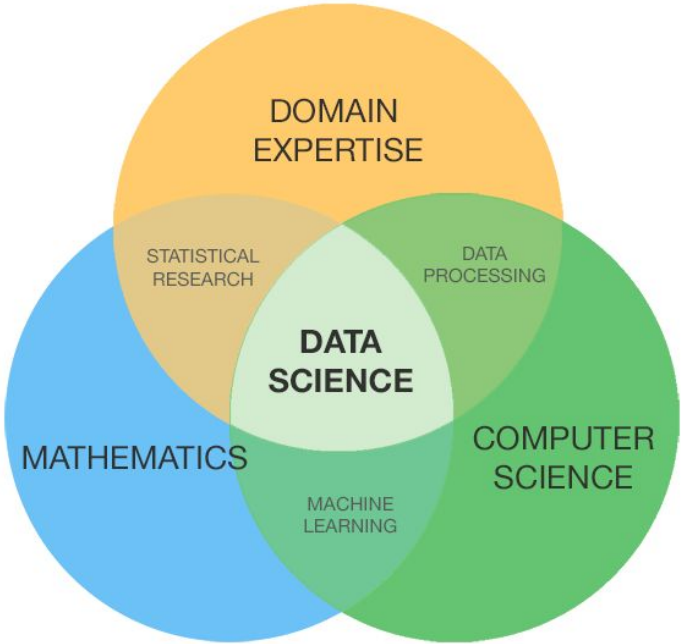
Aprendizaje Automático

Fabián Villena

Ciencia de datos

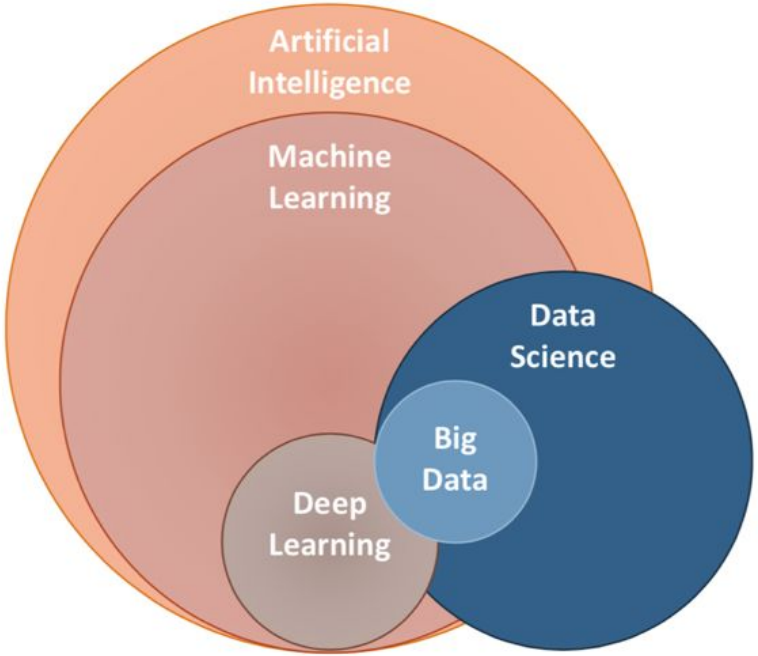
La ciencia de datos es un concepto que contempla el uso de métodos estadísticos, aprendizaje de máquinas y el **conocimiento específico de un área** para entender fenómenos desde los datos.



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

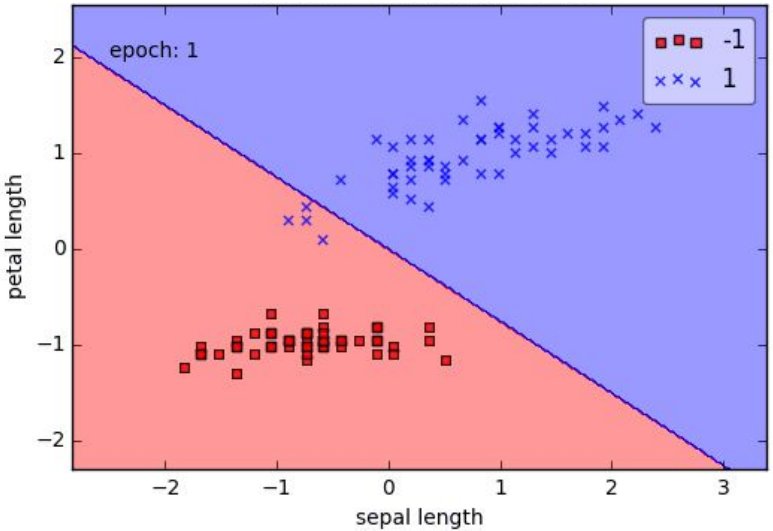
Inteligencia Artificial

- Es la clasificación más general de un conjunto de metodologías para la **generación de modelos**.
- Estos modelos pueden ser utilizados para la toma de decisiones.
- El método más básico es la generación de una serie de reglas para modelar un fenómeno.



Aprendizaje Automático

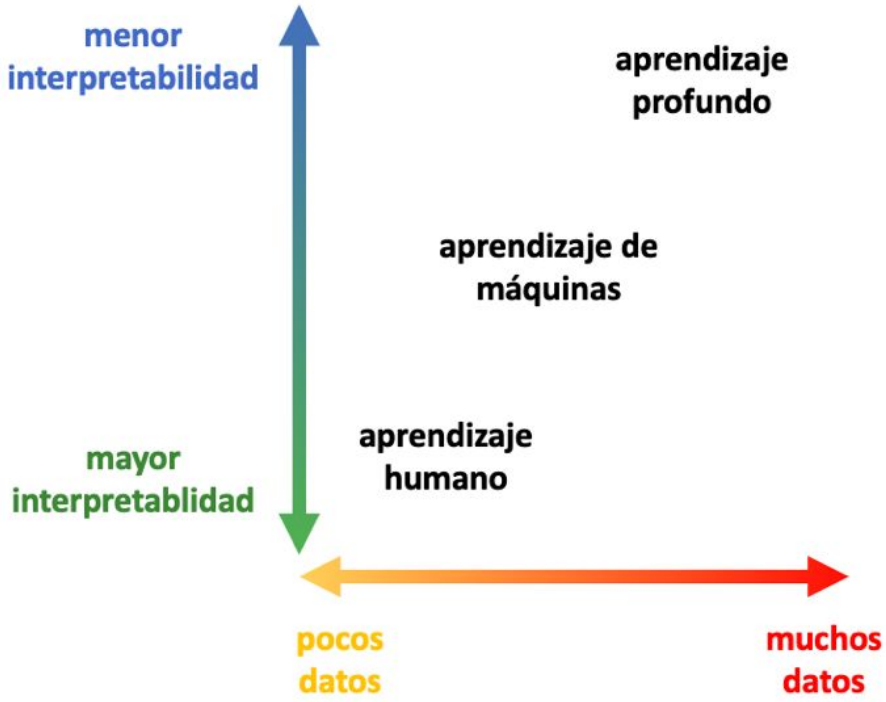
- El aprendizaje automático es el estudio de algoritmos que automáticamente **mejoran su rendimiento a través de la experiencia**.
- Estos algoritmos construyen modelos basados en datos de muestra con la intención de realizar predicciones sin ser explícitamente programados para hacerlo.



Interpretabilidad y cantidad de datos

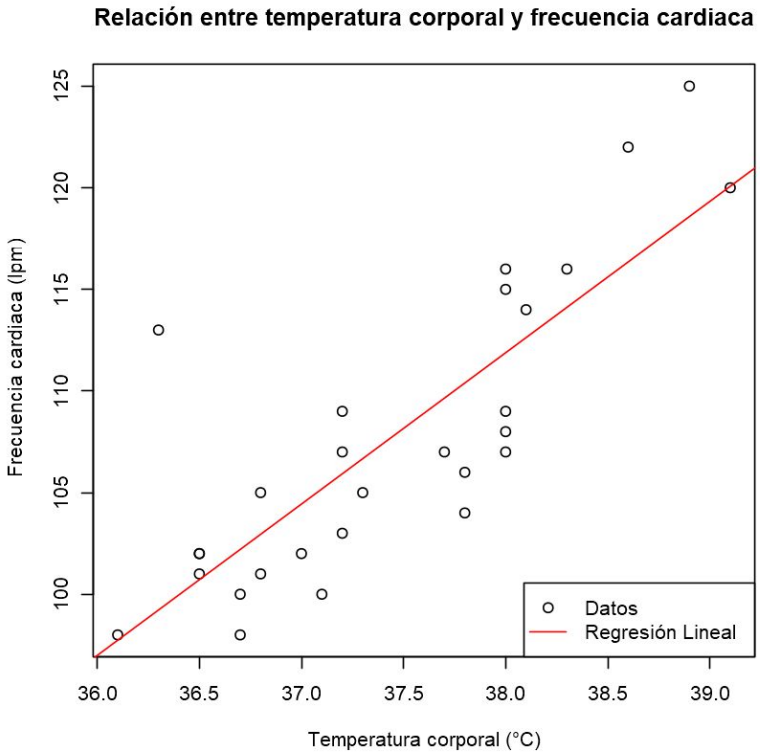
La interpretabilidad es la posibilidad de comprender un modelo y presentar las decisiones que se toman en una forma entendible por los humanos.

Modelos más potentes tienden a necesitar más datos y además su interpretabilidad decrece.



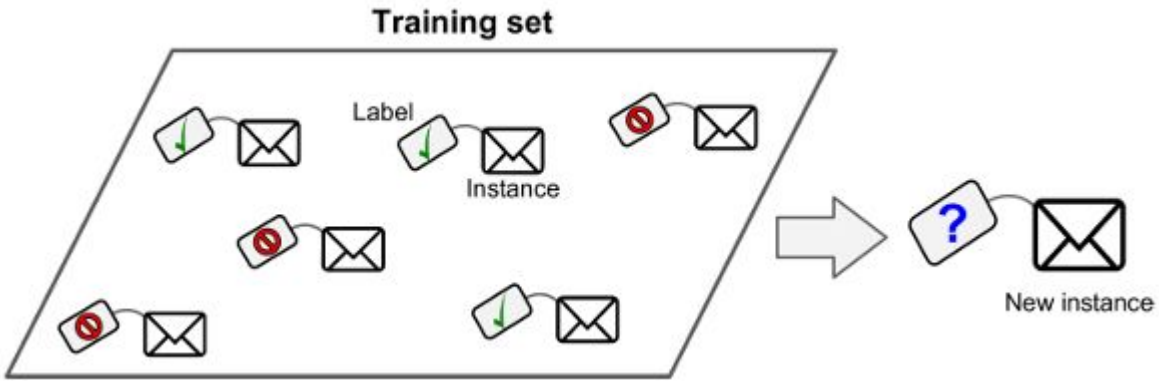
Aprendizaje supervisado

Este tipo de aprendizaje tiene la tarea de **aprender una función** que estime una salida dada una serie de características. Se infiere una función desde **datos de entrenamiento previamente etiquetados**.



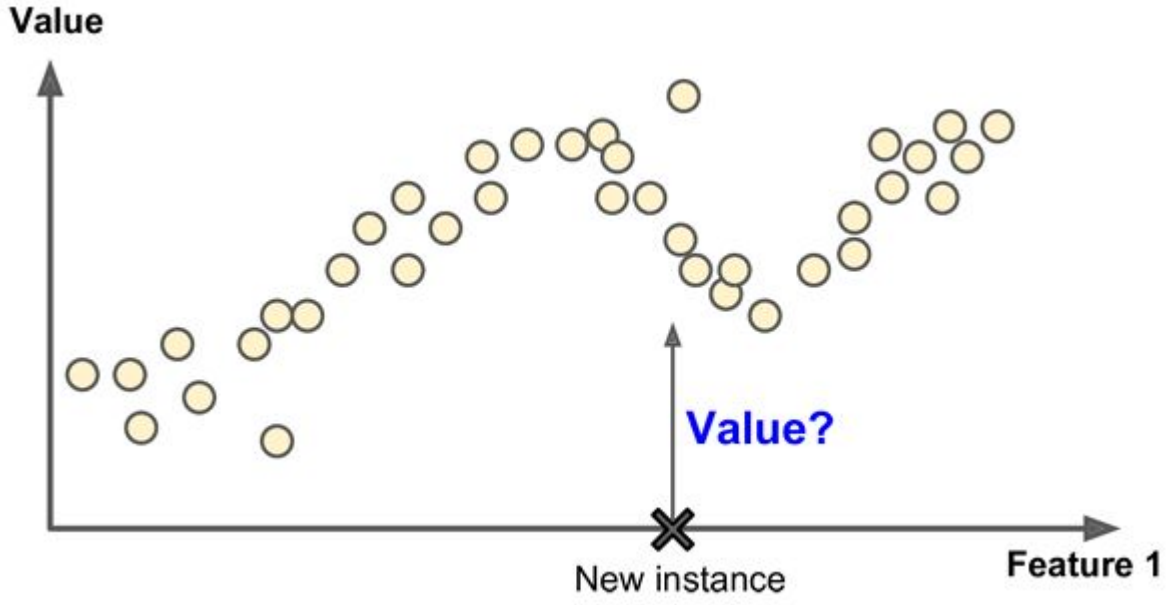
Clasificación

Esta tarea es una de las más comunes junto a la regresión. En esta tarea **buscamos predecir la clase** a la cual pertenece un objeto.



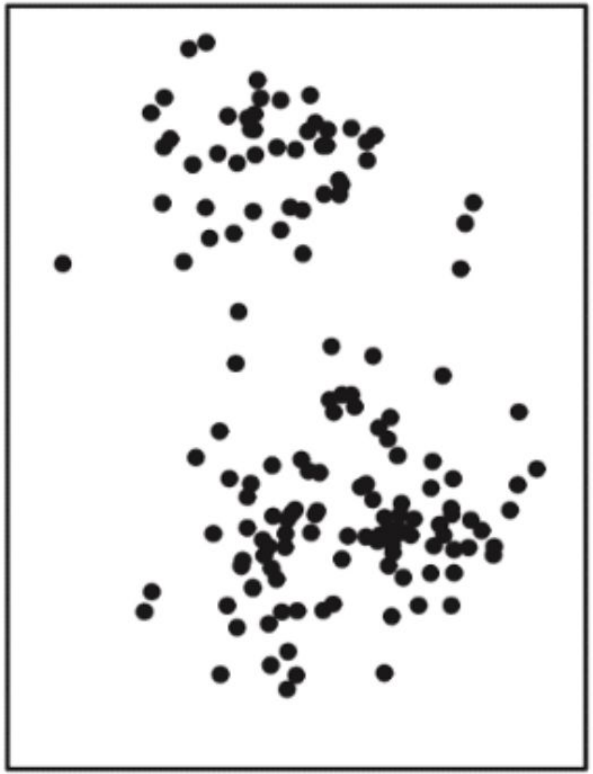
Regresión

Esta tarea es una de las más comunes junto a la clasificación. En esta tarea **buscamos predecir un valor numérico** asociado al objeto.



Aprendizaje no supervisado

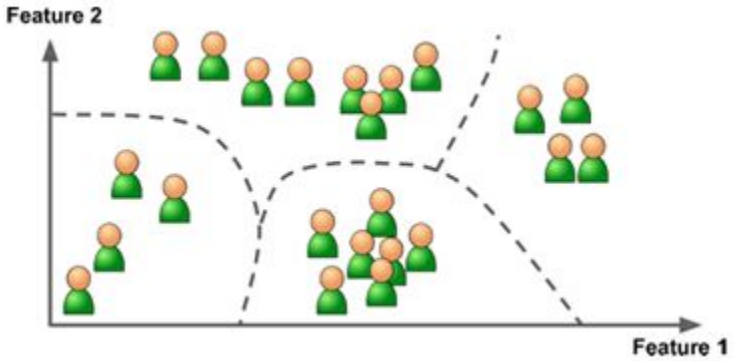
En este tipo de aprendizaje de máquinas se busca **detectar patrones en conjuntos de datos**, sin tener etiquetas de datos previas. Principalmente se utiliza en etapas exploratorias de investigación.



Clustering

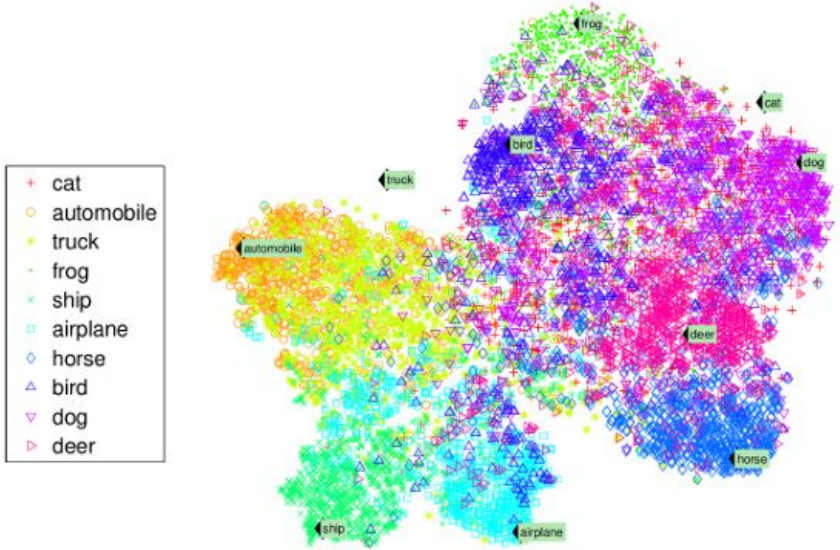
Es la tarea que busca grupos de objetos similares dentro de un conjunto de datos.

En ningún momento se le comunica al algoritmo a qué grupo pertenece cada objeto (porque no lo sabemos), sino que detecta agrupaciones sin supervisión.



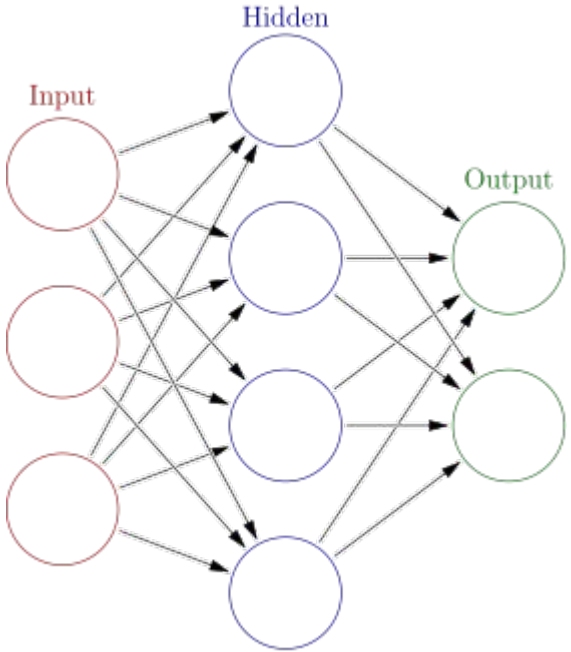
Reducción de dimensionalidad

La tarea de reducción de dimensionalidad tiene como objetivo simplificar los datos sin perder información significativa. Una forma de realizarlo es uniendo varios atributos correlacionados en uno sólo.



Deep Learning

- Estos métodos están dentro del aprendizaje automático.
- Se realiza el ajuste del modelo a través de sus unidades mínimas llamadas **neuronas**, las cuales están conectadas entre sí.
- Estas conexiones están determinadas por pesos que se ajustan para optimizar el rendimiento.



Datos

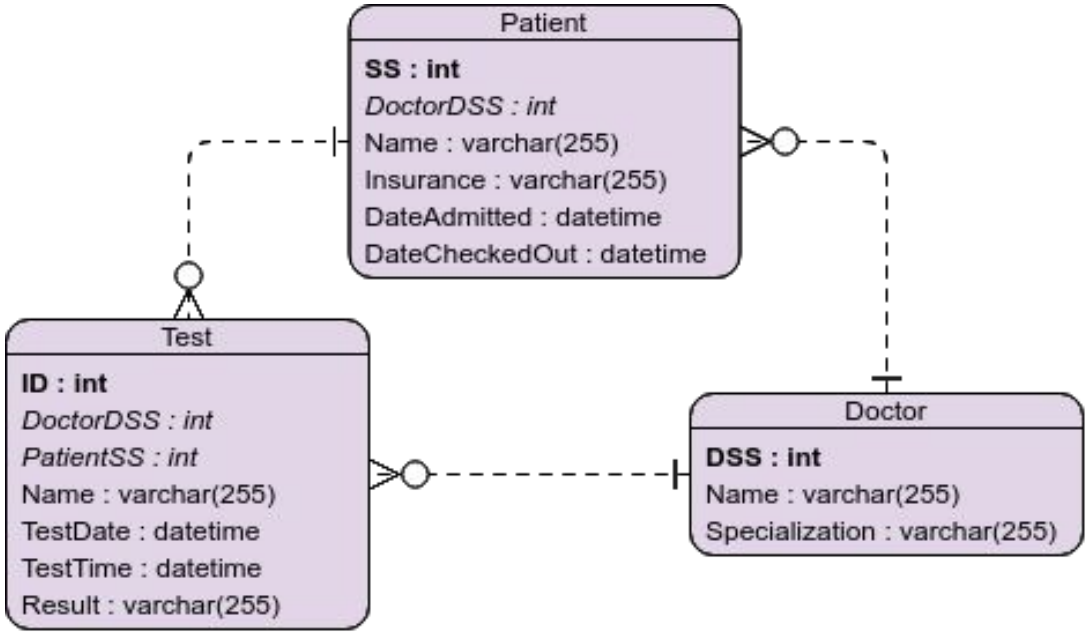
- Los datos son características de objetos coleccionadas a través de la observación.
- **No son información**, esta es producto de un análisis.
- Un conjunto de datos típicamente se operacionaliza como un conjunto de instancias de datos de una misma clase. Estas instancias cuentan con atributos y valores.

Paciente:
    id: <número>
    Nombre: <texto>
    Fecha de Nacimiento: <fecha>
    Salario: <número>
    COPD: <número>

id	**Nombre**	**...**	**COPD**
1	Juan	...	10
2	Pedro	...	8
...			

Datos estructurados

Con datos estructurados, nos referimos a que bajo el conjunto de datos **existe un modelo abstracto que estandariza** (modelo de datos) los valores de cada uno de los atributos y cómo se relacionan entre sí.

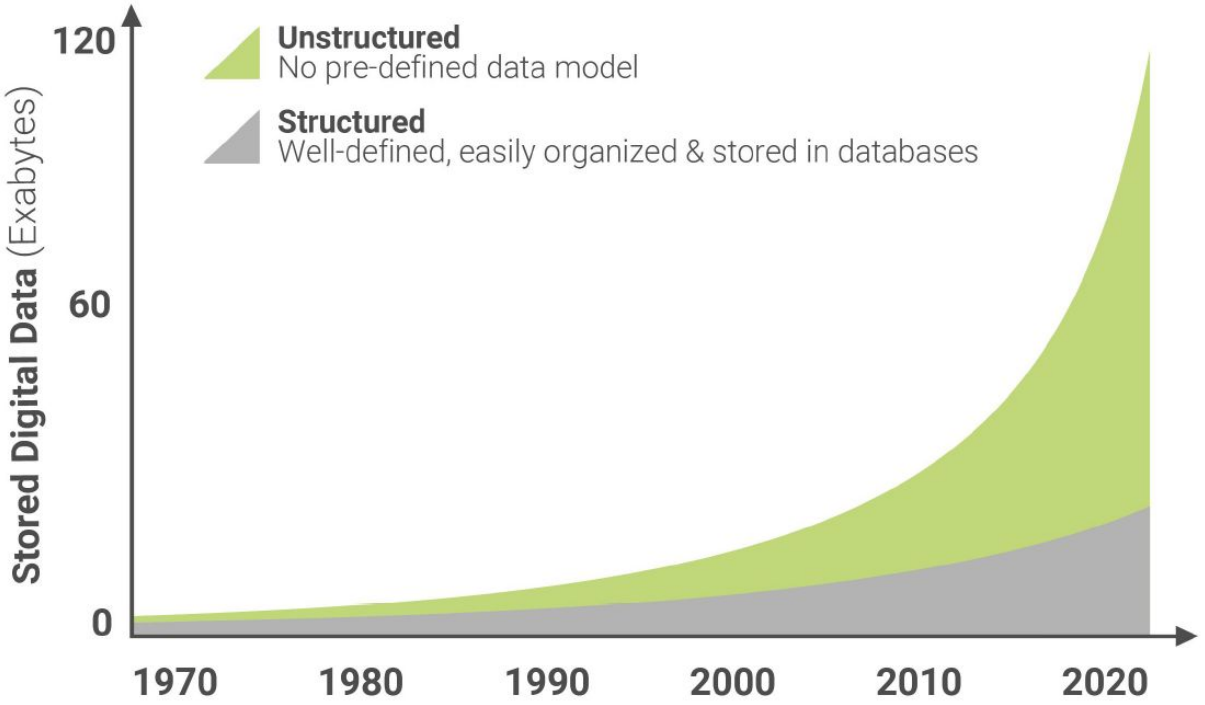


Datos no estructurados

Los datos no estructurados son aquellos que **no cuentan con un modelo de datos predefinido** o no están organizados de una manera predefinida. Esta característica genera irregularidades o ambigüedades que dificultan la extracción de información.

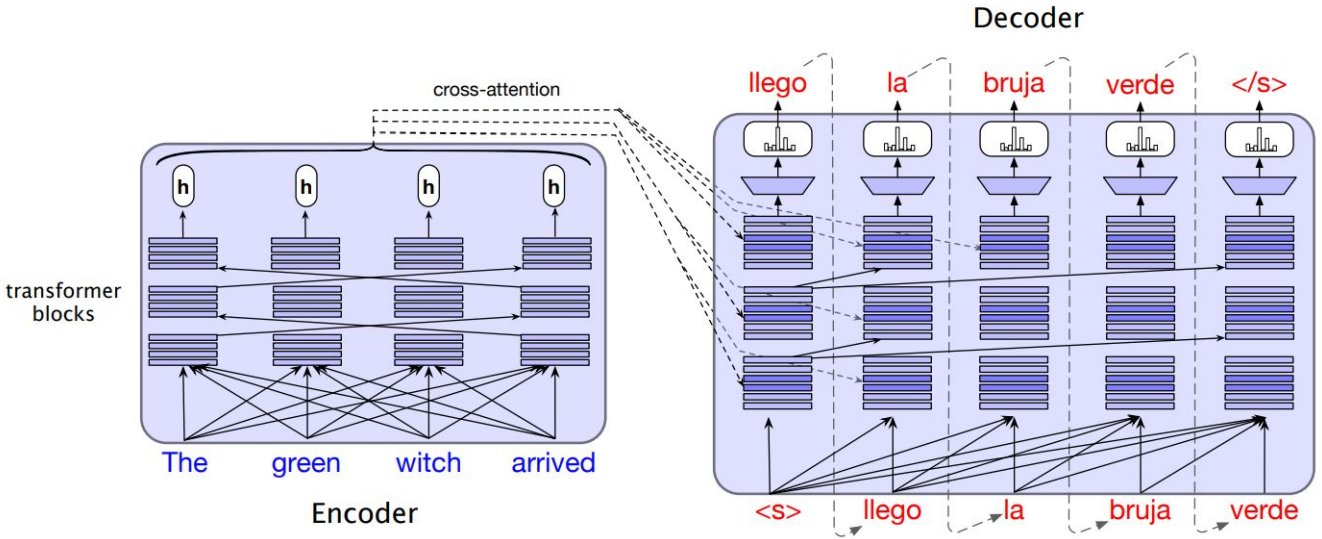
PCTE CON CUADROS DE PERICORONITIS RECURRENTE EN ZONA PZA 3.8 SEMIERUPCIONADA, SE RUEGA EVALUACION PARA EVENTUAL CIRUGIA DE EXODONCIA PZA 3.8 Y POSIBLEMENTE PZA 4.8

Evolución de la cantidad de datos disponibles



Para datos no estructurados usamos Deep Learning

En general, los modelos basados en Deep Learning se comportan mejor que modelos clásicos en tareas que utilizan datos no estructurados, tales como texto, audio, imágenes, grafos, etc.



Desafíos del Aprendizaje Automático

Un proyecto de aprendizaje automático podría fallar si el modelo seleccionado es incorrecto o si los datos de entrenamiento son deficientes.

Datos de entrenamiento no representativos

Cantidad insuficiente de datos de entrenamiento

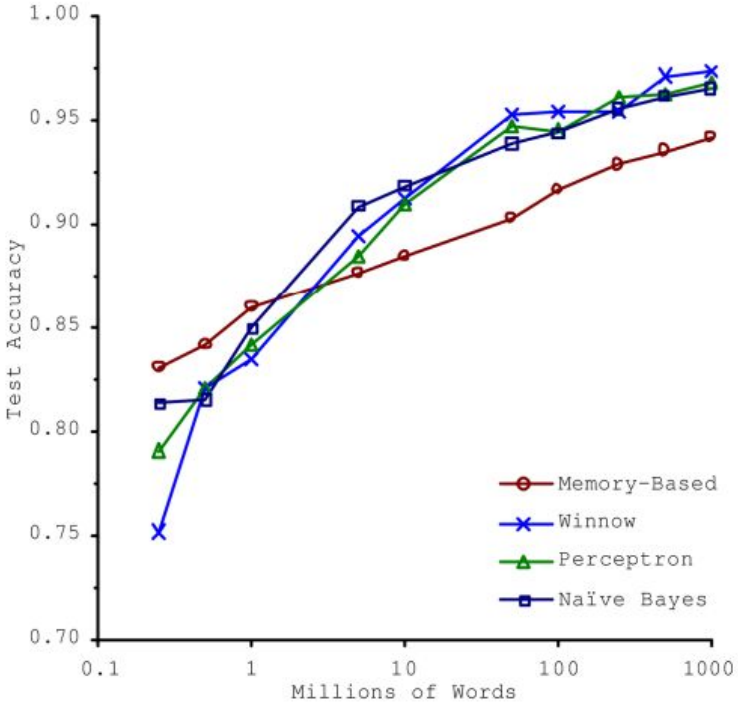
Características irrelevantes

Datos de mala calidad

Sobreajuste a los datos de entrenamiento

Subajuste a los datos de entrenamiento

Cantidad y representatividad de datos para entrenar

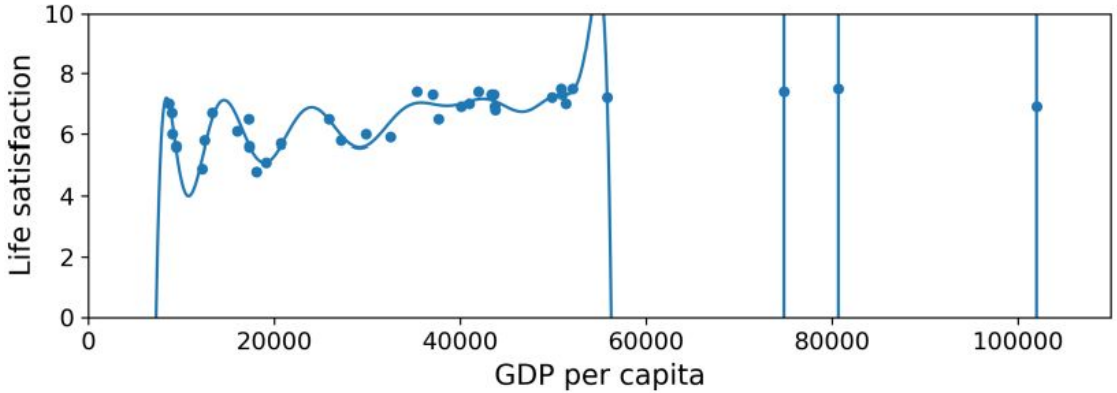


- Se ha visto que modelos simples se comportan de manera similar que modelos complejos al entrenarlos con suficientes datos.
- Para poder generalizar de manera correcta es necesario que los datos de entrenamiento sean representativos de los datos a los cuales quieres generalizar.

Sobreajuste y subajuste a los datos de entrenamiento

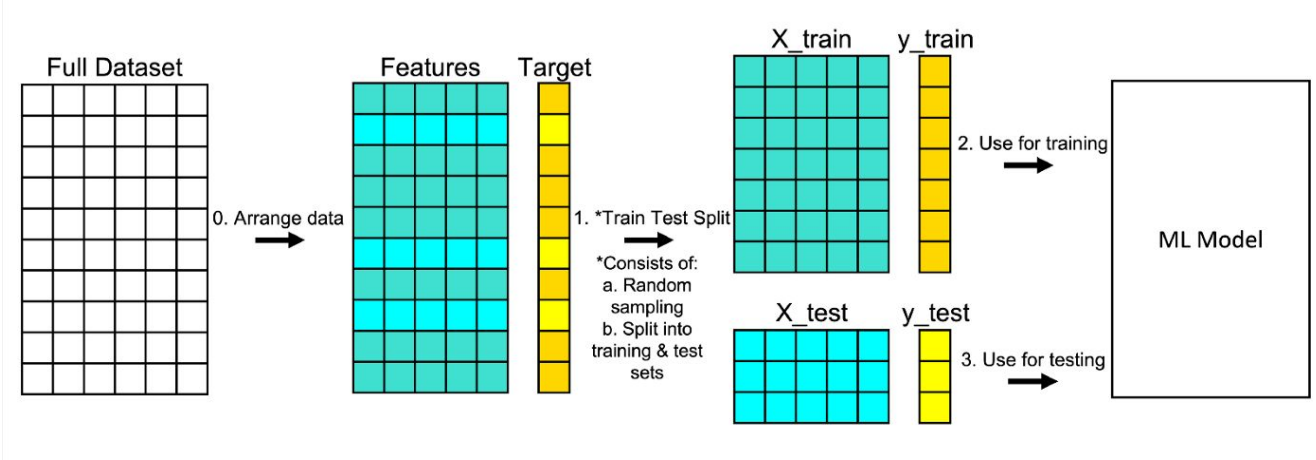
El sobreajuste significa que el modelo se comporta de manera correcta en el conjunto de entrenamiento, pero se comporta muy mal en el conjunto de entrenamiento (no generaliza correctamente). Normalmente se da porque el algoritmo es muy flexible

El subajuste significa que el modelo es demasiado simple para ajustarse a la estructura subyacente de los datos.



Prueba y validación

Para saber qué tan bien generaliza un modelo hay que dividir el conjunto de datos en dos subconjuntos: Un subconjunto de entrenamiento, al cual ajustaré mi modelo y un subconjunto de prueba, con el cual se evaluará la generalización.



Compromiso entre el sesgo y la varianza

El error de generalización puede ser expresado como la suma de diferentes errores:

Sesgo (*bias*): Se debe a suposiciones incorrectas, como asumir que los datos se comportan de manera lineal cuando realmente se comportan de manera cuadrática.

Varianza: Se debe a la excesiva sensibilidad del modelo a pequeñas variaciones en los datos de prueba. Un modelo muy complejo podría tener mucha varianza, por lo tanto se podría sobreajustar.

Aumentar la complejidad del modelo típicamente aumentará su varianza y reducirá su sesgo, por el contrario, reducir la complejidad aumentará su sesgo y disminuirá la varianza. Por eso es un compromiso.

Selección de modelos

Para evaluar el rendimiento de un modelo, debemos medir su capacidad de generalización.

Si nosotros tenemos un conjunto de modelos con sus hiperparámetros (parámetros determinados por el usuario) y medimos cada rendimiento en un único conjunto de prueba, estaríamos adaptándonos sólo a ese conjunto, por lo que debemos disminuir este sesgo de selección de ejemplos.

Para solucionar este problema podemos usar validación cruzada y generar múltiples particiones de subconjuntos de entrenamiento y prueba, promediando sus rendimientos.

Python para ciencia de datos

El lenguaje de programación Python cuenta con múltiples bibliotecas para ayudarnos en todas las etapas de un proyecto de ciencia de datos.

Bibliotecas fundamentales

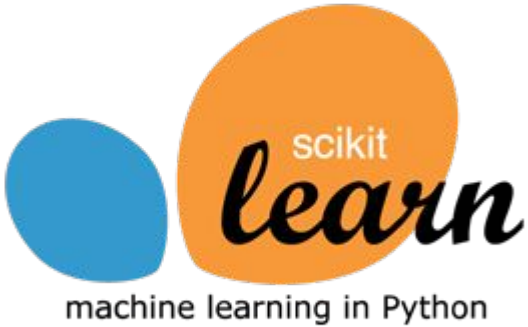
Las bibliotecas fundamentales para realizar ciencia de datos son las siguientes:

- Numpy: Algoritmos de álgebra lineal.
- SciPy: Algoritmos matemáticos ya implementados.
- Matplotlib: Visualización de datos.
- Pandas: Procesamiento de datos.

Scikit-learn

Esta es la biblioteca más utilizada para realizar aprendizaje de máquinas en Python.

Tiene funciones implementadas desde el preprocesamiento de datos hasta la validación de los modelos.



Tensorflow / PyTorch

Estas bibliotecas son las más utilizadas para realizar deep learning con python.

Tienen todos los algoritmos necesarios para diseñar y ajustar un modelo basado en redes neuronales.