¡Ni se te ocurra!

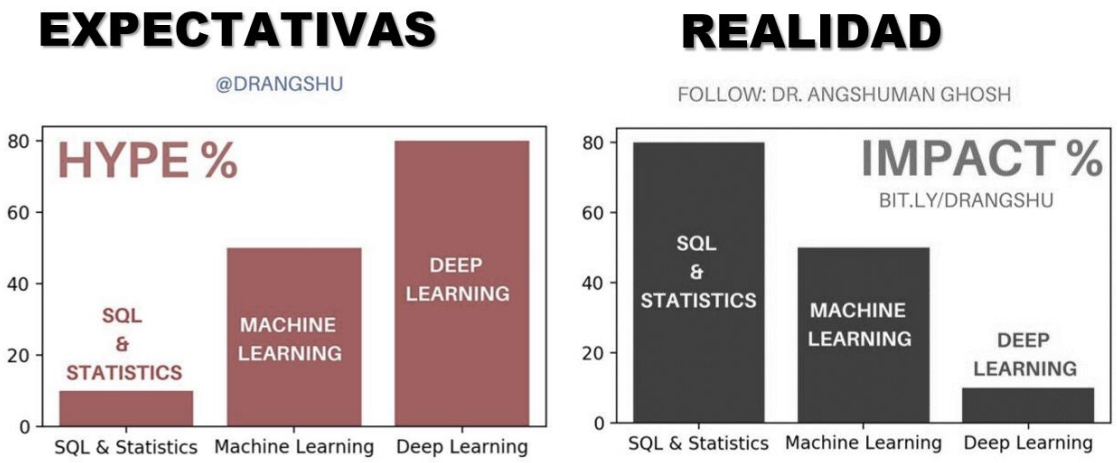
Sabemos que el Machine Learning es una de las últimas innovaciones que llama poderosamente la atención y, después de haber cogido experiencia aplicando tus modelos es muy probable que ya estés en el camino del aprendizaje para saber cómo implementarlo en tu caso real o en una empresa.

También es muy probable que, en este camino de la implementación, te estés topando con algunos obstáculos que no esperabas, ¿verdad?

Esto nos ha pasado a todos, sin embargo, y a pesar de estas dificultades, no tienes que preocuparte. En esta guía se pretende que vayas aclimatándote y que poco a poco vayas aprendiendo de ella.

1. **Empieza por algo sencillo.**

Esto, aunque no lo creas, suele suceder en muchas empresas, las cuales tienen como factor común no apostar por la inclusión de un proyecto predictivo de Machine Learning. Por ello, antes de coger un proyecto es muy importante tener una meta-estructura de cuanto te va a llevar, el gasto que va a tener y la viabilidad del mismo.



2. **Ser precavido**

Por ejemplo, si tu jefe te dice que vais a coger un proyecto X, antes de dar el sí o hablar con el cliente, es recomendable hacer un estudio previo para saber dónde nos estamos metiendo.

Además, sería recomendable que el cliente nos pudiera pasar una demo de sus datos o unos psudo-datos ya que puede que tengas un increíble modelo, si los datos de entrada son malos, no se pueden hacer milagros.

Hasta ahora normalmente hemos trabajado con datos realmente bien procesados, pero…hay de todo en la tierra del señor.

3. **Alternativas a Excel**

Un error bastante típico es trabajar en formatos de fichero cuya lectura es muy lenta. Por ejemplo, Excel. La diferencia entre leer el mismo fichero en excel o csv crece a medida que el fichero es más grande.

Por darte unos números, un fichero excel de unas 100 000 filas 10 columnas puedes tardar en leerlo unos 45 segs - 1 min con un laptop normalito. El mismo fichero en csv 1-3 segs.

4. **No comentar el código**

Otro que no tiene misterio.

5. **No buscar lo suficiente en internet**

Lo más probable es que nuestra idea ya esté en internet. Puede que no esté ensamblada o sea específicamente lo que buscas, pero una buena investigación previa nos ahorrará muchos quebraderos de cabeza. link

"En internet hay de todo menos lo que justo estás buscando en ese momento."

…entonces traza un boceto de qué es lo que quieres hacer. Descompone las distintas funciones y busca las más complicadas por separado.

Si, es un poco más de trabajo, pero a la larga es mejor y te sirve como un desarrollo para enseñarle a tus superiores si no se llega al tiempo o al trabajo propuesto.

6. **No hace falta big data**

La verdad es que sí depende de los datos, pero lo que no necesariamente requiere es trabajar con el Big Data.

Mucha gente piensa que para poder crear modelos de ML es necesario una arquitectura informática media dentro de una organización. Si, es cierto también, pero no hace falta que tengan una fábrica informática para minar bitcoins. Depende de los requerimientos o el objetivo que nos pidan, las exigencias de datos serán proporcionales.

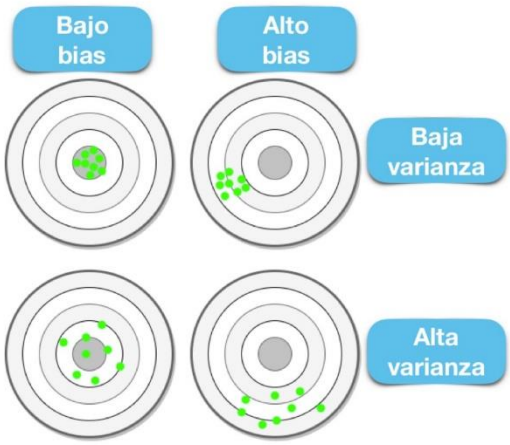
7. **El overfitting**

Pero debemos saber que podemos "caer" en el overfitting de diversas maneras, a veces sin darnos cuenta. Dos de sus caras son:

- Bias (ó sesgo): es la tendencia a aprender -equivocadamente- algo falso.
- Varianza: es la tendencia a aprender algo random no relacionado con la realidad.

Es difícil evitar el Overfitting, se puede utilizar Regularización ó la validación cruzada (cross validation), u otras técnicas, pero ninguna nos asegura evitarlo del todo.

Al querer corregir la varianza, caemos en el Bias o sesgo… y viceversa. Lograr evitar ambos en simultáneo es "el dilema" que tenemos.



8. **Evitar las dimensiones paralelas**

Cuando tenemos muchas características (columnas), por ejemplo 100 ó más, puede que alguno de nuestros algoritmos de aprendizaje "se vuelvan locos" … es decir, que no logre generalizar ó que lo haga mal ya que tiene muchos datos y tiende al sobreajuste.

A esto, nuestro compañero Juan Ignacio Bagnato le llama "la maldición de la dimensionalidad".

link

9. **El dato vs el modelo**

Más muestras superan a un algoritmo complejo. Fin de la discusión.

10. **La importancia del Feature Ingeniering**

Seguramente pasemos mucho más tiempo seleccionando los features, transformando, preprocesando… que el tiempo dedicado a ejecutar el algoritmo.

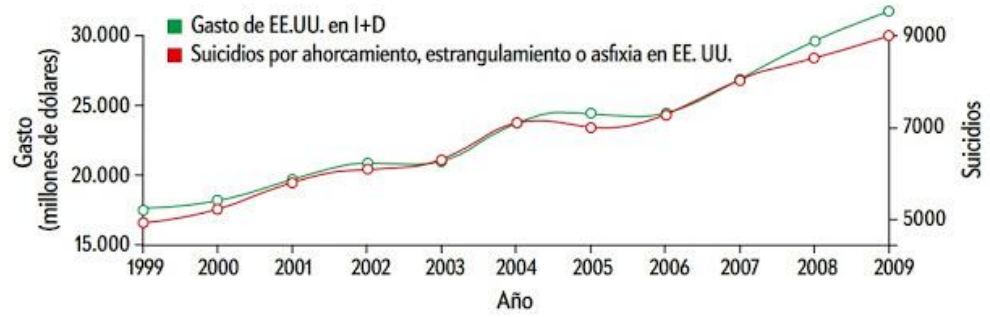
Si tenemos pocas dimensiones ser creativos y poder generar nuevas y útiles características va a ser el reto.

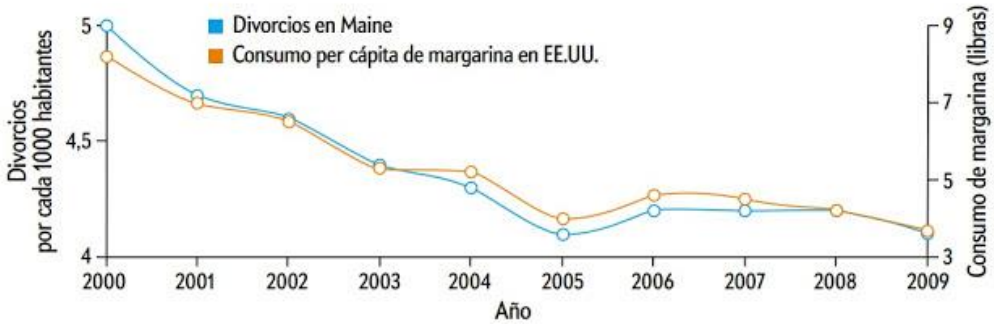
Por otro lado, si tenemos muchas features, podemos evaluarlas individualmente y pensar que algunas no aportan demasiado valor. Sin embargo, esas mismas características puede que sean imprescindibles si las consideramos en combinación con otras.

11. **Correlación no implica causalidad**

No nos fiemos de las correlaciones que podamos encontrar entre dos variables. Puede que no signifique nada. Mucho ojo con los mapas de calor que hagamos. Será importante investigar una posible correlación o un porqué.

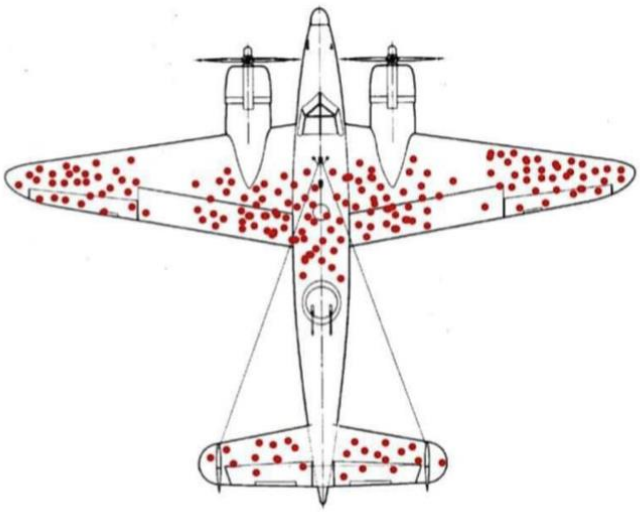
link





12. **Tener una visión global del problema**

Un error muy común es ver las cosas desde una perspectiva equivocada. Ejemplo de esto puede ser el estudio que hicieron la armada británica. En ella, analizaron los aviones que volvían de la guerra en la década de los 40 y analizaron las partes en las que los aviones estaban siendo más perjudicado.



Se tardaron meses en analizar dichas áreas para reforzar la armadura en esas zonas y se invirtió una gran cantidad de dinero y tiempo en la producción de aviones con las nuevas características en un tiempo récord.

¿El resultado? - Estadísticamente seguían volviendo el mismo número de aviones, pero mejor compuestos.

El problema era que los ingenieros aeronáuticos se estaban fijando en los datos equivocados.

"En vez de fijarse en las características de los aviones que no volvían, se enfocaron en los que sí."

Antes de ponernos a analizar los datos, tenemos que interceptar el problema.

13. **No todo es posible**

Una de las cosas más frustrantes es saber de que no todo es posible de realizar o por lo menos, aún no. Parece algo fácil y de sentido común, pero cuando entras dentro del mundo de la IA y escuchas las noticias, parece que sí.

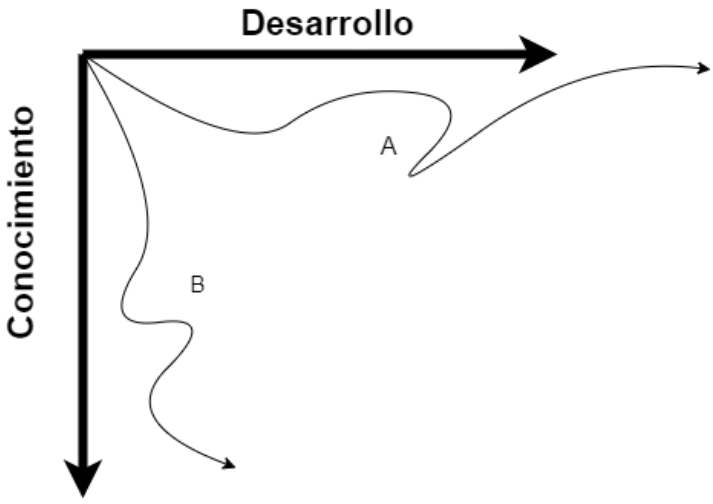
Al estar trabajando con series temporales, por ejemplo, vamos a ver muchas ecuaciones y aproximaciones, pero todas ellas están bien acotadas y no siempre funcionan (90 % del tiempo).

Por ello, hay que ir con pies de plomo y no prometer o creer en resultados que puede que nunca lleguen. Hay que intentar buscar el equilibrio entre tener los pies en la tierra y poder crear. No es fácil.

14. **Hay que aprender a nadar**

Entender la matemática está bien, el de la programación, la filosofía, la optimización, el de la IA…también. La lista se puede alargar hasta el infinito.

A la hora de desarrollar un proyecto, vamos a tener que buscar el equilibrio entre profundizar en un tema y finalizar el objetivo. Depende el área y la empresa el paradigma cambia por lo que los ejemplos A y B pueden bifurcarse. Es muy importante tenerlo claro si no nos queremos ahogar. No es sencillo.

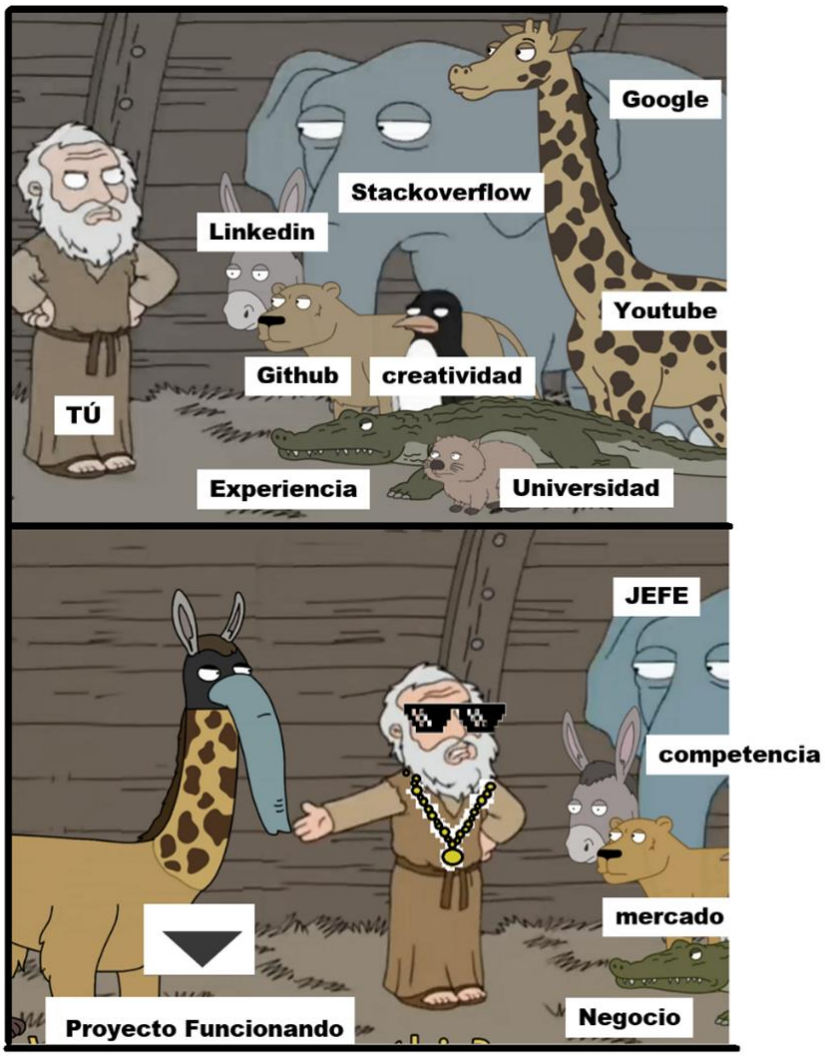


15. **Soy idiota**

Una de las bellezas de la IA es la comunidad que se ha creado. A diferencia de otras áreas el flujo de conocimiento al ser una tecnología nueva y en desarrollo es muy trasparente.

No dudes en pedir ayuda tanto en Github, stackoverflow, TheEgg, Linkedin…es muy gratificante ver que **NO ESTAS SOLO.** Te aseguro que te podrías sorprender con la de gente interesante y predispuesta a ayudar y a crear nos hemos encontrado.

Sigue a algún referente. Prueba escribirle si tienes alguna duda. Procesa el resultado.

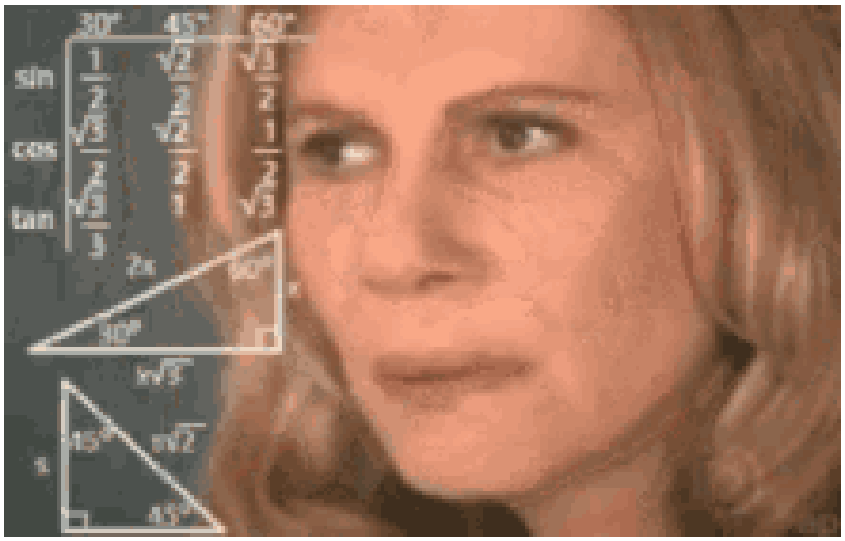


16. **Keep it Simple…**

Se que se ha repetido muchas veces, pero creo que no está de más. Antes de ir a por resultados complejos, probar lo básico. Una regresión logística o una clasificación. Una vez conseguidos unos resultados decentes con los datos de entrada y ver donde se puede mejorar, ya se mejorará.

Lo complejo se aplicará si lo básico no es suficiente, pero se tendrá una referencia de la que partir, si no…poder avanzar en horizontal en un suelo incierto será complicado.

link



Conclusión

Esperamos que esta información os sea de utilidad para que lo tengáis en cuenta en futuros proyectos. No es fácil adaptarlas a uno mismo, pero es importante que las vayamos asimilando y aplicando poco a poco.