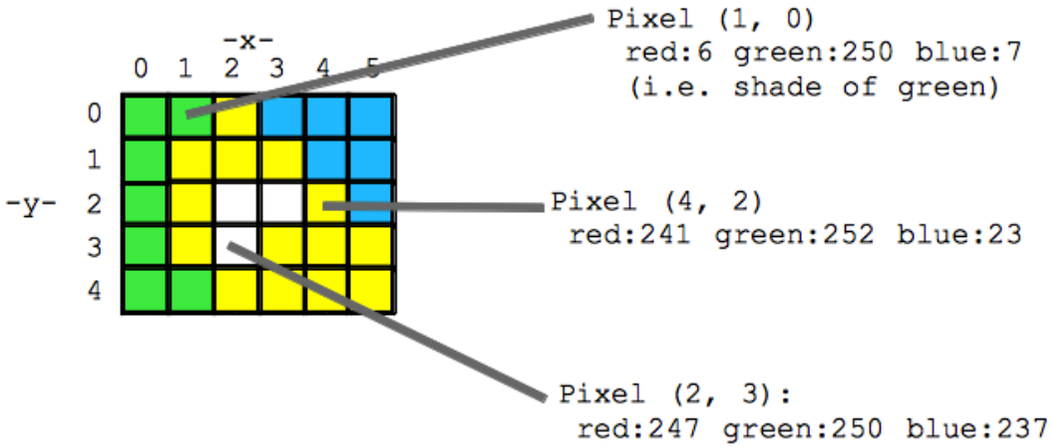
Big Data, IA y Machine Learning

GOBIERNO DE ESPAÑA
VICEPRESIDENCIA PRIMERA DEL GOBIERNO
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL
red.es
Centro de Referencia Nacional en Comercio Electrónico y Marketing CRN Digital
UNIÓN EUROPEA
Barrabés
The Valley
"El FSE invierte en tu futuro"
Fondo Social Europeo

Índice

1. Big Data

R0 G0 B0	R0 G0 B0	R55 G160 B146	R48 G182 B135	R39 G183 B119	R17 G119 B128	R11 G68 B134	R31 G43 B113	R0 G0 B0	R0 G0 B0
R0 G0 B0	R79 G132 B188	R58 G205 B142	R53 G240 B124	R86 G246 B142	R24 G42 B219	R1 G20 B199	R3 G69 B121	R23 G99 B66	R0 G0 B0
R77 G51 B216	R8 G231 B42	R133 G250 B178	R166 G208 B220	R145 G130 B249	R67 G57 B244	R7 G10 B228	R0 G90 B114	R4 G133 B39	R33 G71 B55
R68 G45 B228	R48 G185 B127	R164 G218 B211	R200 G192 B250	R167 G152 B251	R76 G65 B245	R9 G13 B231	R0 G91 B115	R0 G148 B21	R19 G89 B43
R45 G32 B226	R76 G118 B210	R8 G231 B42	R166 G156 B248	R141 G126 B249	R65 G55 B244	R7 G10 B227	R0 G90 B114	R0 G148 B21	R15 G96 B39
R12 G13 B209	R22 G22 B237	R61 G148 B174	R133 G250 B178	R64 G85 B220	R15 G16 B235	R0 G5 B209	R0 G55 B126	R6 G83 B75	R29 G64 B81
R10 G14 B178	R1 G6 B212	R7 G120 B133	R8 G231 B42	R6 G192 B68	R33 G156 B131	R0 G10 B176	R1 G6 B150	R20 G47 B117	R41 G67 B85
R25 G29 B133	R3 G6 B175	R0 G54 B146	R0 G179 B36	R0 G190 B27	R0 G144 B48	R1 G27 B131	R14 G37 B122	R40 G82 B105	R55 G64 B70
R0 G0 B0	R19 G21 B132	R4 G9 B146	R1 G104 B58	R0 G115 B46	R6 G29 B122	R20 G47 B118	R40 G83 B107	R56 G77 B83	R0 G0 B0
R0 G0 B0	R0 G0 B0	R40 G40 B94	R19 G34 B102	R15 G20 B112	R26 G43 B107	R42 G66 B94	R58 G69 B75	R0 G0 B0	R0 G0 B0



1.1 Big Data y Computer Science

Antes de hablar de conceptos como big data, machine learning, inteligencia artificial o ciencia de datos, es imprescindible conocer el término de ‘computer science’ o ‘ciencias de la computación’.

En este sentido, según la Wikipedia:

*“Las **ciencias de la computación** o **ciencias de la informática** son las ciencias formales que abarcan las bases teóricas de la información y la computación, así como su aplicación en sistemas computacionales. El cuerpo de conocimiento de las ciencias de la computación es frecuentemente descrito como el estudio sistemático de los procesos algorítmicos que describen y transforman información: su teoría, análisis, diseño, eficiencia, implementación, algoritmos sistematizados y aplicación”*

Así pues, como veremos durante las próximas diapositivas, el Big data no es más que un campo dentro de todo lo que conocemos como ciencias de la computación o computer science

1.2 Redefiniendo Big Data

Ya hemos hablado de la importancia de 3 elementos básicos que conforman el Big Data:

- Velocidad a la que se consume la información
- Variedad de información
- Volumen de información

Pero todo esto, no nos sirve de nada si no incorporamos un componente de:

VALOR

Si los datos no sirven para aprender, descubrir o analizar, todo lo anterior no sirve de nada.

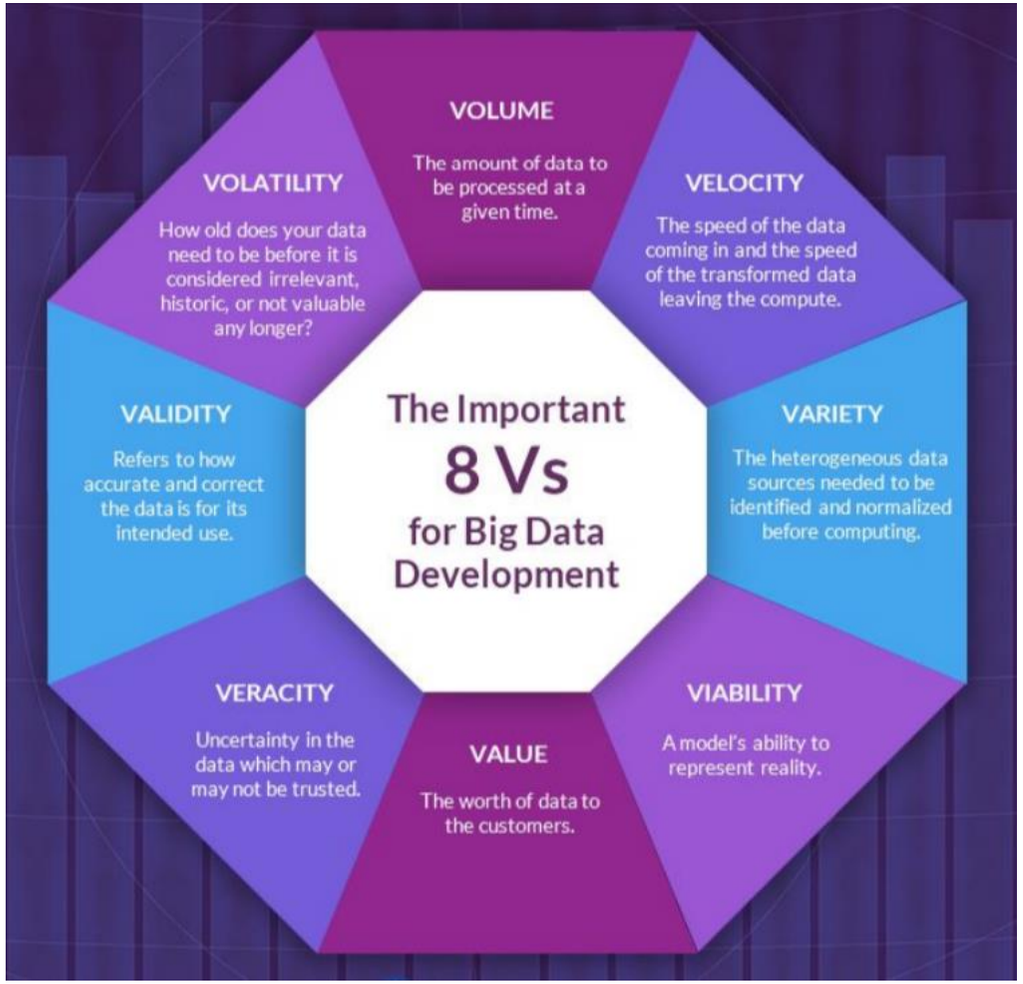
Recoger información es importante pero por sí solos, los datos no nos ayudarán en nada. Es importante no caer en el *hype* del marketing y la innovación asociado al término de Big Data. El big data es algo mucho más amplio y profundo que solamente extraer y almacenar datos.

1.2 Redefiniendo Big Data

¿4 V's u 8 V's?

- ✓ Volumen
- ✓ Variedad
- ✓ Velocidad
- ✓ Valor

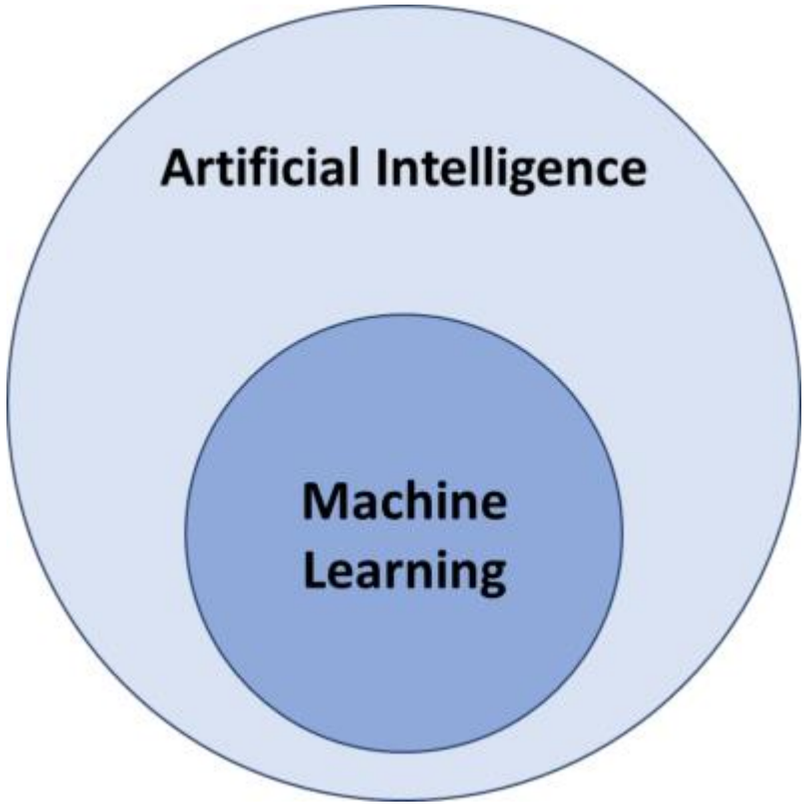
- ✓ **Viabilidad**
- ✓ **Veracidad**
- ✓ **Validez**
- ✓ **Volatilidad**



2. Data Science: AI y Machine Learning

2.1 AI y ML: ¿Son lo mismo?

Técnicamente diríamos que el machine learning y la inteligencia artificial son prácticamente lo mismo. Realmente, conviven en distintos niveles uno dentro del otro. A grandes rasgos podríamos concluir que:



Inteligencia artificial

Entendemos como IA el hecho de que las máquinas imiten comportamientos parecidos al de un ser humano. Puede abarcar muchísimas formas y aplicaciones.

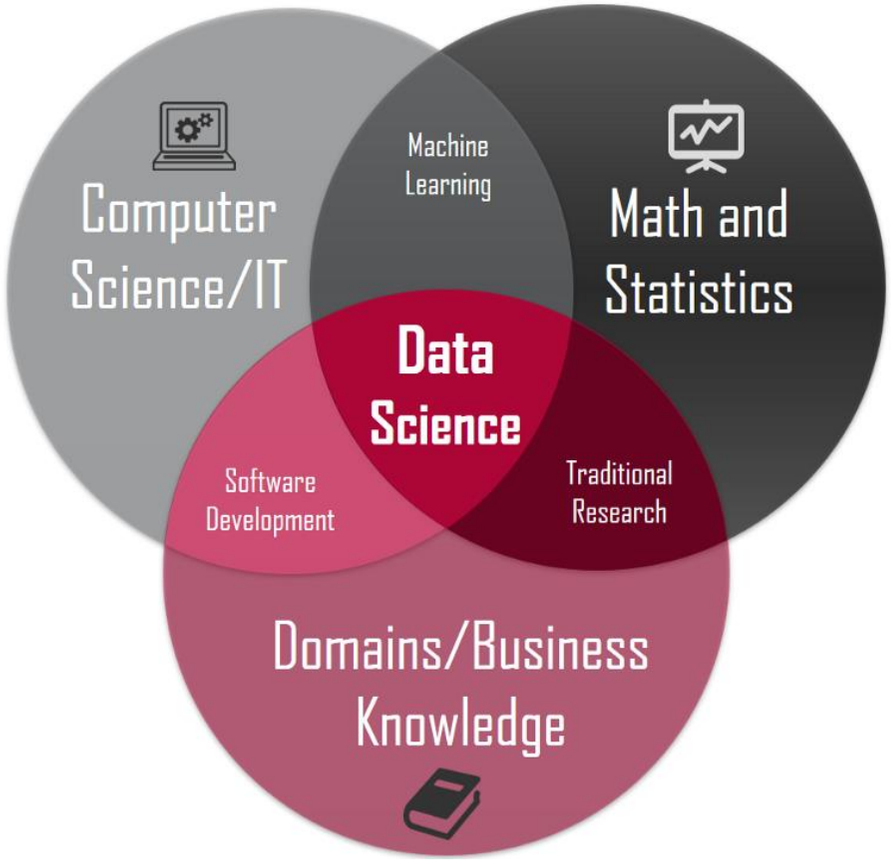
Machine learning

Lo entendemos como un subconjunto de inteligencia artificial en el cual las máinas aprenden y se entrenan para comportarse como humanos y hacer predicciones.

Deep learning

2.1 AI y ML: ¿Son lo mismo?

Entonces…¿cómo se relacionan estos términos con las ciencias de la computación, el software, la matemática, la estadística o la ciencia de datos? Fijémonos en la siguiente imagen (*cotejar con lo visto en la sesión 1*):



2.2 ¿Puede un ordenador pensar?

A menudo escuchamos o nos hacemos preguntas de este tipo cuando nos relacionamos u oímos términos como inteligencia artificial. Y a esta pregunta podemos añadir varias reflexiones:

UN ORDENADOR **NO** PUEDE PENSAR COMO UN HUMANO.
UN ORDENADOR **SÍ** PUEDE IMITAR EL COMPORTAMIENTO DE UN HUMANO

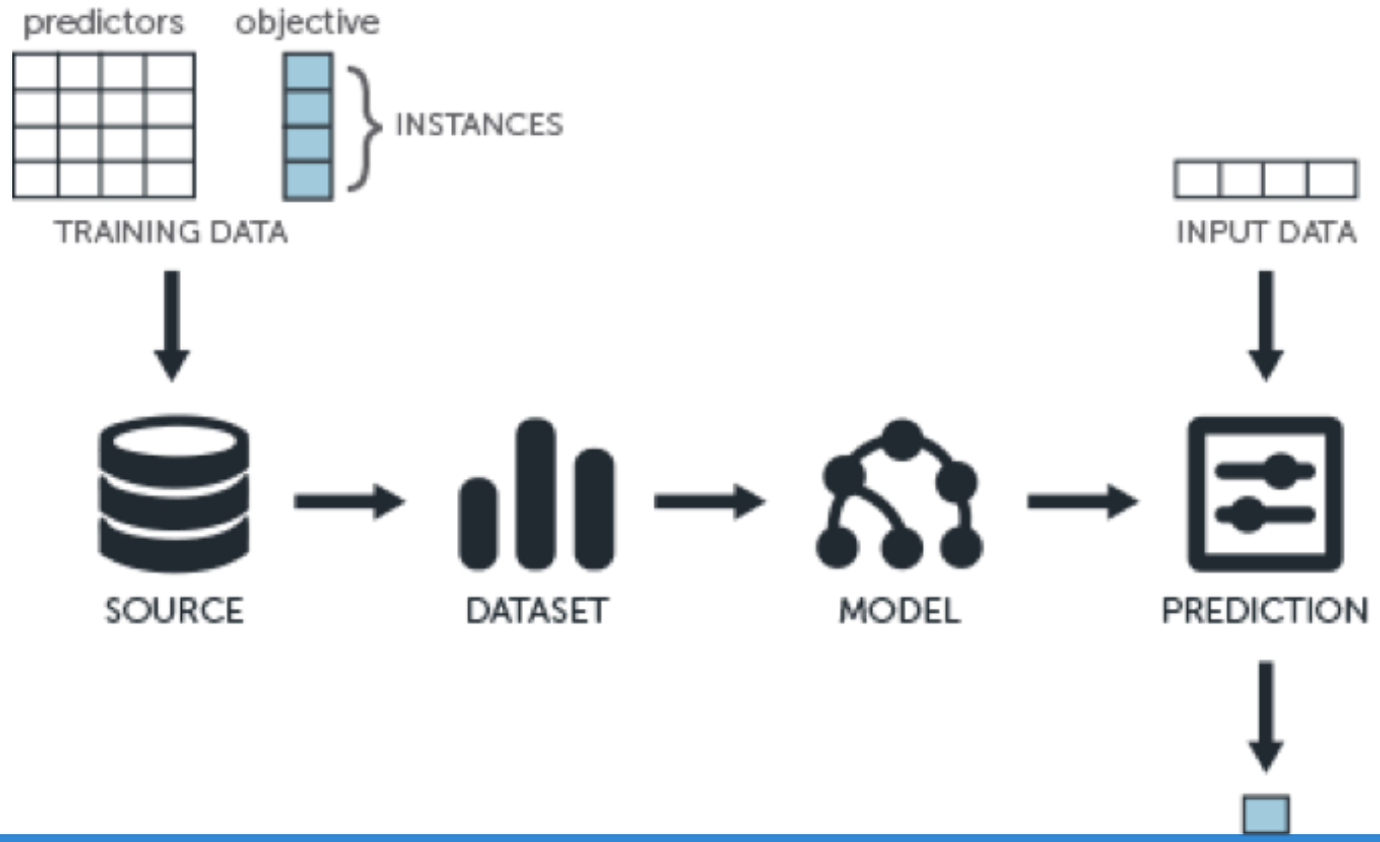
En este sentido, la pregunta clave es:
¿REALMENTE IMPORTA SI UN ORDENADOR PUEDA PENSAR COMO UN HUMANO?

- ¿Puede un submarino bucear tal como lo hace un humano o un pez?
- ¿Puede un avión volar tal como lo hace una golondrina?
- ¿Puede un transatlántico nadar como un deportista profesional?

LO QUE IMPORTA NO ES SI PUEDE PENSAR. LO IMPORTANTE ES SI LO PUEDE CONSEGUIR

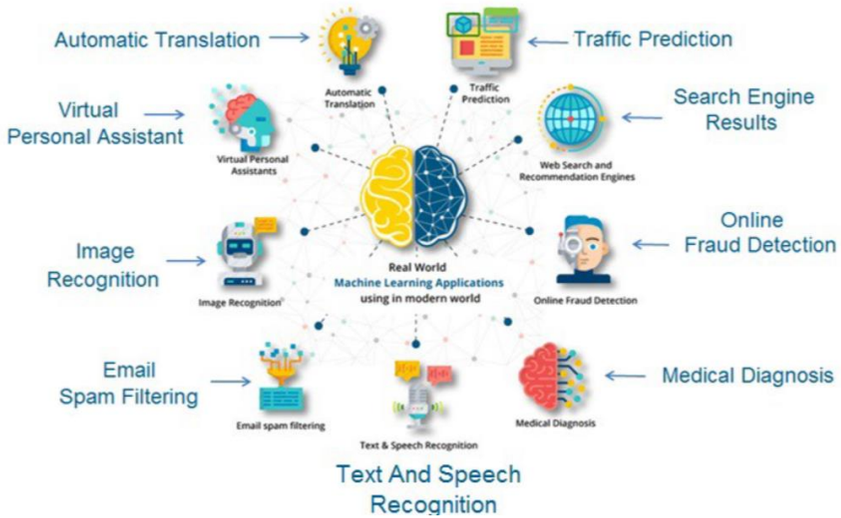
2.3 ML: Carácter predictivo

Cuando hablamos de machine learning, lo que realmente aporta valor e importancia a su existencia y desarrollo no es más que su capacidad para poder aplicar **modelos predictivos**. Esa es la auténtica revolución de esta disciplina, enmarcada dentro de la amplia inteligencia artificial.



2.4 Aplicaciones ML en el día a día

- Coches autónomos
- Reconocimiento facial
- Motores de recomendación
- Detección de fraude en tarjetas de crédito
- Abandono de clientes
- Filtros de spam en los emails
- Categorización y segmentación
- Análisis de emociones
- Personalización de experiencias de clientes
- Predicciones meteorológicas
- Diagnósticos médicos
- Reconocimiento de voz

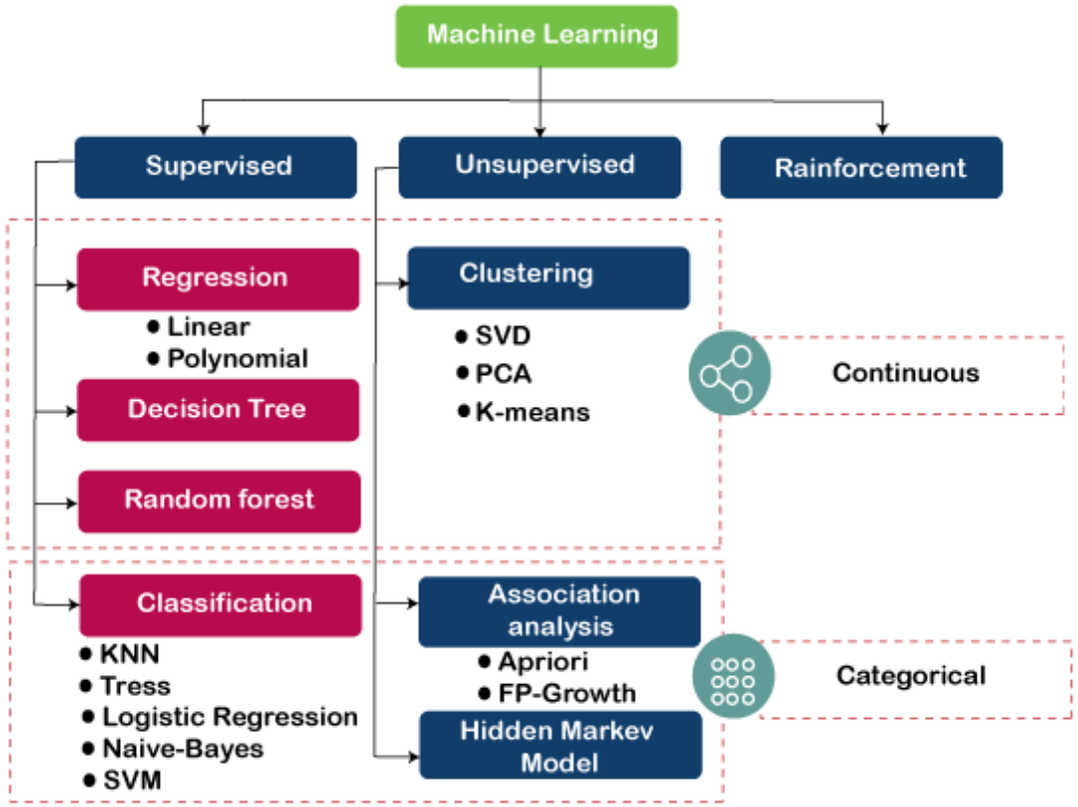


2.5 La importancia de los algoritmos

El machine learning basa prácticamente todo su potencia en la aplicación de algoritmos para aplicar el aprendizaje automático y desarrollar así modelos predictivos.

En este sentido, es sumamente importante la participación de perfiles con conocimientos matemáticos y estadísticos para diseñar estos modelos de aprendizaje y convertirlos en usables y accesibles para la industria.

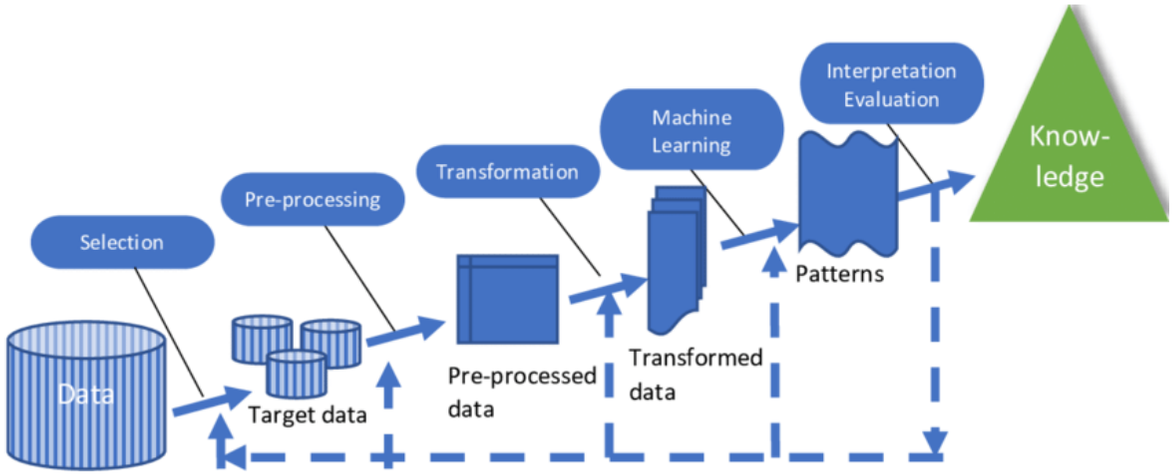
Estos algoritmos pueden estar basados en aprendizaje supervisado o aprendizaje no supervisado como veremos en las siguientes diapositivas.



2.6 Proceso de Machine Learning

Aunque este proceso lo veremos con mayor profundidad en el siguiente capítulo del temario, podríamos decir que todo proyecto de machine learning está basado en la aplicación de estas fases o pasos.

1. **Responder a la pregunta adecuada**
2. **Identificar y preparar los datos correctamente**
3. **Estudiar y aplicar la mejor solución logarítimica**
4. **Evaluar y ajustar el modelo para su correcta eficiencia**
5. **Uso de la solución y presentación del modelo**



2.7 Aprendizaje supervisado vs no supervisado

El aprendizaje supervisado es uno de los tipos o modelos más utilizados dentro del machine learning. Para ello, necesitamos contar con unas entradas y salidas de la información sobre la cual vamos a aplicar el aprendizaje automático.

Normalmente, encontramos dos tipos de aprendizaje supervisado:

Clasificación:
Elige entre una lista de opciones previamente definidas y limitada. Por ejemplo elegir un número entre el 0 y el 9. Se puede utilizar en los filtros de spam de los emails.

Regresión:
El objetivo de este algoritmo es predecir números reales o números con infinitas posibilidades. Por ejemplo, lo podríamos aplicar para predecir el precio de un inmueble.

2.7 Aprendizaje supervisado vs no supervisado

Por su parte, en el aprendizaje NO supervisado dentro del machine learning, los algoritmos se aplican para aprender de datos con elementos no etiquetados **buscando** patrones o relaciones entre ellos. En este caso no necesitarían delimitar el número de entradas y salidas.

Existen dos tipos de algoritmos para Machine Learning no supervisado:

- **Clustering**: clasifica en grupos los datos de salida. Es el caso de las segmentaciones de clientes según qué hayan comprado.

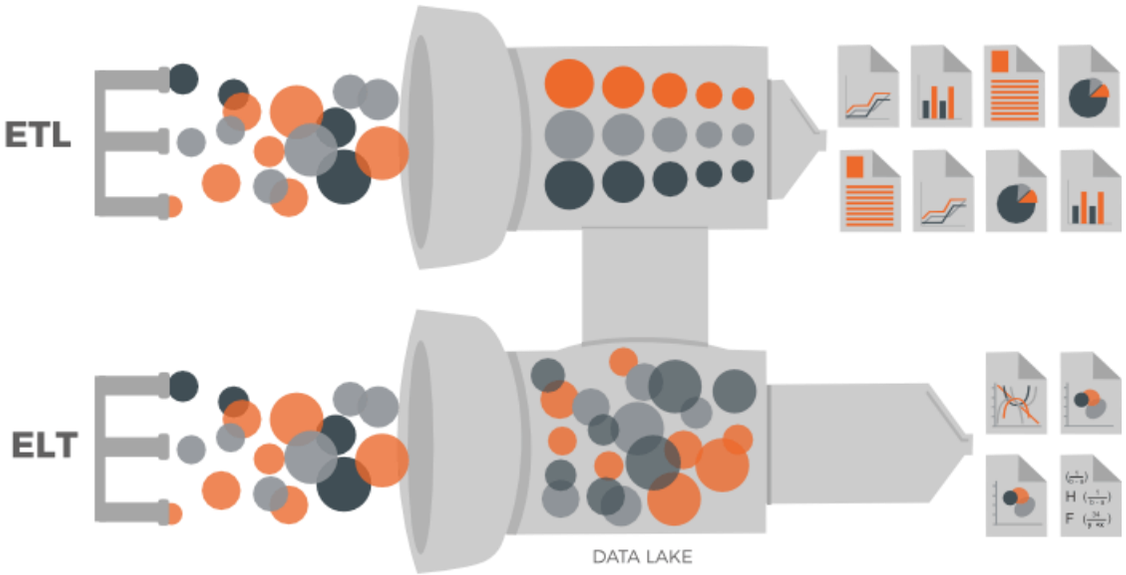
- **Asociación**: descubre reglas dentro del conjunto de datos. Por ejemplo, aquellos clientes que compran un coche también contratan un seguro, por lo que el algoritmo detecta esta regla.

Este tipo de aprendizaje es el que se utiliza por ejemplo en sistemas de recomendación de plataformas de streaming bajo demanda (netflix, hbo, amazon, etc.)

2.8 ELT y Data Lakes

ELT (Extraer, cargar, transformar) es un método diferente de acercarse al flujo de datos, en el que los datos extraídos se cargan primero en el sistema de destino. Las transformaciones se realizan después de que carguemos los datos en el almacén de datos.

Los datos primero se copian en el **data lake** y luego se transforman in situ. Funciona bien cuando el sistema objetivo es lo suficientemente potente como para manejar transformaciones a gran escala. ELT generalmente se usa con bases de datos NOSQL, un dispositivo de datos o una instalación en la nube.



VER VÍDEO DATADEMIA

2.9 Ejemplos y reflexiones

THISPERSONDOESNOTEXIST

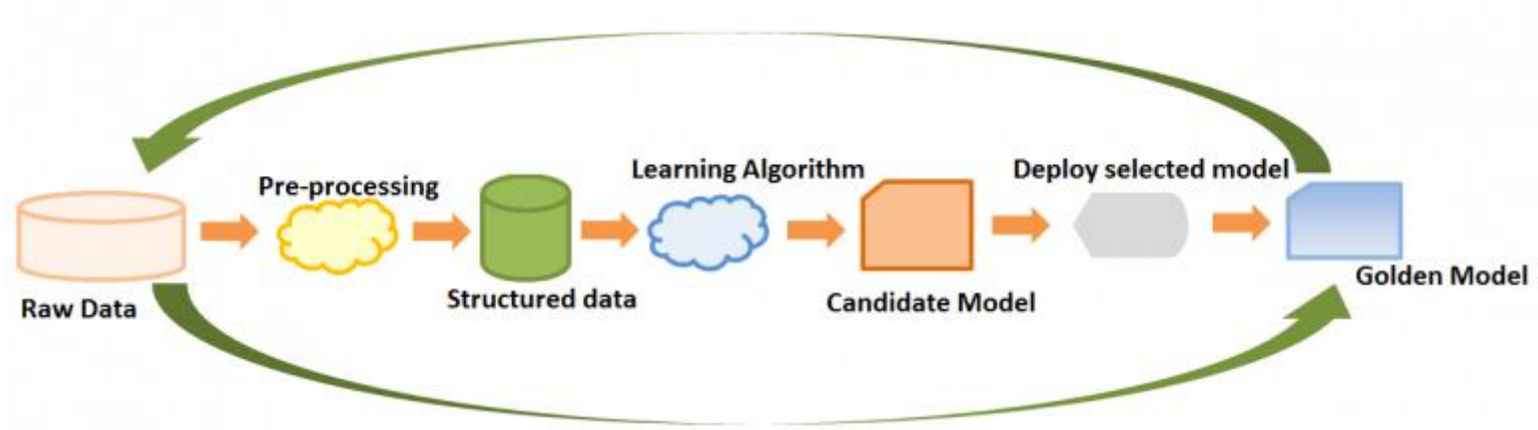
ALGORITMO DE NETFLIX ¿QUÉ TIENE EN CUENTA?

MEDICINA PREDICTIVA

2.10 Qué resolvemos con ML

Llegados a este punto debemos tener en cuenta que el machine learning:

- Se aplica a problemas en los que se ha de aprender a base de datos o ejemplos.
- Se aplican modelos de clasificación, regresión o clustering entre otros.
- No es necesario que los datos sean obligatoriamente estructurados.
- Un elemento clave en la utilización del machine learning es la elección del algoritmo que va a encargarse de aprender y dotar de sentido su aplicación
- Si no tienes los inputs necesarios y las herramientas necesarias no serás capaz de sacar datos relevantes ni resolver el problema que se plantea.

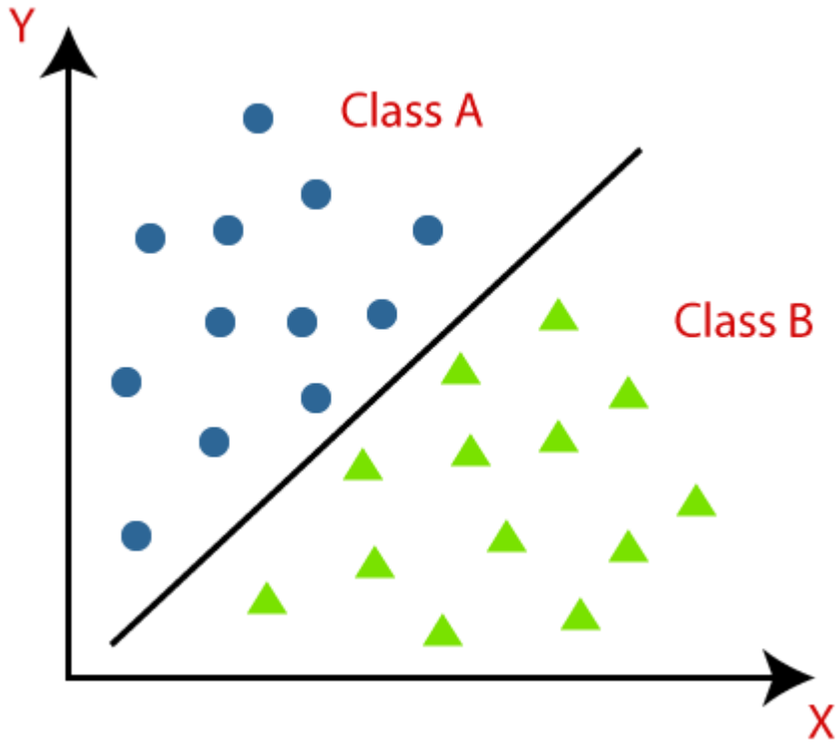


2.11 Clasificación

La clasificación es una subcategoría del aprendizaje supervisado en la que el objetivo es predecir las etiquetas de clase categóricas (discreta, valores no ordenados, pertenencia a grupo) de las nuevas instancias, basándonos en observaciones pasadas.

Clasificación Binaria: Es un tipo de clasificación en el que tan solo se pueden asignar dos clases diferentes (0 o 1). El ejemplo típico es la detección de email spam, en la que cada email es: spam → en cuyo caso será etiquetado con un 1 ; o no lo es → etiquetado con un 0.

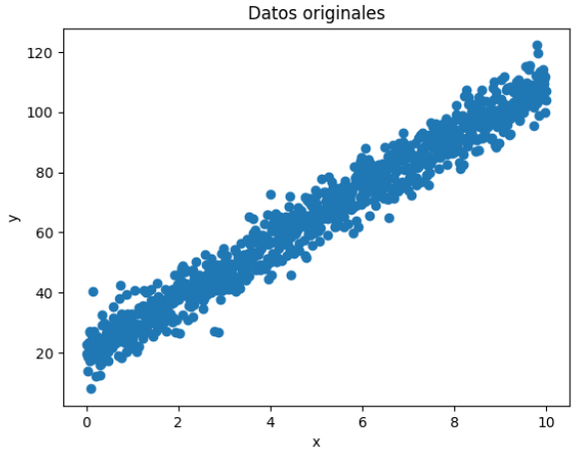
Clasificación Multi-clase: Se pueden asignar múltiples categorías a las observaciones. Como el reconocimiento de caracteres de escritura manual de números (en el que las clases van de 0 a 9).



2.12 Regresión lineal

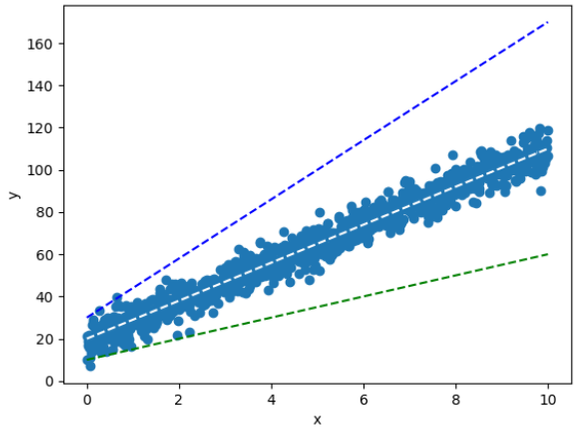
La Regresión Lineal es un procedimiento que permite encontrar la línea recta que mejor representa la relación entre dos variables.

Supongamos que tenemos una serie de datos como la mostrada en la figura, en donde la variable independiente es *x* y la variable dependiente es *y*:



Se observa en esta gráfica que la relación entre las dos variables es lineal, es decir que un incremento en la variable *x* genera un incremento proporcional en la variable *y*.

El objetivo de la Regresión Lineal es entonces encontrar la línea recta que mejor se ajusta a los datos.



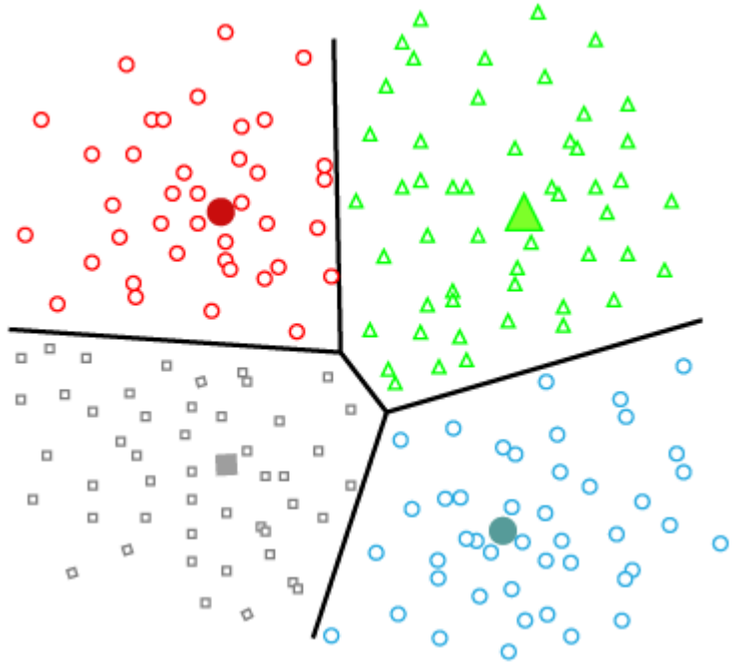
2.13 Clustering

Formalmente, podemos decir que **la agrupación en clústeres el proceso de dividir todos los datos en grupos, también conocidos como clústeres, basados en los patrones de los datos**.

El Clustering o agrupamiento es una de las formas de aprendizaje no supervisado más utilizada.

Es una gran herramienta para dar sentido a los datos no etiquetados y para agrupar datos en grupos similares.

Un algoritmo de agrupamiento puede descifrar estructuras y patrones en un conjunto de datos que no son aparentes para el ojo humanos.



3. ML: Entendiendo el proceso

3.1 La pregunta adecuada

Antes de aplicar cualquier proceso de machine learning o, en general, cualquier proyecto que involucre la extracción, procesamiento y análisis de datos, deberíamos hacernos algunas preguntas que no son en absoluto baladí.

- ¿Qué queremos resolver y qué esperamos obtener de ello?
- ¿Qué recursos tecnológicos tenemos? ¿Contamos con los perfiles y equipo adecuado?
- ¿Cuáles son nuestras fuentes de datos? ¿Son estructuradas o no estructuradas?
- ¿Estamos cumpliendo la legalidad vigente en el uso de estos datos? ¿Nos enfrentamos a información personal sensible? ¿Necesito anonimizar o aplicar una máscara a mis datos?
- ¿Estamos aplicando un sesgo discriminatorio en nuestro enfoque?
- ¿Qué va a definir que el modelo es exitoso o no? ¿Qué porcentaje de error podemos permitirnos?

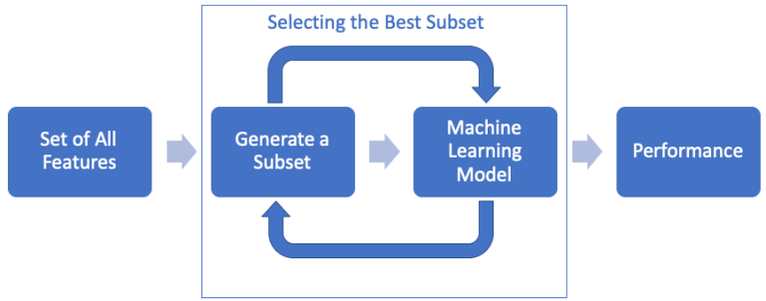
3.2 Identificar los datos y prepararlos

El segundo paso consiste en conocer los datos o la fuente de datos a utilizar para poder organizarla y prepararla conscientemente para su utilización en el modelo de machine learning.

¿Qué coste tiene acceder a los datos? ¿Vamos a necesitar datos de fuentes externa o son de nuestra propiedad? ¿Qué velocidad de obtención y procesamiento podemos alcanzar?
¿Qué variedad de datos vamos a manejar y cuántos formatos vamos a tener que procesar?

A la hora de preparar nuestros datos:

¿Están todos los datos recogidos o hay datos extraviados?
¿Los datos están correctamente estructurados? ¿Hay datos repetidos?
¿Los datos son consistentes?



3.3 Aplicación del algoritmo correcto

El éxito de tu modelo de machine learning va a depender **enormemente** del acierto que se tenga a la hora de relacionar el algoritmo elegido con los datos existentes.

Árboles de decisión

No necesita mucho entrenamiento. Pueden ser de clasificación o de regresión.

Naive Bayes

Es un modelo puramente probabilístico. ¿Qué probabilidad tengo de ver la opción A cuando ocurre B?

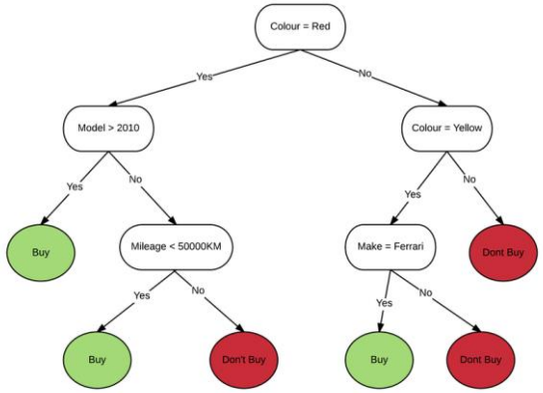
Redes neuronales (VIDEO).

Funcionan como un sistema de neuronas en el cerebro. Todas tienen entradas y salidas y están conectadas entre ellas distribuyendo un "**peso**" determinado entre cada conexión.

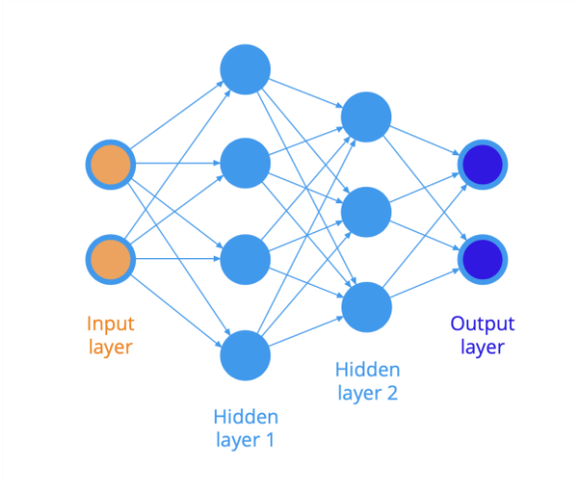
KNN

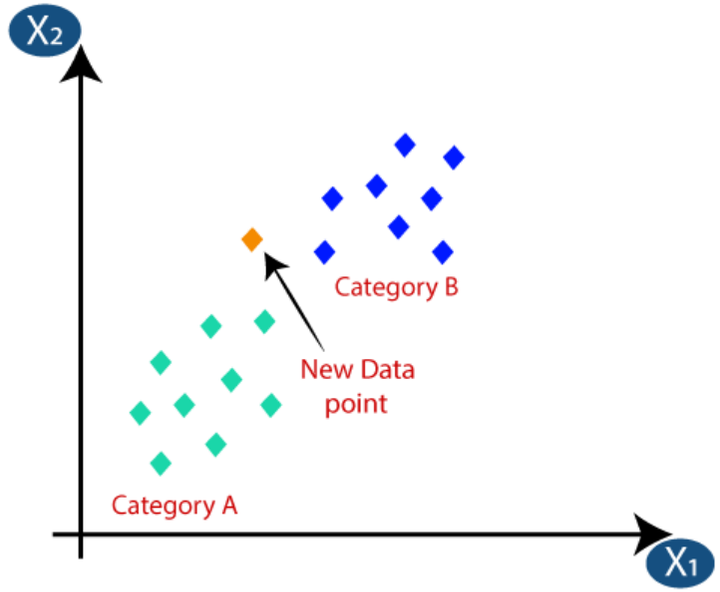
Sirve esencialmente para clasificar valores buscando los puntos de datos "más similares" (por cercanía) aprendidos en la etapa de entrenamiento

3.3 Aplicación del algoritmo correcto



$$p(C|F_1,\ldots,F_n) = \frac{p(C)\,p(F_1,\ldots,F_n|C)}{p(F_1,\ldots,F_n)}.$$





3.4 Evaluación y ajuste de modelo

Es el momento de evaluar el rendimiento de tu modelo de machine learning, y en este momento es especialmente crucial la revisión y la corrección del modelo atendiendo a las preguntas que has planteado y respondido en el primer paso del proceso.

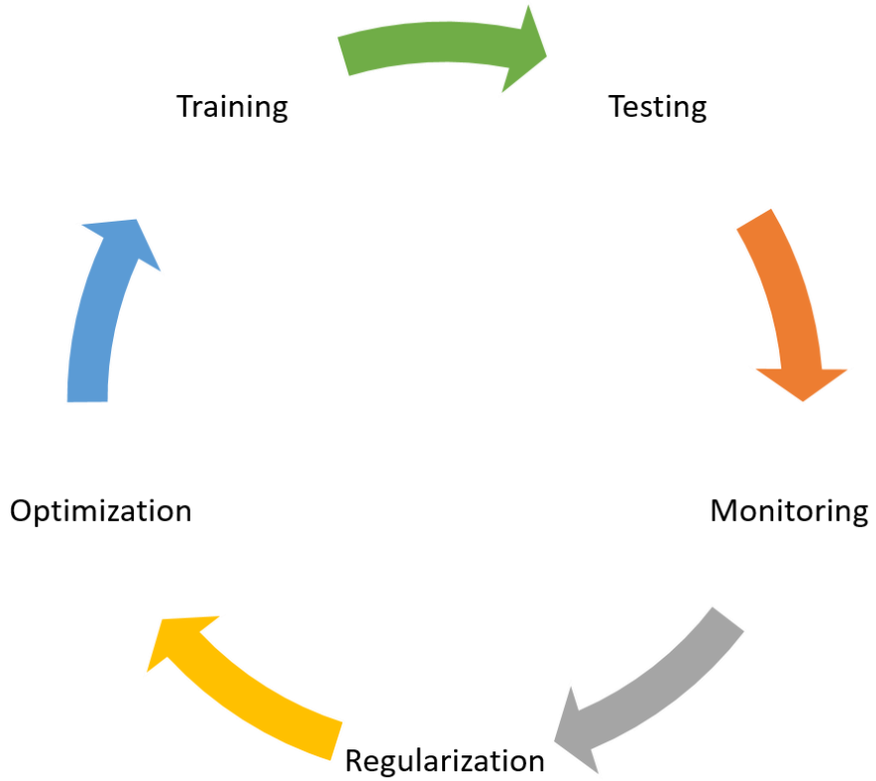
- ¿Necesito ajustar los datos que estoy dando a mi modelo?
- ¿El algoritmo que he elegido satisface y soluciona mi problema?
- ¿Es escalable a la inserción y el análisis de un mayor volumen de datos en el futuro?

- ¿Me está dando resultados muy simples? ¿Puedo percibir la inteligencia aplicada?
- ¿Los resultados son muy complejos? ¿Se aleja demasiado de la realidad y no es comprensible o parecido a un proceso de aprendizaje humano?

SI NO RESPONDES TU PRINCIPAL PREGUNTA O TUS RESULTADOS NO SON LOS ESPERADOS, ES EL MOMENTO DE VOLVER ATRÁS Y MODIFICAR LOS PASOS 2 Y 3.

3.5 Uso y presentación de modelo

Es el momento de empezar a utilizar tu modelo y presentarlo a tu cliente o implementarlo en tu organización. La perspectiva y el juicio humano será la que finalmente confirmará que el modelo funciona y soluciona el problema que se había planteado.



4. Ejercicios

4. Ejercicios

E.1
Visionado de vídeos en los anexos

E.2
Con la ayuda de la extensión ‘Data Miner’, haz un scraping y descarga los datos de las casas en venta en tu ciudad existentes en la web pisos.com. Ahora, haz una regresión múltiple en una hoja de Excel para predecir un precio en base a m2 y/o número de habitaciones.

E.3
Introducción a Deep Learning y Python: Consulta y explicación del anexo Google Collab para convertir Celsius a Fahrenheit.

E.4
Separar la clase en dos grupos y aportar al menos 5 argumentos en contra y a favor de usar el big data en el sector de los seguros.

E.5
Examen final

"El FSE invierte en tu futuro"

Fondo Social Europeo