UNIVERSIDAD NACIONAL DE LA MATANZA

INTELIGENCIA DE NEGOCIOS

Introducción a la Minería de Datos y la Explotación de Información

Docentes: ING. LORENA R. MATTEO

Autores ppt orig.: Lic. HUGO M. CASTRO / MG. DIEGO BASSO

Fecha última actualización.: 26/05/2025

Ejemplo de un Problema Motivador

- Un banco trata de evitar que sus clientes se vayan a otros bancos. Por eso quieren detectar los *signos tempranos de deserción*, es decir las actitudes que toma un cliente antes de irse del banco.
- Esto le permitirá identificar qué clientes están haciendo eso mismo, e intentar retenerlos mientras son clientes .
- El banco tiene información histórica sobre movimientos de los clientes (entre ellos la baja), y quiere usar esa información para detectar cuáles son estos signos, para luego identificar qué clientes presentan los mismos "comportamientos".

La Solución con OLAP

- Se formulan hipótesis:
  - ✓ *Los clientes que no renovaron plazos fijos tienen tendencia a irse.*
  - ✓ *Los clientes que disminuyeron sus operaciones de cajero automático tienen tendencia a irse.*
  - ✓ *Los clientes que cerraron cuentas tienen tendencia a irse.*
- Utilizando las variables adecuadas del DW se analiza qué porcentaje de clientes en esas condiciones se fueron del banco.
  - Esto confirma o rechaza las hipótesis.
- Puede ocurrir que no haya respuestas satisfactorias.

Existe otra técnica…

- … que permite extraer el conocimiento necesario de los datos históricos para resolver el problema:

DATA MINING o Minería de Datos

- La minería es un término que caracteriza el proceso de encontrar piedras preciosas o un material valioso, proveniente de una gran cantidad de materia prima.

¿Qué es Minería de Datos?

- Descubrir automáticamente información útil en grandes repositorios de datos.

¿Cuál es la diferencia entre …?

Inteligencia Artificial (Artificial Intelligence)

- es la capacidad de las máquinas para realizar tareas que normalmente requerirían inteligencia humana.
- se basa en una amplia variedad de técnicas, incluyendo el ML, el Procesamiento del Lenguaje Natural y la Visión Artificial (análisis de imágenes y vídeos).

Ciencia de Datos (Data Science)

- es un campo interdisciplinario que intenta transformar los datos en información relevante y oportuna aplicando el Método Científico y generando Conocimiento.
- evoluciona de la DM y utiliza como herramientas las Ciencias de la Computación, Probabilidad, Estadística y ML englobando métodos, procesos y sistemas.

Minería de Datos (Data Mining)

- es el proceso de descubrir patrones en grandes conjuntos de datos.
- se utiliza para encontrar relaciones y patrones ocultos en los datos que pueden ser utilizados para tomar decisiones informadas.

Aprendizaje Automático (Machine Learning)

- es una técnica de AI que permite a los sistemas aprender de los datos sin ser programados explícitamente.
- se basa en algoritmos que analizan los datos y hacen predicciones o toman decisiones basadas en patrones que han sido descubiertos en los datos.
- se busca generalizar, aprender conceptos a partir de un conjunto de ejemplos y sus características.
- cuantos más ejemplos, probablemente más fácil sea la tarea.

Aprendizaje Profundo (Deep Learning)

- es una técnica específica dentro del ML que se basa en redes neuronales artificiales para realizar tareas complejas.
- el término "profundo" se refiere a la complejidad de estas redes, que están compuestas por múltiples capas de neuronas interconectadas.

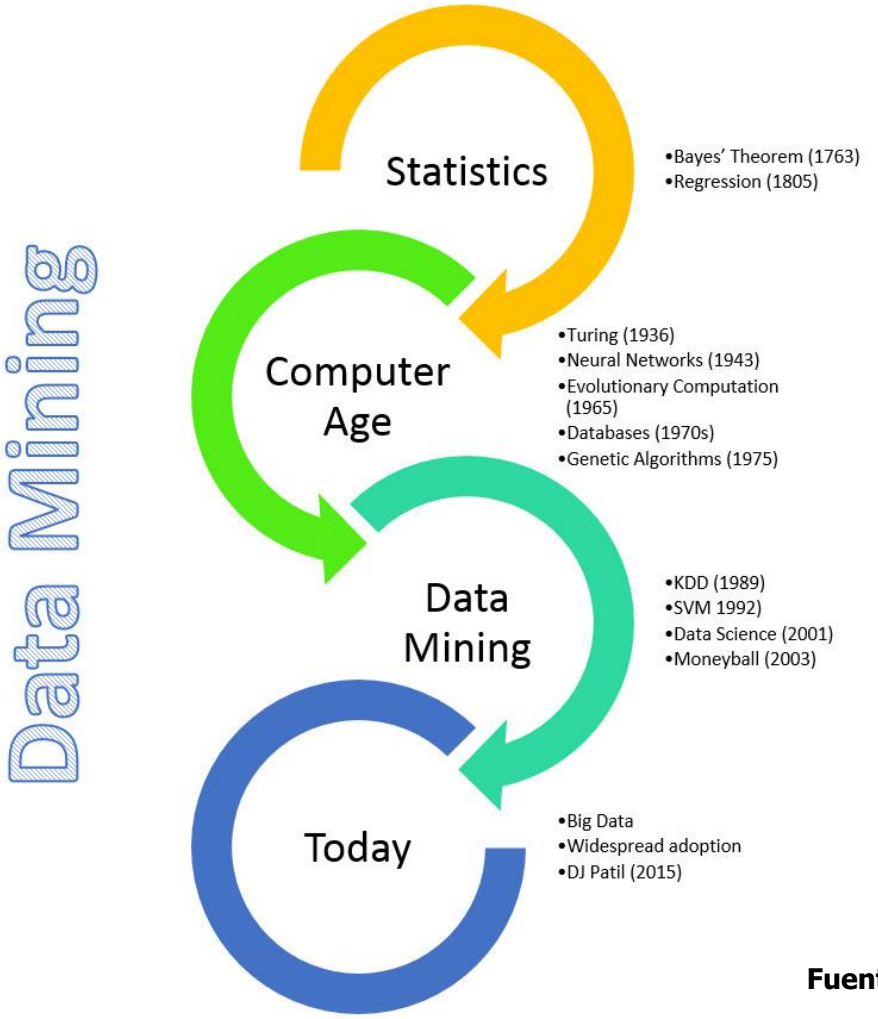
Datos Masivos (Big Data)

- son datos cuyo volumen, diversidad y complejidad requieren nueva arquitectura, técnicas, algoritmos y análisis para gestionar y extraer valor y conocimiento oculto en ellos.
- generalmente macrodatos no estructurados
- sus principales características son el volumen, la velocidad, la variedad, la veracidad y el valor.

Historia Visual de la Inteligencia Artificial

Fuente: @Parisa__Rashidi | High-res version here: https://bit.ly/3QN5xc5

Orígenes de la Minería de Datos

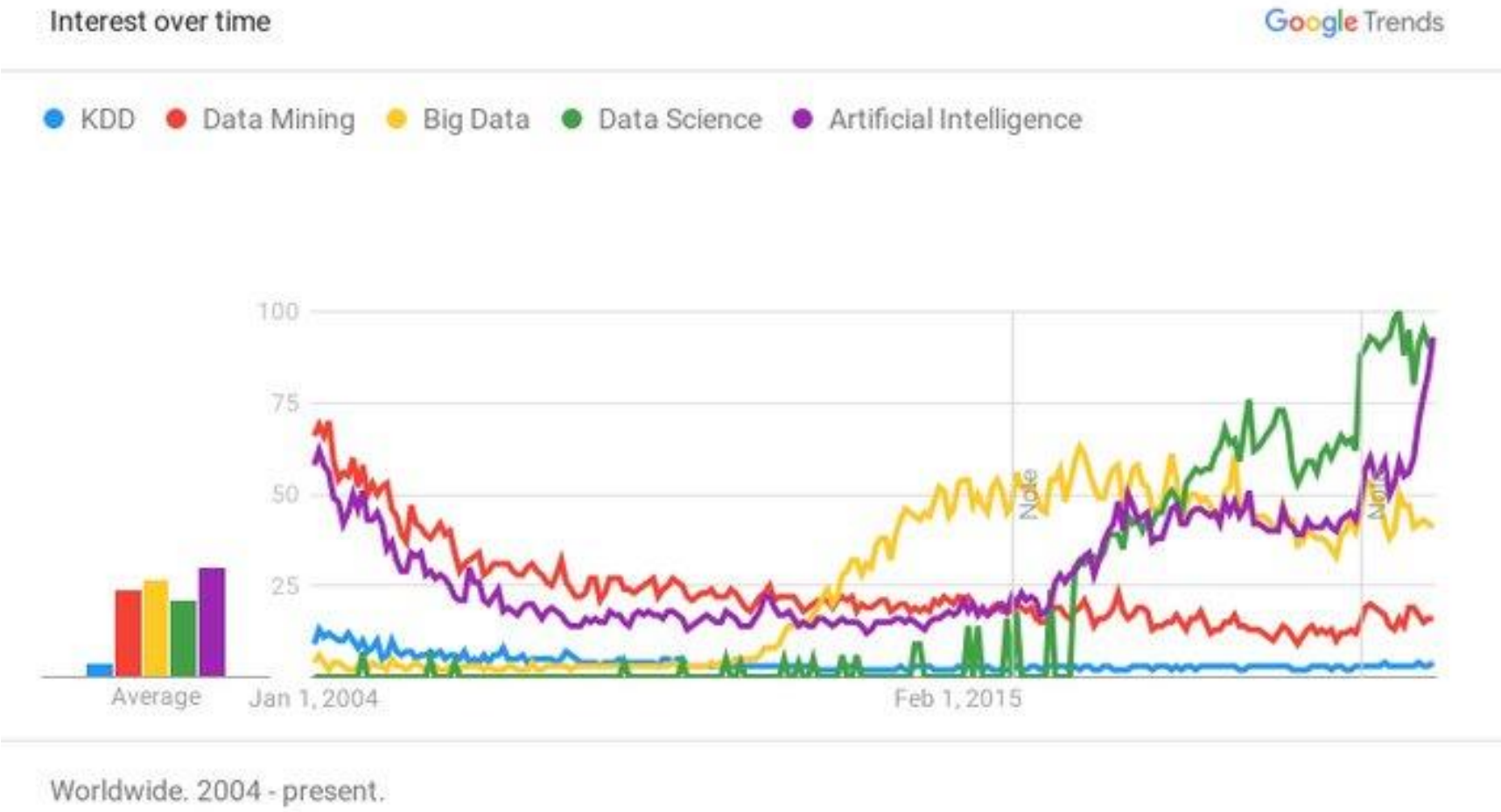


- Extrae ideas de la AI, ML, reconocimiento de patrones, estadística y sistemas de BD.

- Existen diferencias en términos de:
  - datos utilizados
  - los objetivos

Fuente: https://dataconomy.com/2016/06/16/history-data-mining/

Evolución de algunos conceptos



¿Por qué es importante entender estas diferencias?...
...porque este curso se centrará en la **Minería de Datos**.

Motivación de la Minería de Datos

- Necesidad de analizar grandes volúmenes de datos para obtener información desconocida que sea útil para tomar decisiones.
  - Volumen y variedad de información informatizada que **desborda la capacidad humana**.
  - Uso de técnicas que imiten la cualidad humana del **aprendizaje**, es decir, con capacidad de extraer nuevo conocimiento a partir de experiencias (ejemplos).
  - Las **decisiones** se basan en la información de **experiencias** pasadas extraídas de fuentes muy diversas.
  - Se cuenta con información histórica que es útil para predecir.

Minería de Datos

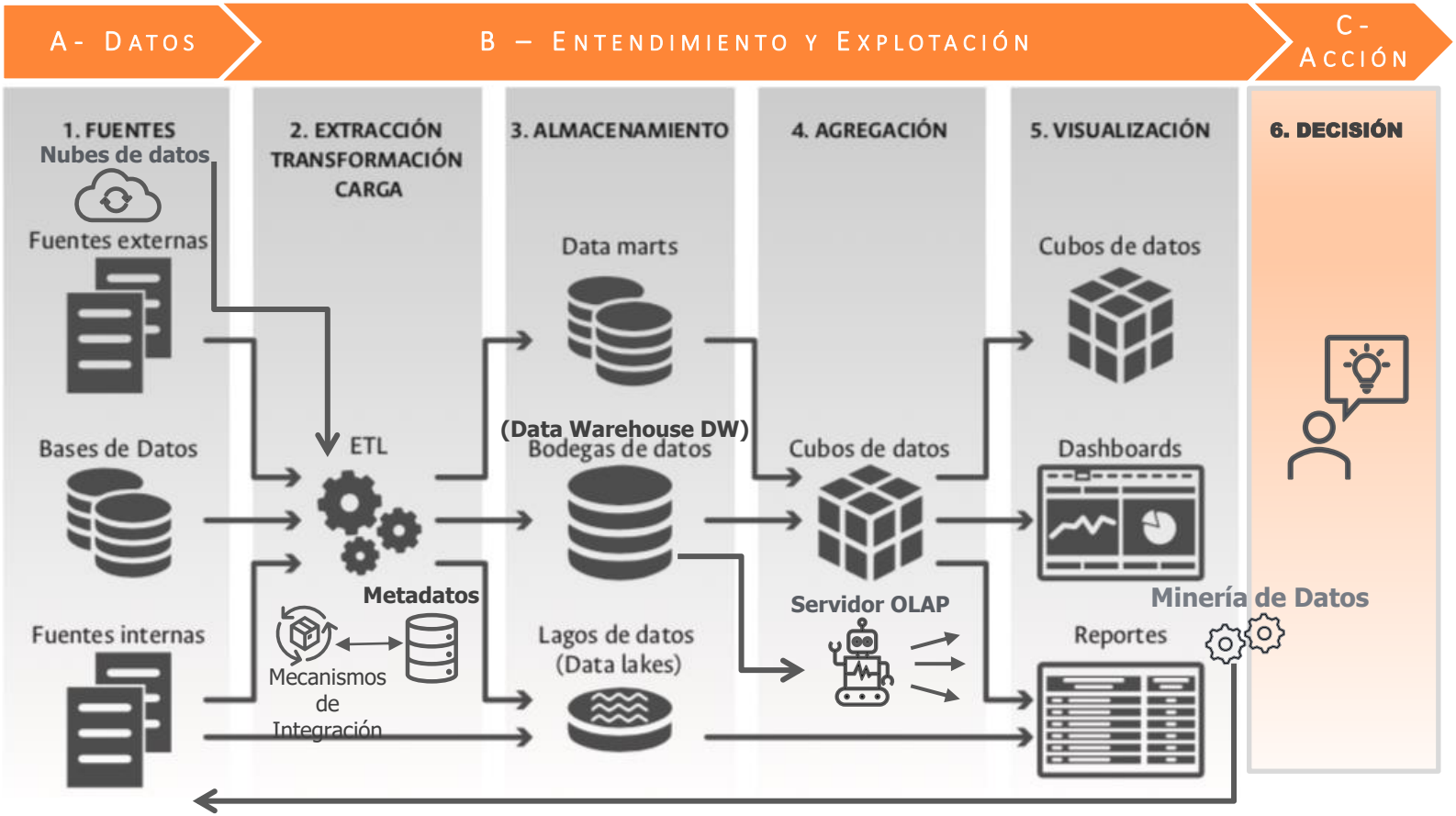
- Proceso automático que permite extraer y descubrir **patrones de conocimiento interesantes**, **no triviales**, **previamente desconocidos** y **potencialmente útiles** de los datos y descubrir relaciones entre variables.
- Sirve de ayuda en el proceso de toma de decisiones, formando parte del conjunto de tecnologías aplicables a la Inteligencia de Negocio (BI).
- Fase esencial del proceso de *"Descubrimiento de Conocimiento a partir de Bases de Datos"* (KDD, del inglés Knowledge Discovery from Databases), aunque los términos suelen ser usados como sinónimos.

Fuente: U. Fayyad et al. (1996) https://dl.acm.org/doi/10.1145/240455.240464

¿ Qué es y qué no es Minería de Datos?

Que sí es…	Que no es…
Grupos de documentos/usuarios similares.	Localizar un número telefónico en el directorio. *No, es una simple consulta a la BD.*
Preferencias de compras de los usuarios – e-Commerce.	Consultar en un buscador información acerca de un tópico en particular.
Zonas de mayor criminalidad - prevención.	Dividir a los clientes de una empresa según su rentabilidad. *No, este es un cálculo contable, seguido de la aplicación de un umbral. Sin embargo, predecir la rentabilidad de un nuevo el cliente sería la minería de datos.*
Predicciones meteorológicas - agro.	Extraer la frecuencia de una onda sonora. *No, porque eso es procesamiento de señales.*

ARQUITECTURA BI



- **A - Datos**: Centralizar información de múltiples fuentes en un DW.
- **B - Entendimiento y Explotación**: Herramientas de BI y DM para analizar y mejorar el entendimiento del negocio.
- **C - Acción**: Actuar sobre los hallazgos realizados en el análisis.

Fuente: https://estadisticaun.github.io/L_Conceptual/2-4-inteligencia-de-negocios.html

OLAP vs Minería de Datos

Herramientas OLAP	Minería de Datos
Facilidad para **manejar** y **transformar** datos.	Extrae **patrones** a partir de los datos, se construyen **modelos**, descubre **relaciones** entre atributos, **tendencias**, etc.
Producen **información** (datos agregados y combinados, medidas derivadas)	Produce **patrones** de conocimiento a partir de **reglas**.
Permite al usuario **analizar** los datos desde diferentes vistas.	Analiza los datos y ayuda al usuario a **tomar decisiones** a partir del **conocimiento** descubierto.

OLAP vs Minería de Datos

- El análisis OLAP puede responder a preguntas como:
  - ¿Han subido las ventas en el mes de Abril?
  - ¿Las ventas del producto X bajan cuando se promociona el producto Y?
  - ¿Venden más las sucursales del Gran Buenos Aires o del Interior?
- La minería de datos puede responder a preguntas como:
  - ¿Qué factores influyen en la venta del producto X?
  - ¿Cuál será el producto más vendido si se abre una sucursal en Córdoba?
  - ¿Cuándo un cliente compra el producto Y, qué otro/s producto/s suele comprar mayormente?

Requerimientos Minería de Datos

- ¿Qué se necesita para hacer minería de datos?
  - Herramientas de SW
  - Datos, digitalizados y de buena calidad
  - RRHH especialistas: técnico, analítico y de negocios

Áreas de Aplicación

- Comercio / Marketing
  - Identificar patrones de compra de los clientes.
  - Buscar asociaciones entre clientes y características demográficas.
  - Predecir respuesta a campañas de mailing.
- Análisis de canasta de compra

.

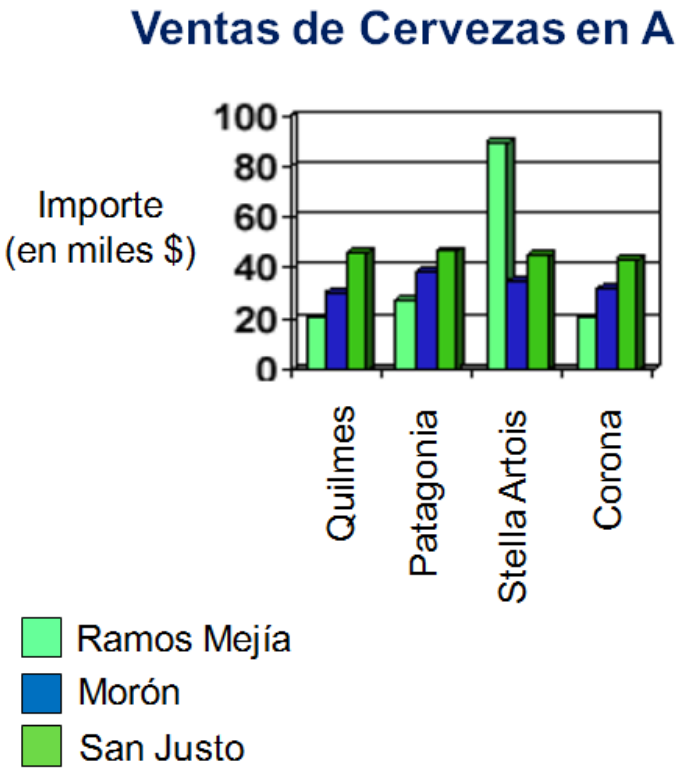
Áreas de Aplicación

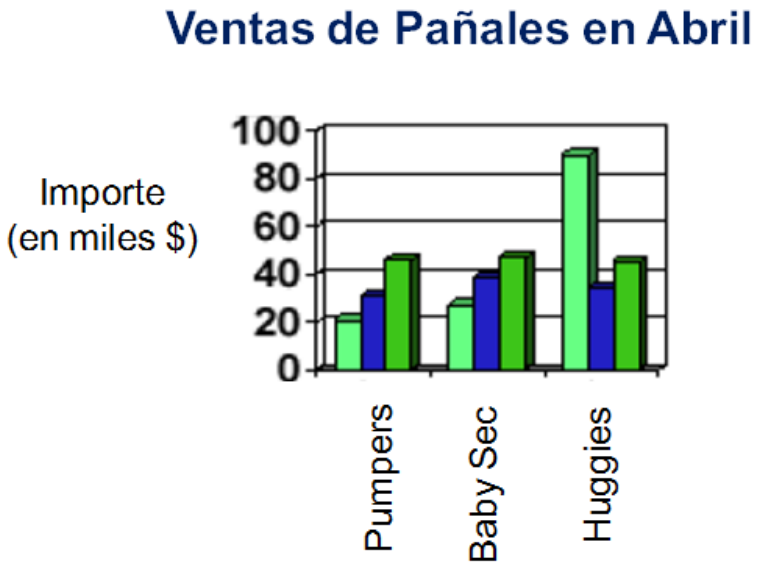
- Bancos
  - Detectar patrones de uso fraudulento de tarjetas de crédito.
  - Identificar clientes leales.
  - Predecir clientes con probabilidad de darse de baja.
  - Determinar gasto en tarjetas de crédito por grupos.
  - Encontrar correlaciones entre indicadores financieros.

Áreas de Aplicación

- Salud Privada
  - Identificar patrones de comportamiento de pacientes con alto riesgo.
  - Análisis de procedimientos médicos.
- Medicina
  - Segmentación de pacientes para una atención más inteligente según su grupo.
  - Estudio de factores (genéticos, neurológicos, alimenticios, etc.) de riesgo/salud en distintas patologías.
  - Identificación de terapias médicas satisfactorias para diferentes enfermedades.

Caso de Estudio: Marketing-Ventas





- Si se realiza sólo la toma de decisión en función de los informes (datos) de ventas de cervezas y pañales.

¿Qué información aporta?

Caso de Estudio: Marketing-Ventas

- *Objetivo*: determinar grupos de ítems que tienden a ocurrir juntos en una misma transacción de compra.
- Utilizando minería de datos se puede descubrir información como:
  - Los clientes que compran cervezas también compran papas fritas y leche. *¡Para eso no es necesario el uso de técnicas de DM!*
  - Los viernes por la tarde, con frecuencia, quienes compran pañales también compran cerveza.
- ¿Qué significa esto? ¿A qué se debe?
- ¿Qué acciones debemos realizar?

Caso de Estudio: Marketing-Ventas

- Algunas explicaciones probables:
  - Se acerca el fin de semana
  - Hay un bebé en casa
  - No quedan pañales
  - Los padres compran pañales al salir de trabajar.
  - ¡Los padres no pueden salir!
  - Se compra cerveza para ver un partido/película
- Aparecen asociaciones:

  Pañales → Cerveza

  Pañales → Cerveza, Leche **[sup=5%, conf=75%]**

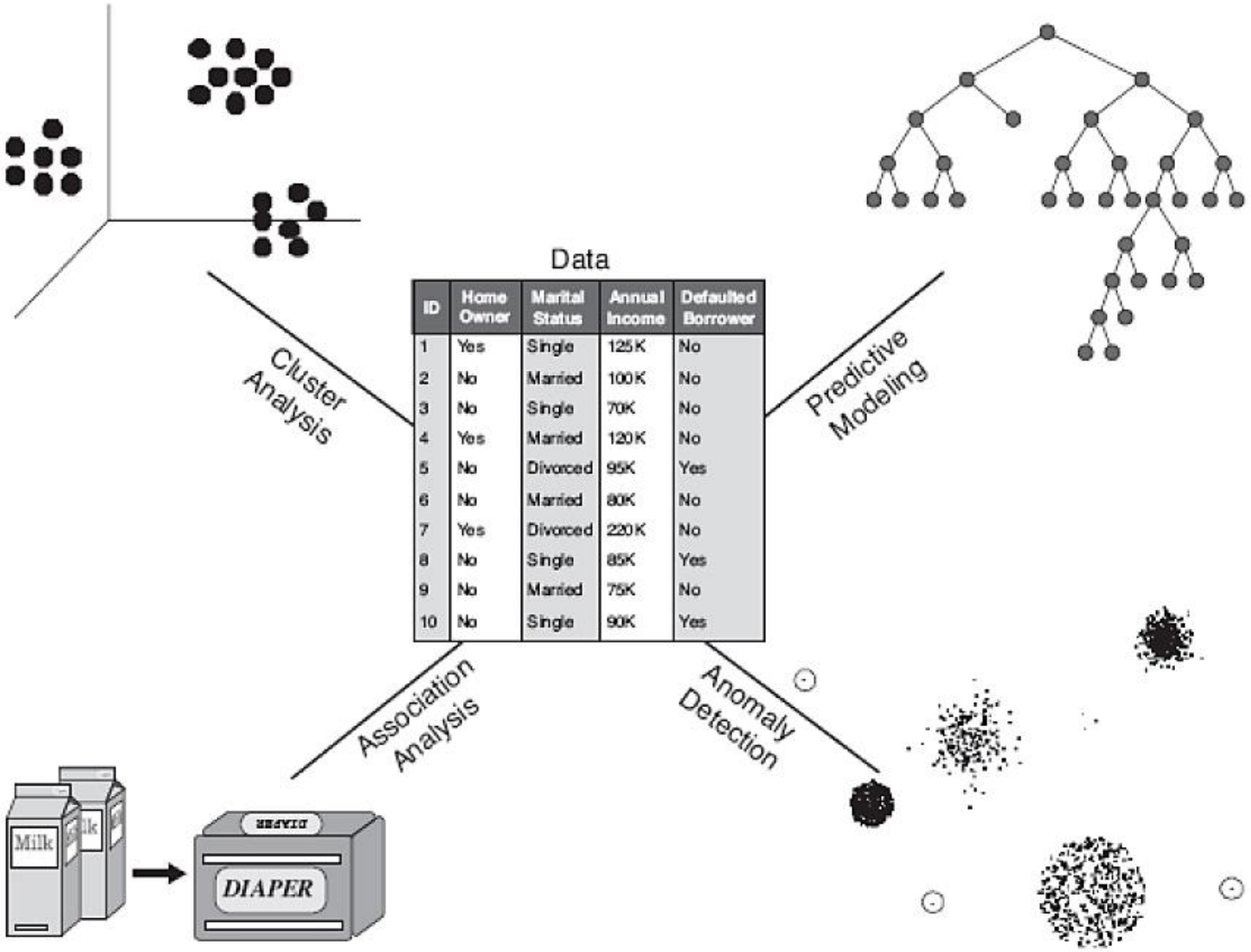
  Regla de Asociación:

  "El 75% de los clientes que compran Cerveza y Leche también compran Pañales.
  Y el 5% de los clientes compran todos estos productos juntos."

Caso de Estudio: Marketing-Ventas

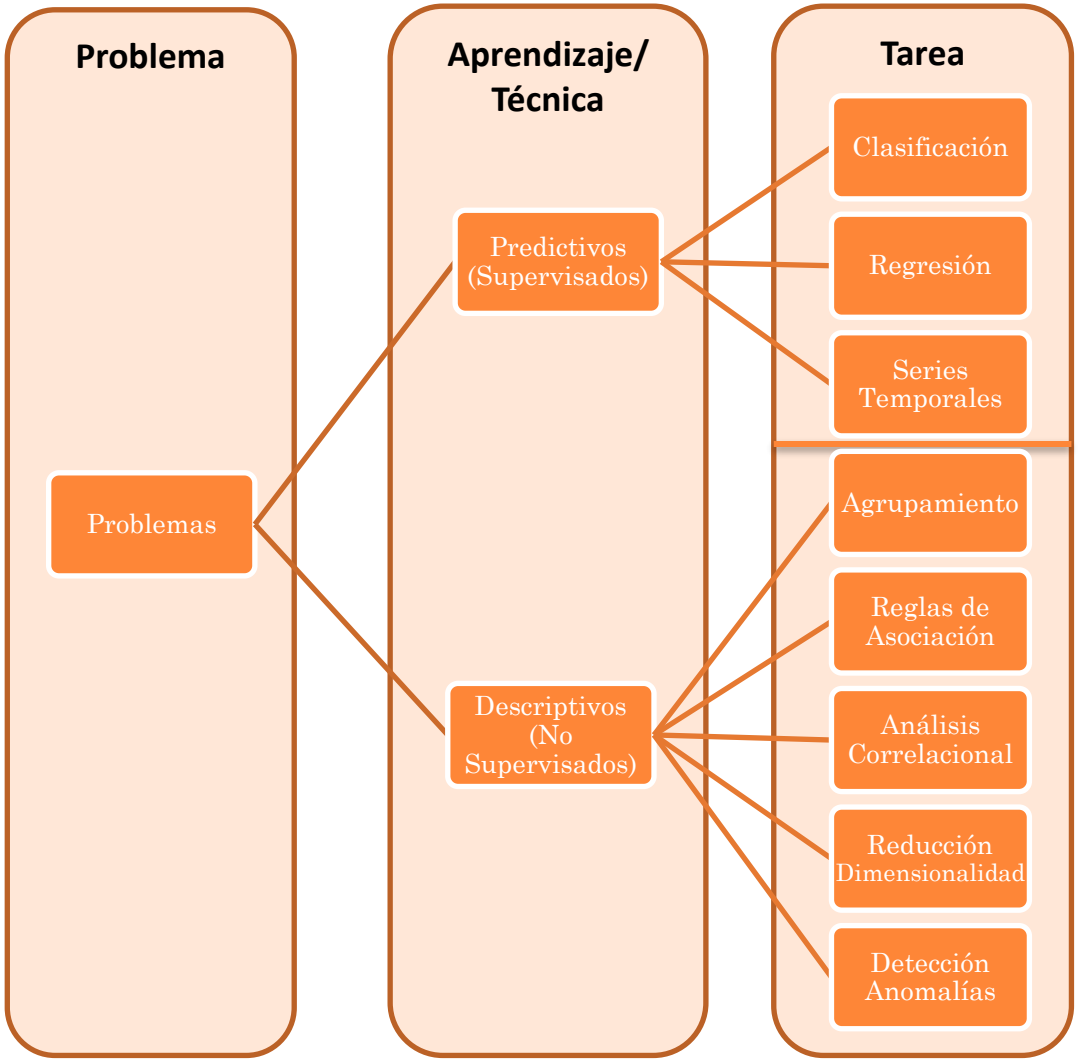
- Acciones a realizar:
  - Planificar la disposición de los productos en las góndolas:
    - Las leches al lado de los alimentos lácteos para bebés y niños
    - Las cervezas frente a la góndola de snacks.
  - Poner los aperitivos que más margen dejan entre los pañales y las cervezas.
  - Poner ofertas de pañales.
  - Poner productos de bebés en oferta y cerca de las cervezas.
  - Ofrecer cupones de descuento para el producto "complementario", cuando uno de los productos se venda por separado.

Tareas de Minería de Datos



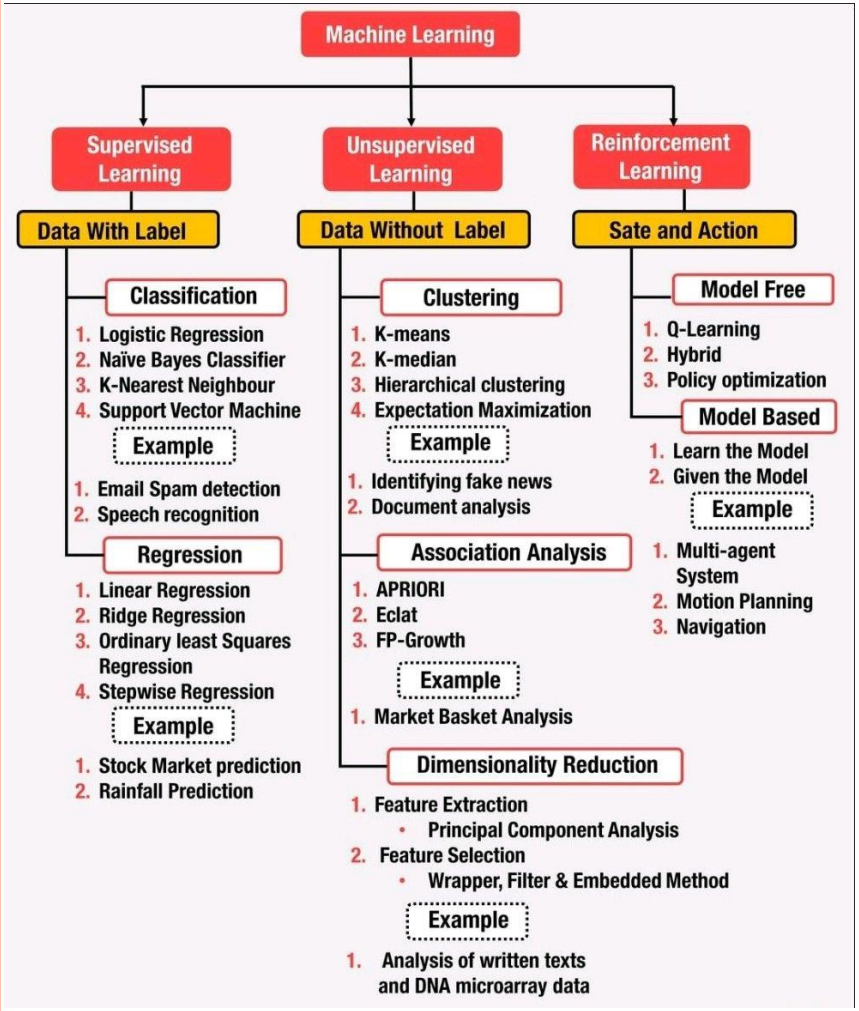
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	80K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

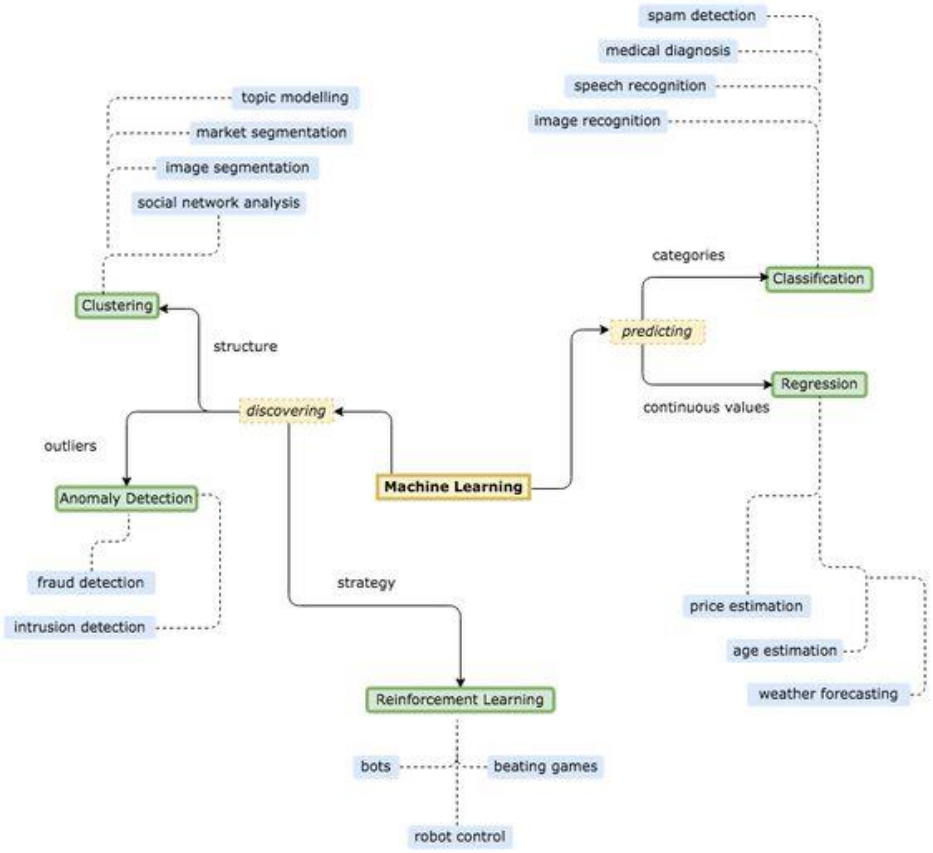
Técnicas de Minería de Datos



Las técnicas de minería de datos son herramientas que facilitan el descubrimiento de conocimiento.

Técnicas de Minería de Datos - Detallado





Fuente: https://www.kaggle.com/discussions/getting-started/350099

Técnicas de Minería de Datos

- Luego de aplicar procesos de limpieza, normalización y análisis de datos (usando conceptos de probabilidad y estadística con diferentes métodos de tratamiento sobre datos y sus fallas)
  ➔ conjunto de datos listo ➔ para abordar algoritmos de Minería de datos.

Objetivo: entender la naturaleza del problema y patrones subyacentes a ser asimilados o "aprendidos" para realizar:

- una clasificación o una predicción basada en un determinado número de ejemplos, en caso de supervisión,
- o agrupando por similitud, en casos no supervisados

Técnicas de Minería de Datos

- Se busca generalizar, aprender conceptos a partir de un conjunto de ejemplos y sus características. Cuantos más ejemplos, probablemente sea más fácil la tarea.
- Son robustos sistemas de regresión, capaces de ajustarse a una altísima dimensionalidad y una enorme complejidad, difícil de entender.
- El Aprendizaje Inductivo consiste en construir un modelo general a partir de información específica (instancias).

Como principio metodológico, ante igualdad de condiciones (por ejemplo, igual desempeño), debemos elegir al modelo más simple porque esperamos que generalice mejor.

Modelos y Algoritmos

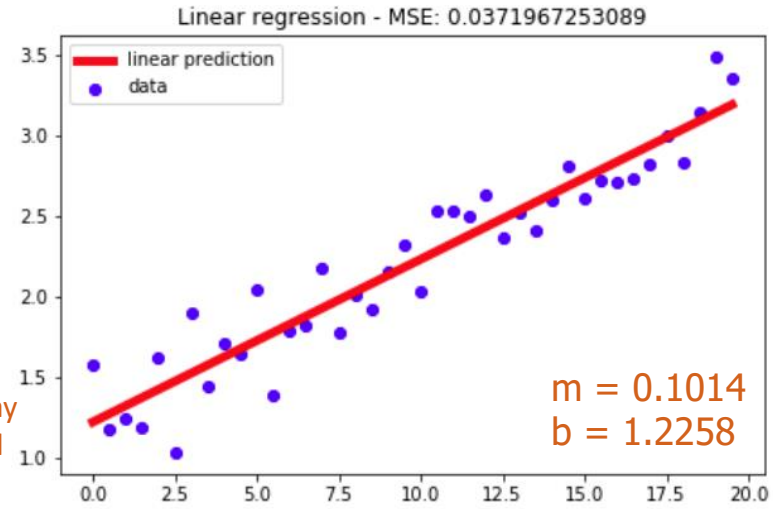
Minería de Datos convierte datos en modelos, extrayendo conocimiento de los datos.

Utilizando algoritmos en dos etapas:

- Aprendizaje | Inducción (Entrenamiento).
- Resolución | Deducción (Cálculo de la predicción o clasificación)

¿Cómo aprende el Modelo?

Encontrando los valores de los hiperparámetros del modelo que minimicen el error.



Regresión:
Las entradas (valores de los atributos) y la salida son todas numéricas. La salida es la suma de los valores ponderados de los atributos de entrada. Se busca encontrar buenos valores para las ponderaciones Hay diferentes maneras de hacerlo: la más utilizada es minimizar el error al cuadrado.

Trabajo con Algoritmos

Luego del proceso de limpieza de datos, ya con el set de datos confiable, continuar con los siguientes pasos:

1- Selección del Algoritmo

- Elección de que va utilizarse y testear por diferentes criterios.

2 – Entrenamiento

- Verificación de los resultados del entrenamiento, conforme al algoritmo elegido y los datos disponibles

3 - Evaluación de Calidad

- Utilización de métricas y métodos para decidir si el algoritmo es el adecuado o se debe cambiar, también si es preciso ajustar sus hiperparámetros.

4 – Ajuste de hiper parámetros

- Modificación según situación, los datos y métricas obtenidas durante y luego del entrenamiento.
- Volver a Paso 2.

5 - Objetivos y Métricas

- Si se está satisfecho ➔ Modelo entrenado FIN
  Si no ➔ Volver a Paso 1.

Técnicas de Minería de Datos

Métodos Predictivos – Aprendizaje Supervisado

- aprenden a predecir la clase/etiqueta de una nueva instancia mediante el entrenamiento usando ejemplos pre-etiquetados (ya clasificados).
- una empresa cuenta con un dataset de datos crediticios de clientes y una de las columnas dice si es viable ofrecerle un préstamo o no (variable objetivo/clase/etiqueta).
- el algoritmo deberá encontrar qué tienen en común ambos grupos (aprender), para poder predecir automáticamente si conviene o no dar un préstamo a un nuevo cliente.
- se intenta "minimizar la función coste", reducir el error.

Ejemplos

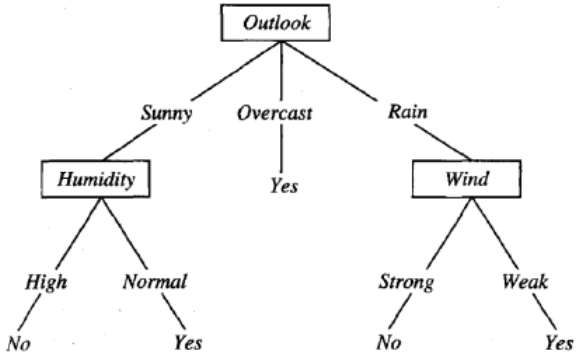
- Clasificación: *¿Esta transacción es fraudulenta? |Se necesita predecir las células tumorales como benignas o malignas. | ¿Esta imagen es un gato? (Reconocimiento Imagen) | ¿Quién nos está llamando? (Reconocimiento Voz)*
- Regresión: *¿Qué valor de franquicia es más probable que contrate el cliente Carlos Gómez para su auto? | ¿En qué dosis el Fármaco A afecta menos a la presión arterial?*

- Series temporales: *Análisis de tendencias ¿Cuál será la producción en Kg de la cosecha cada semestre? | Se necesita predecir el valor de las acciones de una empresa en la bolsa minuto a minuto.*

Técnicas de Minería de Datos

- Tareas de Clasificación
  - Predicen un valor discreto
    - SI / NO
    - Alto / Mediano / Bajo

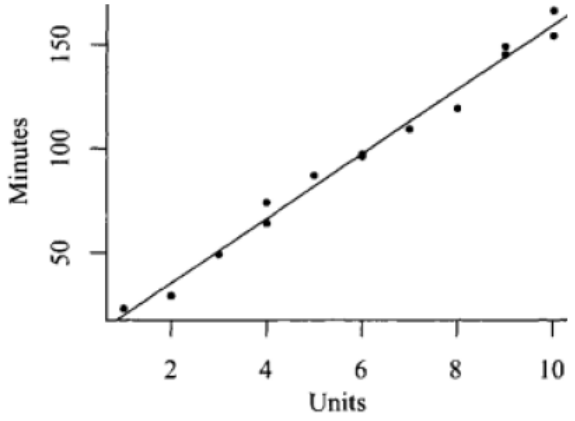
Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes



- Tareas de Regresión
  - Predicen un valor continuo
    - Importes
    - Cantidades

Row	Minutes	Units
1	23	1
2	29	2
3	49	3
4	64	4
5	74	4
6	87	5
7	96	6

$$\text{Minutes} = 4.162 + 15.509 \cdot \text{Units}$$



Técnicas de Minería de Datos

Métodos Descriptivos – Aprendizaje No Supervisado

- no hay conocimiento a priori sobre el problema, no hay instancias etiquetadas, no hay supervisión sobre el procedimiento.
- describen el comportamiento de los datos (encuentran patrones: relaciones entre los datos y sus características) de forma que sea interpretable por un usuario experto.
- el algoritmo deberá agrupar las instancias por sí solo, extrayendo nuevas variables que expliquen los datos. Se busca explorar el dato.
- **Ejemplos**
  - Agrupamiento/Segmentación: *Identificar grupos de viviendas de acuerdo a su tipo, valor o situación geográfica | Los clientes sin televisión y con bicicleta tienen características muy diferenciadas del resto. | Clientes habituales o Clientes ocasionales con presupuesto limitado.*
  - Reglas de Asociación: *Los clientes que compran pañales suelen comprar cerveza. | ¿Cómo afecta la demografía de la vecindad en la compra de los clientes?*
  - Análisis Correlacional: *El tabaco y el alcohol son los factores que más inciden en la enfermedad X.*

Técnicas de Minería de Datos

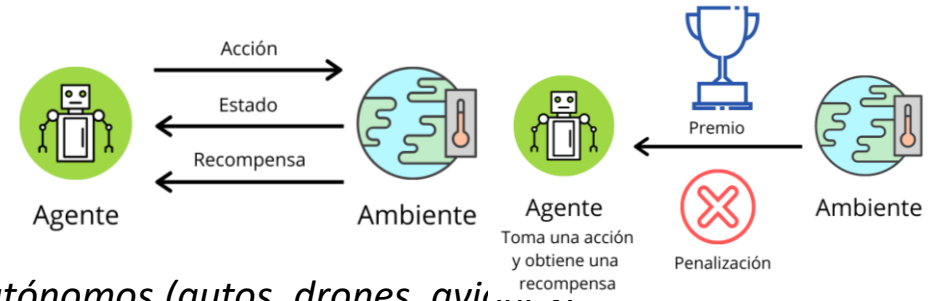
- Tareas de Asociación
  - Descubren por medio de reglas de asociación hechos que ocurren en común dentro de un determinado conjunto de datos.
  - Utilizado en análisis de canasta (market basket analysis).
    - {cebollas, vegetales} $\Rightarrow$ {carne}
    - {cerveza} $\Rightarrow$ {leche, pañales}
- Tareas de Segmentación (Clustering)
  - Agrupamiento jerárquico o no jerárquico de datos de acuerdo con un determinado criterio.
    - Jerárquico: Puede ser aglomerativo o divisivo.
    - No Jerárquico: N° Grupos determinados de antemano.

Aprendizaje por Refuerzo

- no tenemos una "etiqueta de salida ➔ no es de tipo supervisado
- tampoco son de tipo no supervisado, en donde se intenta clasificar grupos teniendo en cuenta alguna distancia entre muestras.
- estos algoritmos también aprenden por sí mismos, intentan que la máquina aprenda basándose en un esquema **de "premios y castigos"** -cómo con el perro de Pávlov- en un **entorno** en donde hay que **tomar acciones** y que está afectado por **múltiples variables** que **cambian con el tiempo.**
- se intenta "maximizar la recompensa", a pesar de cometer errores o de no ser óptimos.

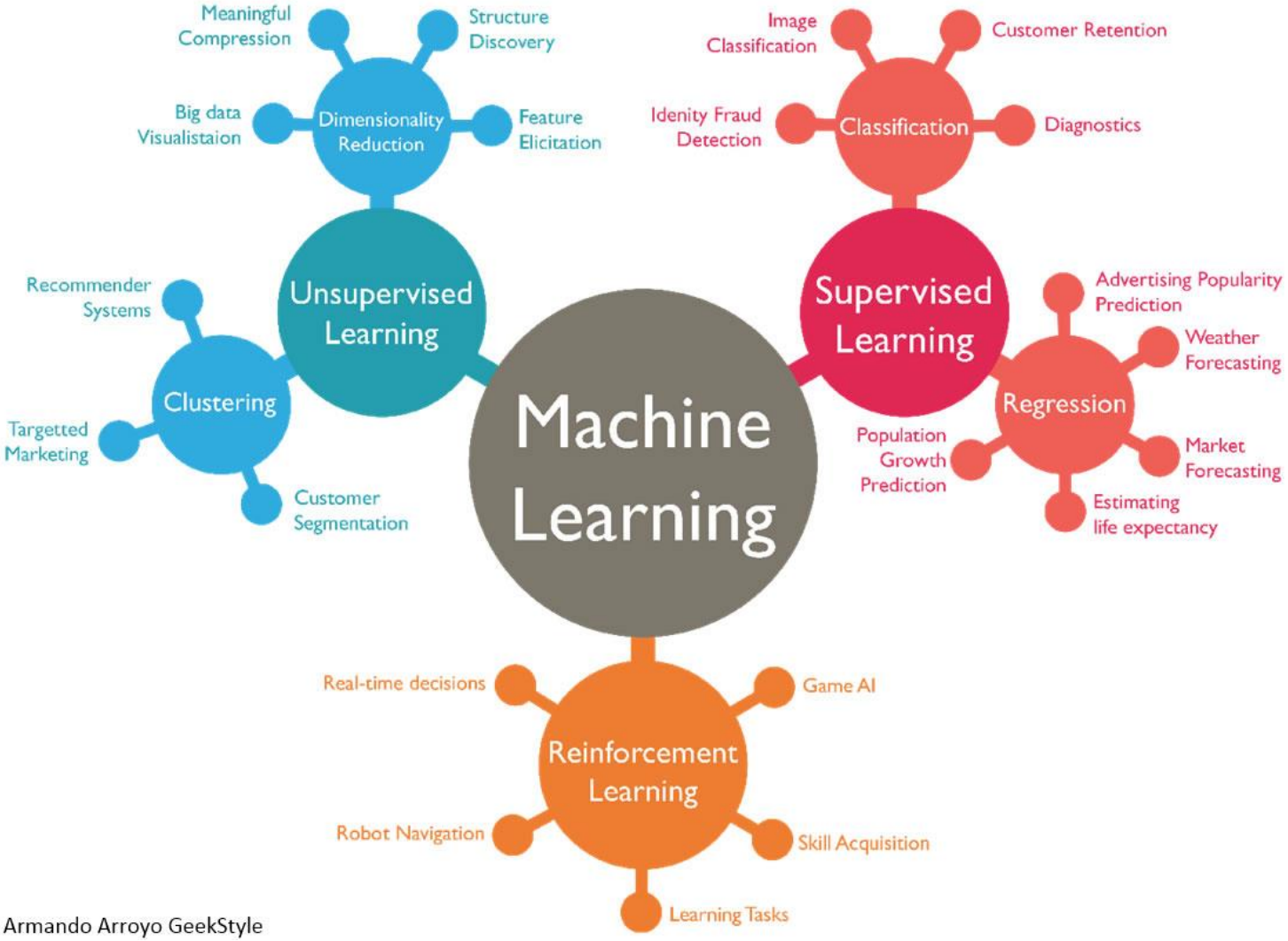
Ejemplos

- *juegos de mesa y videojuegos*
- *robótica, máquinas industriales*
- *medicina y la biología*
- *sistemas de navegación vehículos autónomos (autos, drones, aviones)*
- *reducción del consumo energético*



Fuente: https://www.aprendemachinelearning.com/aprendizaje-por-refuerzo/
https://github.com/jbagnato/machine-learning/blob/master/Ejercicio_Reinforcement_Learning_Pong.ipynb

Técnicas de Minería de Datos



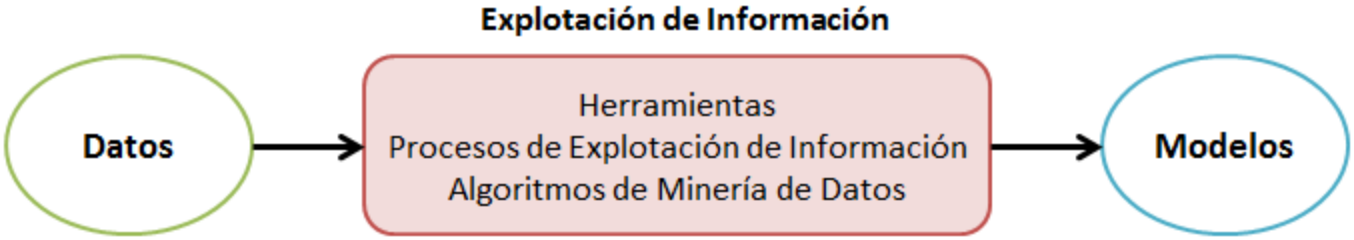
Armando Arroyo GeekStyle

Explotación de Información

- La Explotación de Información es la sub-disciplina informática que aporta a la Inteligencia de Negocios las herramientas (procesos y tecnologías) para la transformación de información en conocimiento.
- Utiliza la Minería de Datos.
- Aborda la solución a problemas de predicción, clasificación y segmentación.
- La minería de datos y la explotación de información no son conceptos equivalentes.

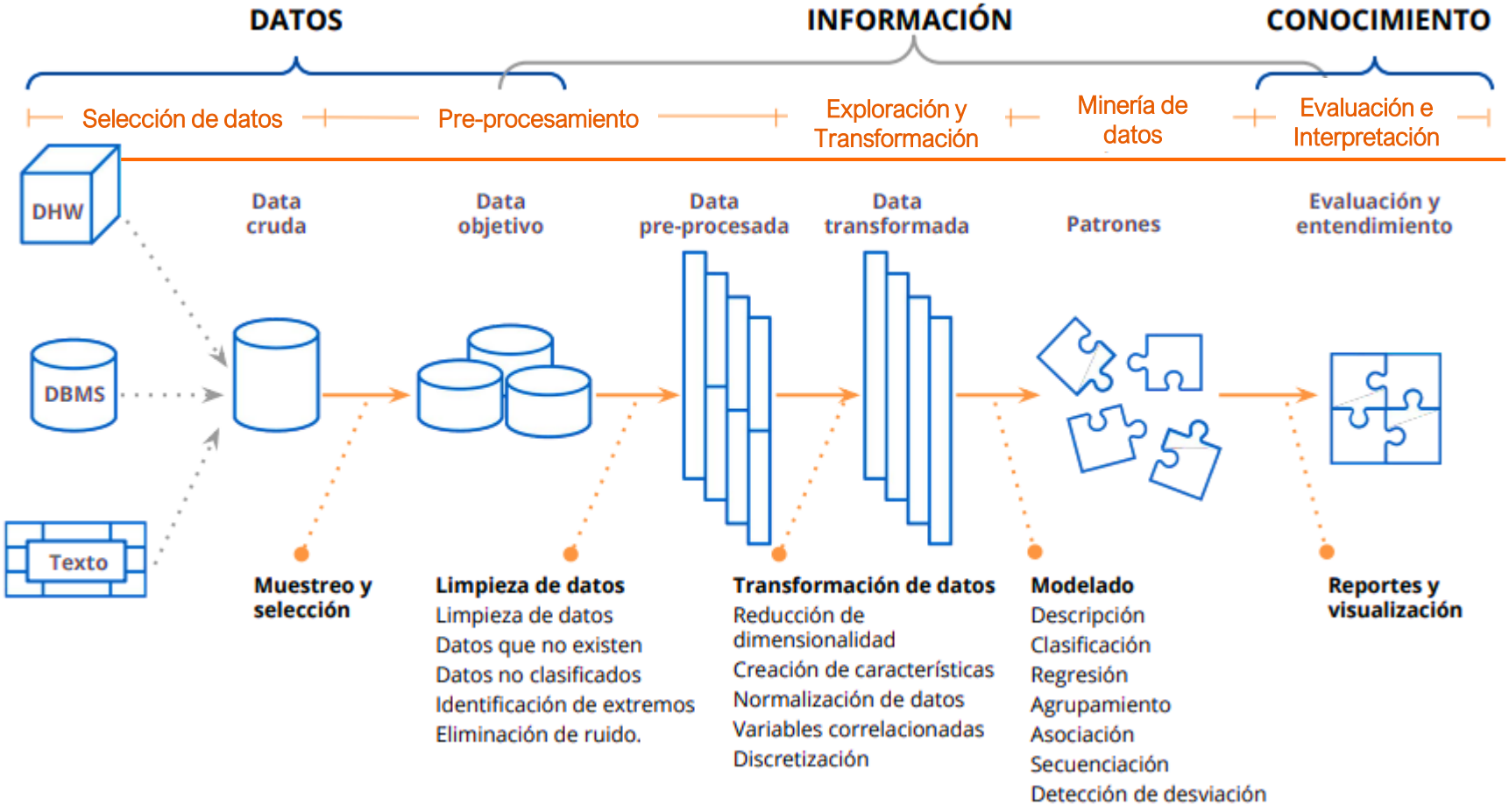
Explotación de Información

- La minería de datos está relacionada a la programación, a los algoritmos para resolver un problema de inteligencia de negocios.
- La explotación de información está relacionada a tareas de la Ingeniería de Software, a la aplicación de técnicas y procesos ingenieriles para construir la solución de un problema de inteligencia de negocios.
- La minería de datos describe la tecnología que da soporte a la explotación de la información.



Proceso Descubrimiento de Conocimiento

- También conocido como KDD, del inglés Knowledge Discovery in Databases.



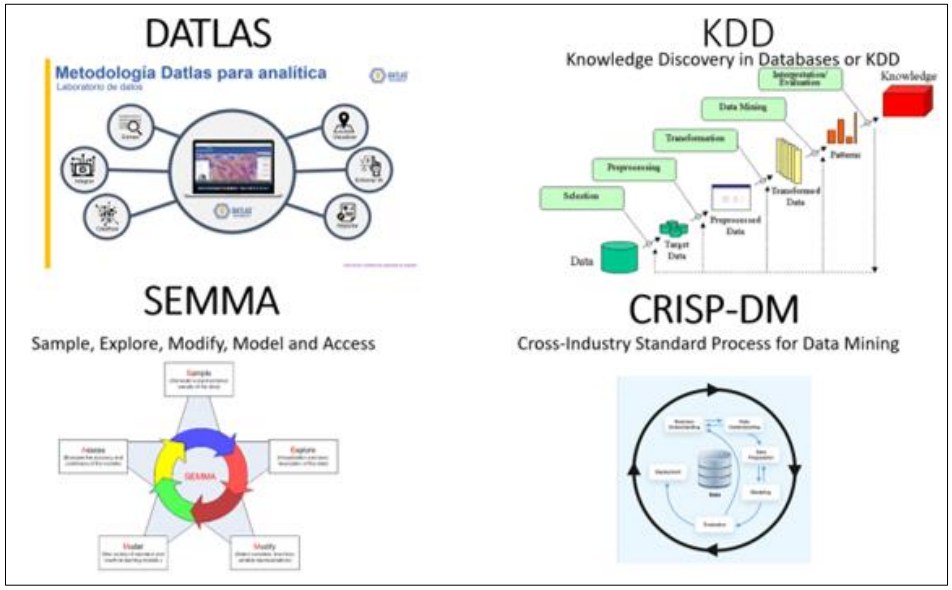
Proceso Descubrimiento de Conocimiento

- **Selección de datos**: Datos sobre los que se trabajará.
- **Preprocesamiento**: Preparación y limpieza de los datos. Estrategias para manejar datos faltantes o nulos, datos inconsistentes o que están fuera de rango.
- **Transformación**: Tratamiento preliminar de los datos, transformación, agregación, normalización y generación de nuevas variables a partir de los datos existentes.
- **Minería de Datos**: Construcción de modelos con técnicas de minería de datos y procesos de explotación de información para extracción de patrones de conocimiento.
  - Técnicas Predictivas
  - Técnicas Descriptivas
- **Evaluación e interpretación**: Evaluación del modelo construido, del conocimiento obtenido y validación si los resultados son satisfactorios en el dominio del problema.

Metodologías de Explotación de Información

- Conjunto de actividades organizadas que tienen como objetivo la realización de un proyecto de explotación de información.
- Para cada actividad se define, las entradas, las salidas y la forma en la que debe llevarse a cabo.

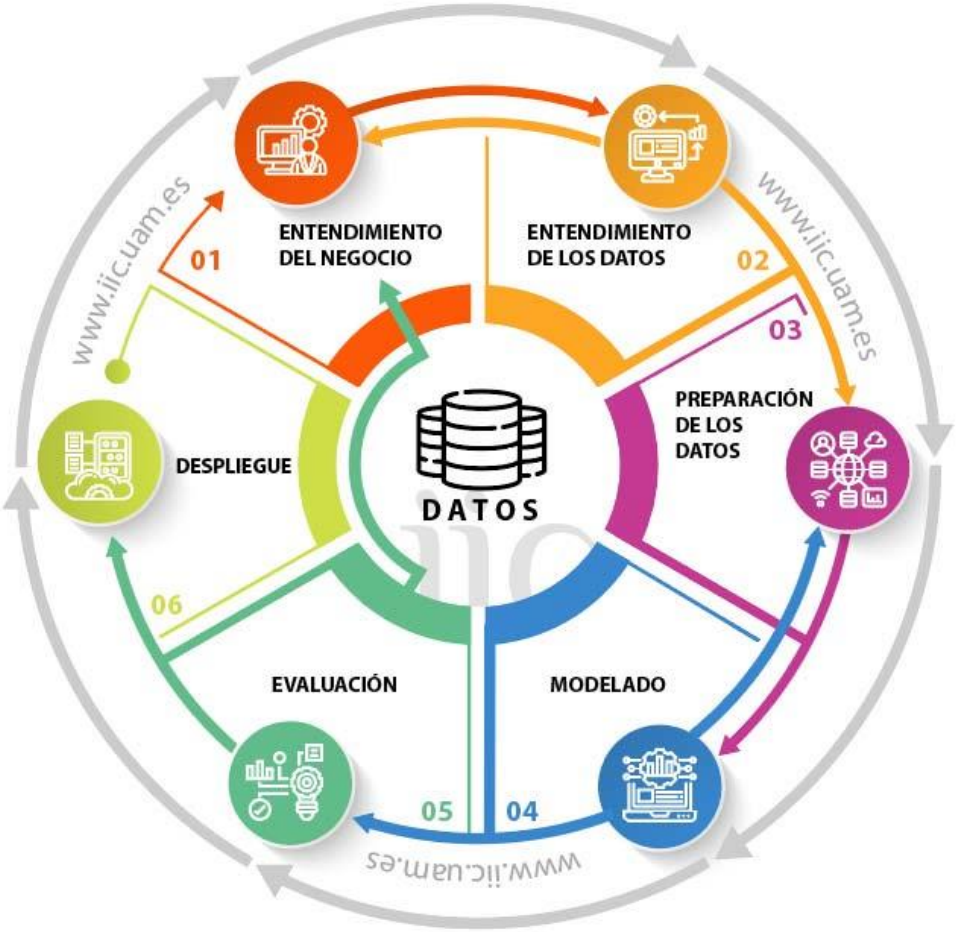
- Metodologías probadas por la comunidad científica:
  - ✓ CRISP-DM
  - ✓ SEMMA
  - ✓ P^{3}TQ
  - ✓ DATLAS



Fuente: https://blogdatlas.wordpress.com/2020/02/16/4-metodologias-para-proyectos-de-data-science-datlas-research/

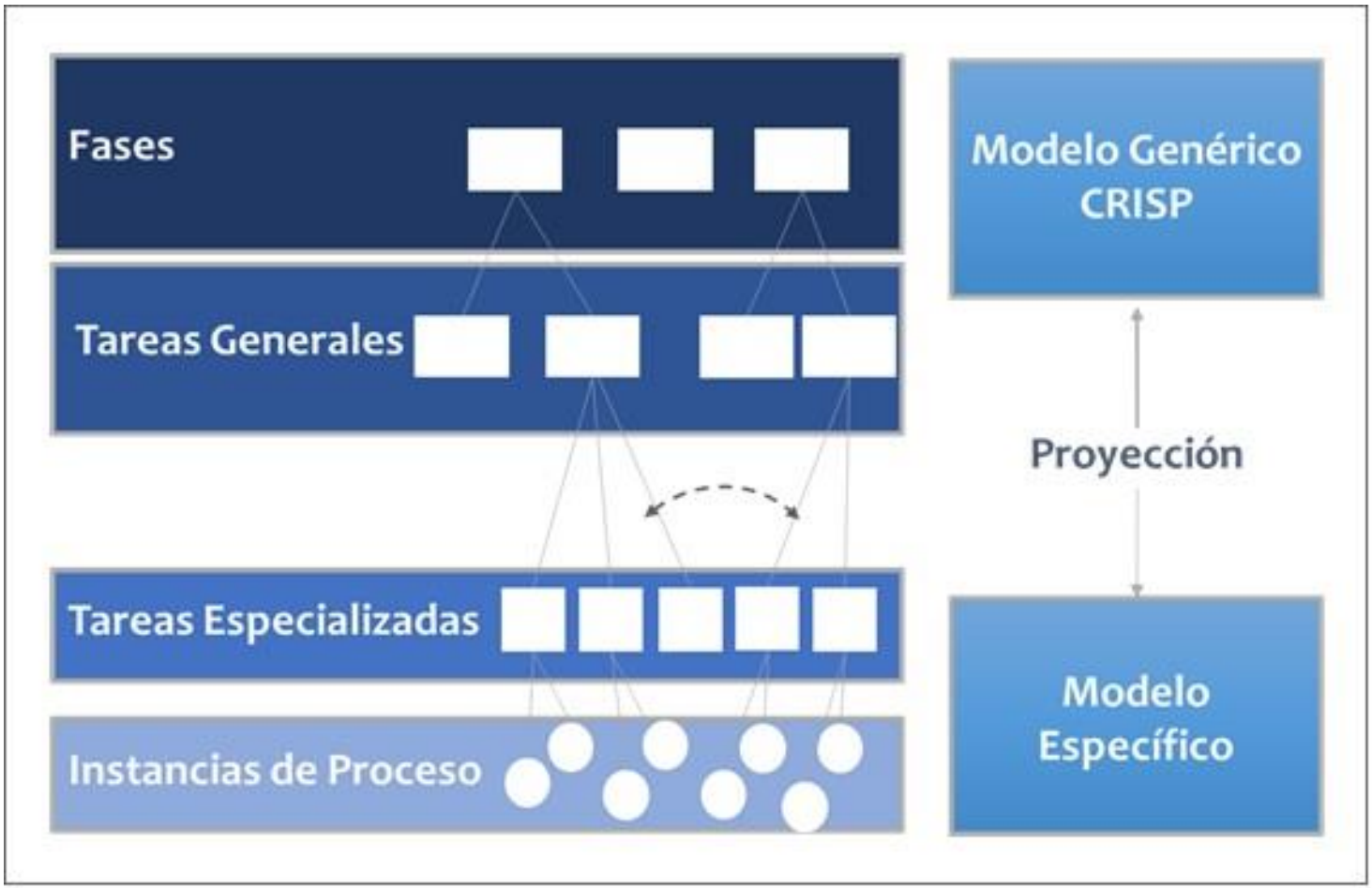
METODOLOGÍA CRISP-DM

CRISP-DM, del inglés de Cross-Industry Standard Process for Data Mining, proporciona un proceso estándar no patentado y de libre acceso para adaptar la minería de datos a la estrategia general de resolución de problemas de una unidad comercial o de investigación.



Fuente: https://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/

METODOLOGÍA CRISP-DM



Esquema de los cuatro Niveles de Abstracción de CRISP-DM

Fuente: http://cic-smartbase.blogspot.com/2016/05/metodologia-crisp-dm-parte-i.html

METODOLOGÍA CRISP-DM

Comprensión del Negocio

- Se determinan los objetivos y requerimientos del proyecto desde una perspectiva del negocio, definiendo el problema de minería y el plan de trabajo.
  - Objetivos de negocio y criterios de éxito
    - Detectar fraude con tarjetas de crédito
    - Captar nuevos clientes bancarios
    - Detectar signos tempranos de algún padecimiento clínico
    - Etc.
  - Análisis del problema
  - Objetivos de minería de datos

METODOLOGÍA CRISP-DM

Comprensión de los Datos

- Se recolectan los datos que se utilizarán y se analizan las características de los mismos. Surgen las primeras hipótesis acerca de la información que podría estar oculta.
- Atributos Nominales
  - Llamados Categóricos o Discretos
  - Número finito de valores, no tienen orden.
    - Ejemplo: género, color de ojos, sucursales, booleanos, etc.
- Atributos Ordinales
  - Llamados Numéricos o Continuos
  - Número finito de valores (enteros o reales), tienen orden
    - Ejemplo: puntuación, rangos, altura, importes, temperaturas, fechas, etc.

METODOLOGÍA CRISP-DM

Preparación de los Datos

- Comprenden actividades de tratamiento de los datos o conjunto de datos final sobre el cual se aplicarán procesos de explotación de información y minería de datos.
  - Selección, Limpieza y Transformación
- Análisis de la calidad de los datos
  - ¿Qué tipos de problemas de calidad podemos encontrar?
    - Valores anómalos (ruido, outlier)
    - Valores faltantes o nulos
    - Datos duplicados
  - ¿Cómo podemos detectarlos en los datos?
  - ¿Qué podemos hacer al respecto?

METODOLOGÍA CRISP-DM

- **Preprocesamiento de los datos**
  - Transformaciones de los datos necesarias en función del análisis previo, con el objetivo de prepararlos para aplicarles explotación de información según el problema de negocio.
    - Agregación
    - Seleccionar conjunto de atributos
    - Creación de atributos
    - Discretización
    - Transformación de atributos

METODOLOGÍA CRISP-DM

- **Modelado**: se aplican procesos de explotación de información y algoritmos de minería sobre el conjunto de datos para obtener información oculta y patrones de conocimiento.
- **Evaluación**: se analizan los patrones obtenidos en función de los objetivos organizacionales. Se determina si se ha omitido algún objetivo importante del negocio y si el nuevo conocimiento será implementado.
- **Implementación**: se comunica e implementa el nuevo conocimiento, el cual debe ser representado de forma entendible para el usuario.

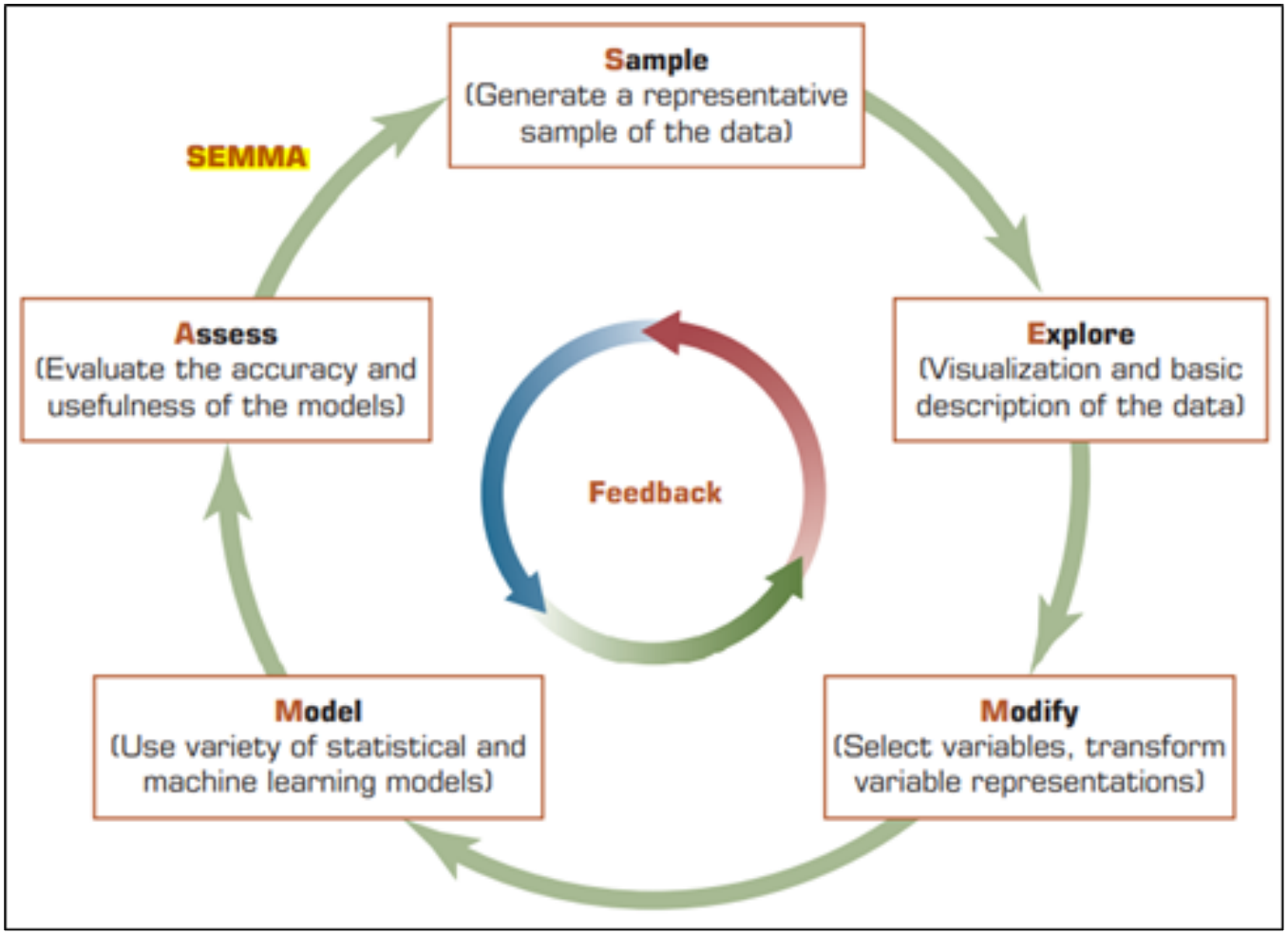
METODOLOGÍA SEMMA

SEMMA (Sample, Explore, Modify, Model and Assess) Desarrollado por SAS Institute

- **"Sample" o Muestro:** de la base de datos principal.
- **"Explore" o Exploración:** se exploran los datos para ganar entendimiento e ideas, así como definir nuestro proceso de búsqueda de anomalías, patrones y tendencias.
- **"Modify" o Modificación:** se enfoca en crear, seleccionar y transformar variables (proceso de selección). En esta etapa también se buscan anomalías y reducir el número de variables.
- **"Model" o Modelado:** se aplican distintos métodos estadísticos evaluando fortalezas y cumplimiento de objetivos.
- **"Assess" o Evaluar:** se evalúa la confiabilidad y utilidad de los hallazgos, en especial la performance.

Al igual que en KDD y CRISP-DM, si no se logran los objetivos en una primera iteración, tendremos que repetir el proceso.

METODOLOGÍA SEMMA



Fuente: https://sis.binus.ac.id/2021/09/30/data-mining-semma/

Comparación de Metodologías

KDD

- Selection
- Preprocessing
- Transformation
- Data Mining
- Interpretation/ Evaluation

SEMMA

- Sample
- Explore
- Modify
- Model
- Asssess

CRISP-DM

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

Fuente: https://blogdatlas.wordpress.com/2020/02/16/4-metodologias-para-proyectos-de-data-science-datlas-research/

Herramientas para Minería de Datos

- **Licenciadas**
  - SAS (Analytics, Enterprise Miner)
  - SPSS (IBM SPSS Statistics, IBM SPSS Modeler – ex Clementine)
  - Microsoft Azure Machine Learning
- **Libres**
  - KNIME - https://www.knime.com/
  - WEKA - http://www.cs.waikato.ac.nz/ml/weka/
  - ORANGE - https://orangedatamining.com/
  - Rapid Miner - https://rapidminer.com/
  - Python - https://www.python.org/
  - R - https://www.r-project.org/
  - Otros

Herramientas para Minería de Datos

Comparación Herramientas:

	Características	Lenguaje de programación	Sistema operativo	Precio/Licencia
WEKA	Muchos métodos de clasificación	Java	Windows, macOS, Linux	Software libre (GPL)
Orange	Crea una visualización de datos atractiva sin que se requieran muchos conocimientos previos para ello	Núcleo del software: C++, ampliación y lenguaje de entrada: Python	Windows, macOS, Linux	Software libre (GPL)
KNIME	Software de data mining de código abierto que ha democratizado el acceso a los análisis predictivos	Java	Windows, macOS, Linux	Software libre (GPL) (a partir de la versión 2.1)
RapidMiner	Apto para todos los procesos. Destaca en el análisis predictivo	Java	Windows, macOS, Linux	Freeware, diferentes versiones de pago
SAS	Caro, pero potente para grandes empresas	Lenguaje SAS	Windows, macOS, Linux	Freeware limitado a instituciones públicas, el precio se establece tras solicitud, diferentes modelos disponibles

Fuente: https://www.ionos.es/digitalguide/online-marketing/analisis-web/software-de-data-mining-las-mejores-herramientas/

Caso de Estudio – Créditos Personales

- Un banco dispone de una muestra de 144 clientes históricos a los que se les otorgó un crédito personal.
- Las muestras contienen los siguientes atributos:
  - Nivel de ingresos
  - Composición familiar
  - Tipo de vivienda
  - Servicios que posee
  - Antecedente de otros créditos
  - Resultado del otorgamiento de crédito
- El banco quiere lanzar una línea de créditos y necesita analizar la información, en base a las siguientes necesidades:
  - Identificar criterios de otorgamiento de créditos
  - Identificar y caracterizar grupos de clientes en orden a estudiar líneas de crédito diferenciales por grupo.
  - Identificar los factores de incidencia en cada grupo de clientes con ingresos superiores a $ 15.000.

Caso de Estudio – Créditos Personales

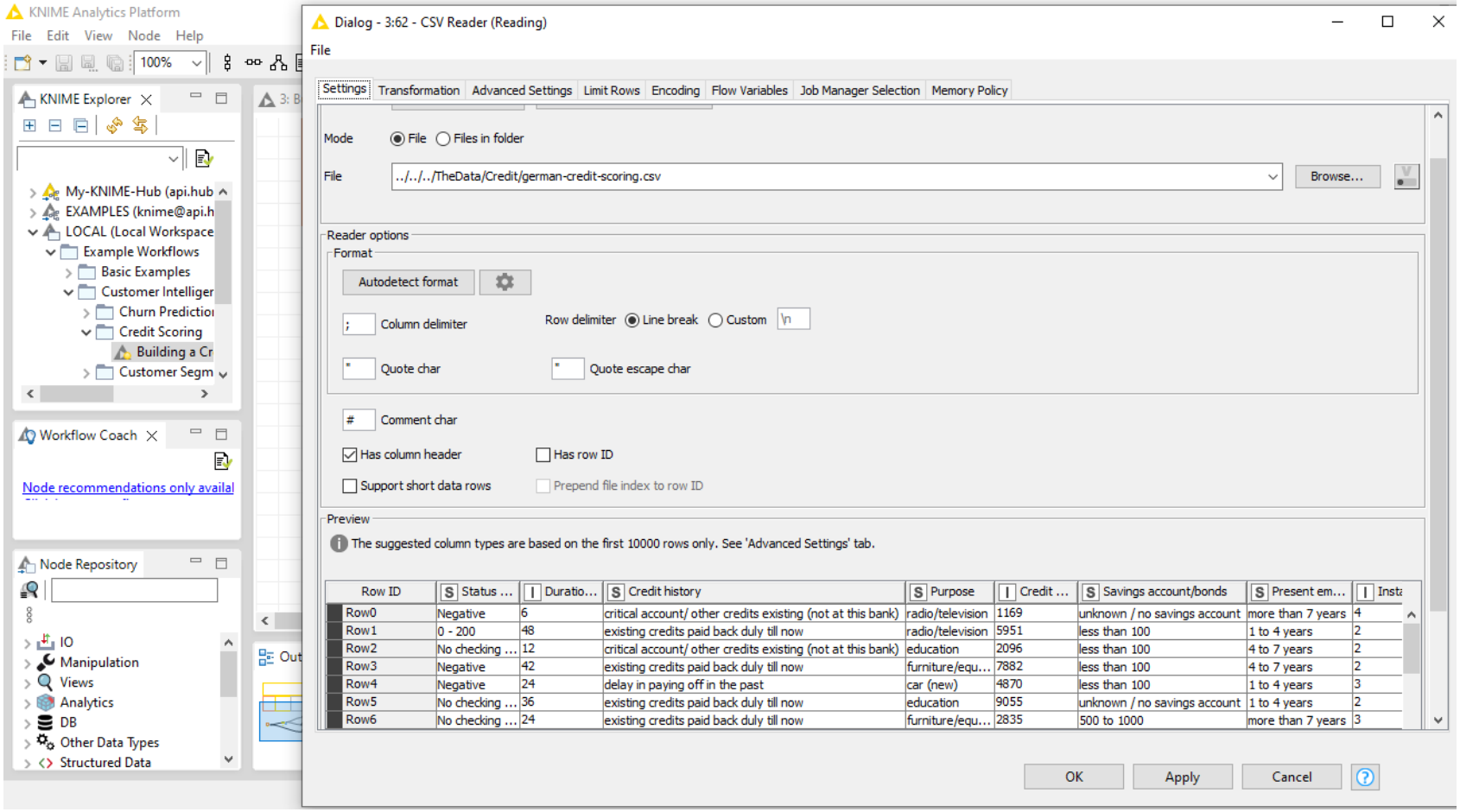
- Comprensión de los datos

Atributo	Valor	Descripción
Ingreso	1	Entre $ 8.000 y $ 15.000
	2	Más de $ 15.000
Composición familiar	1	Soltero
	2	Casado sin hijos
	3	Casado con un hijo
	4	Casado con dos hijos
Vivienda	1	Alquila
	2	Propia
Servicios	1	Básicos
	2	Básicos y TV por cable
	3	Básicos, TV por cable y celular
Otros créditos	1	Un crédito
	2	Dos créditos
	3	Tres créditos
Otorga Crédito	Sí	Préstamo otorgado
	No	Préstamo rechazado

Caso de Estudio – Créditos Personales

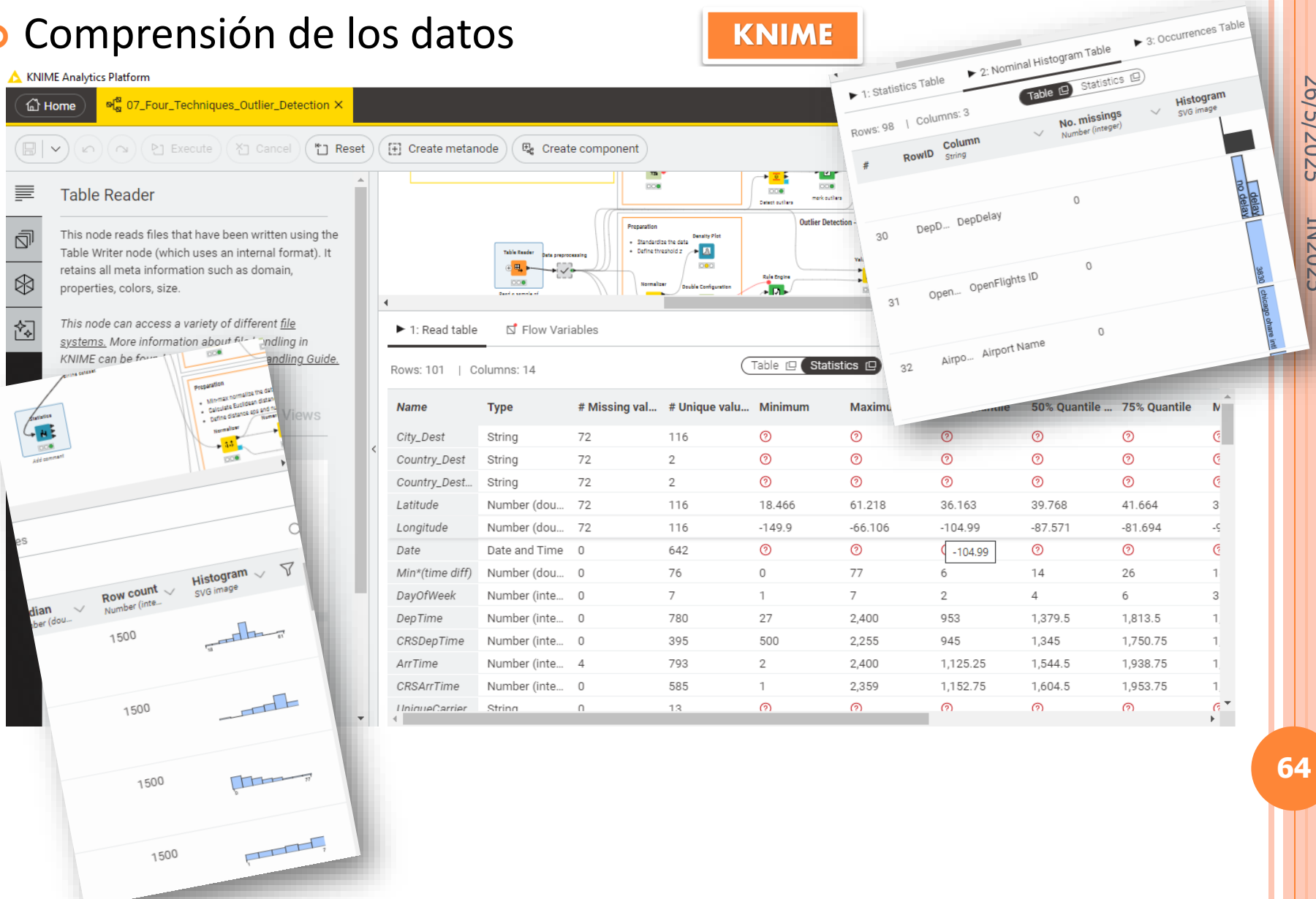
- Comprensión de los datos

KNIME

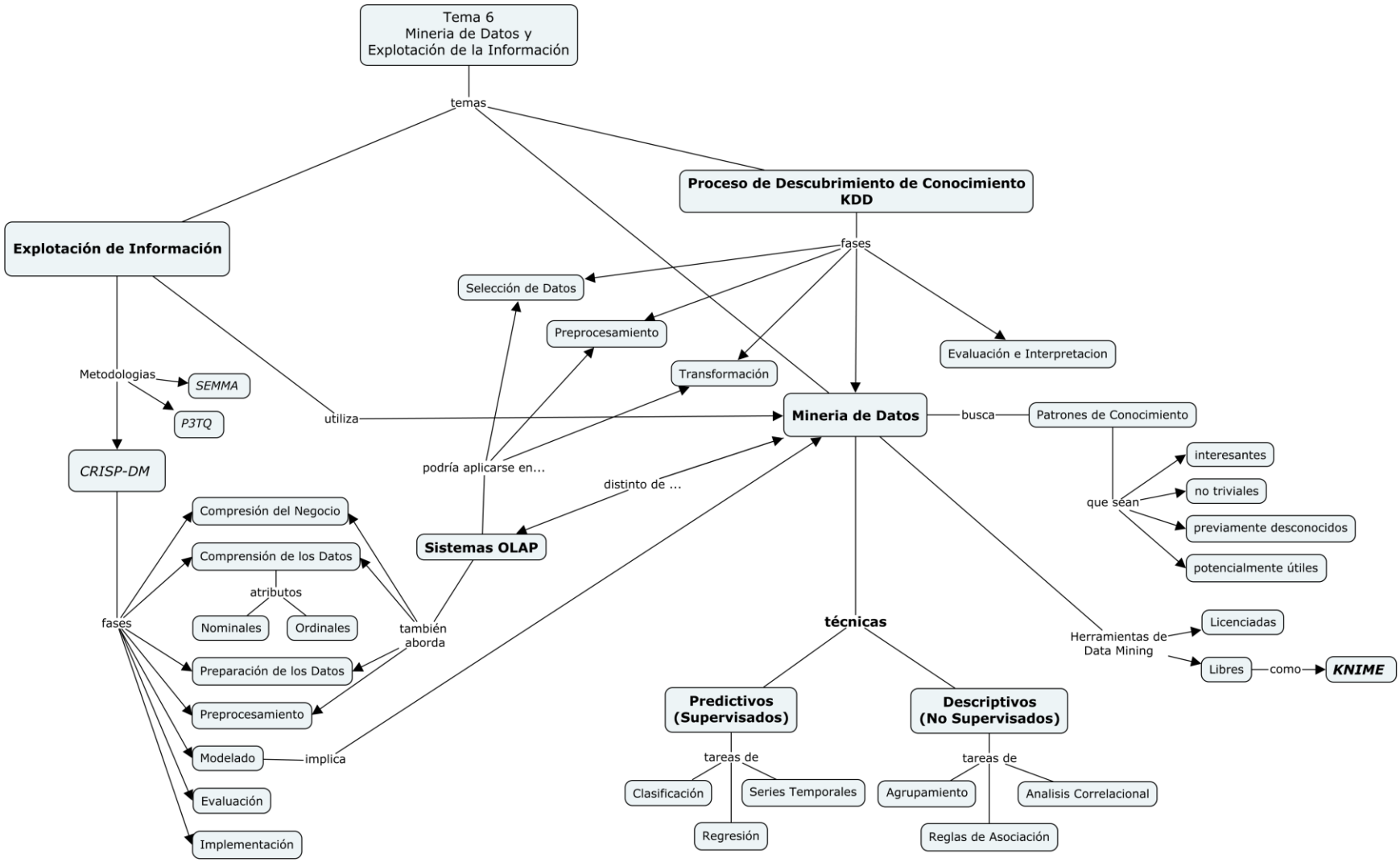


Caso de Estudio – Créditos Personales

- Comprensión de los datos



RESUMEN TEMA 6



Referencias

- **Fuentes:**
  - Introducción a la Minería de Datos Instituto de Computación – CPAP Universidad Uruguay
  - Introduction to Data Mining, 2nd Edition Tan, Steinbach, Karpatne, Kumar
  - Curso Machine Learning | Educación IT, Buenos Aires, Argentina
  - Jose Martinez Heras | European Space Agency (ESA)

- **Libros recomendados:**
  - "Introduction to Data Mining", 2nd Edition (2019): Tan, Steinbach, Karpatne, Kumar: Introduction to Data Mining (umn.edu)
  - "Data Mining: Concepts and Techniques", 3rd Edition (2012): Jiawei Han, Micheline Kamber & Jian Pei: Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems) (sabanciuniv.edu)

- **Material Complementario - Cursos ML Google y Microsoft**
  - Google Básico Machine Learning | Google for Developers
  - Google Avanzado Machine Learning | Google for Developers
  - Microsoft GitHub - microsoft/ML-For-Beginners: 12 weeks, 26 lessons, 52 quizzes, classic Machine Learning for all

Tareas Previas Trabajo Práctico de Minería de Datos

1) **Instalar la herramienta KNIME**
   - **Enlaces de Descarga Software para cada S.O.:** Windows; Linux; MAC (Intel) o MAC (Apple Silicon) *No requiere registro, solo aceptar los términos y condiciones.*

     Download KNIME Analytics Platform | KNIME
   - **Documentación oficial KNIME:**

     KNIME Documentation

2) **Consultar Material de Lectura KNIME disponible en MIeL**
   - "Introducción al Análisis de Datos - Prácticas con Power BI, R y KNIME"
   - “Practicing Data Science – The Data Science Case Study Collection”
   - “Cheat Sheet: Machine Learning with KNIME Analytics Platform“

3) **Consultar Tutoriales KNIME**
   - KNIME Learning Center | KNIME
   - KNIME 5.2 | KNIME
   - Intro to KNIME Analytics Platform Version 5 – YouTube
   - KNIME Base nodes — NodePit

Trabajo Práctico de Minería de Datos
Caso: Créditos Bancarios
Entrega 18/06/2025

- Utilice la herramienta KNIMNE para desarrollar los procesos de explotación de información identificados en el Caso de Estudio; incluyendo tareas de Preprocesamiento, Modelos Predictivos, Descriptivos y Evaluación.

- Entregue un informe que contenga resultados, conclusiones obtenidas, gráficos, una tabla comparativa de métodos aplicados, y las recomendaciones que daría, de acuerdo con lo requerido en el enunciado del trabajo práctico.

Links de Interés:

- Página oficial: https://www.knime.com/
- Documentación: https://www.knime.com/documentation
- Comunidad: https://www.knime.com/knime-community
- Tutoriales: https://www.knime.com/resources
- Videos: https://www.youtube.com/user/KNIMETV