Tema 1: Introducción al proceso KDD

Minería de Datos

viu Universidad Internacional de Valencia

De:
Planeta Formación y Universidades

1. Proyectos de minería de datos y clasificación de los sistemas y tipos de modelos de data mining.

2. La Inteligencia Artificial, el Aprendizaje automático (Machine Learning), el aprendizaje estadístico, la Minería de Datos y el nuevo reto del Big Data.

3. El proceso KDD (Knowledge Discovery in Databases)

1. Elegir un origen de datos, como un cubo, una base de datos o incluso archivos de texto o de Excel, que contenga los datos sin formato que utilizará para generar los modelos.
2. Defina un subconjunto de los datos del origen de datos que se usarán para el análisis y guárdelos como vista del origen de datos.
3. Defina una estructura de minería de datos para el modelado.
4. Agregue modelos de minería de datos a la estructura, elija un algoritmo y especifique el modo en que el algoritmo controlará los datos.
5. Entrene los modelos rellenándolos con los datos seleccionados o con un subconjunto filtrado de los datos.
6. Explore, pruebe y genere modelos.

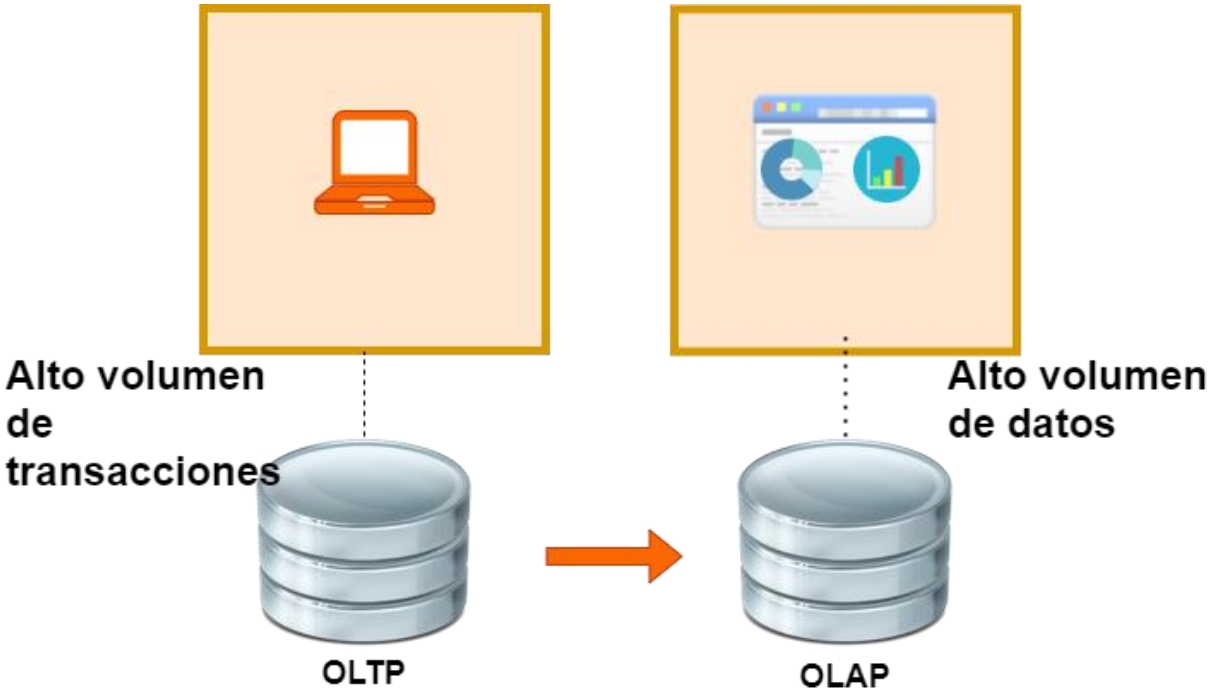
Cuando el proyecto esté completo, puede implementarlo para que los usuarios lo examinen o lo consulten,

o puede proporcionar acceso mediante programación a los modelos de minería de datos en una aplicación, para permitir las predicciones y el análisis.

Diferencias entre OLTP y OLAP

OLTP	OLAP
Sistema transactional y gestiona modificaciones	Sistema recuperación y análisis de datos
Insertar, actualizar y eliminar información	Consulta información
Las transacciones son la Fuente original de datos	La base de datos OLTP es la Fuente de datos
Tiempo transacción bajo	Tiempo de transacción elevado
Consultas simples	Consultas complejas
Tablas normalizadas (3NF)	No normalizadas
Mantiene integridad de datos	La integridad de datos no se ve afectada porque no hay insercciones

Diferencias entre OLTP y OLAP

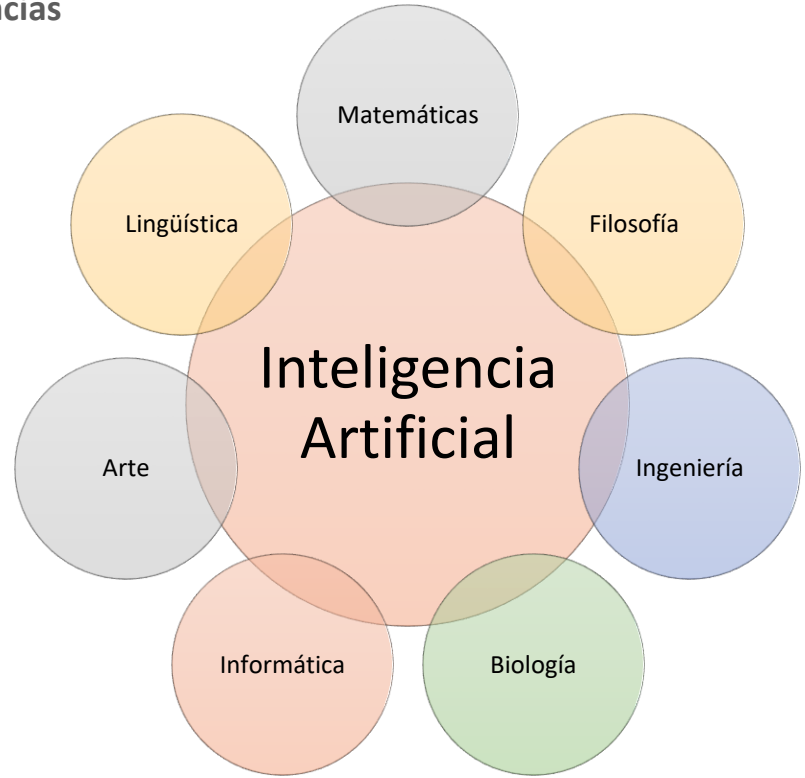


- El **modelo** de minería de datos define el algoritmo o método de análisis que utilizará en los datos. Para cada estructura de minería de datos, agrega uno o varios modelos de minería de datos.
- Según el proyecto se pueden requerir:
  - Combinar varios modelos
  - Crear varios proyectos para cada tarea analítica
- **Entrenar** el modelo con los datos almacenados
- Explorar visualmente el modelo y realizar consultas para la **obtención de detalles**
- **Explotar** el modelo: implementar un modelo:
  - Asegurarse de que las opciones de procesamiento
  - Actualizar el modelo con nuevos datos si procede

- **Reglas de Asociación**: detecta asociaciones comunes entre elementos.
- **Clustering**: busca elementos afines dentro de un conjunto
- **Árboles de decisión**: clasificación de elementos
- **Series temporales**: para predecir una magnitud en función del tiempo
- **Naive Bayes**: para explorar datos, puede usarse para buscar correlaciones entre atributos
- **Redes neuronales**: regresión y clasificación pero es difícilmente descriptible

- Inteligencia Artificial es una rama de la **Informática**: (https://aitopics.org/search)
- Meta: conseguir que sistemas no naturales resuelvan (o ayuden a resolver) los mismos problemas que resolvemos los humanos
  [de la misma manera que nosotros]
- Por tanto, estudia y resuelve problemas situados en la frontera de la Informática
- Se basa en dos ideas fundamentales:
  - Representación del conocimiento explícita y declarativa
  - Resolución de problemas heurística

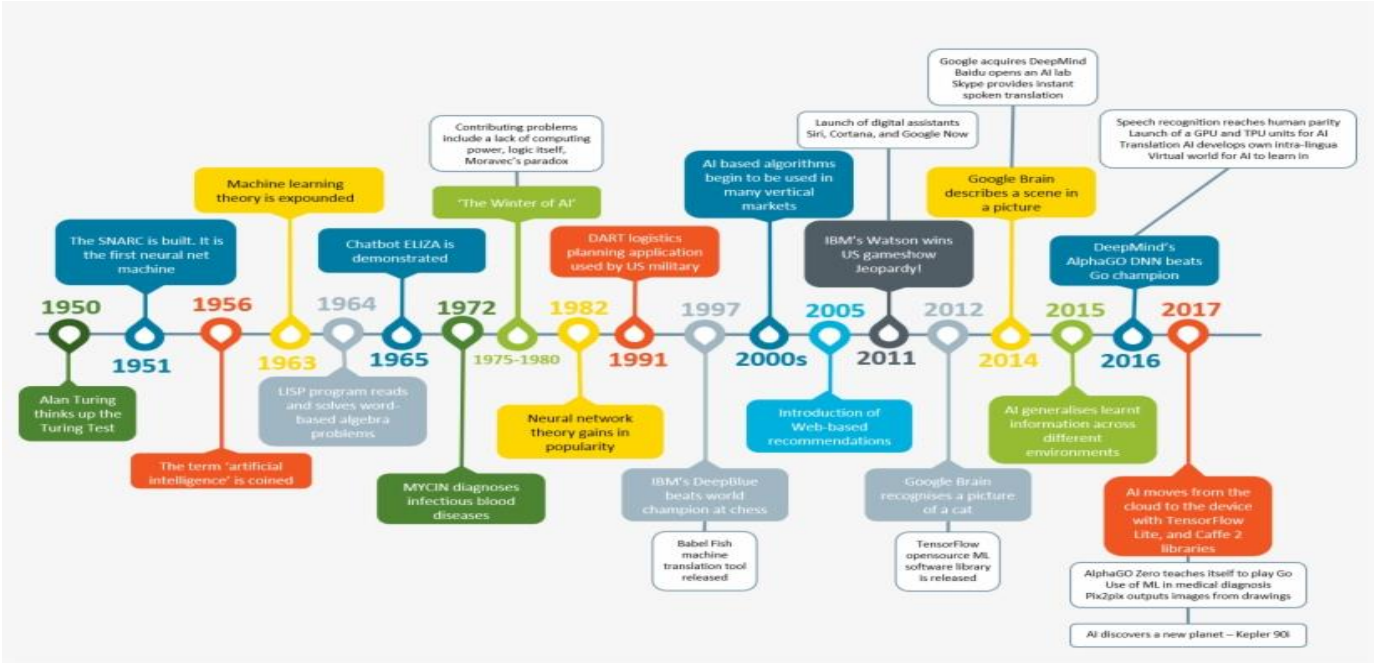
Se trata de la union de varias ciencias



¿Qué es Inteligencia?

- Consciencia
- Comunicación
  - Test de Turing: https://es.wikipedia.org/wiki/Prueba_de_Turing
- Sentimientos, sociabilidad
  - Agentes sociales
  - Kismet:
    www.youtube.com/watch?v=EP8zN0CKQnI
    www.youtube.com/watch?v=3GkI374ZkM4
- Creatividad
- Percepción: visión, reconocimiento

- Historia de la Inteligencia Artificial: **https://aitopics.org/misc/brief-history**



¿Qué es?

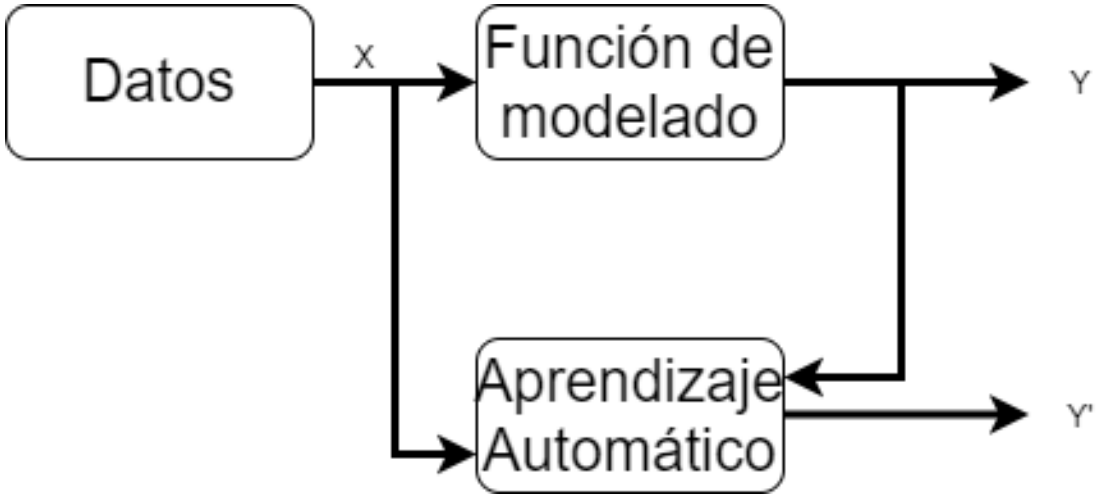
- Es una rama de la Inteligencia Artificial que tiene como objetivo desarrollar técnicas que permitan a las computadoras **aprender**.
- Se trata de crear **algoritmos** capaces de generalizar comportamientos y reconocer patrones a partir de una información suministrada en forma de ejemplos.
- Es un proceso de **inducción del conocimiento**: métodos que permite obtener por generalización un enunciado general a partir de enunciados que describen casos particulares.

Tipos de aprendizaje

- **Aprendizaje supervisado**: establece una correspondencia entre las entradas y las salidas del sistema, donde la base de conocimientos del sistema está formada por **ejemplos etiquetados** a priori.
- **Aprendizaje no supervisado**: el proceso de modelado se lleva a cabo sobre un conjunto de ejemplos formados únicamente por **entradas** al sistema, sin conocer su clasificación correcta
- **Aprendizaje por refuerzo**: el algoritmo aprende observando el mundo que le rodea y con un continuo flujo de información en las dos direcciones realizando un proceso de ensayo-error, y **reforzando** aquellas acciones que reciben una respuesta positiva en el mundo.

- El aprendizaje estadístico se refiere a un conjunto de **herramientas** para **modelar** y comprender conjuntos de **datos** complejos.
- Es un área desarrollada en **estadística** y combina desarrollos paralelos en informática y, en particular, **aprendizaje automático**.

- Estas herramientas se pueden clasificar como supervisadas o no supervisadas (ya que está dentro del aprendizaje automático).
- Pero se refiere a un conjunto de técnicas o enfoques para estimar una función **f**.



- El **data mining** (minería de datos), es el conjunto de técnicas que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto.

  - Determinación de los objetivos
  - Preprocesamiento de datos
  - Determinación del modelo
  - Análisis de los resultados

- Las 3 V del Big Data:

1. Volumen → No hablamos de GB, si no de TB, PB, EB o incluso ZB
2. Velocidad → Información generada a gran velocidad, como la de la bolsa
3. Variedad → Nuestra información puede ser estructurada o no serlo

¿Qué dificultades encontramos para ello?

- Complejidad → Parte de la información puede no ser entendida, debido a la no estructuración de la misma

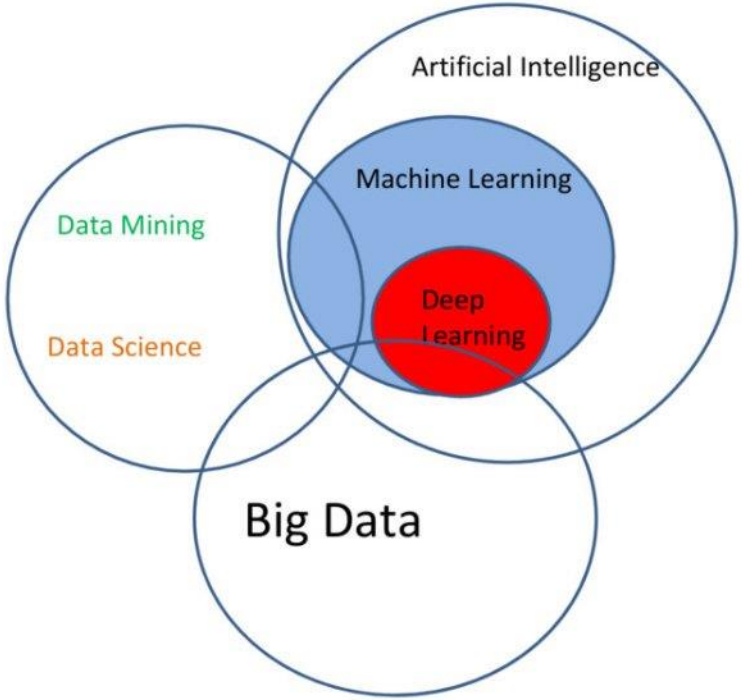
- Almacenamiento → Hablamos de cantidades enormes de información, ¿Cómo gestionar tal cantidad de datos a nivel de almacenamiento?

- Rendimiento → ¿Cómo se puede procesar de manera eficiente la información para mejorar el rendimiento?

¿De dónde viene la información que se procesa en Big Data?

- Sensores meteorológicos
- Geográfica
- Textos
- Chats de redes sociales
- Sensores de tráfico
- Imágenes de satélite

- Todas esta información no puede ser gestionado por aplicaciones normales o **datawarehouses**, y en su mayoría se producen por sistemas automáticos
- En ocasiones, es información no estructurada, y crece a gran velocidad



1. Proyectos de minería de datos y clasificación de los sistemas y tipos de modelos de data mining.

2. La Inteligencia Artificial, el Aprendizaje automático (Machine Learning), el aprendizaje estadístico, la Minería de Datos y el nuevo reto del Big Data.

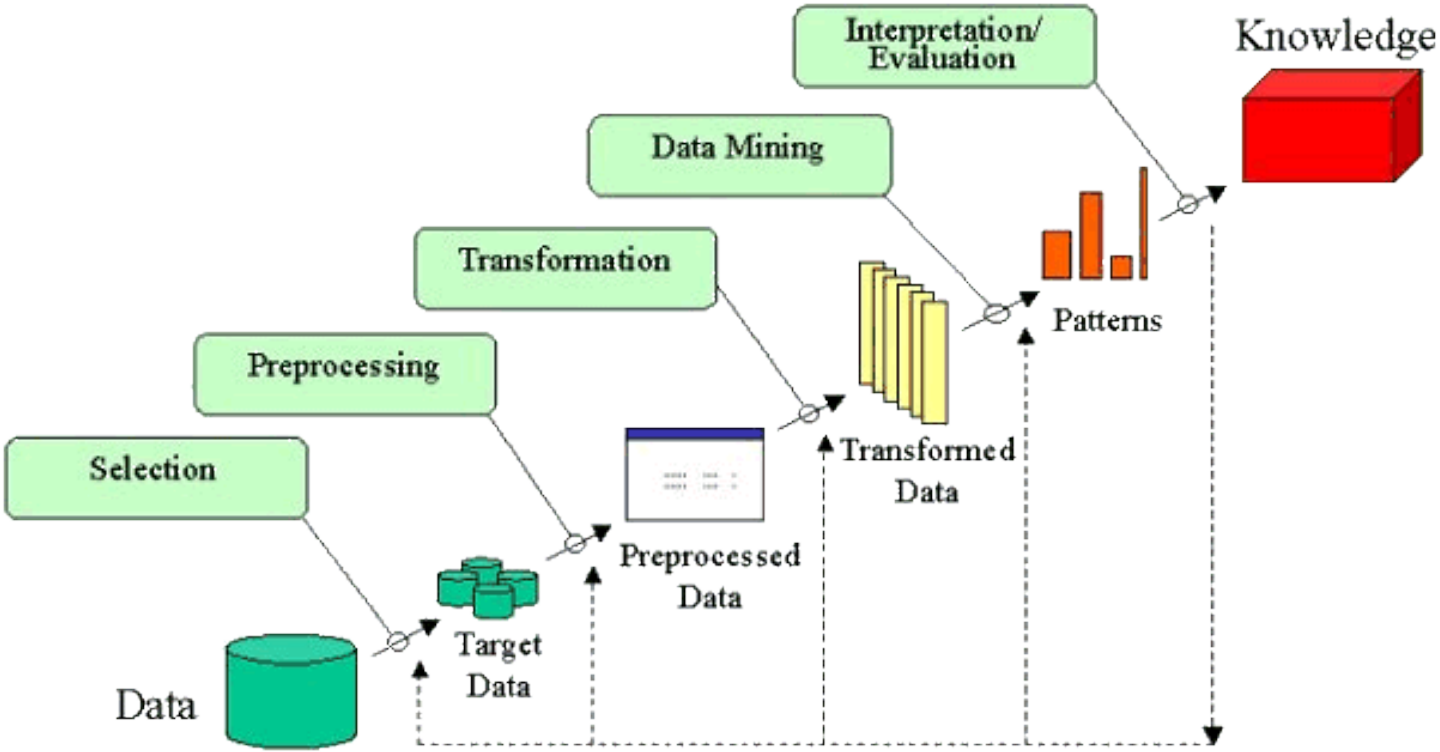
3. El proceso KDD (Knowledge Discovery in Databases)

¿Qué es el proceso KDD?

- “El proceso de identificación de patrones y relaciones ocultas dentro de los datos"
- La minería de datos ayuda a los usuarios finales a extraer información comercial útil de una gran base de datos

¿Por qué se usa?

- Pepitas escondidas de información valiosa enterradas en lo profundo de una montaña de datos que de otra manera no serían nada interesantes.
- Datos omnipresentes
- Buscar la ventaja competitiva


Interpretation/
Evaluation
Knowledge
Data Mining
Transformation
Patterns
Preprocessing
Selection
Transformed
Data
Preprocessed
Data
Target
Data
Data

La importancia del proceso

- Un **usuario de negocios** estará interesado en la eficiencia y los resultados, la validez puede no ser tan importante.

- Un **investigador** estará claramente interesado en un tipo diferente de resultados, y la validez será importante.

- **Data Scientist**: un informático puede estar interesado en introducir nuevos algoritmos o enfoques computacionales y lograr mejores resultados o un procesamiento más eficiente.

Conocimientos necesarios:

- Estadísticas
- Visualización
- La inteligencia artificial
- Aprendizaje automático
- La tecnología de la base de datos
- Redes neuronales
- Reconocimiento de patrones
- Sistemas basados en el conocimiento
- Adquisición de conocimientos
- Recuperación de información
- Computación de alto rendimiento
- Y así sucesivamente…

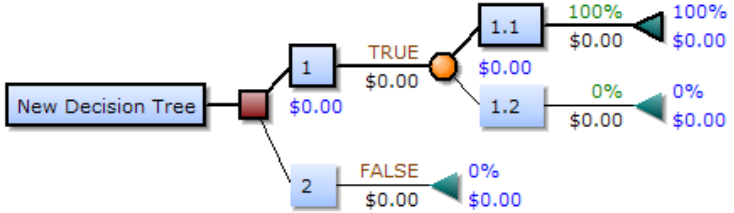
¿Qué se necesita?

- Datos adecuados
- Poder de computación
- Software de **minería de datos**
- Operador cualificado que conoce tanto la naturaleza de los datos como las herramientas de software
- Razón, teoría o corazonada

Aplicaciones típicas basadas en KDD

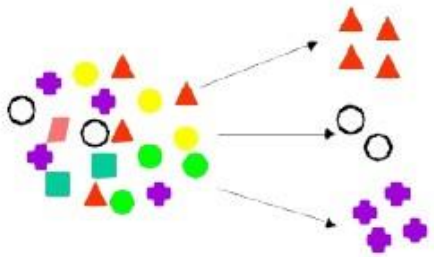
- Servicios financieros
  - Aprobación de crédito
  - Detección de fraude
  - Marketing
- Cuidado de la salud
  - Análisis epidemiológico: incidencia y prevalencia de la enfermedad en grandes poblaciones y detección del origen y la causa de las epidemias de enfermedades infecciosas
  - Conocimientos para la financiación
  - Política, programas

Dos aproximaciones:



- Supervisadas
  - Dependen de la variable de salida

- No supervisadas
  - Técnicas de análisis de patrones o tendencias
  - Estudios de nichos de mercado
  - Típicamente se emplea el uso de clustering



Automatización del proceso

- La capacidad de apuntar una herramienta a algunos datos y pulsar un botón
- Algunos métodos de KDD/Minería de datos son más adecuados para la automatización que otros

Preparación de los datos

- Este paso tiene que ver con la transformación de los datos en bruto que se recogieron en una forma que pueda ser utilizada en el modelado.

- Las técnicas de preprocesamiento de datos generalmente se refieren a la adición, eliminación o transformación de los datos del conjunto de capacitación.

Data Cleaning (Limpieza de datos)

- Tratamiento de **missing values** (valores desconocidos o perdidos)
- Esto también implica limpiar datos **outliers** (fuera de rango)

Data Cleaning (Limpieza de datos)

- Los valores nulos o perdidos no proporcionan información para un modelo, y por eso necesitamos encontrar una forma de tratarlos.
- Podemos eliminar toda la característica si faltan más de un **80%**de los valores.
- Tener este tipo de característica explicará muy poco mientras se entrena el modelo, y no tendrá un gran efecto durante la clasificación o la regresión.
- Podemos usar los valores disponibles de esa característica en particular y construir un modelo de regresión para predecir los valores que faltan.
- Esto es aconsejable si los valores perdidos son inferiores al **10-15%**.

Data Integration (Integración de datos)

- Esto incluye el agrupamiento de datos similares (estudiar correlación entre variables)
- Implica acceder a diferentes bases de datos y extraer información de varias fuentes de confianza y unirlas.
- Este proceso dará significado a los datos no estructurados enmarcándolos de manera que puedan ser entendidos.

Data Transformation (Transformación de datos)

- Por lo general, los datos tendrán enormes diferencias en los números y una computadora dará peso a los valores que sean más altos.
- El escalado de datos es obligatorio para asegurarse de que el modelo da igual peso a todos los valores.
- Un modelo ve las variables como simples números, mientras que nosotros los interpretamos. Por lo tanto, para dar igual peso a ambas características, necesitamos escalar los valores.
- Este proceso también se llama **normalización.**

Edad

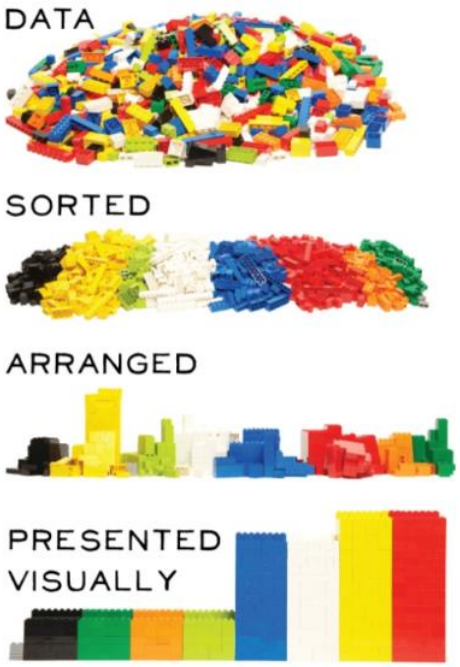
Sexo

Data Reduction or Dimension Reduction (Reducción de datos / dimensiones)

- Este proceso implica la reducción de características basadas en el grado de información que proporciona.
- Esto implica dos métodos.
  - **Selección de características**, donde algunas de las características son descuidadas mientras se entrena un modelo porque dan muy poca información.
  - Extracción de características, donde diferentes métodos como el **PCA (Análisis de Componentes Principales)** mapean las características en una dimensión alta con menos número de características.
- En este procedimiento se pueden utilizar tanto métodos lineales como no lineales. Las ventajas de la reducción de la dimensión es que la correlación de los datos puede ser altamente reducida con una baja varianza.

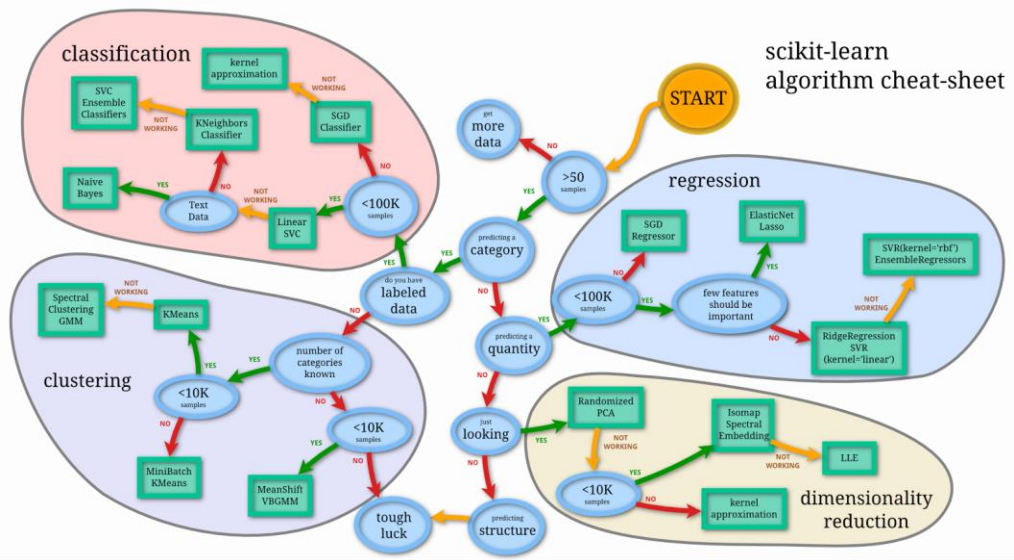
Finalizando el preprocesado de datos

- Una vez que hayamos computado los datos y los hayamos analizado.
- Debemos centrarnos en seleccionar el modelo que mejor se adapte a nuestros datos.
- Esto depende únicamente del tipo de datos con los que estamos tratando.
- Hay muchos modelos de Machine Learning



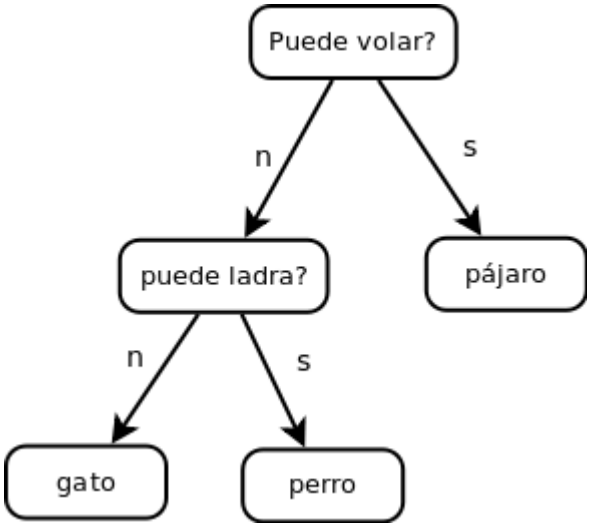
Principales modelos a usar

- Árboles de decisión
- Redes neuronales (artificiales)
- Grupo/vecino más cercano KNN
- Algoritmos Genéticos/Computación Evolutiva
- Redes Bayesianas
- Estadísticas
- Híbridos



Árboles de decisión

- Representaciones gráficas de las relaciones con los datos
- Excelencia en los modelos de clasificación y predicción

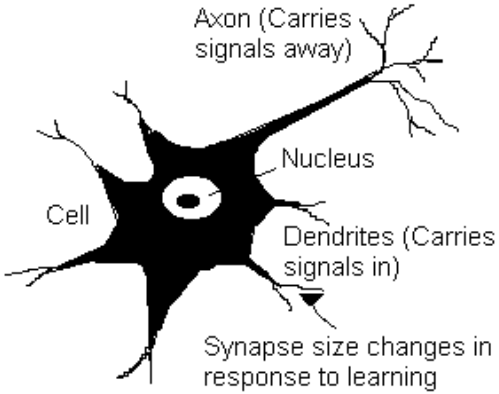


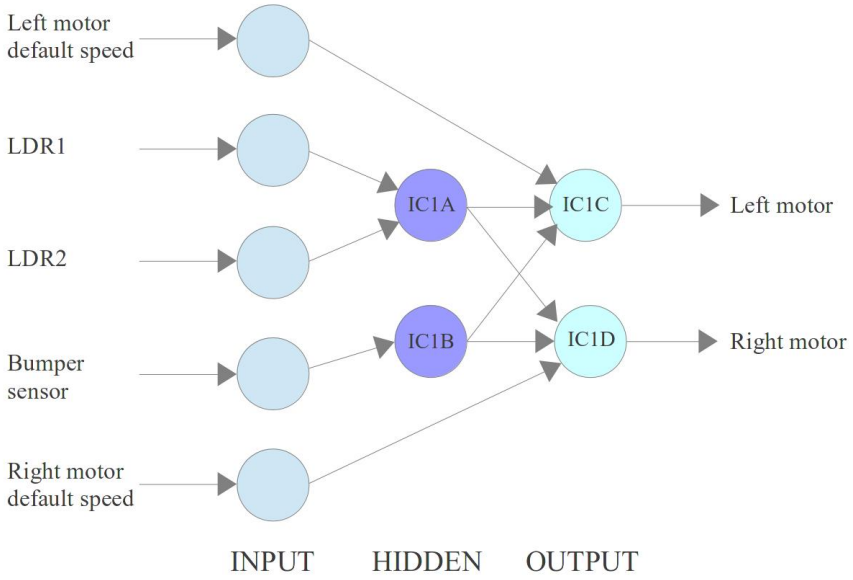
Árboles de decisión

Ventajas	Desventajas
Fácil de entender e interpretar	Los grandes árboles con un gran número de variables se vuelven difíciles de entender
Representar la complejidad en una forma compacta	Los datos faltantes deben ser manejados apropiadamente en la construcción y uso de los modelos
Manejar bien los datos no lineales	
Relativamente bien adaptado a la automatización.	

Redes de neuronas

- Derivado de la investigación de la inteligencia artificial
- Modelado en la Neurona Humana





Redes de neuronas

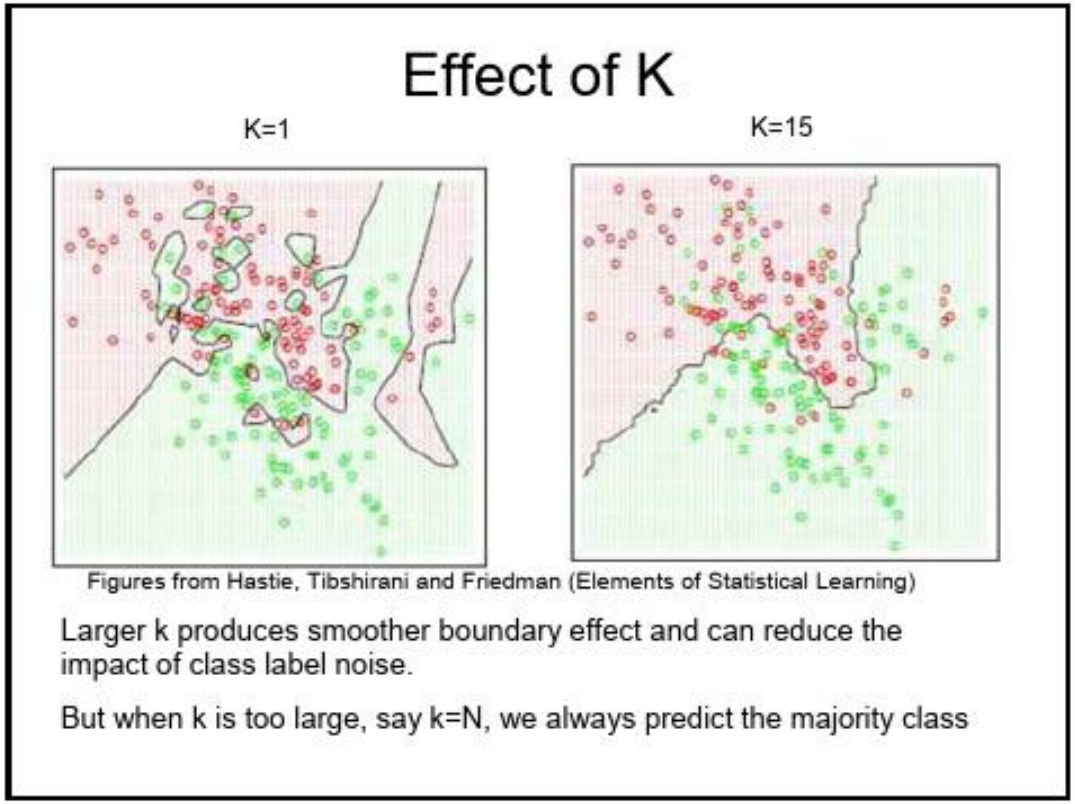
- Se trata de un modelo hipotético que podría utilizarse para predecir.
- Este modelo es supervisado
- Las variables de entrada en la parte inferior, como entradas a una capa oculta de nodos
- Los pesos se fijan y ajustan durante el entrenamiento;
- La convergencia del algoritmo utilizando arquitecturas de retroalimentación
- Algoritmos para optimizar el tiempo de computación que puede ser extenso: múltiples pases de entrenamiento

Redes de Neuronas

Ventajas	Desventajas
Buena precisión	Propenso al sobreajuste (overfitting)
Rendimiento robusto con una amplia variedad de tipos de datos	Poca claridad del modelo

Clustering/Nearest Neighbour

- El objetivo es asignar registros "parecidos" a un grupo
- Grupos asignados según alguna variable o criterio objetivo
- El vecino más cercano utilizado para la predicción

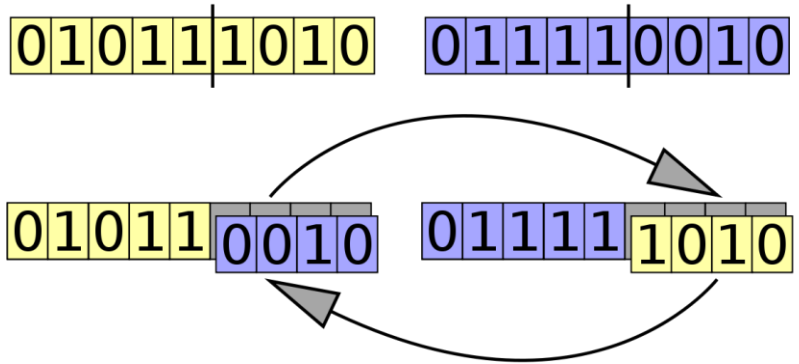


Clustering/Nearest Neighbour

Ventajas	Desventajas
Fácil de interpretar	Datos complejos que no se adaptan bien a la automatización (se requiere mucho preprocesamiento)
Fácil de aplicar en situaciones básicas	

Algoritmos Genéticos/Computación Evolutiva

- Basado en Darwin - aplicado usando las matemáticas
- Requiere una forma de representar una solución a un problema
- una forma de probar lo buena que es la solución (función de fitness)
- Las soluciones están matemáticamente "mutadas"
- Las soluciones más eficaces sobreviven
- Convergencia



Algoritmos Genéticos/
Computación Evolutiva

2 El Algoritmo Genético Simple

```
BEGIN /* Algoritmo Genetico Simple */
    Generar una poblacion inicial.
    Computar la funcion de evaluacion de cada individuo.
    WHILE NOT Terminado DO
    BEGIN /* Producir nueva generacion */
        FOR Tamaño poblacion/2 DO
        BEGIN /*Ciclo Reproductivo */
            Seleccionar dos individuos de la anterior generacion,
            para el cruce (probabilidad de seleccion proporcional
            a la funcion de evaluacion del individuo).
            Cruzar con cierta probabilidad los dos
            individuos obteniendo dos descendientes.
            Mutar los dos descendientes con cierta probabilidad.
            Computar la funcion de evaluacion de los dos
            descendientes mutados.
            Insertar los dos descendientes mutados en la nueva generacion.
        END
        IF  la poblacion ha convergido THEN
            Terminado := TRUE
    END
END
```

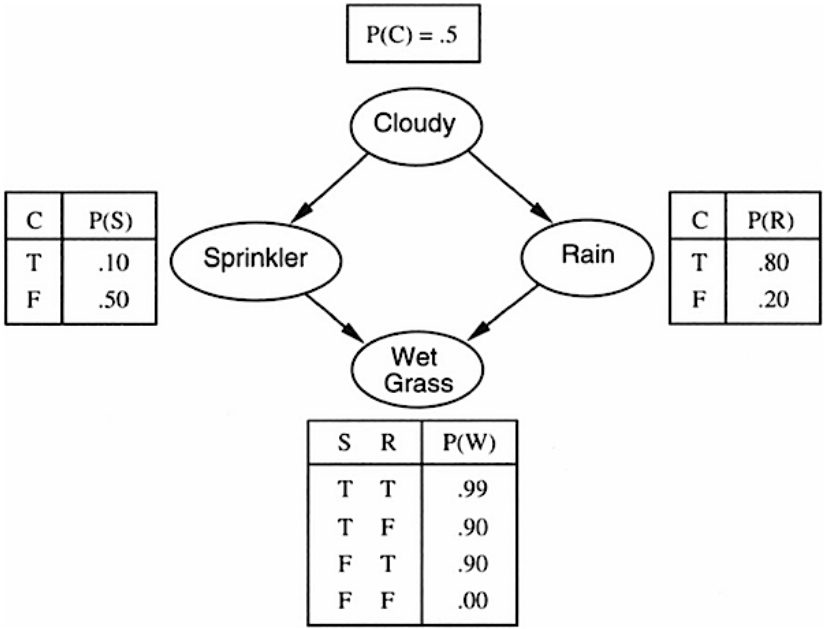
Algoritmos Genéticos/Computación Evolutiva

Ventajas	Desventajas
Se adapta a nuevos problemas más facilmente	No es fácil de automatizar
Adecuado cuando los datos no son muy buenos	Costoso de implementar
Puede ser útil cuando se se pueden aplicar otras técnicas	

Redes Bayesianas

- Puede construir redes de eventos vinculados, cada uno con probabilidades previas
- Basada en la teoría de Bayes

$$P(a|b) = P(b|a) * P(a) / P(b)$$



Redes Bayesianas

Ventajas	Desventajas
Claridad de los modelos resultantes	Dificil de construir y mantener
Buena precisión en los resultados	
Se adapta fácilmente a nuevas probabilidades siempre que los datos no se alejen de los entrenados	Mala predicción para eventos raros

Estadisticos

- Con un resultado o variable dependiente:
  - Correlaciones
  - ANOVA
  - Regresión
- Utilizados por sí mismos o para confirmar los resultados de otro método

Ventajas	Desventajas
Usado por el método científico	Limita los hallazgos a las técnicas que se aplican y sus limitaciones asociadas (normalidad, linealidad, etc.)

Híbridas

- Técnicas utilizadas en combinación
- Ejemplo: uso de un algoritmo genético para identificar las variables objetivo para su inclusión en un modelo de red neuronal

Bases de datos

Las normas de las bases de datos permiten a cualquier paquete de extracción de datos acceder a los datos en bruto

- Los principales proveedores de productos de gestión de bases de datos/datos (IBM, SPSS, Oracle PeopleSoft, SAS, etc.)
- Amazon
- Google
- Añadido como un componente de los paquetes llave en mano
- Puede incorporar varios métodos (SAS Enterprise Miner)
- Método único (TreeAge Software Inc.: un producto de árbol de decisión dedicado)

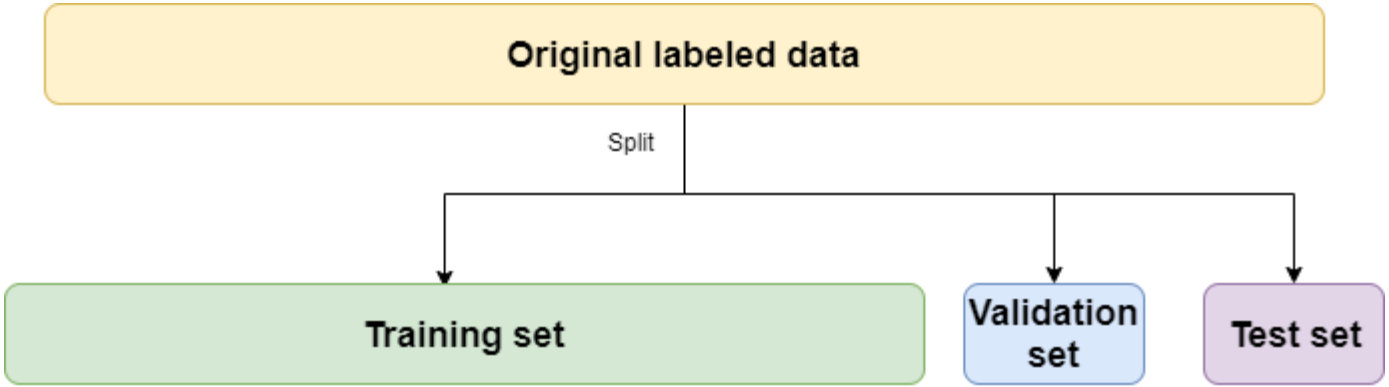
¿Cómo evaluamos el modelo?

- El propósito de la evaluación es probar un modelo con datos diferentes a los que se le entrenó.
- Esto proporciona una estimación imparcial del rendimiento del aprendizaje.

  - División conjunto de entramiento y test
  - Validación Cruzada

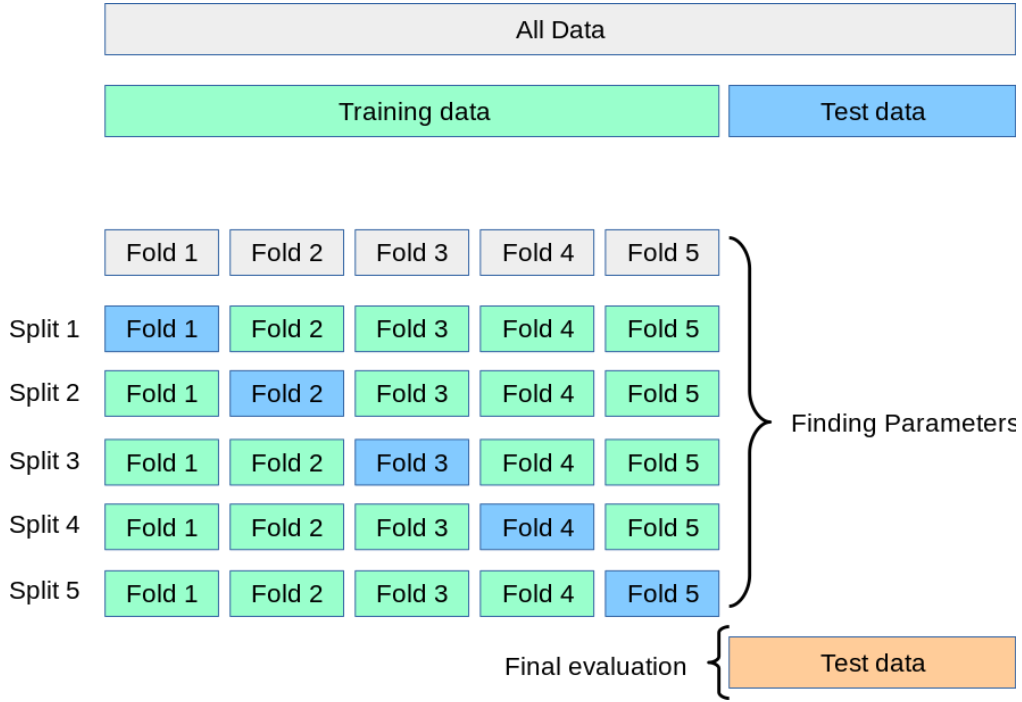
Conjunto de entrenamiento y test

- En este método, el conjunto de datos se divide aleatoriamente en dos subconjuntos:
  - El conjunto de entrenamiento es un subconjunto del conjunto de datos utilizado para construir modelos predictivos.
  - El conjunto de validación es un subconjunto del conjunto de datos utilizado para evaluar el rendimiento del modelo construido en la fase de entrenamiento.



Validación cruzada

- Esta validación se realiza en K fases:
  - Se genera tantos conjuntos como valor de K.
  - Dejando uno para k-1 para entrenamiento y 1 para test
  - Se repite el proceso K veces



Gracias

viu Universidad Internacional de Valencia

De:
Planeta Formación y Universidades