*Algoritmos de Agrupacion I I*

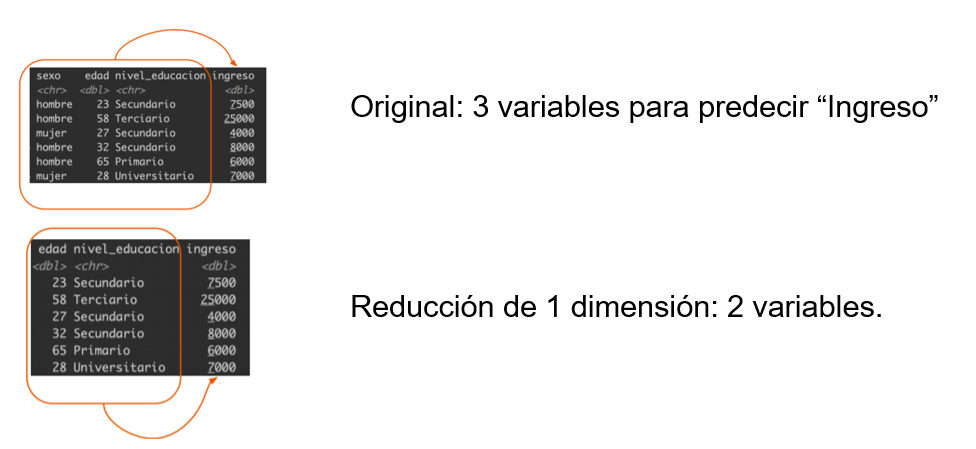
***Repaso de Sesiones Anteriores – Reducción de la Dimensionalidad***

Buscamos reducir la cantidad de features de un dataset, pero reteniendo la mayor cantidad de “información” posible.

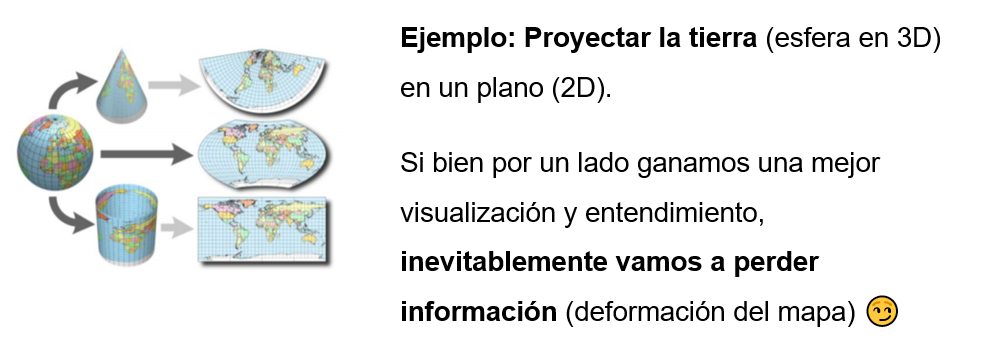
Para alcanzar este objetivo tenemos 2 opciones:

* Caso 1) Eliminar variables del dataset.
* Caso 2) Aplicar alguna transformación matemática.

*Caso 1: Eliminar variables del dataset*

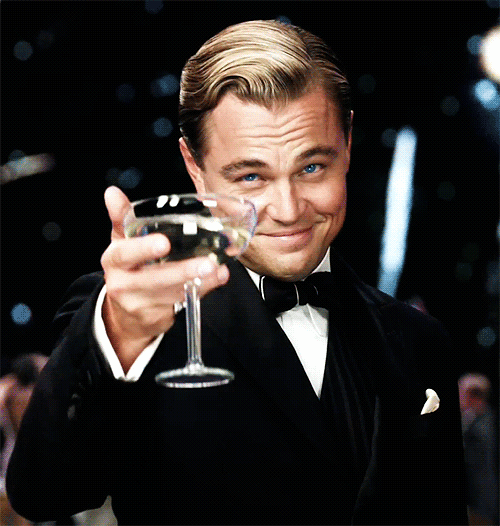


*Caso 2: Aplicar alguna transformación matemática*



*En los métodos de Reducción de la Dimensionalidad, siempre vamos a perder información*

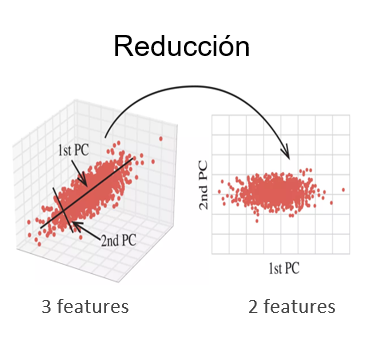
***¡Pero claramente el objetivo es perder lo menos posible!***

******

***¿Qué hace el PCA?***

* *El método gira los datos de forma que, desde un punto de vista estadístico, no exista una correlación entre las características rotadas pero que conserven la mayor cantidad posible de la varianza de los datos originales.*
* *Es decir, el PCA reduce la dimensionalidad de un conjunto de datos proyectándose sobre un subespacio de menor dimensionalidad.*

Por ejemplo, **datos con dos características** (dispuestos en un plano) **pueden ser proyectados sobre una única línea.**  Por otro lado un conjunto de datos de tres características (dispuestos en un espacio de tres dimensiones) pueden ser proyectados en un plano (de dos dimensiones). **Incluso los datos resultantes en el plano podrían ser reducidos a una única línea es decir pasar de 3 dimensiones a 1.**



***Detección de Outliers***

Valor que no se corresponde con el patrón general de nuestros datos. Puede ser bueno, malo o simplemente un error de datos pero en todos esos casos tenemos que realizar un análisis.

***¿Por qué es importante tratar los outliers?***

En términos generales, tratar los outliers suele mejorar los modelos de ML. Muchos modelos avanzados son sensibles a los valores extremos y además, siempre es preferible realizar una buena preparación de datos antes que complejizar los modelos.