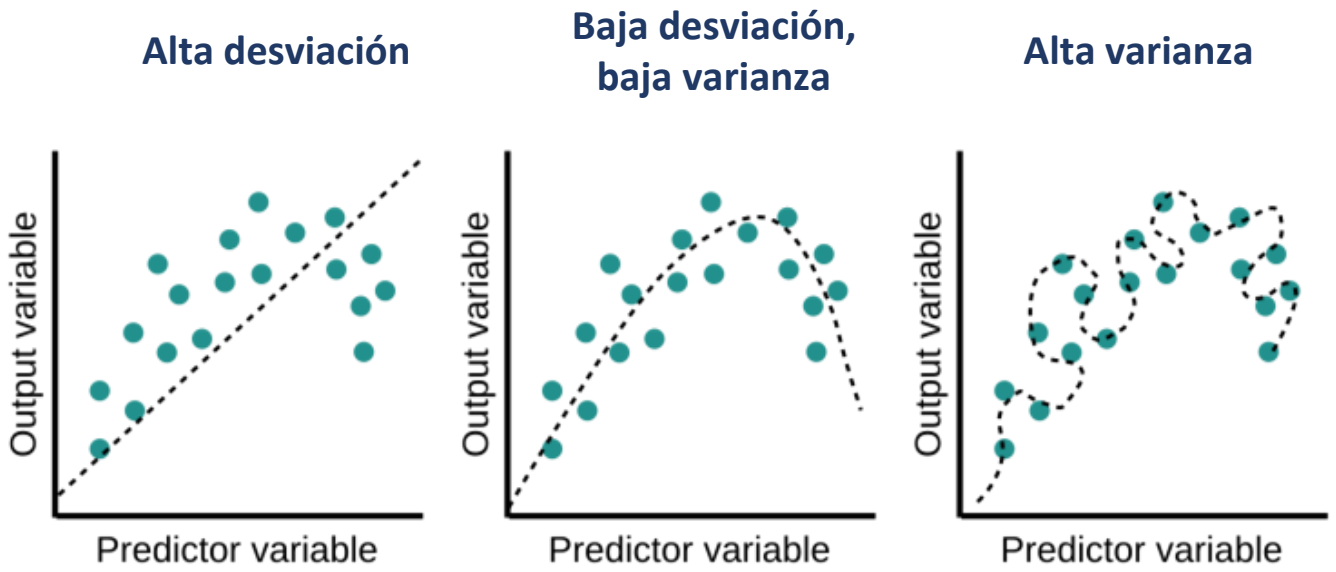
Comprender la desviación y la varianza es fundamental, pero lo que realmente importa es el error general.



Un buen modelo es aquel que encuentra un **buen equilibrio entre desviación y varianza** de manera de minimizar el error total.

Desviación v/s Varianza

Un modelo con ajuste apropiado tiene baja desviación y baja varianza.



KIBERNUM

Validación Cruzada

¿Qué es la validación cruzada?

La validación cruzada (cross validation) es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba. Se utiliza en entornos en donde el objetivo principal es la predicción y se quiere estimar la precisión de un modelo que se llevará a cabo a la práctica. Esta técnica es muy utilizada en inteligencia artificial y aprendizaje de máquina.

En la elaboración de un modelo predictivo, se corre el riesgo que el modelo sea demasiado dependiente de los datos con el cual fue confeccionado, los cuales podrían no representar del todo la realidad. Es por esto que se utiliza validación cruzada, para verificar que los modelos tendrán un buen poder de generalización, es decir no tienen una fuerte dependencia de los datos con los cuales fueron entrenados.

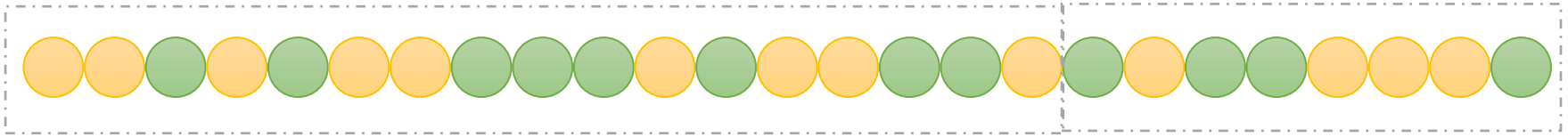
¿Qué es la validación cruzada?

Esta técnica puede ser aplicada de muchas maneras, las más conocidas son las siguientes:

- Método de retención.
- Validación cruzada de k-iteraciones.
- Validación cruzada aleatoria.
- Validación cruzada dejando uno afuera.

Método de retención

Es el mecanismo de validación cruzada más simple de todos, en donde tomamos el set de datos disponible y lo dividimos en un set de entrenamiento y en otro set de validación.



Set de Entrenamiento

70% - 80% aprox.

Set de Validación

20% - 30% aprox.

Validación Cruzada de k-iteraciones

La muestra se divide en K subconjuntos. Uno de los conjuntos se usa como set de test y el resto para entrenamiento.

KIBERNUM

Dudas y consultas

KIBERNUM

Fin Presentación

KIBERNUM