**¿Cómo afectan nuestros features a los modelos de Machine Learning?**

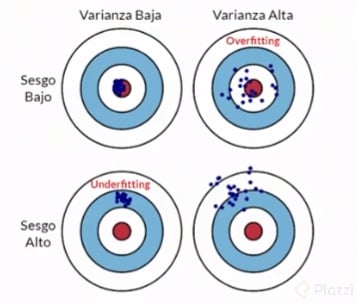
* ¿Qué son los features? Son los atributos de nuestro modelo que usamos para realizar una interferencia o predicción. Son las variables de entrada.

**Más features simpre es mejor, ¿verdad?**  
*La respuesta corta es: NO*  
En realidad si tenemos variables que son irrelevantes pasarán estas cosas:

* Se le abrirá el paso al ruido.
* Aumentará el costo computacional.
* Si introducimos demasiados features y estos tienen valores faltantes, se harán sesgos muy significativos y vamos a perder esa capacidad de predicción.  
  *Nota: Hacer una buena selección de nuestro features, hará que nuestros algoritmos corran de una manera mas eficiente.*\_

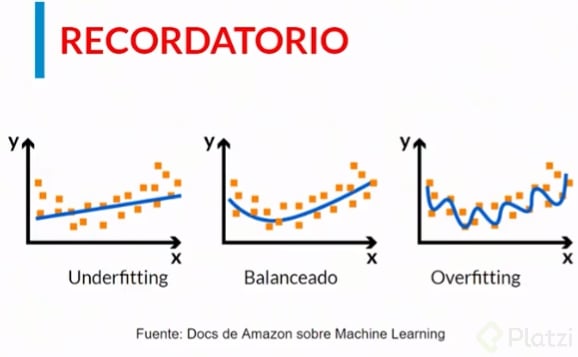
**Una de las formas de saber que nuestros features han sido bien seleccionados es con el sesgo y la varianza.**

* Una mala selección de nuestro features nos puede llevar a alguno de esos dos escenarios indeseados.



**Algo que debemos que recordar es que nuestro modelo de ML puede caer en uno de 2 escenarios que debemos evitar:**

* Uno es el Underfitting: Significa que nuestro modelo es demasiado simple, en donde nuestro modelo no está captando los features y nuestra variable de salida, por lo cual debemos de investigar variables con mas significado o combinaciones o transformaciones para poder llegar a nuestra variable de salida.
* Por otro lado está el Overfitting: Significa que nuestro modelo es demasiado complejo y nuestro algoritmo va a intentar ajustarse a los datos que tenemos, pero no se va a comportar bien con los datos del mundo real. Si tenemos overfiting lo mejor es intentar seleccionar los features de una manera mas critica descartando aquellos que no aporten información o combinando algunos quedándonos con la información que verdaderamente importa.



**¿Qué podemos hacer para solucionar estos problemas?**

* Aplicar técnicas reducción de la dimensionalidad. Utilizaremos el algoritmo de PCA.
* Aplicar la técnica de la regulación, que consiste en penalizar aquellos features que no le estén aportando o que le estén restando información a nuestro modelo.
* Balanceo: Se utilizará Oversampling y Undersampling en problemas de rendimiento donde tengamos un conjunto de datos que está desbalanceado, por ejemplo en un problema de clasificación donde tenemos muchos ejemplos de una categoría y muy pocos de otra.