**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.

A picture containing text, electronics, screenshot, clipart

Description automatically generated

Accurracy: sesgo, que tan lejos del blanco

Precision: varianza, que tan agrupados o disperse (varianza alta: overfiting)

Sesgo bajo, varianza baja

**EVITAR:**

* 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.
* 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.

Chart

Description automatically generated

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

* Aplicar técnicas reducción de la dimensionalidad. Utilizaremos el algoritmo de PCA.
* Técnica de la regulación: penalizar 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.