

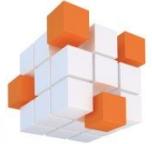
UNIVERSIDAD NACIONAL DE LA MATANZA

INTELIGENCIA DE NEGOCIOS

**Introducción a la Minería de Datos y
la Explotación de Información**

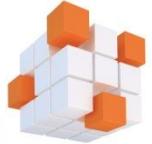
Docentes: ING. LORENA R. MATTEO

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



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:

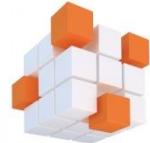
DATA MINING o Minería de Datos

- 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 videos).

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)

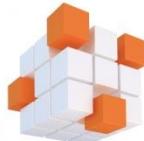
- es una técnica de AI que permite a los sistemas aprender de los datos sin ser programados explícitamente.
- se basa en algoritmos que analizan los datos y hacen predicciones o toman decisiones basadas en patrones que han sido descubiertos en los datos.
- se busca generalizar, aprender conceptos a partir 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)

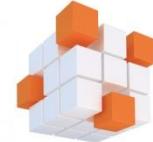
- son datos cuyo volumen, diversidad y complejidad requieren nueva arquitectura, técnicas, algoritmos y análisis para gestionar y extraer valor y conocimiento oculto en ellos.
- generalmente macrodatos no estructurados
- sus principales características son el volumen, la velocidad, la variedad, la veracidad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



HISTORIA VISUAL DE LA INTELIGENCIA ARTIFICIAL

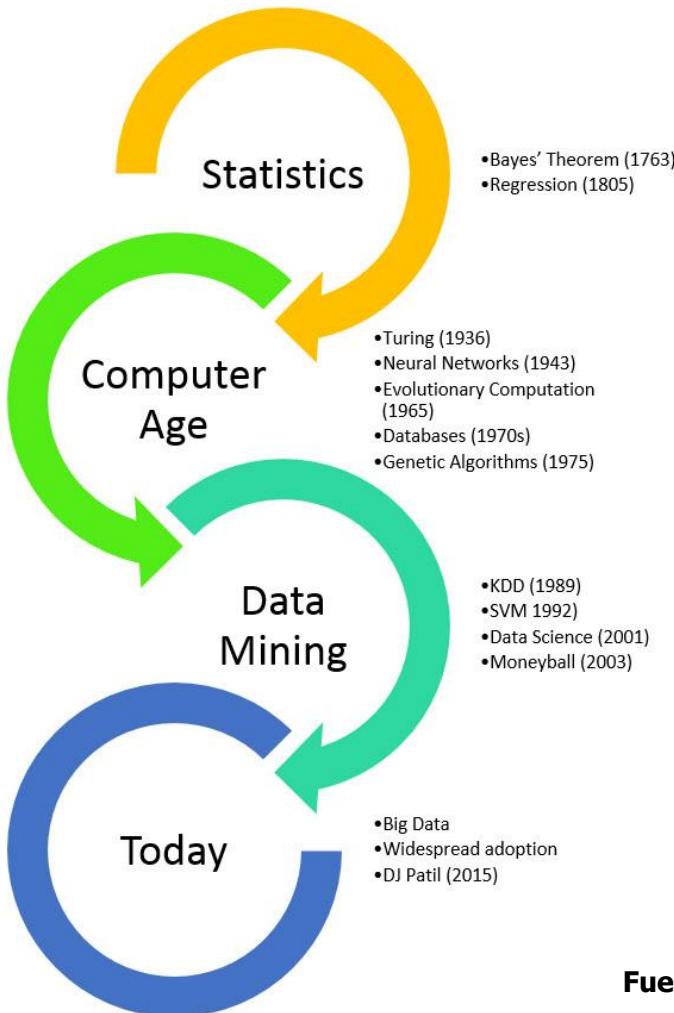
1943	1950	1952	1956	1956	1957	1958	1959	1960	1964
First mathematical model of a neuron. $\Phi(x) = \begin{cases} 1 & \text{if } x \geq \theta \\ -1 & \text{otherwise} \end{cases}$	Turing test is proposed.	One of the first computer board game.	Symbolic AI: The first automated reasoning program.	The conference that started it all.	Weights automatically learned.	One of the first AI programming languages.	Reasoning as search.	The first cost function.	An early natural language processing conversation program.
Electronic Brain by McCulloch & Pitts	Turing Test by Alan Turing	Checkers Program by Arthur Samuel	Logic Theorist by Newell, Simon, Shaw	Dartmouth Summer Research Project organized by John McCarthy	Perceptron by Frank Rosenblat	LISP by John McCarthy	General Problem Solver by Newell, Simon, Shaw	ADALINE by Widrow & Hoff	ELIZA by Joseph Weizenbaum
1964	1964	1964	1965	1965	1966	1966	1968	1969	1970
Algorithmic probability & formal induction is introduced. $P_M(x) = \sum_{i=1}^{\infty} 2^{- p_i(x) }$	Solving algebra word problem, an early NLP accomplishment.	The term Fuzzy Logic is introduced.	One of the first Expert System: organic chemistry.	SRI International	ALPAC report shuts down NLP research.	SHRDLU by Terry Winograd	Perceptron cannot learn XOR.	INTERNIST-I by Myers, Miller, Pope	Backpropagation & automatic differentiation.
Universal Bayesian Methods by Ray Solomonoff	STUDENT by Daniel G. Bobrow	Fuzzy Logic by Lotfi Zadeh	DENDRAL by Feigenbaum, Buchanan, Lederberg, Djerassi	SRI International	ALPAC Report by ALPAC	XOR Problem by Minsky & Papert	INTERNIST-I by Myers, Miller, Pope	Automatic differentiation by Seppo Linnainmaa	
1972	1973	1974	1979	1982	1983	1986	1987	1989	1992
One of the first logic programming languages. <code>animal(X) :- cat(X).</code>	Lighthill Report by James Lighthill	UK ended support for AI research.	An early backward chaining expert system for medical diagnosis.	Neocognitron by Kunihiko Fukushima	Foundation of graphical probabilistic models.	SOAR by Laird, Newell, Rosenbloom	Backpropagation is popularized.	LeNet by Yann LeCun	Almost champion-level backgammon, using reinforcement learning.
PROLOG by Colmerauer & Kowalski	MYCIN by Shortliffe, Buchanan, Cohen	BLIE-BASED EXPERT SYSTEMS	DeepBlue by IBM	Bayesian Networks by Judea Pearl	SOAR by Laird, Newell, Rosenbloom	Backpropagation in MLP by Rumelhart, Hinton, Williams	Subsumption by Rodney Brooks	LeNet by Yann LeCun	TD-Gammon by Gerald Tesauro
1994	1995	1995	1995	1996	1997	2006	2009	2012	2013
Chinook, draughts player: the first program to win the world champion title against humans.	Soft-margin SVM is introduced.	NIST MNIST is born.	DeepBlue beats Kasparov in chess.	LSTM for addressing vanishing gradients.	Modern deep learning is born.	Deep Boltzmann Machine by Salakhutdinov & Hinton	ImageNet, a large-scale image dataset is introduced.	AlexNet by Krizhevsky & Hinton	Word2Vec by Tomas Mikolov
Chinook by a Team led by Jonathan Schaeffer	Support Vector Machines by S. Vapnik & Cortes	MNIST by NIST	DeepBlue by IBM	Long Short-Term Memory (LSTM) by Hochreiter & Schmidhuber	Deep Boltzmann Machine by Salakhutdinov & Hinton	DeepMind	ImageNet by Fei-Fei Li	GAN by Ian Goodfellow	
2016	2017	2018	2019	2020	2021	2022	2023		
DeepMind The first program to beat a professional Go player.	Google Transformer network is introduced, paving the way for pretrained language models.	Ethical AI takes off: accuracy disparities in commercial gender classification	Grandmaster level in Starcraft II.	IBM, Amazon, and Microsoft ban sale of facial recognition products to law enforcement.	AlphaFold by DeepMind	OpenAI Predicting a protein's structure from amino-acid sequence	Microsoft Zero-Shot Text to Speech Synthesis		
AlphaGo by DeepMind	Transformer by Google Brain, Research	Gender Shades by Buolamwini & Gebru	AlphaStar by DeepMind	AI Activism by Ethical AI Community	DALL-E 2, ChatGPT	VALL-E			

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



ORÍGENES DE LA MINERÍA DE DATOS

Data Mining



- 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/>

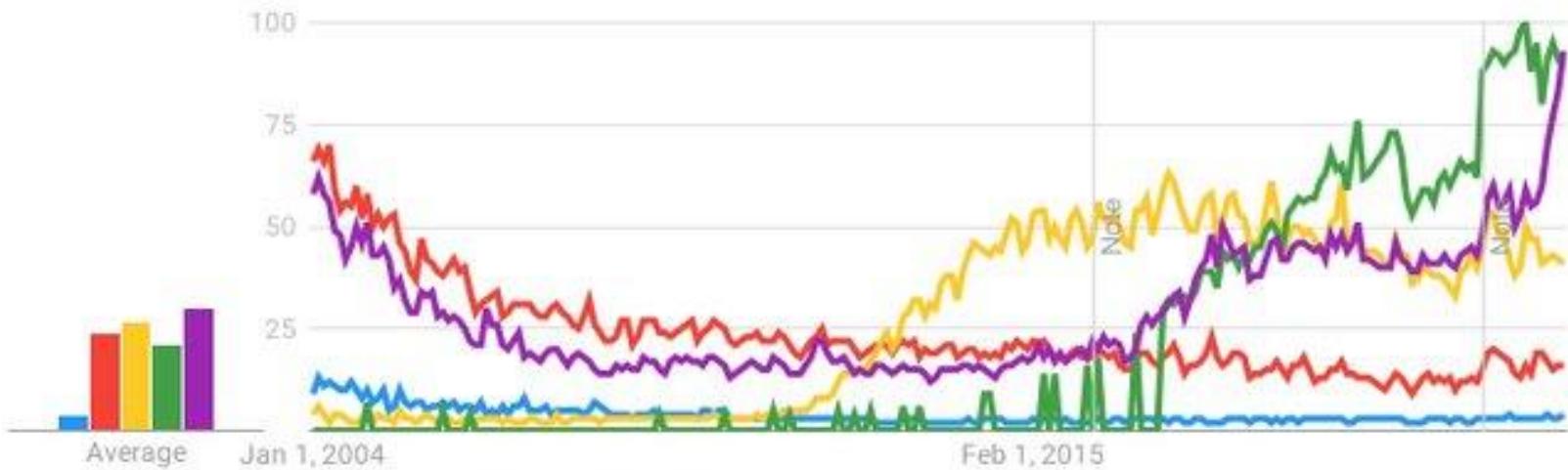


EVOLUCIÓN DE ALGUNOS CONCEPTOS

Google Trends

Interest over time

KDD Data Mining Big Data Data Science Artificial Intelligence



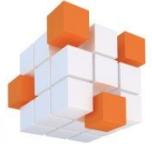
Worldwide. 2004 - present.

¿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 calidad 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.

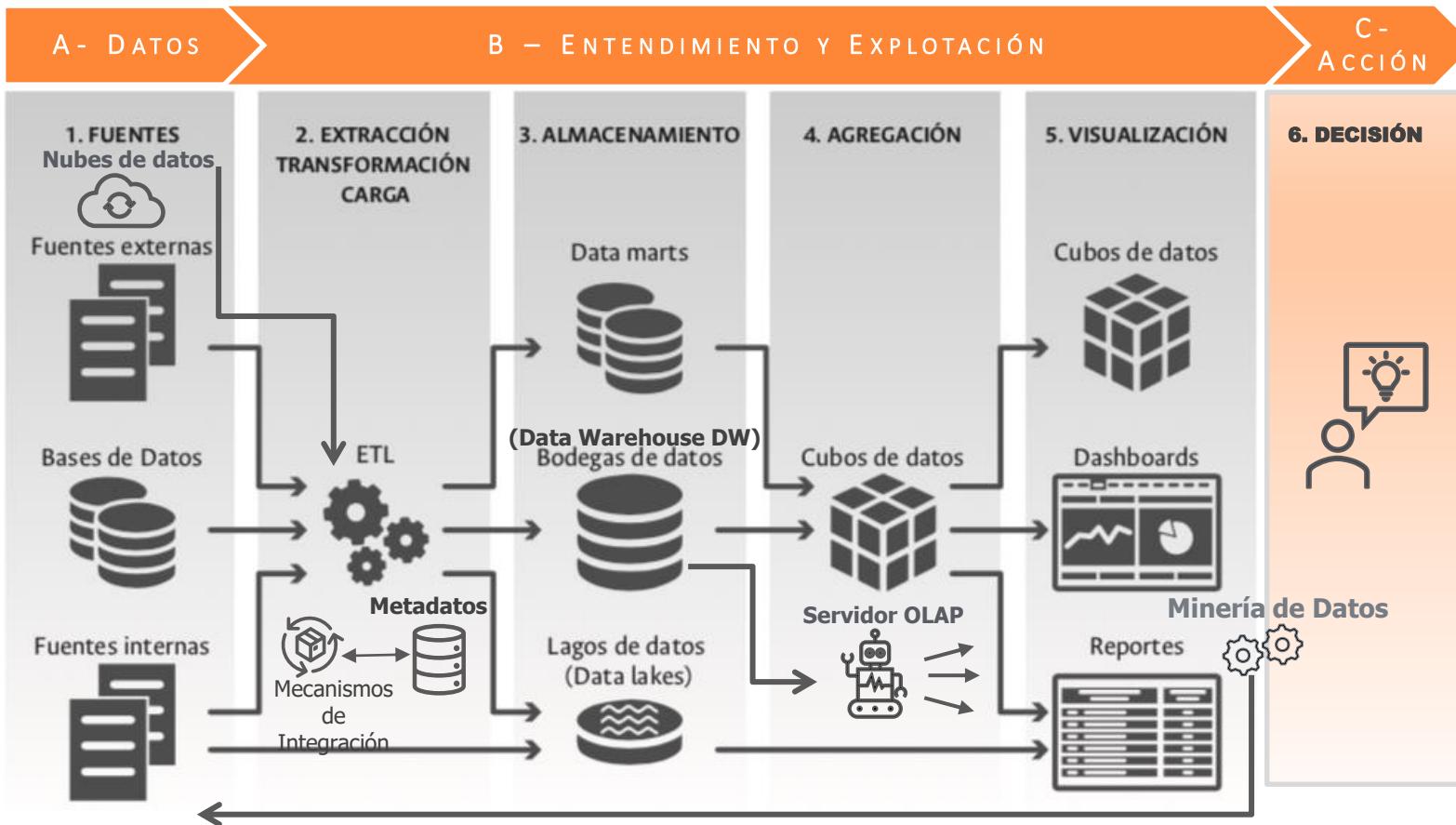


¿ QUÉ ES Y QUÉ NO ES MINERÍA DE DATOS?

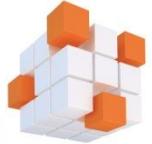
Que sí es...	Que no es...
Grupos de documentos/usuarios similares.	Localizar un número telefónico en el directorio. <i>No, es una simple consulta a la BD.</i>
Preferencias de compras de los usuarios – e-Commerce.	Consultar en un buscador información acerca de un tema en particular.
Zonas de mayor criminalidad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



ARQUITECTURA BI

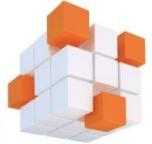


- **A - Datos**: Centralizar información de múltiples fuentes en un DW.
- **B - Entendimiento y Explotación**: Herramientas de BI y DM para analizar y mejorar el entendimiento del negocio.
- **C - Acción**: Actuar sobre los hallazgos realizados en el análisis.



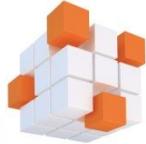
OLAP vs MINERÍA DE DATOS

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.



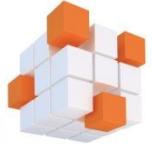
OLAP vs MINERÍA DE DATOS

- 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

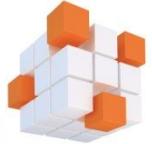
- ¿Qué se necesita para hacer minería de datos?
 - Herramientas de SW
 - Datos, digitalizados y de buena calidad
 - RRHH especialistas: técnico, analítico y de negocios



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



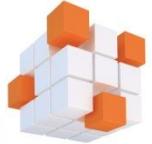


ÁREAS DE APLICACIÓN

○ 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.

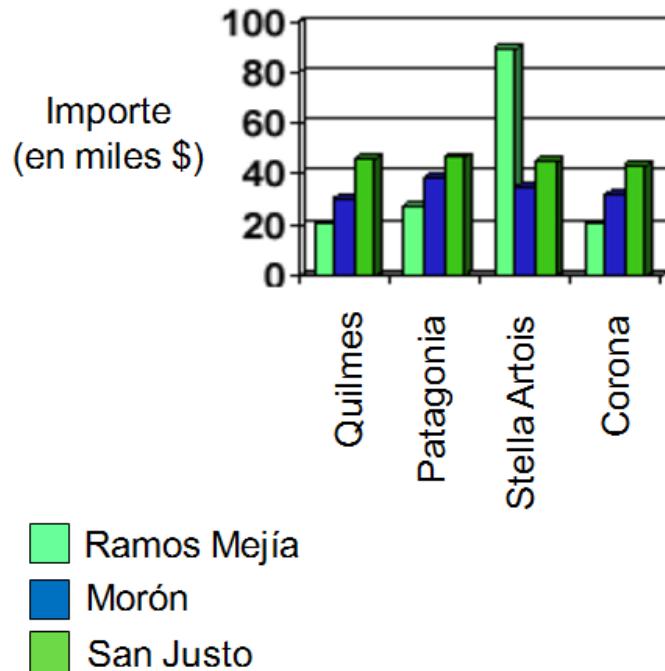
○ Medicina

- 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.

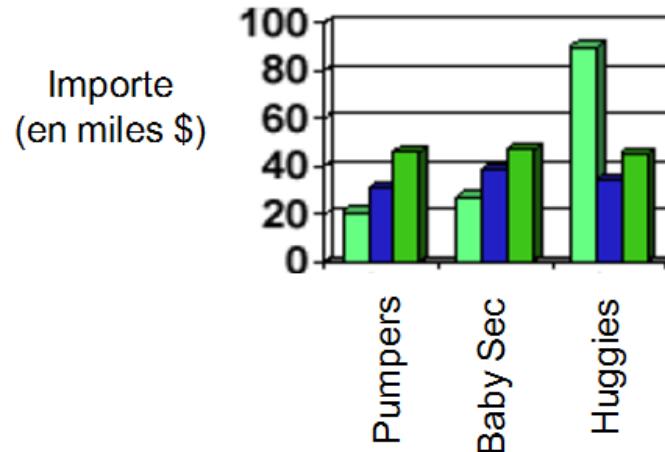


CASO DE ESTUDIO: MARKETING-VENTAS

Ventas de Cervezas en Abril



Ventas de Pañales en Abril



- 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

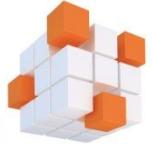
Pañales → Cerveza, Leche [sup=5%, conf=75%]

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.”





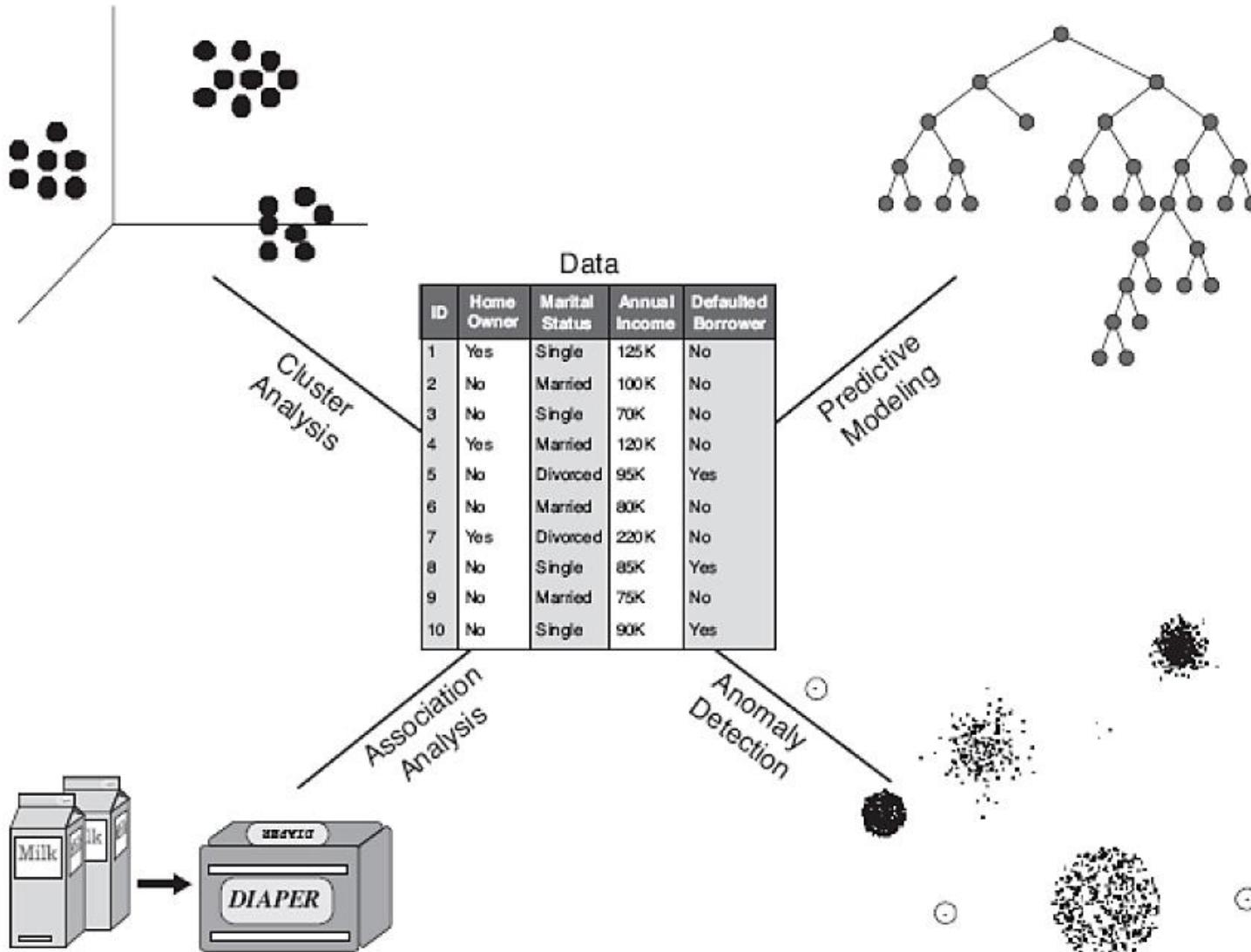
CASO DE ESTUDIO: MARKETING-VENTAS

○ 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.

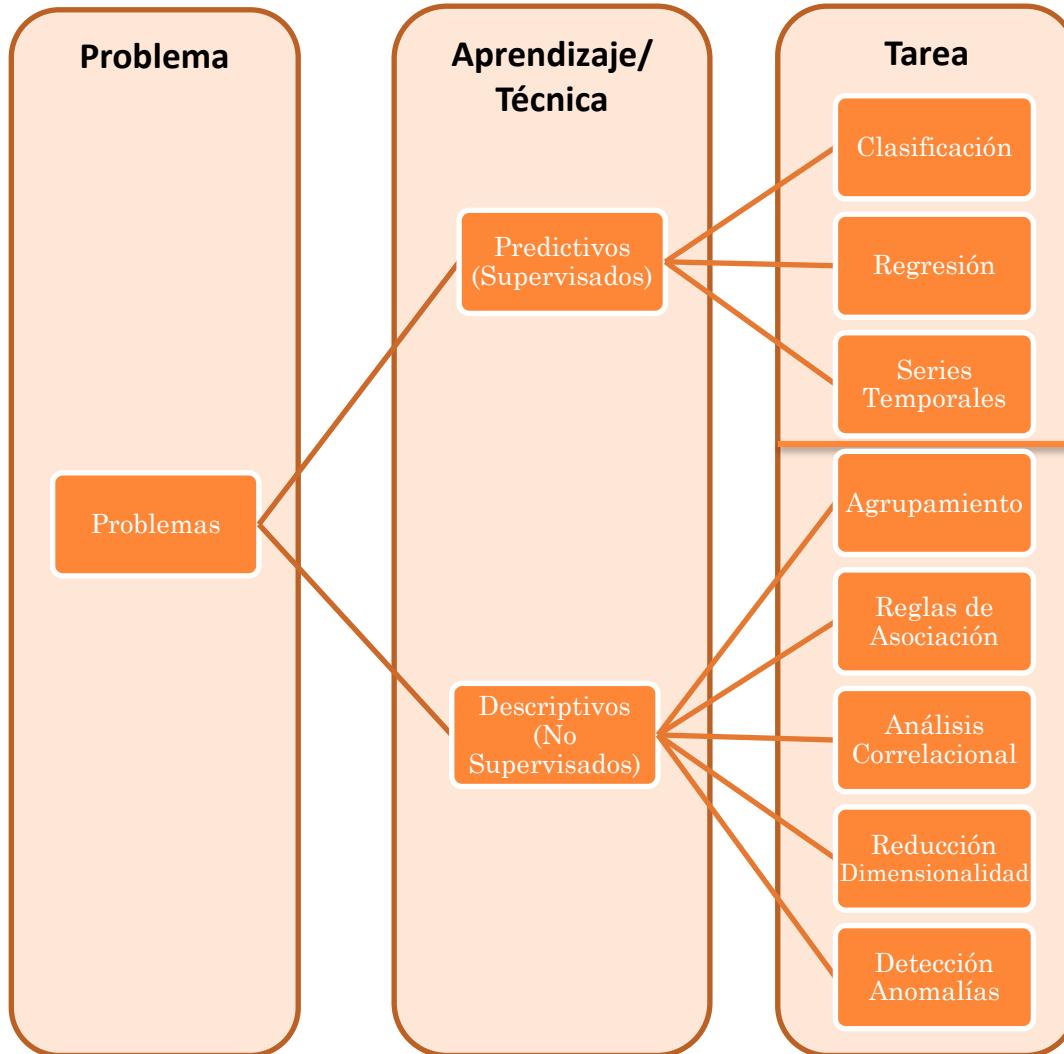


TAREAS DE MINERÍA DE DATOS





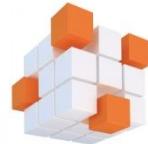
TÉCNICAS DE MINERÍA DE DATOS



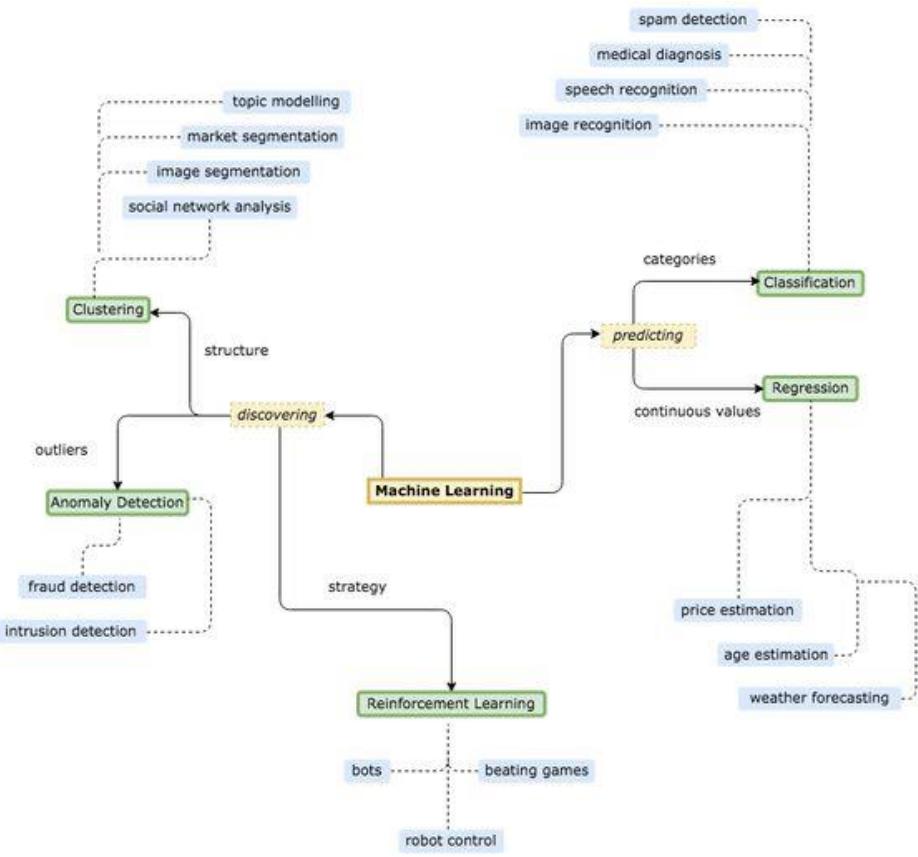
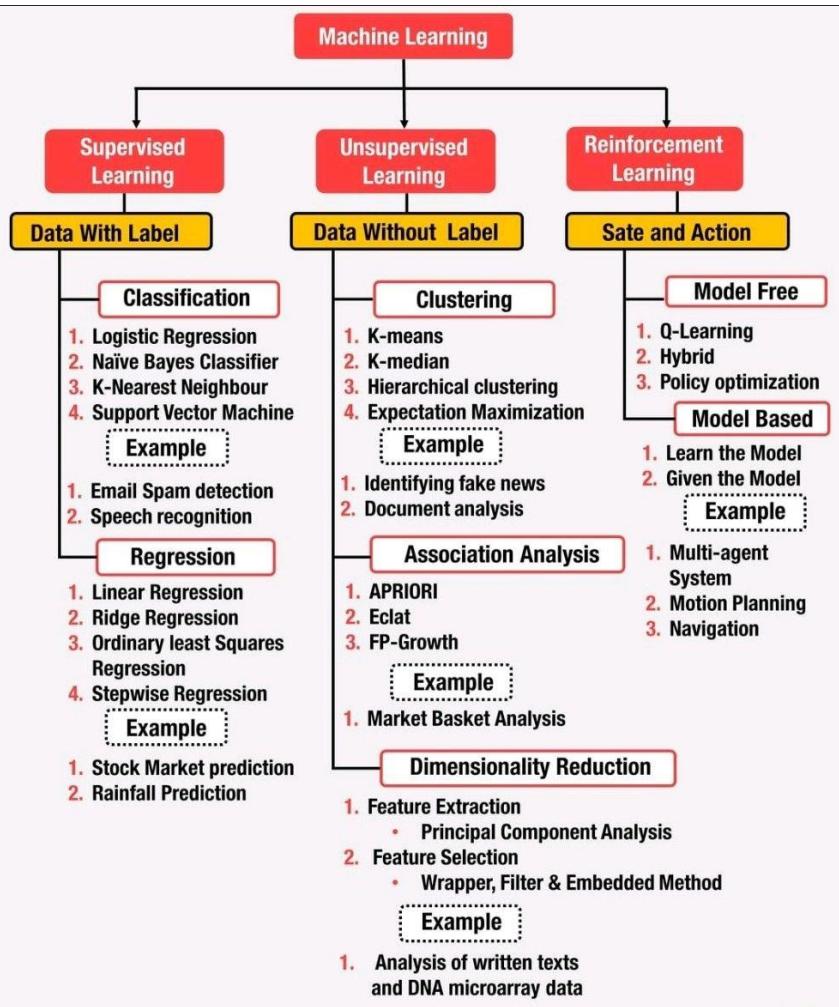
Las técnicas de minería de datos son herramientas que facilitan el descubrimiento de conocimiento.



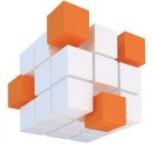
Esta foto de Autor desconocido está bajo licencia CC BY-SA-NC



TÉCNICAS DE MINERÍA DE DATOS - DETALLADO



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

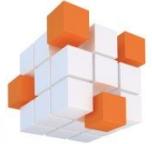


TÉCNICAS DE MINERÍA DE DATOS

- 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:

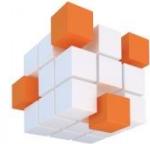
- una **clasificación** o una **predicción** basada en un determinado número de ejemplos, en caso de **supervisión**,
- o **agrupando por similitud**, en casos **no supervisados**



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**.



TÉCNICAS DE MINERÍA DE DATOS

26/5/2025

IN2025

Modelos y Algoritmos

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

Utilizando **algoritmos** en dos etapas:

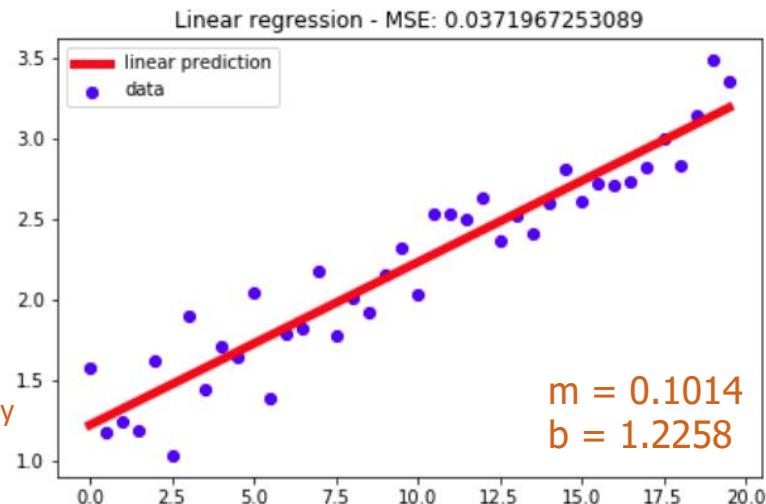
- Aprendizaje | Inducción (Entrenamiento).
- Resolución | Deducción (Cálculo de la predicción o clasificación)

¿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.





TÉCNICAS DE MINERÍA DE DATOS

26/5/2025

IN2025

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

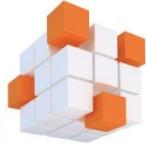
- Utilización de métricas y métodos para decidir si el algoritmo es el adecuado o se debe cambiar, también si es preciso ajustar sus hiperparámetros.

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.



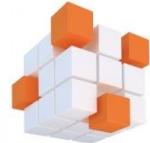
TÉCNICAS DE MINERÍA DE DATOS

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.*

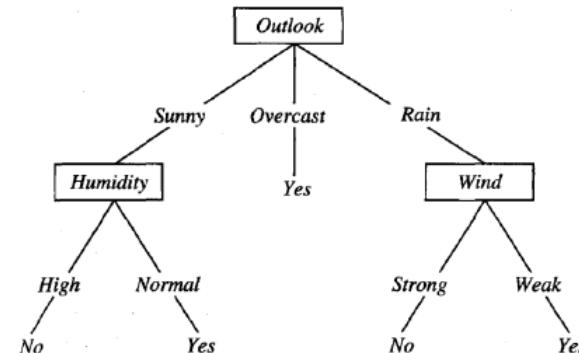


TÉCNICAS DE MINERÍA DE DATOS

○ 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

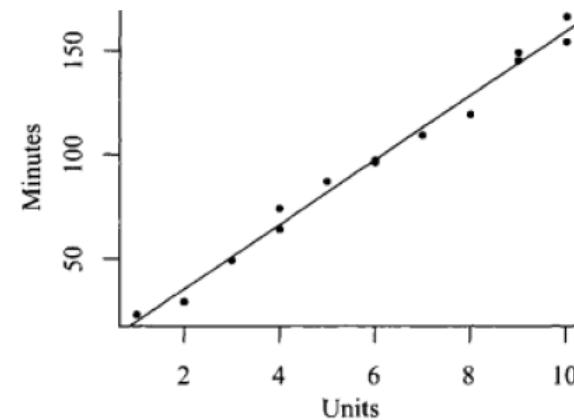


○ Tareas de Regresión

- Predicen un valor continuo
 - Importes
 - Cantidad

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

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





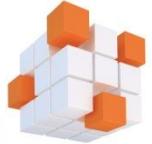
TÉCNICAS DE MINERÍA DE DATOS

26/5/2025

IN2025

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.*



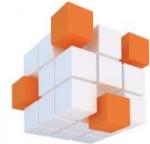
TÉCNICAS DE MINERÍA DE DATOS

○ 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.



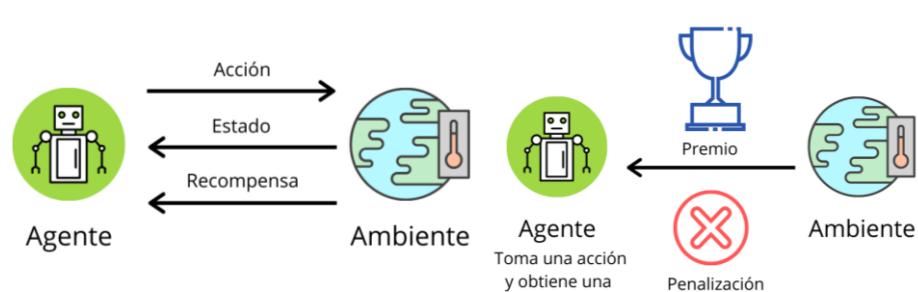
TÉCNICAS DE MINERÍA DE DATOS

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

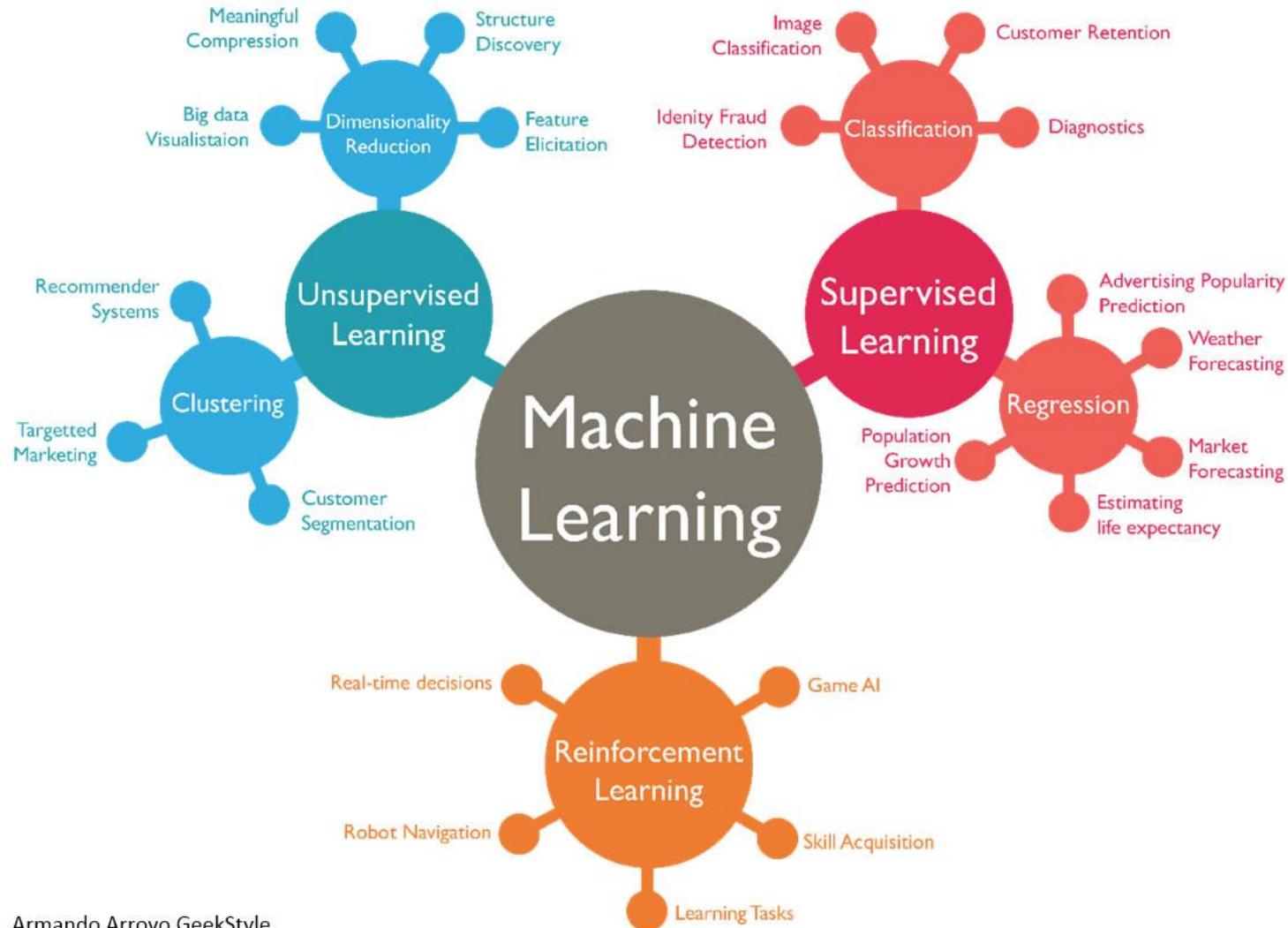
- *juegos de mesa y videojuegos*
- *robótica, máquinas industriales*
- *medicina y la biología*
- *sistemas de navegación vehículos autónomos (autos, drones, aviones)*
- *reducción del consumo energético*

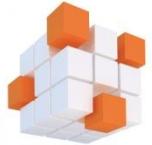


Fuente: <https://www.aprendemachinelearning.com/aprendizaje-por-refuerzo/>
https://github.com/jbagnato/machine-learning/blob/master/Ejercicio_Reinforcement_Learning_Pong.ipynb



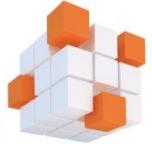
TÉCNICAS DE MINERÍA DE DATOS





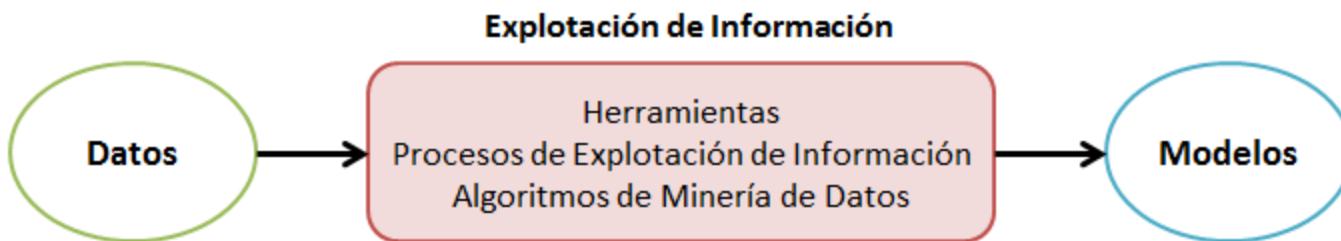
EXPLORACIÓN DE INFORMACIÓN

- La **Exploración 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 exploración de información no son conceptos equivalentes.



EXPLORACIÓN DE INFORMACIÓN

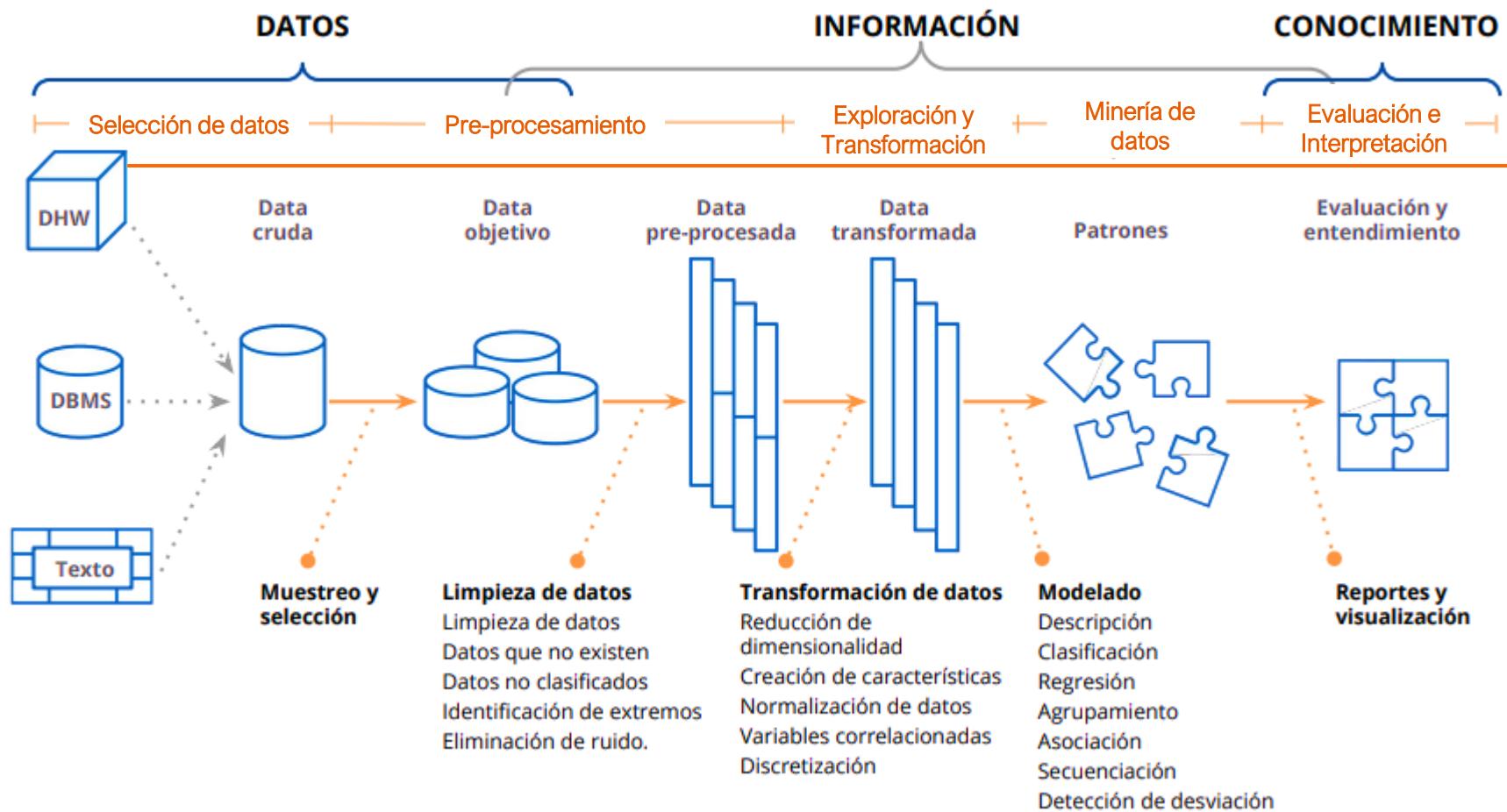
- La **minería de datos** está relacionada a la programación, a los algoritmos para resolver un problema de inteligencia de negocios.
- La **exploración de información** está relacionada a tareas de la Ingeniería de Software, a la aplicación de técnicas y procesos ingenieriles para construir la solución 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.





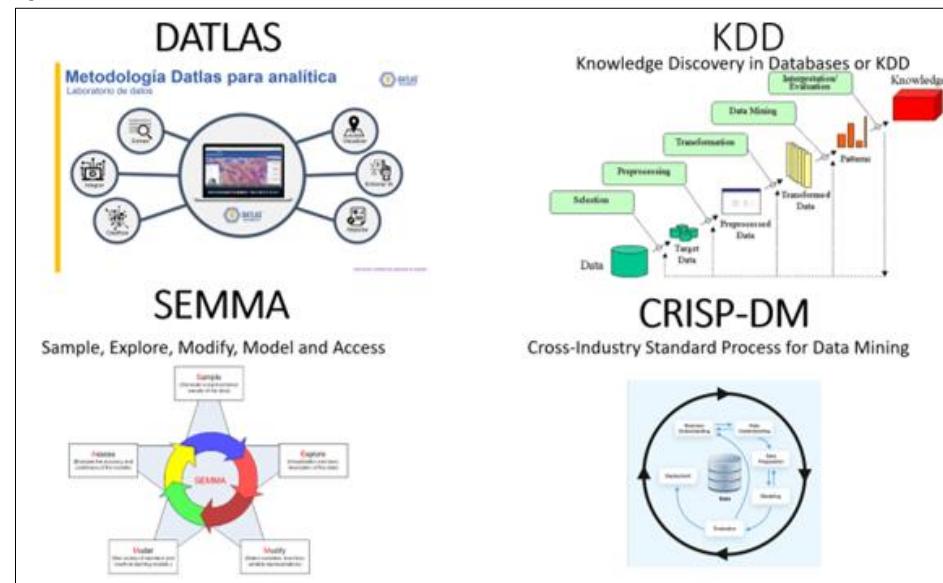
PROCESO DESCUBRIMIENTO DE CONOCIMIENTO

- **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.

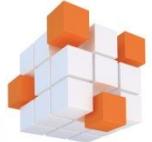


METODOLOGÍAS DE EXPLOTACIÓN DE INFORMACIÓN

- 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
 - ✓ P³TQ
 - ✓ DATLAS

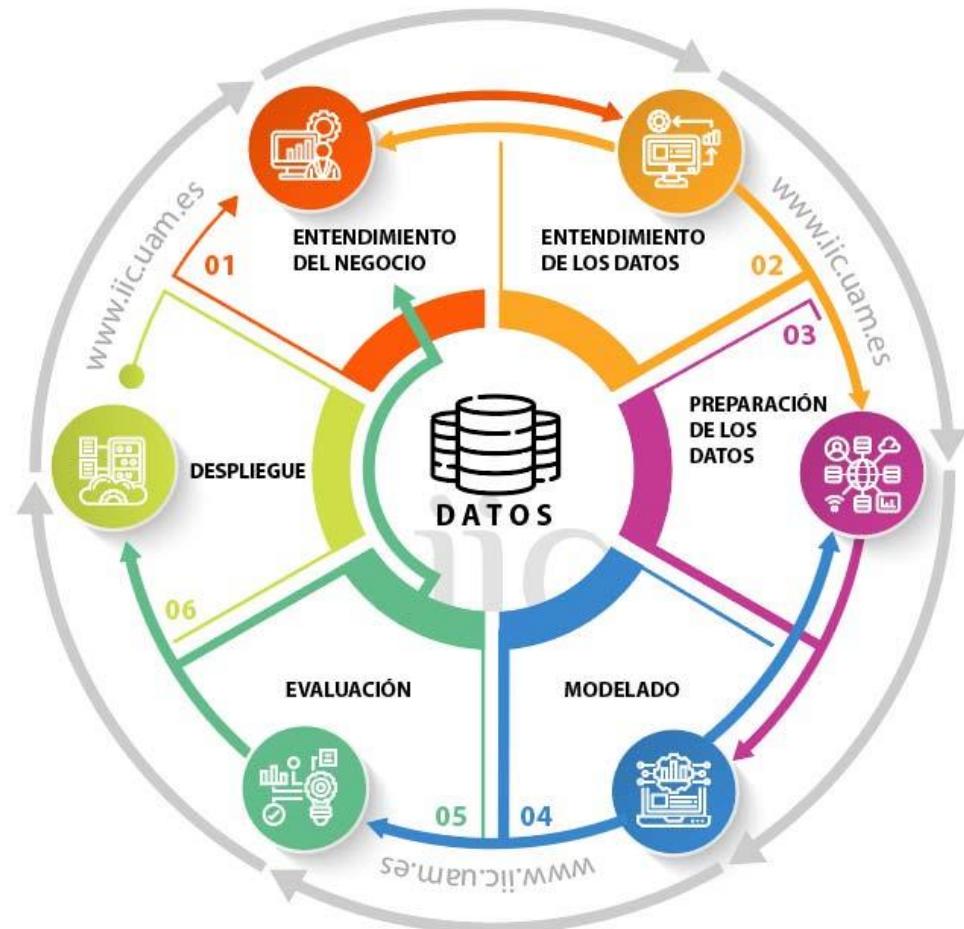


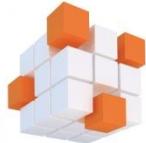
Fuente: <https://blogdatlas.wordpress.com/2020/02/16/4-metodologias-para-proyectos-de-data-science-datlas-research/>



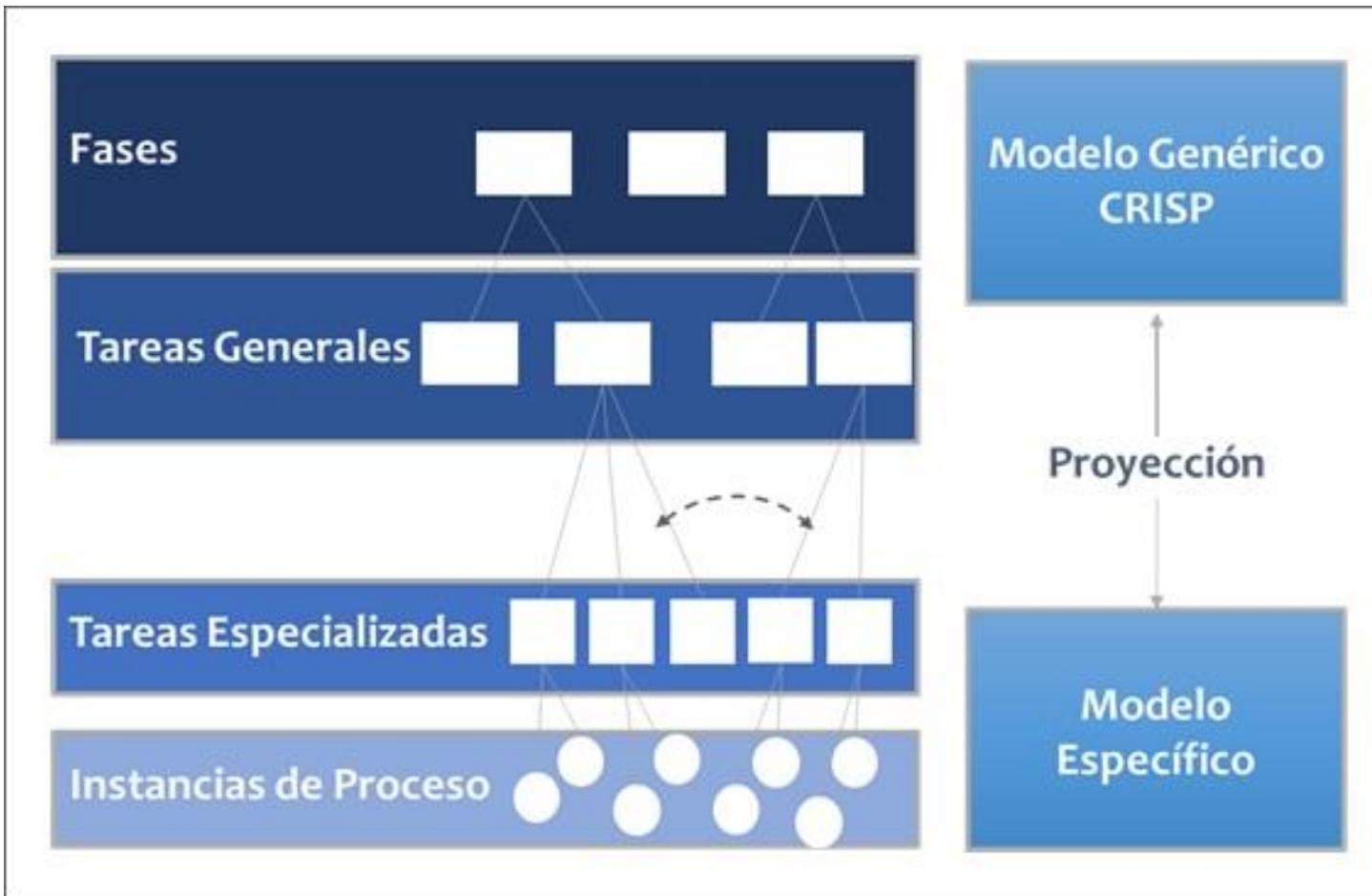
METODOLOGÍA CRISP-DM

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.





METODOLOGÍA CRISP-DM



Esquema de los cuatro Niveles de Abstracción de CRISP-DM



METODOLOGÍA CRISP-DM

Comprendión del Negocio

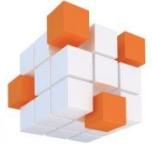
- Se determinan los **objetivos** y **requerimientos** del proyecto desde una perspectiva del negocio, definiendo el problema de minería y el **plan de trabajo**.
 - Objetivos de negocio y criterios de éxito
 - Detectar fraude con tarjetas de crédito
 - Captar nuevos clientes bancarios
 - Detectar signos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



METODOLOGÍA CRISP-DM

Compreensión de los Datos

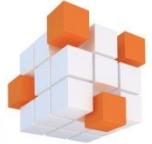
- Se **recolectan** los datos que se utilizarán y se analizan las características de los mismos. Surgen las primeras **hipótesis** acerca de la información que podría estar oculta.
- Atributos **Nominales**
 - Llamados **Categóricos** o **Discretos**
 - Número finito de valores, no tienen orden.
 - Ejemplo: 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

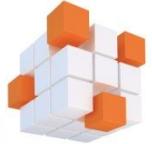
- Comprenden actividades de **tratamiento** de los datos o conjunto de datos final sobre el cual se aplicarán procesos de explotación de información y minería de datos.
 - **Selección, Limpieza y Transformación**
- Análisis de la **calidad** de los datos
 - ¿Qué tipos de problemas de calidad podemos 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

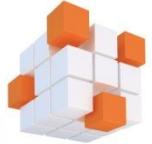
○ Preprocesamiento de los datos

- **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.

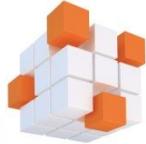


METODOLOGÍA SEMMA

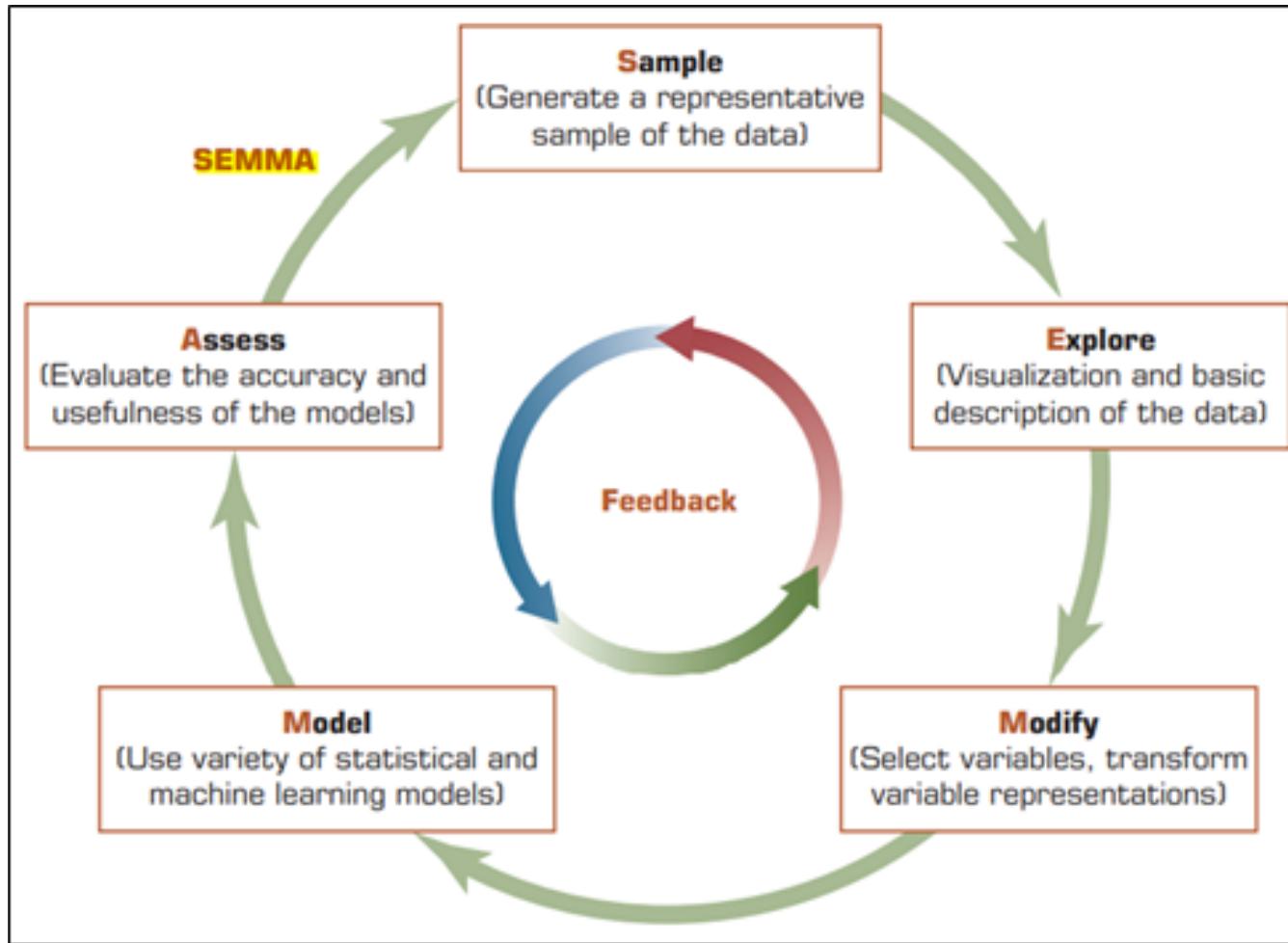
SEMMA (Sample, Explore, Modify, Model and Assess) Desarrollado por **SAS Institute**

- “**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.



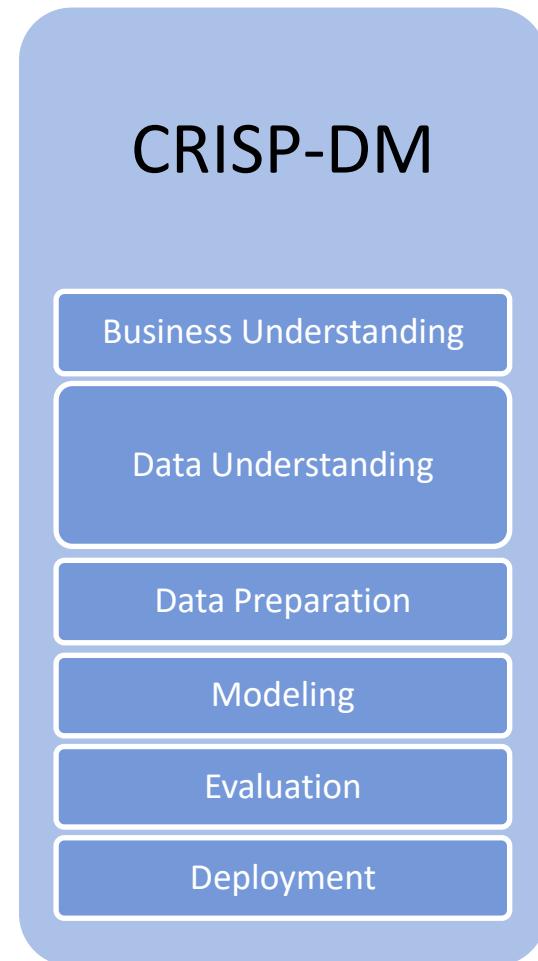
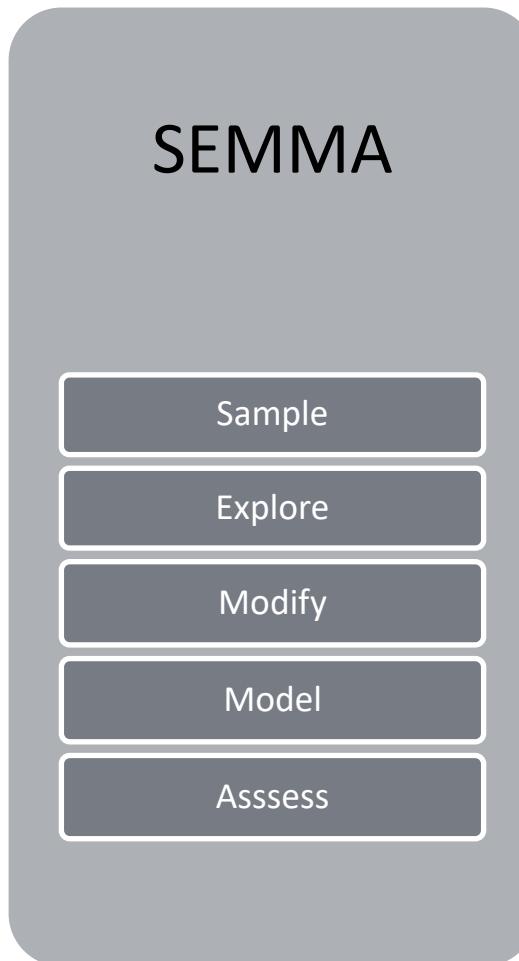
METODOLOGÍA SEMMA



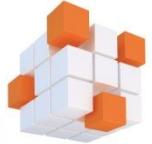
Fuente: <https://sis.binus.ac.id/2021/09/30/data-mining-semma/>



COMPARACIÓN DE METODOLOGÍAS



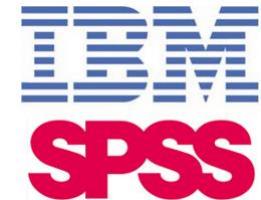
Fuente: <https://blogdatlas.wordpress.com/2020/02/16/4-metodologias-para-proyectos-de-data-science-datlasm-research/>



HERRAMIENTAS PARA MINERÍA DE DATOS

○ Licenciadas

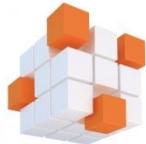
- **SAS** (Analytics, Enterprise Miner)
- **SPSS** (IBM SPSS Statistics, IBM SPSS Modeler – ex Clementine)
- Microsoft Azure Machine Learning



○ 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

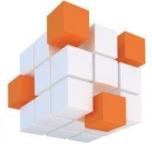




HERRAMIENTAS PARA MINERÍA DE DATOS

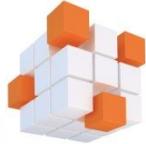
Comparación Herramientas:

	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



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
	1	Alquila
Vivienda	2	Propia
	1	Básicos
Servicios	2	Básicos y TV por cable
	3	Básicos, TV por cable y celular
	1	Un crédito
Otros créditos	2	Dos créditos
	3	Tres créditos
Otorga Crédito	Sí	Préstamo otorgado
	No	Préstamo rechazado



CASO DE ESTUDIO – CRÉDITOS PERSONALES

○ Comprensión de los datos

KNIME

KNIME Analytics Platform

File Edit View Node Help

KNIME Explorer X

- My-KNIME-Hub (api.hub)
- EXAMPLES (knime@api.hub)
- LOCAL (Local Workspace)
 - Example Workflows
 - Basic Examples
 - Customer Intelliger
 - Churn Prediction
 - Credit Scoring
 - Building a Credit Score Card
 - Customer Segmentation

Workflow Coach X

Node recommendations only available

Node Repository

- IO
- Manipulation
- Views
- Analytics
- DB
- Other Data Types
- Structured Data

Dialog - 3:62 - CSV Reader (Reading)

File

Settings Transformation Advanced Settings Limit Rows Encoding Flow Variables Job Manager Selection Memory Policy

Mode: File Files in folder

File: ../../TheData/Credit/german-credit-scoring.csv

Reader options

Format

Autodetect format

Column delimiter: ; Row delimiter: Line break Custom

Quote char: Quote escape char:

Comment char:

Has column header Has row ID

Support short data rows Prepend file index to row ID

Preview

The suggested column types are based on the first 10000 rows only. See 'Advanced Settings' tab.

Row ID	Status ...	Duration ...	Credit history	Purpose	Credit ...	Savings account/bonds	Present empl ...	Instal ...
Row0	Negative	6	critical account/ other credits existing (not at this bank)	radio/television	1169	unknown / no savings account	more than 7 years	4
Row1	0 - 200	48	existing credits paid back duly till now	radio/television	5951	less than 100	1 to 4 years	2
Row2	No checking ...	12	critical account/ other credits existing (not at this bank)	education	2096	less than 100	4 to 7 years	2
Row3	Negative	42	existing credits paid back duly till now	furniture/equ ...	7882	less than 100	4 to 7 years	2
Row4	Negative	24	delay in paying off in the past	car (new)	4870	less than 100	1 to 4 years	3
Row5	No checking ...	36	existing credits paid back duly till now	education	9055	unknown / no savings account	1 to 4 years	2
Row6	No checking ...	24	existing credits paid back duly till now	furniture/equ ...	2835	500 to 1000	more than 7 years	3

OK Apply Cancel ?



CASO DE ESTUDIO – CRÉDITOS PERSONALES

○ Comprendión de los datos

KNIME

The screenshot shows the KNIME Analytics Platform interface. On the left, there's a sidebar with various components like 'Table Reader' and 'Statistics'. The main workspace displays a workflow titled '07_Four_Techniques_Outlier_Detection'. The workflow starts with a 'Table Reader' node, followed by a 'Data preprocessing' section containing a 'Normalizer' node. This leads to a 'Preparation' section with a 'Density Plot' node and a 'Double Configuration' node. Finally, it reaches an 'Outlier Detection' section with a 'Mark outliers' node. To the right of the workflow, there are three tables labeled 1, 2, and 3. Table 1 is a 'Statistics Table' showing rows: 98, columns: 3, RowID: String, Column: String, and No. missings: Number (integer). Table 2 is a 'Nominal Histogram Table' showing rows: 30, DepDelay: 0, OpenFlights ID: 0, and Airport Name: 0. Table 3 is an 'Occurrences Table' showing rows: 31, 32, and 33. Below these tables is a detailed 'Statistics' table for 14 columns:

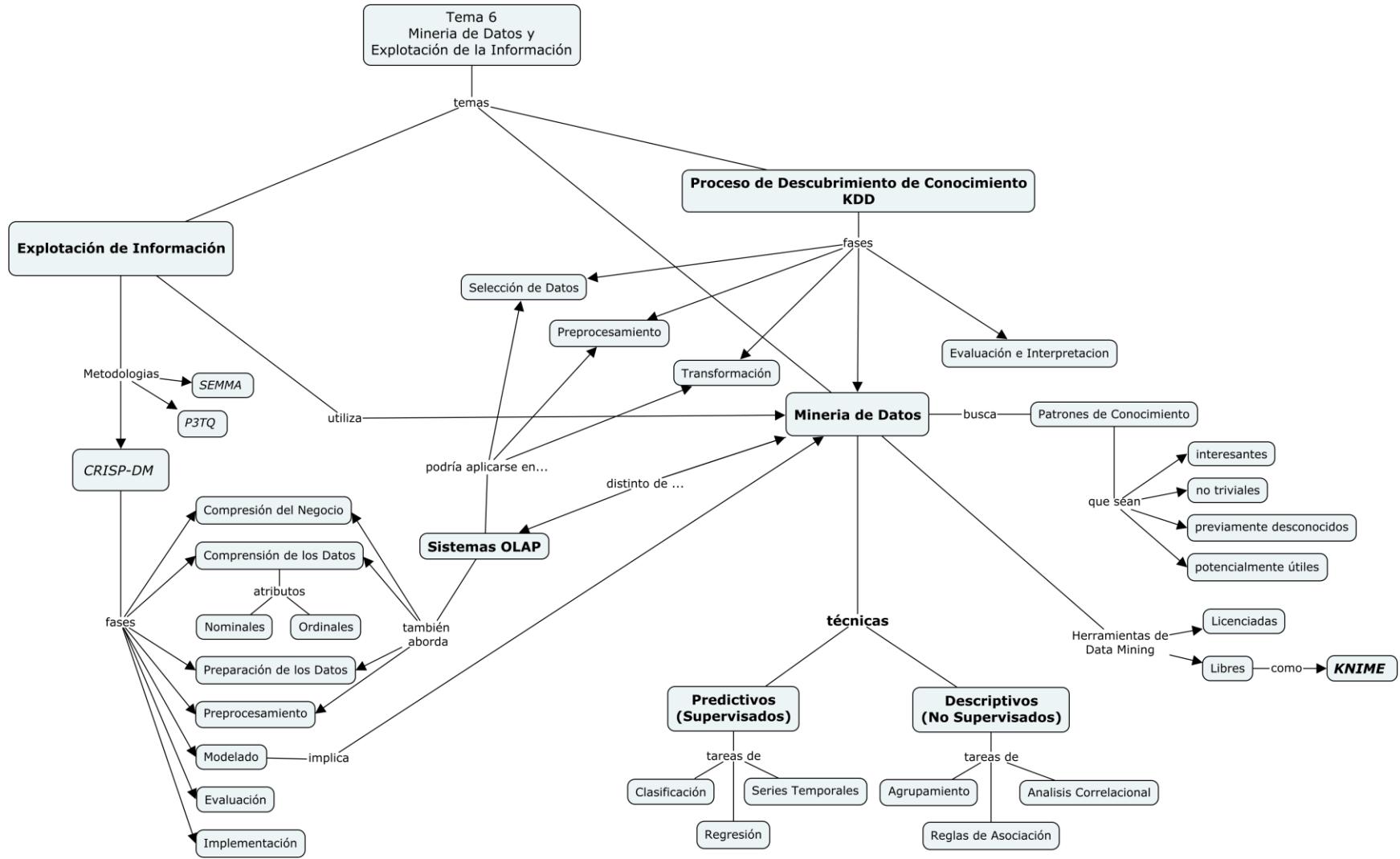
Name	Type	# Missing val...	# Unique val...	Minimum	Maximum	Median	50% Quantile ...	75% Quantile	M
City_Dest	String	72	116	?	?	?	?	?	?
Country_Dest	String	72	2	?	?	?	?	?	?
Country_Dest...	String	72	2	?	?	?	?	?	?
Latitude	Number (dou...	72	116	18.466	61.218	36.163	39.768	41.664	3
Longitude	Number (dou...	72	116	-149.9	-66.106	-104.99	-87.571	-81.694	-5
Date	Date and Time	0	642	?	?	-104.99	?	?	?
Min*(time diff)	Number (dou...	0	76	0	77	6	14	26	1
DayOfWeek	Number (inte...	0	7	1	7	2	4	6	3
DepTime	Number (inte...	0	780	27	2,400	953	1,379.5	1,813.5	1
CRSDepTime	Number (inte...	0	395	500	2,255	945	1,345	1,750.75	1
ArrTime	Number (inte...	4	793	2	2,400	1,125.25	1,544.5	1,938.75	1
CRSArrTime	Number (inte...	0	585	1	2,359	1,152.75	1,604.5	1,953.75	1
UniqueCarrier	String	0	13	?	?	?	?	?	?

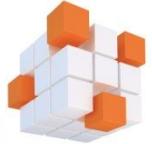


RESUMEN TEMA 6

26/5/2025

IN2025





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:

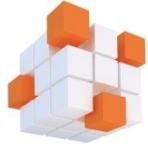
[KNIME Documentation](#)

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

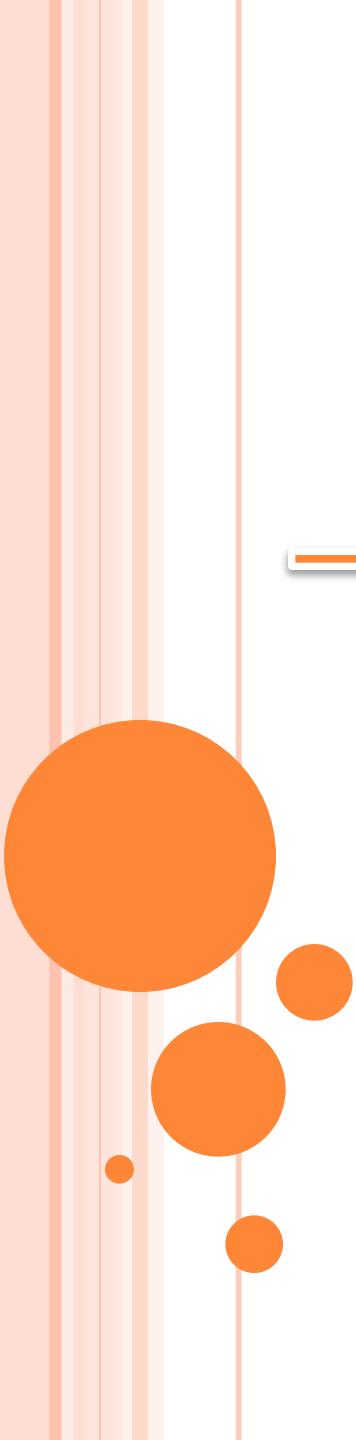
ENTREGA **18/06/2025**

- 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>





UNIVERSIDAD NACIONAL DE LA MATANZA

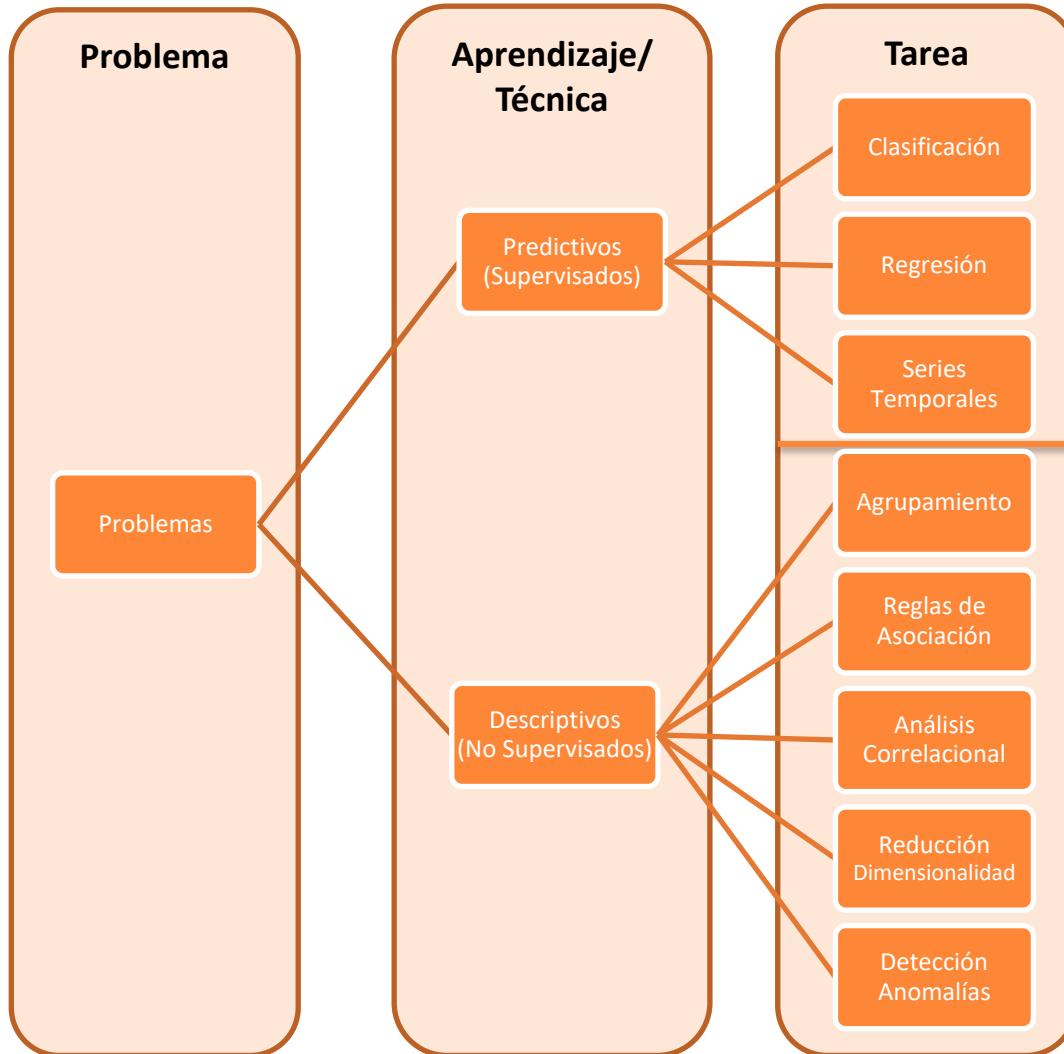
INTELIGENCIA DE NEGOCIOS

**Tecnologías Inteligentes
para Explotación de Información**

**Docentes: ING. LORENA R. MATTEO
Autores ppt orig.: Lic. HUGO M. CASTRO / MG. DIEGO BASSO**



TÉCNICAS DE MINERÍA DE DATOS



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

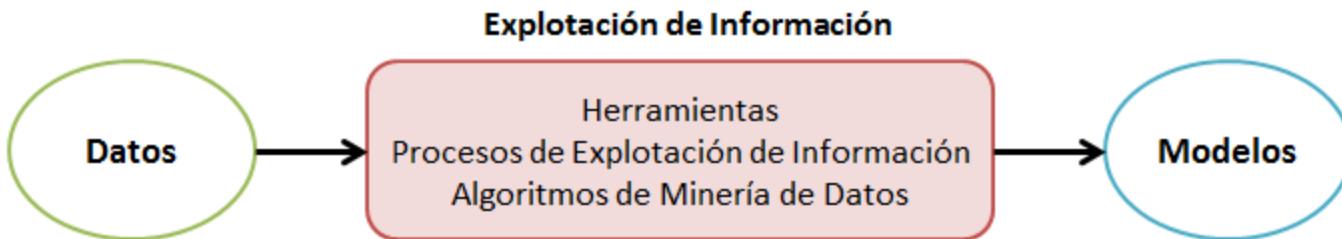


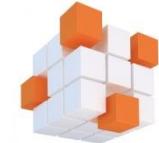
Esta foto de Autor desconocido está bajo licencia CC BY-SA-NC



USO DE LAS TECNOLOGÍAS

- ¿Cómo se usan las **tecnologías** para resolver un **problema**?
 - Tecnologías ⇒ Explotación de Información
 - Problema ⇒ Inteligencia de Negocio
- Construcción de modelos para descubrir conocimiento y soporte a la toma de decisiones:
 - Predictivos o Descriptivos
 - Entrenamiento + Prueba
 - Evaluación del modelo construido





CONSTRUCCIÓN DE MODELOS DE MINERÍA DE DATOS

○ Entrenamiento (Aprendizaje o Inducción)

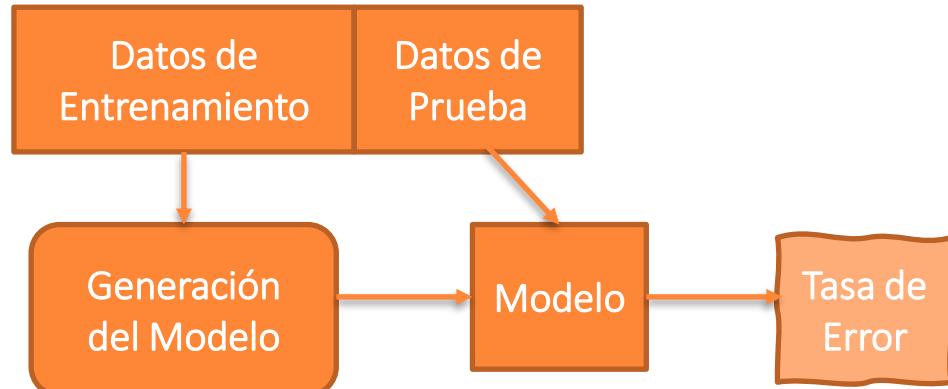
- Supervisado
- No supervisado

○ Prueba

- De los casos históricos disponibles se destina una cantidad para entrenar el modelo y se reserva una porción para probar el modelo
- Se presentan los casos como si fueran nuevos y se coteja la respuesta del modelo con los valores reales

○ Evaluación

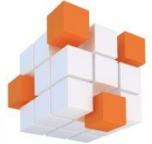
○ Despliegue (Producción): se dispone del modelo apto para su explotación, casos nuevos.





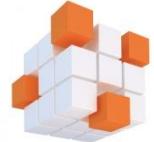
TECNOLOGÍAS DE EXPLOTACIÓN DE INFORMACIÓN

- Basadas en Análisis Estadístico
 - Análisis de varianza
 - Regresión
 - Prueba Chi-cuadrado
 - Análisis de agrupamiento
 - Análisis de determinantes
 - Series de tiempo
- Basadas en Sistemas Inteligentes
 - Algoritmos de inducción TDIDT
 - Redes Neuronales SOM
 - Redes Bayesianas
 - Redes Neuronales Back-Propagation

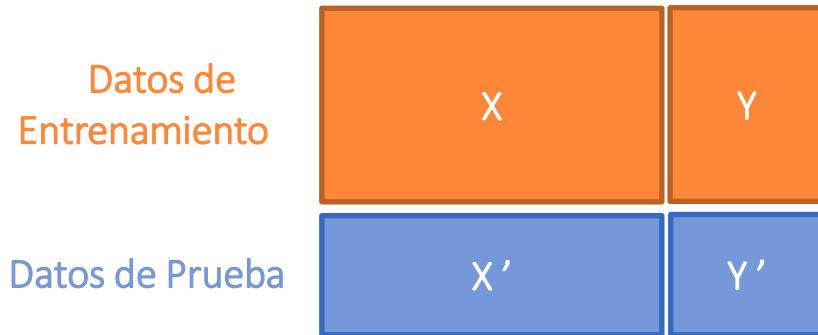


MODELOS PREDICTIVOS

- Son modelos de aprendizaje supervisado que permiten predecir el resultado de variables de interés a partir de los valores de otras variables.
 - Variables de entrada ⇒ Atributos predictores
 - Variable a predecir ⇒ Atributo clase
- Se tiene un conjunto de *casos de entrenamiento* donde cada caso contiene un conjunto de **atributos** y uno de ellos es la **clase** a clasificar.
- Se separa un conjunto de *casos de prueba* para predecir nuevos casos y probar el modelo.
- Los nuevos casos deben ser asignados a su clase con la máxima exactitud y precisión posible.



MODELOS PREDICTIVOS



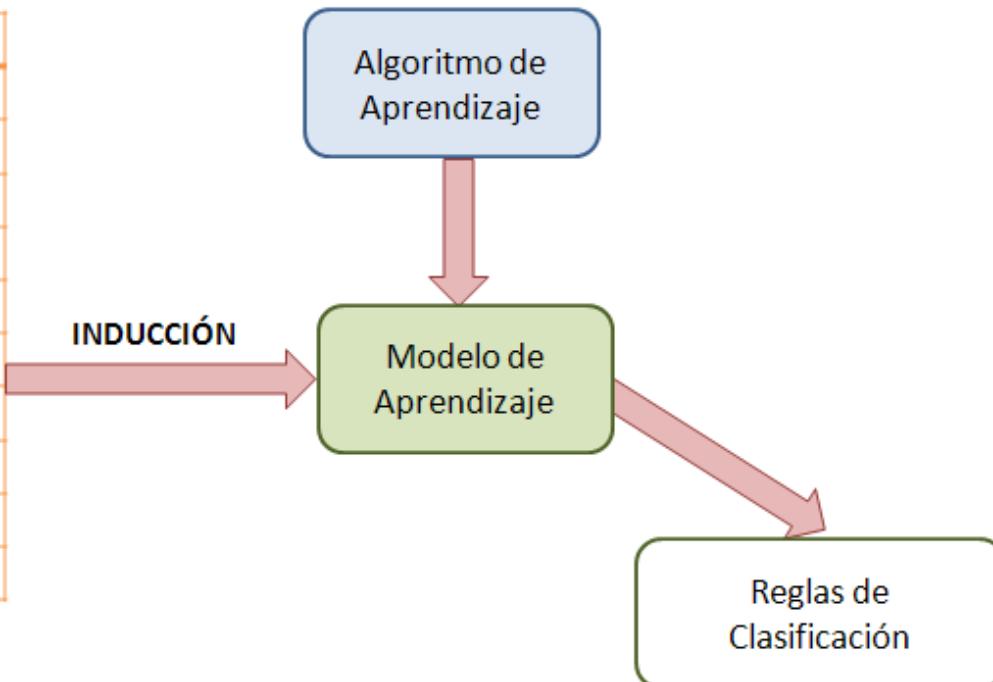
- El entrenamiento busca descubrir las relaciones entre las variables de entrada (X y X') y la variable objetivo (clase) (Y e Y').
- En “producción” se usa ese conocimiento para predecir el valor de la variable objetivo ($Y=?$) de un nuevo caso no incluido en los datos de entrenamiento ni de prueba.



TAREA DE CLASIFICACIÓN

ATRIB 1	ATRIB 2	ATRIB 3	CLASE
Si	Large	125K	No
No	Medium	100K	No
No	Small	70K	No
Si	Medium	120K	No
No	Large	95K	Si
No	Medium	60K	No
Si	Large	220K	No
No	Small	85K	Si
No	Medium	75K	No
No	Small	90K	Si

Casos de Entrenamiento



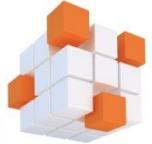
ATRIB 1	ATRIB 2	ATRIB 3	CLASE
No	Small	55K	?
Si	Medium	80K	?
Si	Large	110K	?
No	Small	95K	?
No	Large	67K	?

Casos de Prueba



TECNOLOGÍAS PARA CLASIFICACIÓN

- Árboles de Decisión
 - Algoritmos de inducción TDIDT
 - Métodos basados en reglas
- Redes Bayesianas
 - Naïve-Bayes (Bayes Ingenuo)
- Vecinos más cercanos
 - K-vecinos (CBR)



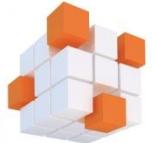
ALGORITMOS DE INDUCCIÓN TDIDT

- La familia TDIDT (*Top Down Induction Trees*) pertenece a los métodos inductivos del Aprendizaje Automático que aprenden a partir de ejemplos preclasificados.
 - Atributos **predictores** ⇒ Se partitionan en diferentes ramas de acuerdo a los valores que el atributo puede tomar.
 - Pueden ser discretos o continuos.
 - Atributo **clase** ⇒ Decide la clase asignada (variable objetivo)
 - Debe ser discretizado.
- Generan árboles y reglas de decisión a partir de ejemplos preclasificados.



ALGORITMOS DE INDUCCIÓN TDIDT

- Se trata de identificar y ubicar en la parte superior del árbol a los atributos que mejor separan los ejemplos o muestras.
- Para encontrar los mejores atributos utiliza la teoría de la información, determinando qué atributo aporta la mayor ganancia de información (o menor pérdida de información) al tomar un determinado valor.
- Algoritmos utilizados ID3, C4.5 y C5

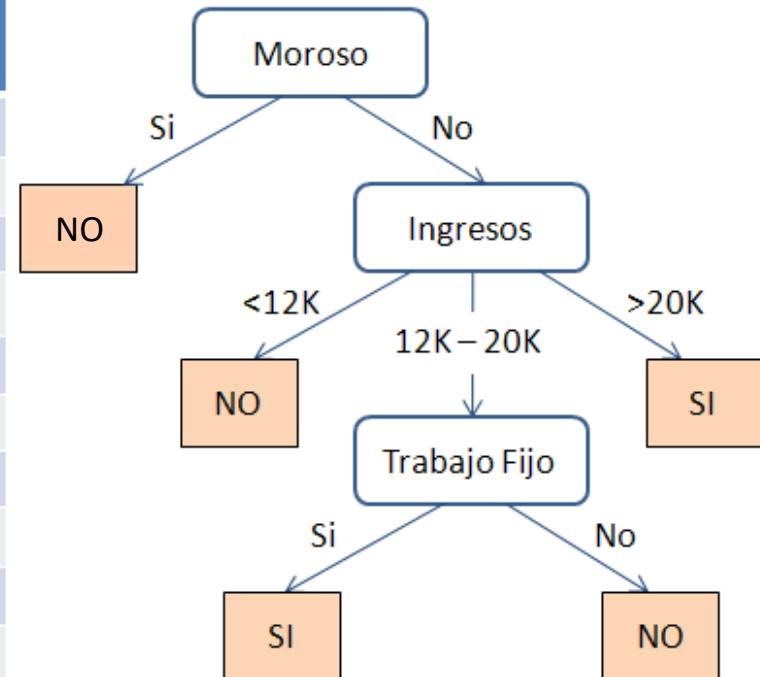


ALGORITMO TDIDT – EJEMPLO 1

- Presentación intuitiva del proceso de inducción.
- Evaluación de otorgamiento de préstamos a clientes
 - Atributo clase: **Otorgar Préstamo**

Cliente	Moroso	Antigüedad	Ingresos	Trabajo Fijo	Otorgar Préstamo
1	Si	> 5	12K – 20K	Si	No
2	No	< 1	12K – 20K	Si	Si
3	Si	1 - 5	> 20K	Si	No
4	No	> 5	> 20K	No	Si
5	No	< 1	> 20K	Si	Si
6	Si	1 - 5	12K – 20K	Si	No
7	No	1 - 5	> 20K	Si	Si
8	No	< 1	< 12K	Si	No
9	No	> 5	12K – 20K	No	No
10	Si	1 - 5	< 12K	No	No

Casos de Entrenamiento

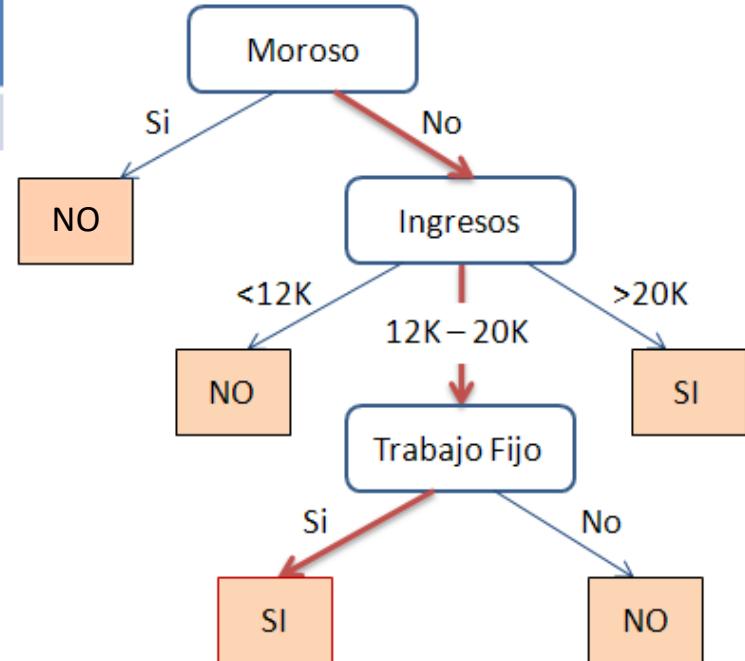




ALGORITMO TDIDT - APLICACIÓN

○ Caso de Prueba

Cliente	Moroso	Antigüedad	Ingresos	Trabajo Fijo	Otorgar Préstamo
11	No	1 - 5	12K – 20K	Si	?

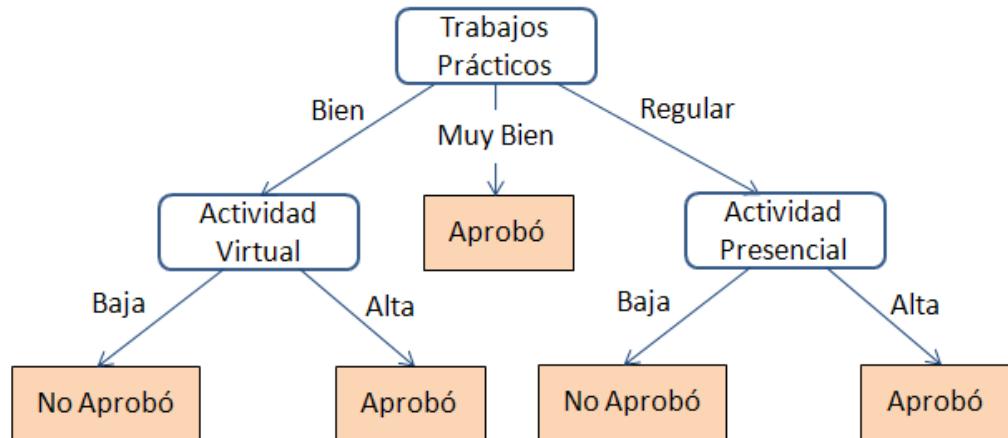


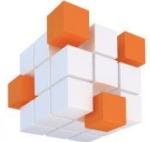


ALGORITMO TDIDT – EJEMPLO 2

- Predicción de resultados de exámenes
 - Atributo clase: Resultado Parcial

Trabajos Prácticos	Actividad Virtual	Actividad Presencial	Resultado Parcial
Bien	Alta	Alta	Aprobó
Bien	Baja	Alta	No Aprobó
Muy Bien	Alta	Alta	Aprobó
Regular	Alta	Alta	Aprobó
Regular	Alta	Baja	No Aprobó
Regular	Baja	Baja	No Aprobó
Muy Bien	Baja	Baja	Aprobó
Bien	Baja	Baja	No Aprobó
Bien	Alta	Baja	Aprobó
Regular	Baja	Baja	No Aprobó
Bien	Alta	Baja	Aprobó
Muy Bien	Alta	Alta	Aprobó
Regular	Baja	Baja	Aprobó
Regular	Alta	Alta	Aprobó





ALGORITMO TDIDT - APRENDIZAJE

- Construcción de reglas del tipo IF-THEN

R₁: IF Trabajos Prácticos = 'Muy Bien' THEN Resultado Parcial = 'Aprobó'

R₂: IF (Trabajos Prácticos = 'Bien') AND (Actividad Virtual = 'Baja')
THEN Resultado Parcial = 'No Aprobó'

R₃: IF (Trabajos Prácticos = 'Bien') AND (Actividad Virtual = 'Alta')
THEN Resultado Parcial = 'Aprobó'

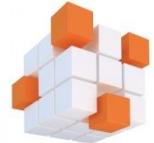
R₄: IF (Trabajos Prácticos = 'Regular') AND (Actividad Presencial = 'Baja')
THEN Resultado Parcial = 'No Aprobó'

R₅: IF (Trabajos Prácticos = 'Regular') AND (Actividad Presencial = 'Alta')
THEN Resultado Parcial = 'Aprobó'

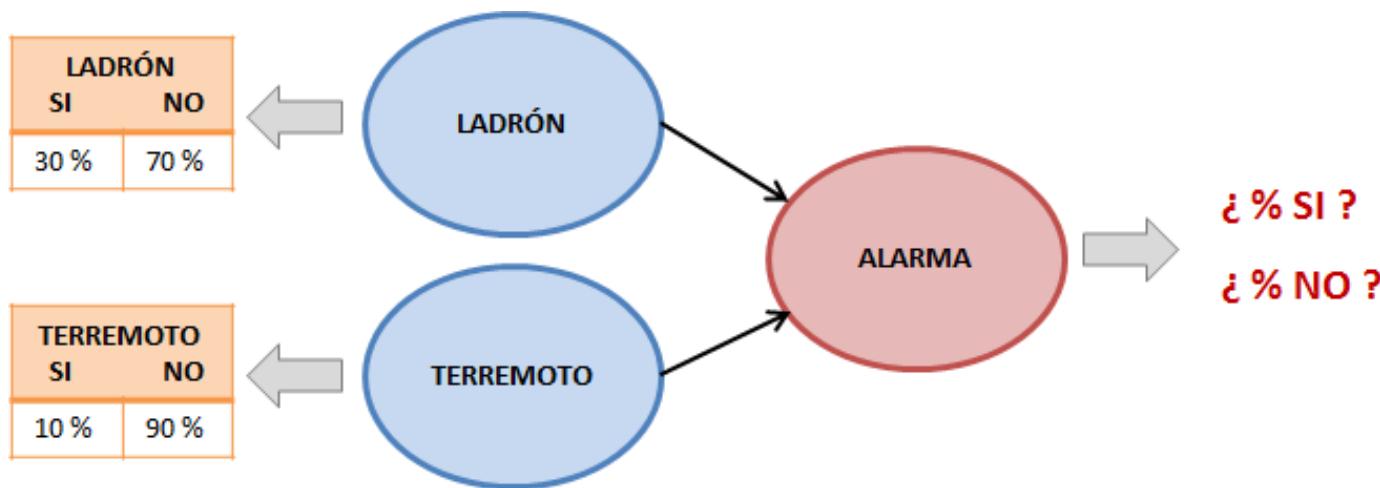


REDES BAYESIANAS

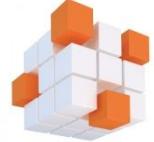
- Una red bayesiana es un grafo acíclico dirigido compuesto de nodos y arcos.
- Los nodos representan las variables aleatorias (o atributos).
- Los arcos representan dependencias probabilísticas de cada variable.
 - El arco entre dos variables significa una influencia directa de una variable sobre otra.
 - Probabilidad condicional (Teorema de Bayes).
- Representan la relación causa-efecto entre atributos.
- Dan a una medida cuantitativa y probabilística de la importancia de los atributos en un problema de clasificación de clases.



REDES BAYESIANAS – EJEMPLO

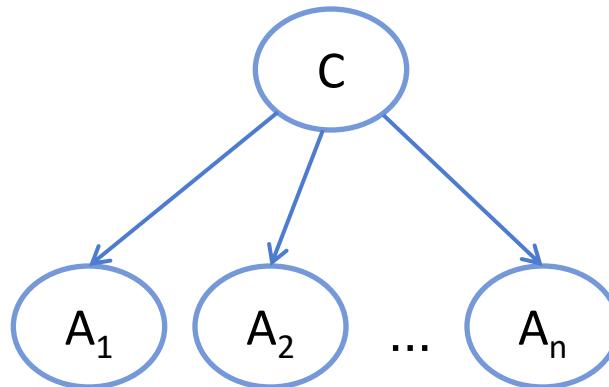


- Las variables *Ladrón* y *Terremoto* son causas para que se dispare una Alarma.
 - Existe una probabilidad a priori para Ladrón y Terremoto.
 - ¿Cuál es la probabilidad de que suene o no la alarma?
- *Ladrón* y *Terremoto* son condicionalmente independientes entre sí dada la variable *Alarma*.

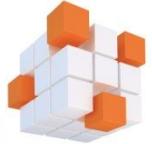


CLASIFICADOR BAYESIANO NAÏVE-BAYES

- Considera que cada atributo predictor A_i y el atributo clase C son variables aleatorias.
- Las relaciones de dependencia entre los atributos A_i son condicionalmente independientes entre sí dado el atributo clase C .



- Dado un registro con atributos A_1, A_2, \dots, A_n el objetivo es predecir la clase C .
- Se busca encontrar el valor de C que maximice la probabilidad $p(C/A_1, A_2, \dots, A_n)$.



REDES BAYESIANAS

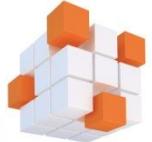
- Obtener una red bayesiana a partir de datos, es un proceso de aprendizaje.
 - Aprendizaje Estructural
 - Aprendizaje Paramétrico
- Proceso de inferencia
 - Predicciones a partir de observaciones



REDES BAYESIANAS – EJEMPLO

- Se tienen los siguientes datos:

	Ambiente	Temperatura	Humedad	Viento	Juega Tenis
1	Soleado	Alta	Alta	Leve	No
2	Soleado	Alta	Alta	Fuerte	No
3	Nublado	Alta	Alta	Leve	Si
4	Lluvioso	Media	Alta	Leve	No
5	Lluvioso	Baja	Normal	Fuerte	No
6	Lluvioso	Baja	Normal	Fuerte	No
7	Nublado	Baja	Normal	Leve	Si
8	Soleado	Media	Alta	Leve	Si
9	Soleado	Baja	Normal	Leve	Si
10	Lluvioso	Media	Normal	Leve	No
11	Soleado	Media	Normal	Fuerte	Si
12	Nublado	Media	Alta	Fuerte	Si
13	Nublado	Alta	Normal	Leve	Si
14	Lluvioso	Media	Alta	Fuerte	No

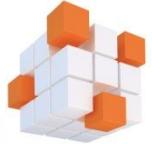


REDES BAYESIANAS – EJEMPLO

- Queremos saber si se jugará al tenis bajo las siguientes condiciones:

Ambiente	Temperatura	Humedad	Viento	Juega Tenis
Soleado	Baja	Alta	Fuerte	?

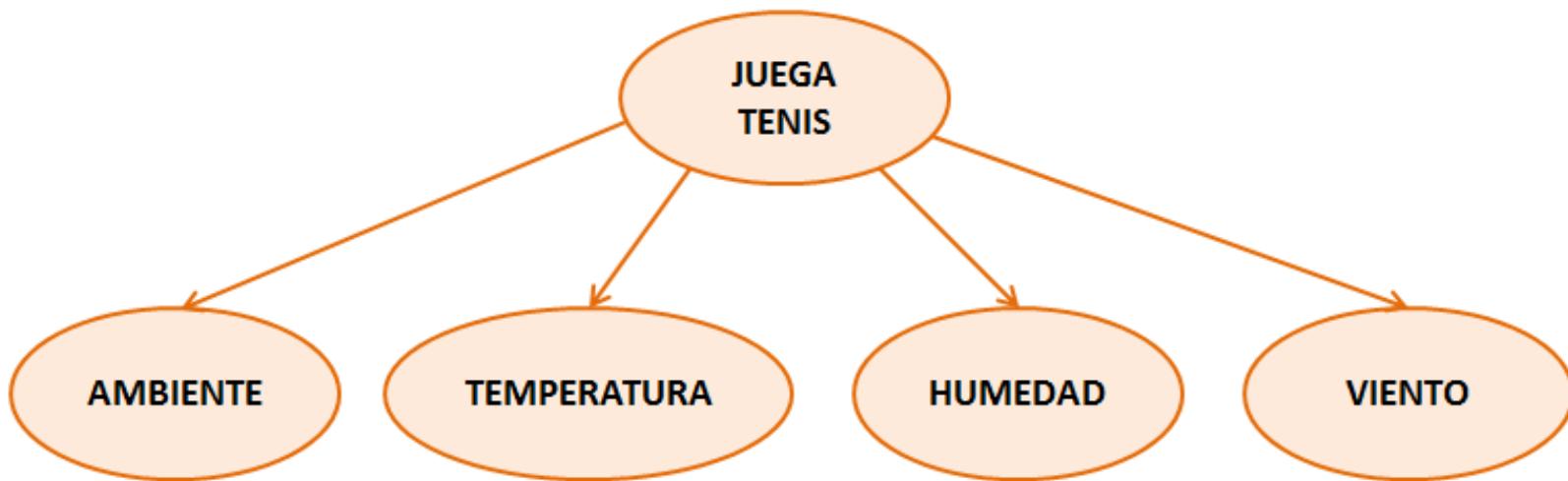
- El atributo clase a predecir es **Juega Tenis** cuyos valores serán **Si o No**.
- El nuevo caso será clasificado como clase C_j si $P(C_j) \prod_{i=1}^n P(A_i | C_j)$ es máximo.

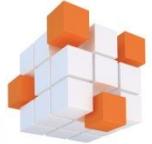


REDES BAYESIANAS – EJEMPLO

Aprendizaje Estructural

- Relaciones de dependencia e independencia





REDES BAYESIANAS – EJEMPLO

Aprendizaje Paramétrico

- Determinar probabilidades a priori de cada clase y las probabilidades condicionales.
- Analizando los 14 casos tenemos:

	Valores que toma	Cantidad de Casos	% casos totales
Ambiente	Soleado	5	35,7%
	Nublado	4	28,6%
	Lluvioso	5	35,7%
Temperatura	Alta	4	28,6%
	Media	6	42,8%
	Baja	4	28,6%
Humedad	Alta	7	50%
	Normal	7	50%
Viento	Leve	8	57,2%
	Fuerte	6	42,8%

Casos Juega Tenis = Si = 7

Casos Juega Tenis = No = 7

$P(\text{Juega Si}) = 0,5 = 50\%$

$P(\text{Juega No}) = 0,5 = 50\%$



REDES BAYESIANAS – EJEMPLO

- Desglosando los casos según si juegan o no al tenis:

Cantidad Casos

	Valores que toma	Clase = Juega Tenis	
		Si	No
Ambiente	Soleado	3	2
	Nublado	4	0
	Lluvioso	0	5
Temperatura	Alta	2	2
	Media	3	3
	Baja	2	2
Humedad	Alta	3	4
	Normal	4	3
Viento	Leve	5	3
	Fuerte	2	4

Casos totales = 14

Casos **Juega Tenis = Si** = 7

Casos **Juega Tenis = No** = 7

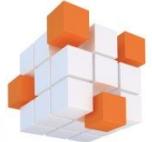


REDES BAYESIANAS – EJEMPLO

- Obtenemos las probabilidades condicionales:

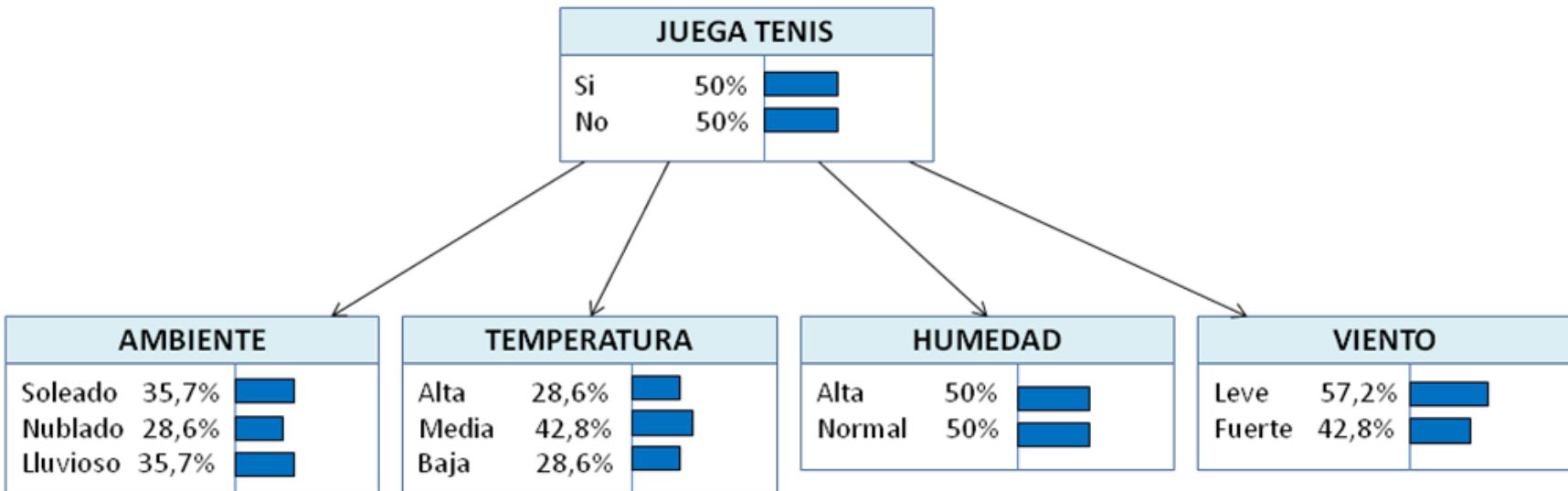
		Probabilidades	
		Clase = Juega Tenis	
		Si	No
Ambiente	Soleado	$3/7 = 42,8\%$	$2/7 = 28,6\%$
	Nublado	$4/7 = 57,2\%$	0
	Lluvioso	0	$5/7 = 71,4\%$
Temperatura	Alta	$2/7 = 28,6\%$	$2/7 = 28,6\%$
	Media	$3/7 = 42,8\%$	$3/7 = 42,8\%$
	Baja	$2/7 = 28,6\%$	$2/7 = 28,6\%$
Humedad	Alta	$3/7 = 42,8\%$	$4/7 = 57,2\%$
	Normal	$4/7 = 57,2\%$	$3/7 = 42,8\%$
Viento	Leve	$5/7 = 71,4\%$	$3/7 = 42,8\%$
	Fuerte	$2/7 = 28,6\%$	$4/7 = 57,2\%$

Casos totales = 14



REDES BAYESIANAS – EJEMPLO

Proceso de Inferencia

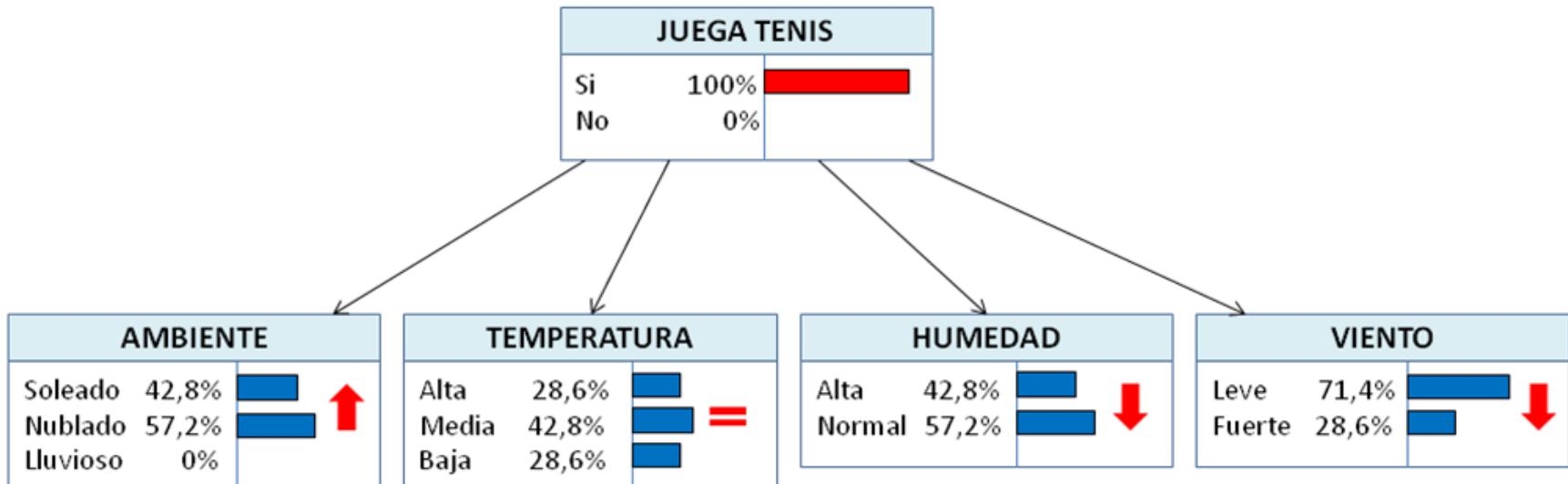




REDES BAYESIANAS – EJEMPLO

Proceso de Inferencia

- Juega Tenis = Si

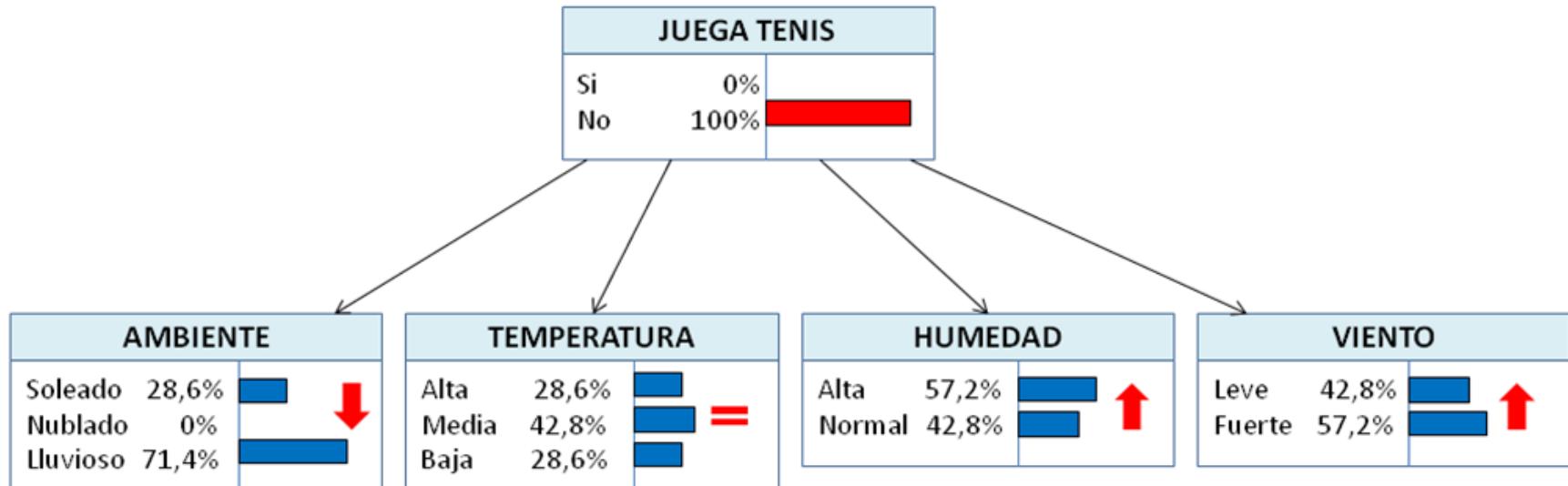


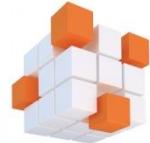


REDES BAYESIANAS – EJEMPLO

Proceso de Inferencia

- Juega Tenis = No





REDES BAYESIANAS – EJEMPLO

- Predicción a realizar:

Ambiente	Temperatura	Humedad	Viento	Juega Tenis
Soleado	Baja	Alta	Fuerte	?

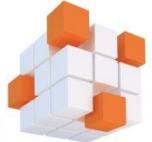
$$P(\text{Juega Si}) = 0,5$$
$$P(\text{Juega No}) = 0,5$$

- $P(\text{Juega Si}) = 0,428 \times 0,286 \times 0,428 \times 0,286 \times 0,5 = 0,0075$



- $P(\text{Juega No}) = 0,286 \times 0,286 \times 0,572 \times 0,572 \times 0,5 = 0,0133$





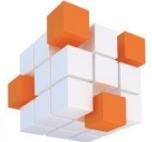
REDES BAYESIANAS – EJEMPLO

- Normalizando:

- $P(\text{Juega Si}) = 0,0075 / (0,0075 + 0,0133) = 36\%$
- $P(\text{Juega No}) = 0,0133 / (0,0075 + 0,0133) = 64\%$

Ambiente	Temperatura	Humedad	Viento	Juega Tenis
Soleado	Baja	Alta	Fuerte	?

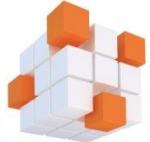
- El clasificador va a predecir que no se juega al tenis con una probabilidad del 64%.



MODELOS DE CLASIFICACIÓN - EVALUACIÓN

Pautas para la evaluación

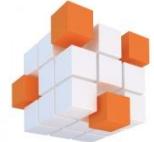
- Exactitud
 - No hay un algoritmo que sea siempre mejor que otro u otros
- Precisión, Recall (Exhaustividad/Sensibilidad/TPR y Especificidad/TNR) + F1-Score, Kappa (Cohen)
 - Útiles cuando el dataset no está balanceado, dan una mejor idea de la calidad del modelo.
- Interpretabilidad
 - Facilidad para interpretar los resultados
- Velocidad
 - Entrenamiento
 - Producción



MODELOS DE CLASIFICACIÓN - EVALUACIÓN

Métricas de Evaluación

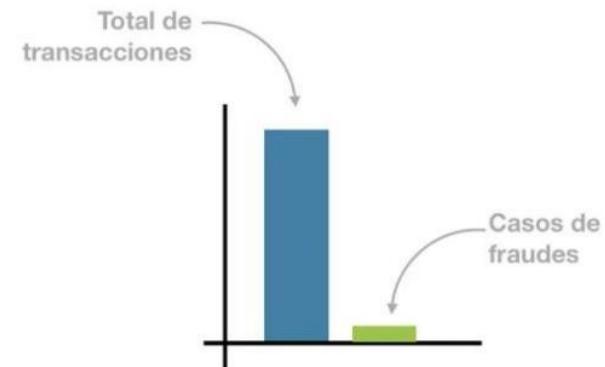
- Se focalizan en analizar la capacidad de predicción y clasificación de clases del modelo construido.
- **Matriz de Confusión:** Permite comparar el resultado obtenido a partir del modelo predictivo construido con los resultados de los datos de prueba del modelo.
- Métricas utilizadas (en entrenamiento y prueba)
 - Exactitud del modelo
 - Precisión del modelo
 - Recall (Exhaustividad/Sensibilidad/TPR y Especificidad/TNR)
 - F1-Score + Coeficiente Kappa
- Otras métricas utilizadas
 - Cobertura de una regla
 - Precisión de una regla

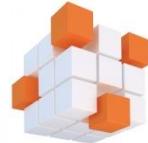


MODELOS DE CLASIFICACIÓN

Dataset desbalanceado

- Es aquel que tiene muchas instancias de una clase y muy pocas de la otra, dificultando así el entrenamiento.
- Algo de desbalance de clases es de esperar y no afecta a nuestro análisis; pero bajo ciertas problemáticas, suelen haber datasets muy desbalanceados:
 - Detección de fraudes.
 - Diagnóstico médico.
 - Falla en cadena de producción.
- Atención en:
 - Cómo se entrena los modelos.
 - Qué métricas se usan para evaluarlos: las siguientes son útiles cuando el dataset no está balanceado, dan una mejor idea de la calidad del modelo,
 - Precisión / Recall (Exhaustividad/Sensibilidad/TPR y Especificidad/TNR) / F1-Score / Kappa (Cohen)
 - Por el contrario, la Exactitud no es buena cuando el dataset está desbalanceado.



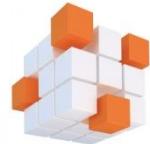


MODELOS DE CLASIFICACIÓN - EVALUACIÓN

- Estructura de una matriz de confusión de 2 clases.

		Clase Clasificada		
		Clase A	Clase B	Total
Clase Real	Clase A	Nº casos clasificados como A y son de clase A (NCVA)	Nº casos clasificados como B pero son de clase A (NCFB)	Total de casos de la clase A
	Clase B	Nº casos clasificados como A pero son de clase B (NCFA)	Nº casos clasificados como B y son de clase B (NCVB)	Total de casos de la clase B
Total	Total de casos clasificados como clase A	Total de casos clasificados como clase B	Número total de casos (NTC)	

- Las métricas NCVA y NCVB representan los valores clasificados correctamente por el modelo (V: Verdadero)
- Las métricas NCFA y NCFB representan los errores (la confusión) entre las clases (F: Falso)

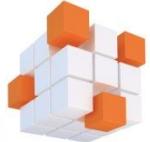


MODELOS DE CLASIFICACIÓN - EVALUACIÓN

- Forma más usual de encontrar un Matriz de Confusión, incluyendo métricas para evaluar data sets desbalanceados

		Clase Clasificada		Total	Exahustividad (Recall, Sensibilidad o TPR) (% casos positivos detectados)	$= TP / (TP+FN)$
Clase Real	Clase A (Positive)	TP (VP ó NCVA)	FN (NCFB)	Total de Casos Reales de la Clase A	Especificidad (TNR) (% casos negativos detectados)	$= TN / (TN+FP)$
	Clase B (Negative)	FP (NCFA)	TN (VN ó NCVB)	Total de Casos Reales de la Clase B	Exactitud (% predicciones positivas correctas) no es util en DS desbalanceados	
Total		Total de Casos Clasificados como Clase A	Total de Casos Clasificados como Clase B	Nro Total de casos (NTC)	F1-Score (A=P)	$= 2*((\text{Precisión} * \text{Recall}) / (\text{Precisión} + \text{Recall}))$
Precisión (A=P) (% predicciones positivas correctas)		Precisión (B=N) (% predicciones negativas correctas)			F1-Score (B=N)	$= 2*((\text{Precisión} * \text{Especificidad}) / (\text{Precisión} + \text{Especificidad}))$
$= TP / (TP + FP)$		$= TN / (TN + FN)$			Coeficiente Kappa	$= 2*(TP * TN - FN * FP) / (TP + FP) * (FP + TN) + (TP + FN) * (FN + TN)$

- Fuente: Mejora del artículo: Telefónica Think Big / Empresas - Cómo interpretar la matriz de confusión: ejemplo práctico Paloma Recuero de los Santos



MODELOS DE CLASIFICACIÓN - EVALUACIÓN

- **Exactitud del modelo:** Proporción de casos clasificados correctamente respecto del número total de casos utilizados. Evalúa la capacidad de generalización del modelo para predecir y clasificar nuevos casos.

$$\text{Exactitud (M)} = \frac{\sum_{i=1}^n \text{NCV}_i}{\text{Nº casos usados}}$$

- ALTA exactitud \Rightarrow Clasificaciones correctas $\geq 70\%$ casos.
- BAJA exactitud \Rightarrow Clasificaciones correctas $< 70\%$ casos. Modelo poco confiable.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$



MODELOS DE CLASIFICACIÓN - EVALUACIÓN

- **Precisión del modelo:** proporción de casos reales de una clase respecto del total de casos clasificados por el modelo en esa misma clase. Evalúa la efectividad del modelo para clasificar casos a una clase particular.

$$\text{Precisión } (C_i) = \frac{NCV_i}{\text{Total Casos Clasificados Positivos}} = \frac{NCV_i}{NCV_i + NCF_i}$$

- ALTA precisión \Rightarrow Modelo efectivo para predecir y clasificar nuevos casos.
- Precisión MEDIA \Rightarrow Modelo inestable. Posible confusión en clasificación y predicción.
- BAJA precisión \Rightarrow El modelo confunde las clases. Modelo poco efectivo.

		predicción	
		0	1
realidad	0	TN	FP
	1	FN	TP

$$\text{precision} = \frac{TN}{TN + FN}$$

$$\text{precision} = \frac{TP}{TP + FP}$$



MODELOS DE CLASIFICACIÓN - EVALUACIÓN

- **Exhaustividad (Recall / Sensibilidad / TPR)**: tasa de verdaderos positivos (True Positive Rate) ó TP; es decir la proporción entre los casos positivos bien clasificados por el modelo, respecto a todos los elementos que en realidad son positivos.

Exhaustividad =

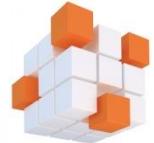
$$\frac{NCV_A}{\text{Total Casos Reales Positivos}} = \frac{NCV_A}{NCV_A + NCF_B}$$

- Expresa qué tan bien el modelo es capaz de detectar a la clase positiva.

$$recall = \frac{TP}{TP + FN}$$

		predicción	
		0	1
realidad	0	TN	FP
	1	FN	TP

Exhaustividad (recall)



MODELOS DE CLASIFICACIÓN - EVALUACIÓN

- **Especificidad (TNR)**: tasa de verdaderos negativos, (True Negative Rate) ó TN, es decir la proporción entre los casos negativos bien clasificados por el modelo, respecto a todos los elementos que en realidad son negativos.

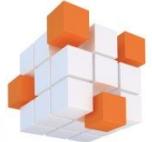
Especificidad =

$$\frac{NCV_B}{\text{Total Casos Reales Negativos}} = \frac{NCV_B}{NCF_A + NCV_B}$$

- Si lo que nos interesa es identificar los verdaderos negativos, (evitar falsos positivos) debemos elegir especificidad alta.

$$Especificidad = \frac{TN}{(TN + FP)}$$

		predicción	
		0	1
realidad	0	TN	FP
	1	FN	TP

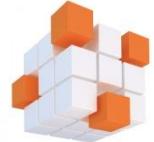


MODELOS DE CLASIFICACIÓN - EVALUACIÓN

- **F1-Score:** se utiliza para combinar las medidas de precisión y Recall en un sólo valor. Esto es práctico porque hace más fácil el poder comparar el rendimiento combinado de la precisión y la exhaustividad entre varias soluciones.
- F1 se calcula haciendo la media armónica entre la precisión y la exhaustividad:

$$F1 - Score_{A(P)} = 2 \left(\frac{(Precision_{A(P)} * Recall_{Sensibilidad})}{(Precision_{A(P)} + Recall_{Sensibilidad})} \right)$$

$$F1 - Score_{B(N)} = 2 \left(\frac{(Precision_{B(N)} * Recall_{Especificidad})}{(Precision_{B(N)} + Recall_{Especificidad})} \right)$$



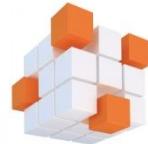
MODELOS DE CLASIFICACIÓN - EVALUACIÓN

- **Coeficiente Kappa - Concordancia entre Predicciones y Realidad:** Este es una métrica que mide el grado de acuerdo entre las predicciones de un modelo y la realidad, más allá del simple azar.
- **Interpretación de Valores Kappa:** Los valores oscilan entre -1 y 1, donde 0 indica un acuerdo únicamente por azar, y 1 representa un acuerdo perfecto entre predicciones y realidad.
- Es especialmente útil cuando las clases en una tarea de clasificación tienen diferentes prevalencias, ya que ofrece una mejor evaluación que la simple precisión.

$$k = \frac{2 * (TP * TN - FN * FP)}{(TP + FP) * (FP + TN) + (TP + FN) * (FN + TN)}$$

Valoración del coeficiente kappa
(Landis y Koch, 1977)

Coeficiente kappa	Fuerza de la concordancia
0,00	Pobre (<i>Poor</i>)
0,01 - 0,20	Leve (<i>Slight</i>)
0,21 - 0,40	Aceptable (<i>Fair</i>)
0,41 - 0,60	Moderada (<i>Moderate</i>)
0,61 - 0,80	Considerable (<i>Substantial</i>)
0,81 - 1,00	Casi perfecta (<i>Almost perfect</i>)



EVALUACIÓN - RESUMEN

Exactitud: (modelo)

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Tasa Error: (modelo)

$$\text{Error} = 1 - \text{Exactitud}$$

Exhaustividad: (para cada clase)

(Recall/Sensibilidad/ TPR)

$$\text{recall} = \frac{TP}{TP + FN}$$

		predicción	
		0	1
realidad	0	TN	FP
	1	FN	TP

F1-Score: (para cada clase)

$$F1 - Score_{A(P)} = 2 \left(\frac{(Precision_{A(P)} * Recall_{Sensibilidad})}{(Precision_{A(P)} + Recall_{Sensibilidad})} \right)$$

Precisión: (para cada clase)

$$\text{precisión}_{(0)} = \frac{TN}{TN + FN}$$

$$\text{precision}_{(1)} = \frac{TP}{TP + FP}$$

		predicción	
		0	1
realidad	0	TN	FP
	1	FN	TP

Especificidad: (para cada clase)

(TNR)

$$\text{Especificidad} = \frac{TN}{(TN + FP)}$$

		predicción	
		0	1
realidad	0	TN	FP
	1	FN	TP

Kappa: (modelo)

$$k = \frac{2 * (TP * TN - FN * FP)}{(TP + FP) * (FP + TN) + (TP + FN) * (FN + TN)}$$



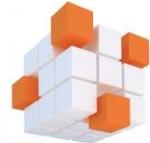
MODELOS DE CLASIFICACIÓN - EJEMPLO

- Consideremos una matriz de confusión con 900 casos de clientes que pueden o no cerrar sus cuentas bancarias:
 - Clase SI – Se va del banco
 - Clase No – No se va del banco

		Clase Clasificada	
		Si	No
Clase Real	Si	455	29
	No	32	384
		487	413

839 predicciones correctas
61 predicciones incorrectas

- Exactitud (M)** = $(455+384)/(455+384+29+32) = 93,2\%$
- Precisión (C_{Si})** = $455/(455+32) = 93,4\%$
- Precisión (C_{No})** = $384/(29+384) = 92,9\%$



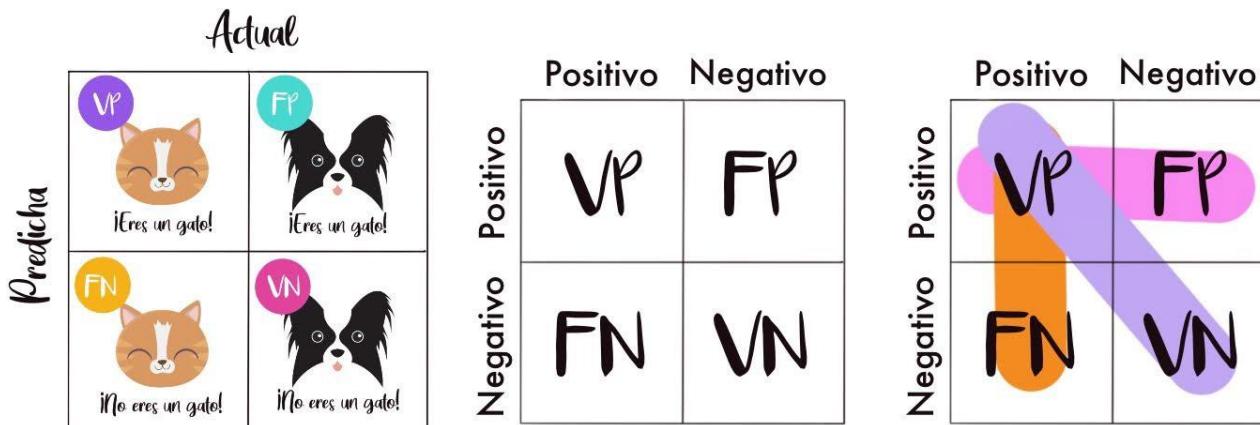
MODELOS DE CLASIFICACIÓN - EJEMPLO

		Clase Clasificada		Total de Casos Reales de la Clase A	484	Exhaustividad (Recall o TPR)	$= TP / (TP+FN)$	94,01%
Clase Real	Clase A (Positive)	Clase A (Positive)	Case B (Negative)					
	Clase B (Negative)	32	384	Total de Casos Reales de la Clase B	416	Especificidad (TNR)	$= TN / (TN+FP)$	92,31%
Total	Total de Casos Clasificados como Clase A	Total de Casos Clasificados como Clase B	Nro Total de casos (NTC)			Exactitud	$= (TN + TP) / (TN + TP + FN + FP)$	93,22%
	487	413				F1-Score (A=P)	$= 2*((Precisión* Recall) / (Precisión+Recall))$	92,64%
	Precisión (A=P)	Precisión (B=N)				F1-Score (B=N)	$= 2*((Precisión * Especificidad) / (Precisión + Especificidad))$	92,64%
	$= TP / (TP + FP)$	$= TN / (TN + FN)$				Coeficiente Kappa	$= 2*(TP * TN - FN * FP) / (TP + FP) * (FP + TN) + (TP + FN) * (FN + TN)$	86,36%
	93,43%	92,98%						

- No todos los errores tienen el mismo costo para el banco.
 - El error en los falsos **No** (29 casos) es mucho más costoso para el banco ya que no se va a tomar ninguna acción y el cliente se va a ir.
 - El banco puede asignar un valor de costo a cada una de las celdas que representan un error en la clasificación para poder comparar soluciones de modelos.



MODELOS DE CLASIFICACIÓN



Recall

De todas las clases positivas cuantas se predijo correctamente

$$\frac{VP}{VP+FN}$$
$$\frac{VN}{VN + FP}$$

Precisión

De todas las positivas que se han predicho correctamente cuántas son realmente positivas

$$\frac{VP}{VP+FP}$$
$$\frac{VN}{VN + FN}$$

Accuracy

De todas las clases cuantas se predijeron correctamente

$$\frac{VP+VN}{Total}$$

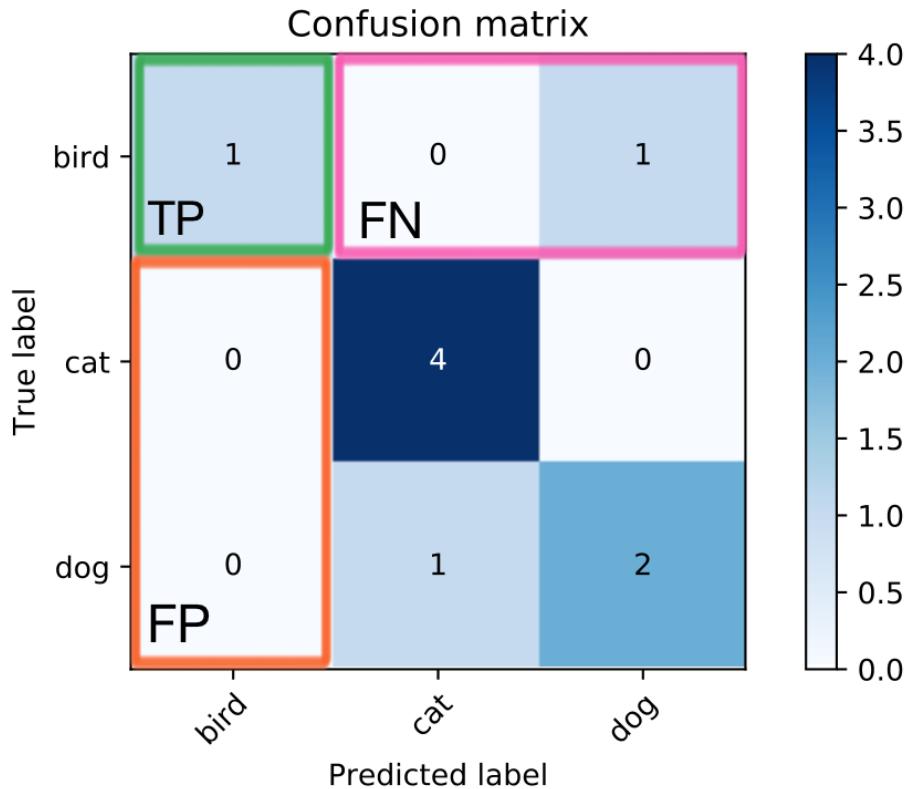
Medida F

Permite comparar dos modelos de baja precisión y alta exhaustividad (recall) utiliza la media armónica para castigar los valores extremos

$$\frac{2 * \text{recall} * \text{precision}}{\text{Recall} + \text{precision}}$$



MODELOS DE CLASIFICACIÓN



Fuentes: <https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion>
<https://www.youtube.com/watch?v=D5iuHDepUFw>



MODELOS DE CLASIFICACIÓN – EJEMPLO MULTICLASE

- Se cuenta con una matriz de confusión para un problema clases múltiples donde se debe predecir si una persona prefiere Facebook, Instagram o Snapchat.
- La matriz de confusión sería de 3×3 .
- El verdadero positivo, verdadero negativo, falso positivo y falso negativo de cada clase se calcularía sumando los valores de celda de la siguiente manera:

		ACTUAL VALUES		
		f	Instagram	Snapchat
PREDICTED VALUES	f	+ve 1	-ve 2	-ve 3
	Instagram	-ve 4	+ve 5	-ve 6
Snapchat	Snapchat	-ve 7	-ve 8	+ve 9

Facebook

$$TP = Cell_1$$

$$FP = Cell_2 + Cell_3$$

$$TN = Cell_5 + Cell_6 + Cell_8 + Cell_9$$

$$FN = Cell_4 + Cell_7$$

Instagram

$$TP = Cell_5$$

$$FP = Cell_4 + Cell_6$$

$$TN = Cell_1 + Cell_3 + Cell_7 + Cell_9$$

$$FN = Cell_2 + Cell_8$$

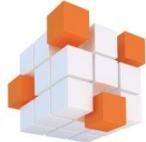
Snapchat

$$TP = Cell_9$$

$$FP = Cell_7 + Cell_8$$

$$TN = Cell_1 + Cell_2 + Cell_4 + Cell_5$$

$$FN = Cell_3 + Cell_6$$



MODELOS DE CLASIFICACIÓN - CONCLUSIONES

Resumen 1:

- No hay que guiarse solamente por la matriz de confusión.
- No solamente cuenta la exactitud y precisión.
 - Proporción de aciertos en la matriz de confusión
 - No hay un algoritmo que siempre sea “mejor” que otros
- No necesariamente un nivel bajo de aciertos en la predicción invalida el uso del modelo.
- Hay que tener en cuenta otros factores dependiendo del caso de estudio.
 - Interpretabilidad
 - Velocidad
 - Entrenamiento
 - Producción



MODELOS DE CLASIFICACIÓN - CONCLUSIONES

Resumen 2:

- Vimos las métricas más extendidas para evaluar el rendimiento de un modelo supervisado en tareas de clasificación.
- La Matriz de Confusión indica qué tipos de errores se cometan
- La métrica Exactitud es engañosa cuando las clases están desbalanceadas, nos hace creer que el modelo es mejor de lo que en realidad es.
- Las medidas de Precisión, Recall y F1 son más representativas y funcionan tanto si las clases están balanceadas como si no:
 - **Precisión** nos da la calidad de la predicción: ¿qué porcentaje de los que hemos dicho que son la clase positiva, en realidad lo son?
 - **Recall** nos da la cantidad: ¿qué porcentaje de la clase positiva hemos sido capaces de identificar? La sensibilidad (+) y la especificidad (-) indican la capacidad del estimador para discriminar los casos positivos, de los negativos.
 - **F1** combina Precisión y Recall en una sola medida.

Fuente: <https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/>



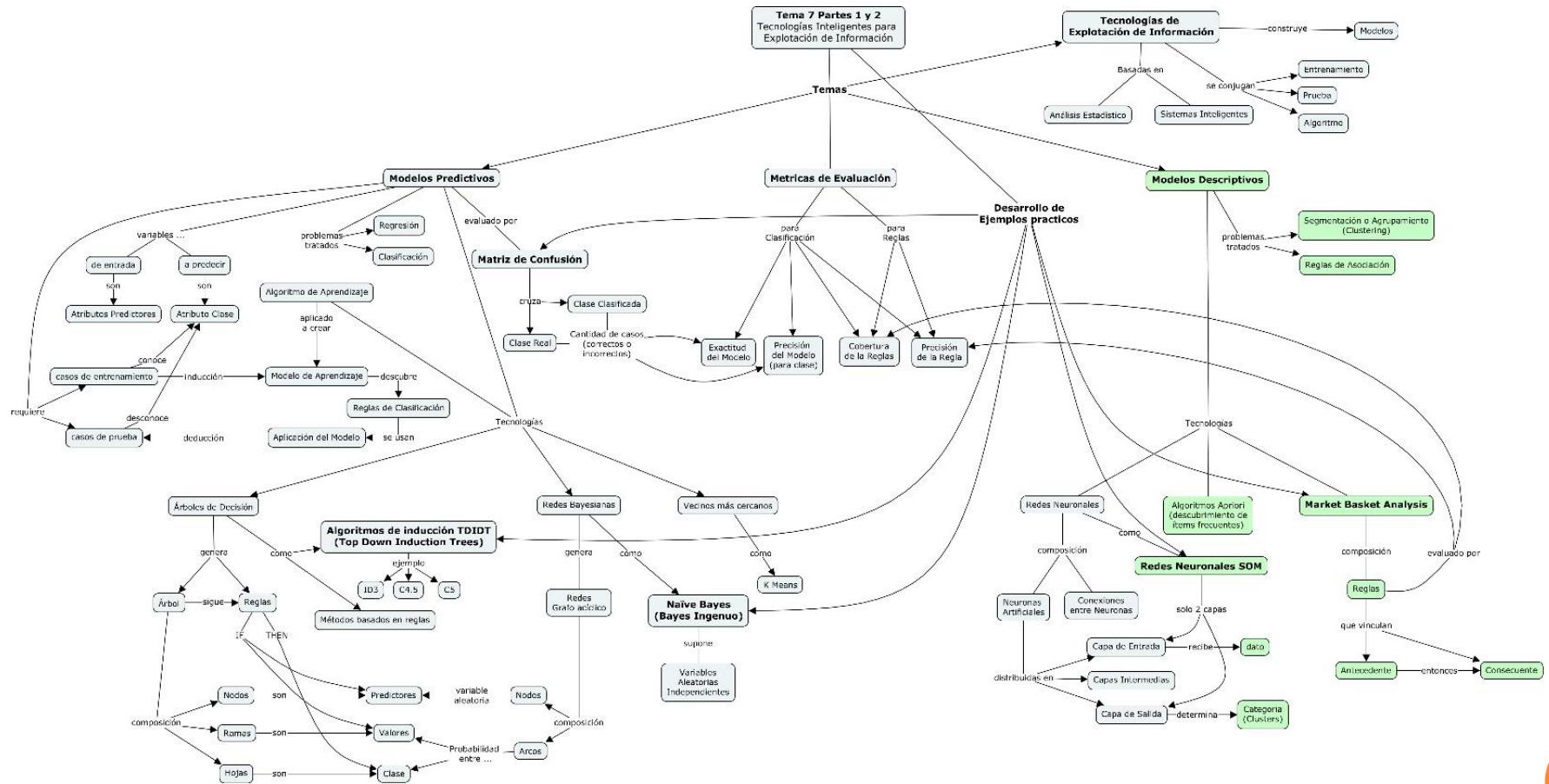
MATERIAL COMPLEMENTARIO

Fuentes:

- **Canal Electronics and Technology**
 - **Lista Machine Learning:** [Machine Learning - YouTube](#)
 - Modelo Naive Bayes: [Modelo Naive Bayes - Machine Learning | aprendizaje automático \(youtube.com\)](#)
 - Evaluación de Modelos: [Evaluación del Modelo - Machine Learning | aprendizaje automático \(youtube.com\)](#)



RESUMEN CLASE





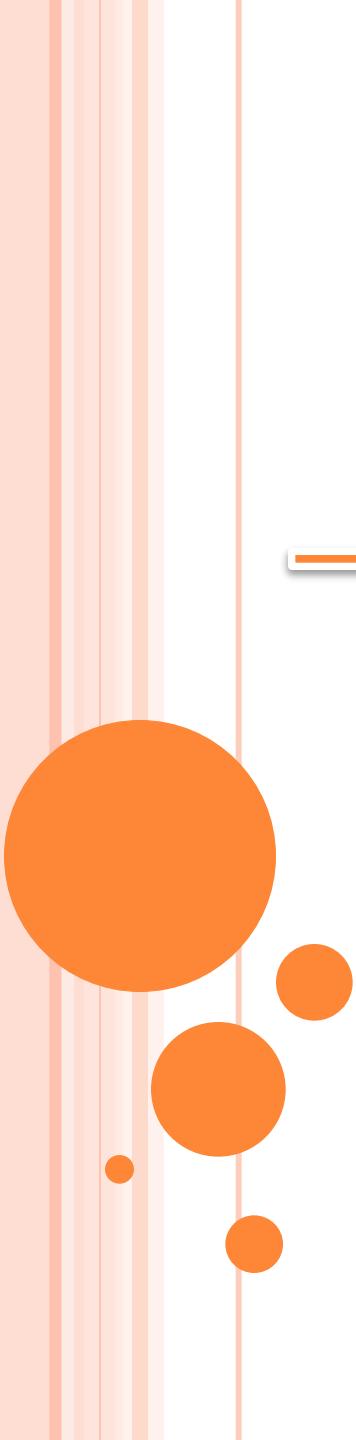
TRABAJO PRÁCTICO DE MINERÍA DE DATOS

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.





UNIVERSIDAD NACIONAL DE LA MATANZA

INTELIGENCIA DE NEGOCIOS

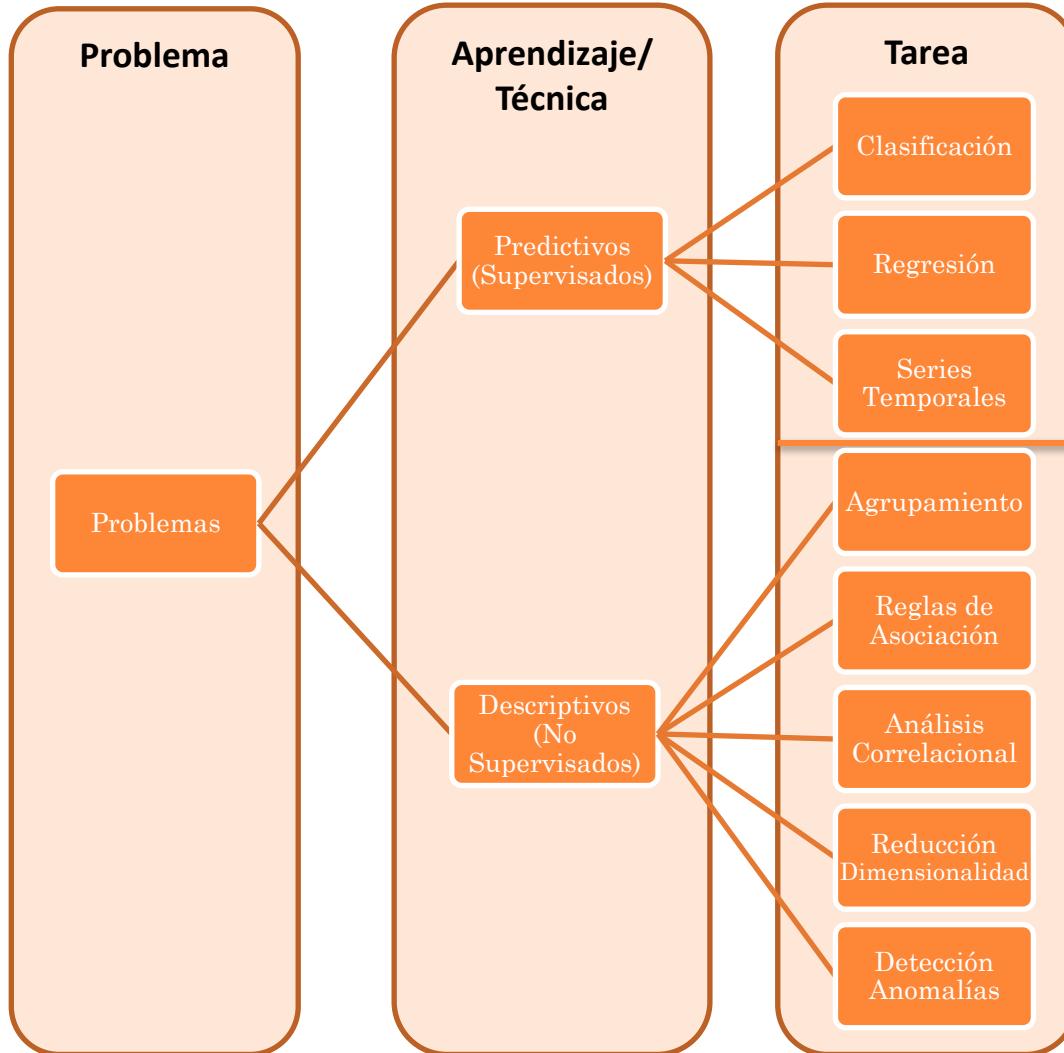
**Tecnologías Inteligentes
para Explotación de Información**

Docentes: ING. LORENA R. MATTEO

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



TÉCNICAS DE MINERÍA DE DATOS



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

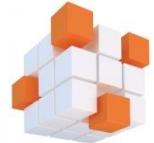


Esta foto de Autor desconocido está bajo licencia CC BY-SA-NC



MODELOS DESCRIPTIVOS

- Tareas de Segmentación (Clustering)
 - Agrupamiento jerárquico o no jerárquico de datos de acuerdo a un determinado criterio.
 - Jerárquico: Puede ser aglomerativo o divisivo.
 - No Jerárquico: N° Grupos determinados de antemano.
- 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}



MODELOS DESCRIPTIVOS

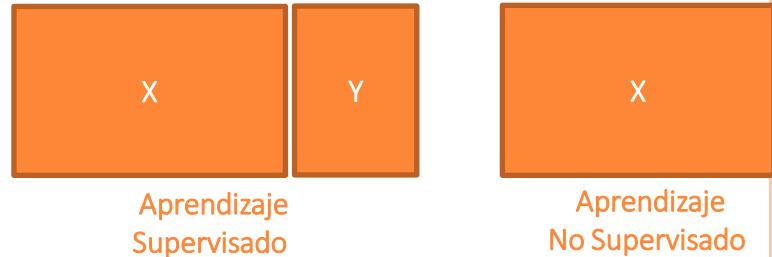
Supervisado vs No Supervisado:

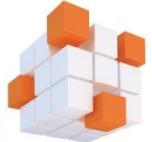
○ Aprendizaje Supervisado:

- X (variables de entrada) e Y (variable objetivo): conocidos
- variable objetivo y sus valores específicos sirven para entrenar modelos y predecir nuevo caso

○ Aprendizaje No Supervisado:

- Solo conozco X (variables de entrada), el conjunto de características del caso no existe una variable objetivo predefinida, podría descubrirse dentro de los ejemplos
- Se busca describir los datos, descubriendo agrupaciones naturales presentes en las variables de entrada.





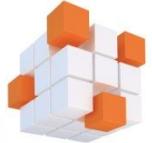
TAREAS DE SEGMENTACIÓN (CLUSTERING)

- ¿Cuándo usar esta tarea?
 - cuando sea necesario dividir los datos en grupos significativos y/o útiles, sin perder la estructura natural de los datos
 - muchas veces es sólo un punto de partida
- ¿Para qué usar esta tarea? Para encontrar....
 - agrupamientos naturales y describir sus propiedades
 - agrupamientos útiles y descubrir posible clase
 - correlación entre las variables
 - representantes para grupos homogéneos, reducción de dim.
 - objetos inusuales (outliers)
 - perturbaciones aleatorias de los datos (noise)



TAREAS DE SEGMENTACIÓN (CLUSTERING)

- ¿Qué aplicaciones típicas tiene esta tarea?
 - **Marketing e Investigación de Mercado:** descubrimiento de clientes con hábitos de compra similares, esquemas de comportamiento, se agrupan → dirigir campañas: ofertas, publicidades, fidelidad, etc.
 - **Seguros:** identificación de grupos de asegurados con características parecidas (siniestros, posesiones, etc.) → ofrecer productos que otros clientes de ese grupo ya poseen.
 - **Fraudes:** detectar puntos fuera del común de los grupos identificados.
 - **Planificación urbana:** identificación de grupos de viviendas de acuerdo a tipo, valor o situación geográfica.
 - **Web Mining:** clasificación de documentos, análisis de logs para descubrir patrones de acceso similares.
 - **Procesamiento de imágenes:** reconocimiento de patrones, búsqueda de objetos en imágenes, áreas con cierta característica de tierra (GIS), ...
 - **Preprocesamiento:** para otras técnicas de DM (maldición de la dimensionalidad, correlación de variables, etc.)



TAREAS DE SEGMENTACIÓN (CLUSTERING)

○ ¿En qué consiste esta tarea?

- Dado un conjunto de datos (puntos heterogéneos) se busca encontrar un número (k) de subgrupos homogéneos (clúster) con características similares, donde los miembros sean:
 - similares a cualquier otro incluido en el mismo grupo (están cerca y relacionados)
 - distintos a los incluidos en otros grupos (están alejados)
- La similitud se define utilizando una medida de distancia:
 - Euclidiana, Minkowski, Jaccard, Coseno, etc.
- En 2 dimensiones parece una tarea sencilla, pero en la realidad estos puntos están en un espacio de alta dimensionalidad.

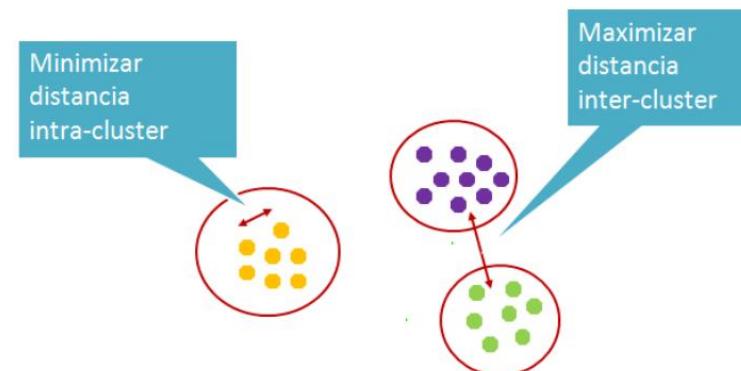


TAREAS DE SEGMENTACIÓN (CLUSTERING)

○ ¿Qué debe tenerse en cuenta?

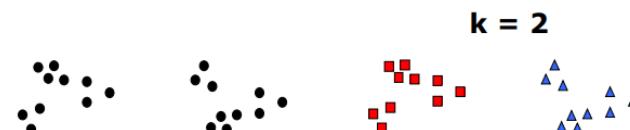
- Un buen método de clustering debe identificar clústeres que sean tanto compactos como separados entre sí:

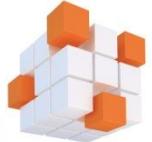
- Alta similaridad intra-cluster
- Baja similaridad inter-cluster



- La decisión del **número de clústeres (k)** es uno de los retos en agrupamiento, existen distintos métodos para su elección:

- Método del Codo
- Conocimiento del campo
- Decisión de negocios

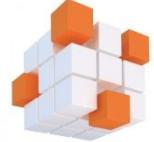




TAREAS DE SEGMENTACIÓN (CLUSTERING)

○ ¿Qué se espera de este método?

- Un buen método de agrupamiento debiera:
 - descubrir algunos o todos los patrones ocultos en los datos
 - ser fácil de interpretar
 - ser escalable
 - ser insensible al orden de los registros de entrada
 - ser válido para registros de alta dimensionalidad
 - brindar la capacidad de incorporar restricciones del usuario
 - permitir dar peso a ciertas variables dependiendo de distintos criterios (relativos a su aplicación, ...)
- La calidad del método de clustering depende de la medida de similitud y de su implementación
- Las funciones de distancia son muy sensibles al tipo de variables.



TAREAS DE SEGMENTACIÓN (CLUSTERING)

- ¿Cómo medir la similitud entre instancias?
 - Las medidas de distancia/similitud dependen del tipo de variable y son no negativas

Variables Numéricas

Distancia Euclídea

Distancia de Manhattan

Distancia Minkowski

Variables Binarias

Coeficiente de Jaccard

Variables Categóricas

Distancia: 1 si los valores son diferentes, 0 si son iguales

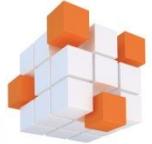


Normalizar/Estandarizar

- z-score
- min-max
- decimal scale

¿Tienen todos los atributos la misma importancia?

- Si no tienen igual importancia, será necesario ponderar los atributos



TIPOS DE SEGMENTACIÓN (CLUSTERING)

○ No Jerárquico (Particional)

Divide los datos en subconjuntos (clústeres) sin solapamiento, \Leftrightarrow cada dato está en un solo subconjunto

- nro de grupos (k) determinado de antemano.
- cada punto pertenece al grupo de "más cercano"
- *Global Optimal, K-Means, K-Modes, K-Medoids (PAM)*

○ Jerárquico

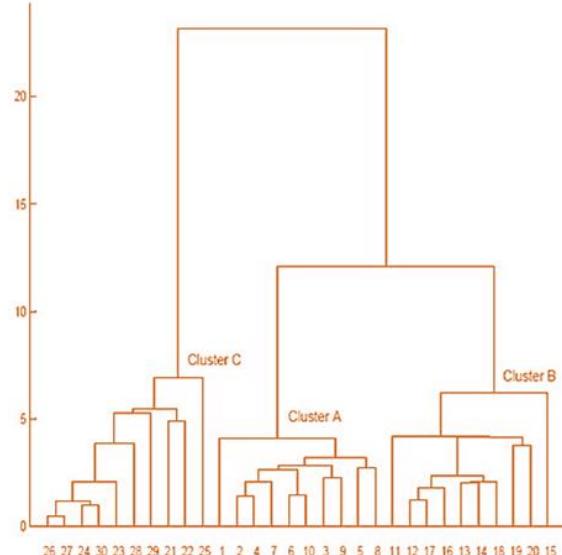
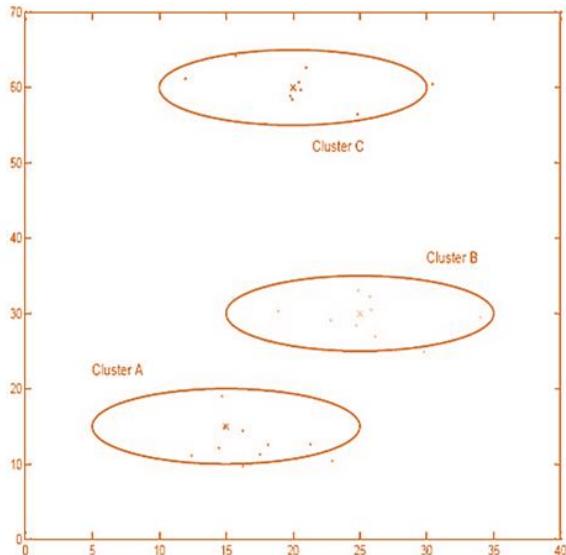
Construye un árbol binario o dendograma a partir de un conjunto de ejemplos

- **Aglomerativo** (bottom-up - de abajo hacia arriba)
 - Inicialmente, cada punto es un clúster
 - Combina repetidamente los dos grupos "más cercanos" en uno, para formar clústeres mayores
 - *AGNES*
- **Divisivo** (top-down – de arriba hacia abajo)
 - Comienza con un clúster y de forma recursiva lo va dividiendo.
 - *DIANA*

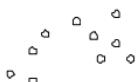


TIPOS DE SEGMENTACIÓN (CLUSTERING)

○ No Jerárquico (Particional) vs Jerárquico

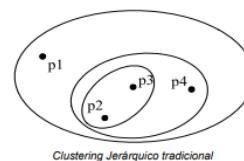


Clustering particional

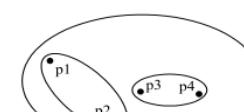


Clustering Particional

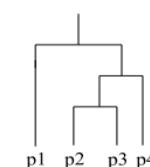
Clustering jerárquico



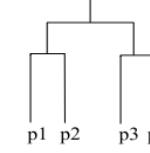
Clustering Jerárquico tradicional



Clustering Jerárquico no tradicional



Dendograma tradicional



Dendograma no tradicional



TIPOS DE SEGMENTACIÓN (CLUSTERING)

○ Basados en densidad

Utilizan funciones de conectividad y densidad

- *DBSCAN, OPTICS, DENCLUE, CLIQUE*

○ Basados en rejillas (Grilla)

Utilizan una estructura de granularidad de múltiples niveles

- *STING, BAN-Clustering/GGRIDCLUS, Wave-Clustering, CLIQUE*

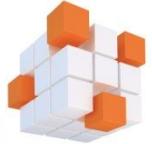
○ Basados en datos categóricos

- *ROCK*

○ Basados en modelos

Suponen un modelo para cada uno de los clústeres con el propósito de encontrar el modelo que mejor ajuste.

- *Redes Neuronales: SOM: Self Organizing Maps*
- *Machine Learning: COBWEB (clustering conceptual)*
- *Estadísticos: Gaussian Mixture Model, Autoclass (Bayesiano)*
- *Lógica Difusa: Fuzzy C-Means*



MODELOS DESCRIPTIVOS - EN RESUMEN...

- No realizan predicciones.
- No hay una variable objetivo identificada en el conjunto de datos. En su lugar, se buscan patrones y estructuras entre todas las variables
- Analizan otros aspectos de los datos.
- Problemas tratados que veremos:
 - Segmentación o Agrupamiento (Clustering)
 - Reglas de Asociación
- Tecnologías utilizadas que veremos:
 - K-Means
 - Redes Neuronales SOM
 - Algoritmo Apriori (descubrimiento de ítems frecuentes)



TAREAS DE SEGMENTACIÓN K-MEANS

¿De qué se trata?

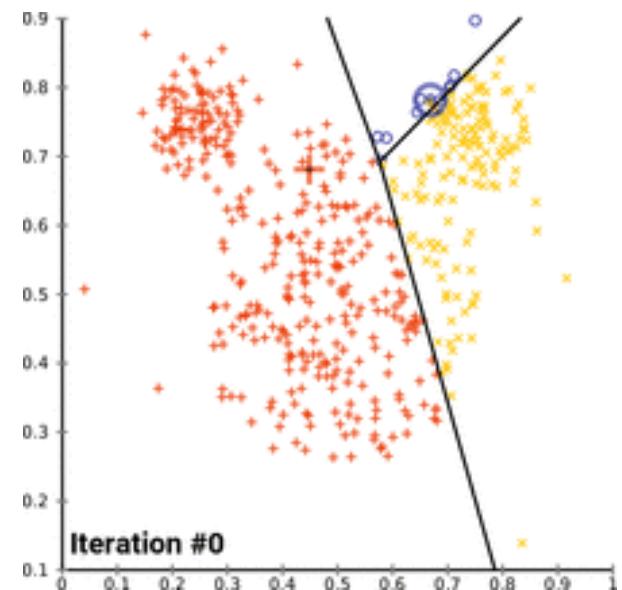
- K-Means o K-Medias es un algoritmo de clustering no supervisado que agrupa datos similares en K clústeres para descubrir patrones ocultos.

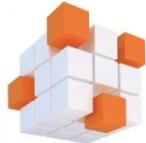
Algoritmo K-Means básico

- 1: **Select k** puntos/objetos como el centroide inicial
- 2: **repeat**
- 3: **From k** clústeres asignando todos los puntos a centroide más cercano.
- 4: Recalcular el centroide de cada clúster
- 5: **until** los centroides no cambien.

Centroides: Puntos centrales de cada clúster.

Clústeres: Grupos de datos alrededor de los centroides.





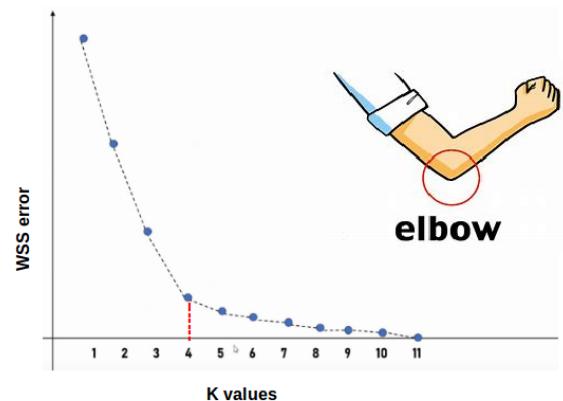
TAREAS DE SEGMENTACIÓN K-MEANS

¿Cómo seleccionar el número correcto de clústeres k?

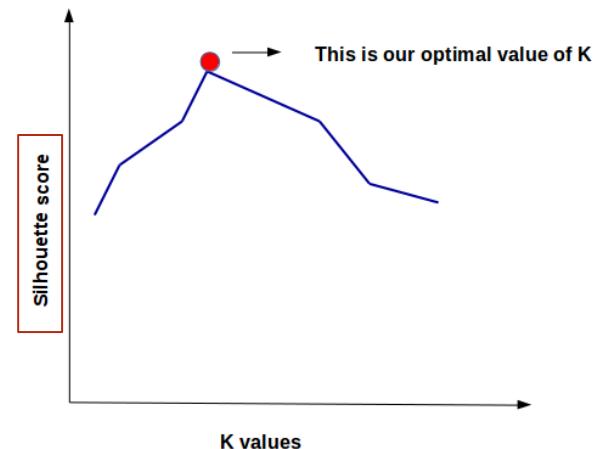
El algoritmo de K-Means es muy sensible a la posición inicial de los centroides de los grupos, su elección requiere un delicado equilibrio. **Métodos para determinar el número óptimo de clústeres:**

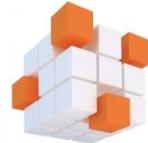
- **Método del Codo:** Se calcula la suma de las distancias al cuadrado dentro de los clústeres (WCSS) para $k=x$, se grafica y busca donde se forma el "codo".
- **Método de la Silueta:** Se la calcula la medida de la silueta para evaluar la calidad de la agrupación y seleccionar el K que maximice esta medida.
- **Conocimiento del Dominio:** Utilizar el conocimiento previo del problema para establecer un rango probable para K y combinarlo con métodos cuantitativos.

Elbow method



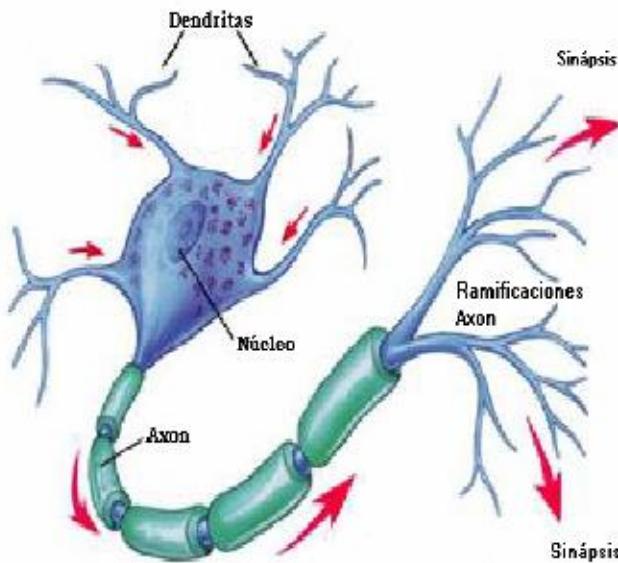
elbow





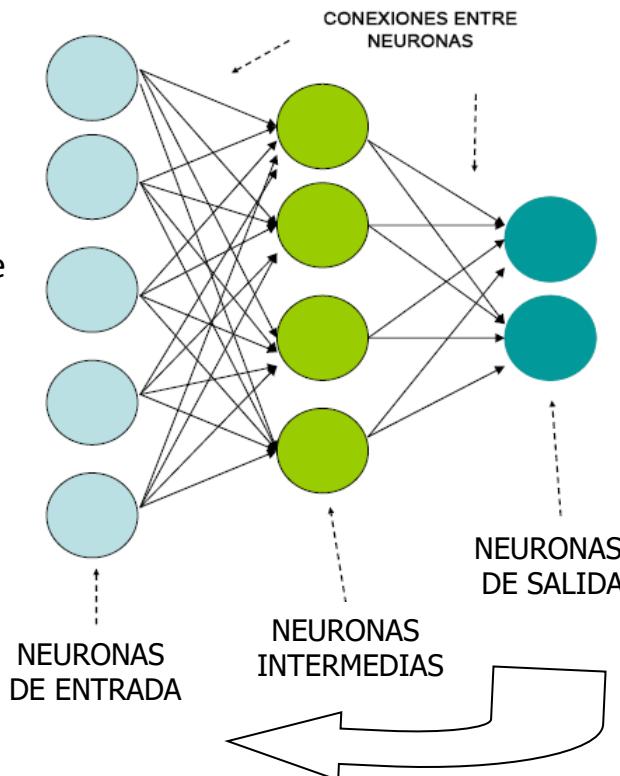
SEGMENTACIÓN REDES NEURONALES

Neurona Biológica



La neurona es la unidad fundamental del sistema nervioso y en particular del cerebro. Cada neurona es una simple unidad procesadora que recibe y combina señales desde y hacia otras neuronas. Si la combinación de entradas es suficientemente fuerte la salida de la neurona se activa.

Neurona Artificial

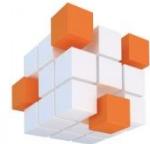


Casos de entrenamiento con sus respuestas
Casos de Prueba

Respuesta
Obtenida

vs.
Respuesta
Deseada

Corrección de
Factores de
Ponderación



REDES NEURONALES

Neuronas Biológicas vs. Artificiales

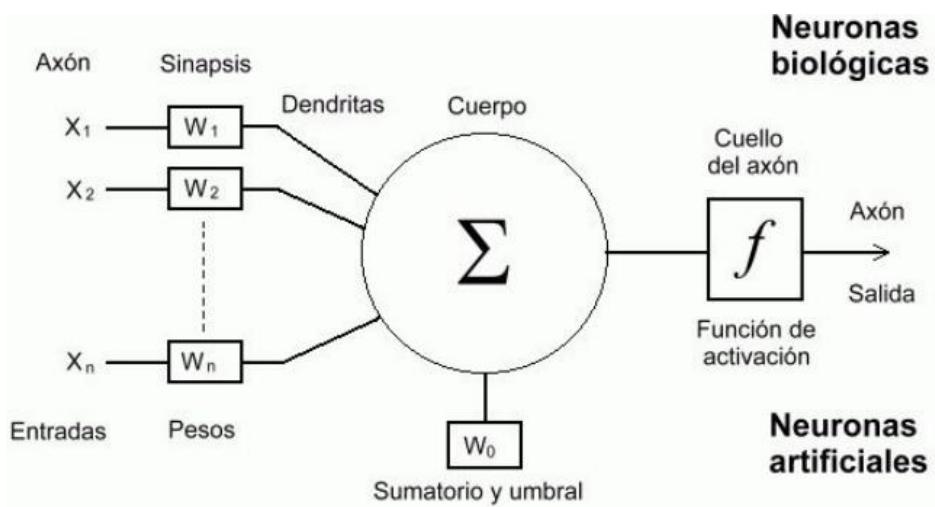
En 1956, se acuñaría el propio término “Inteligencia Artificial” por John McCarthy, Marvin Minsky y Claude Shannon en una conferencia en Dartmouth.

En 1958, Frank Rosenblatt diseña la primera red neuronal artificial, el **Perceptrón**.

Modelo neuronal con n entradas, que consta de:

- Un conjunto de **entradas** x_1, \dots, x_n
- Los **pesos sinápticos** w_1, \dots, w_n , correspondientes a cada entrada
- Una **función de agregación**, Σ
- Una **función de activación**, f
- Una **salida**:

$$Y = f \left(\sum_{i=0}^n w_i x_i \right)$$





SEGMENTACIÓN REDES NEURONALES SOM

○ Mapas Auto-Organizados de Kohonen (SOM)

¿De qué se trata?

- Los SOM son un tipo de red neuronal **no supervisada** desarrollada por Teuvo Kohonen, utilizada para **reducción de dimensionalidad, visualización de datos y segmentación**.

¿Cuál es su propósito?

- Transformar datos de alta dimensión en un mapa de baja dimensión (generalmente 2D) que preserva la topología de los datos originales. Frecuentemente los humanos aprendemos de manera no supervisada.

1. Visualización y Exploración de Datos:

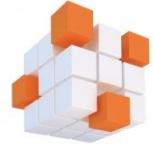
• **Reducción de Dimensionalidad:** Los SOM transforman datos de alta dimensionalidad en un mapa bidimensional, lo que facilita la visualización y el análisis.

• **Preservación Topológica:** Mantienen las relaciones topológicas entre los datos, lo que significa que los datos similares en el espacio original permanecen juntos en el mapa autoorganizado resultante.

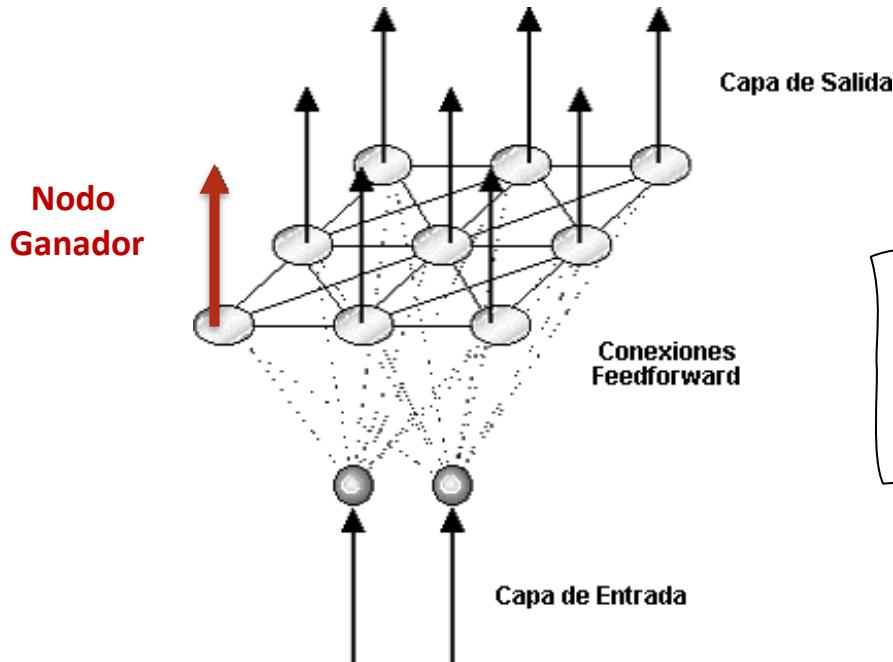
2. Agrupación Natural:

• **Clústeres Naturales:** Durante el entrenamiento, las neuronas de la red SOM se organizan de manera que datos similares se mapean cerca unos de otros en el mapa, formando clústeres naturales.

• **Identificación de Patrones:** Esto permite identificar patrones y segmentos en los datos de manera intuitiva y visual.



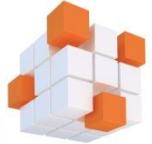
REDES NEURONALES SOM



Neurona: Cada nodo en el SOM es una neurona con un vector de pesos.

Vecindad: Neuronas cercanas en la cuadrícula tienen pesos similares.

- Cada vez que se presenta un **registro de entrada**, las neuronas **“compiten”** y **una** se define como la **ganadora**.
- Si se presenta un registro de **entrada parecido al anterior**, es muy posible que el **ganador** sea el **mismo nodo de salida**.

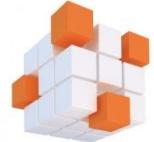


REDES NEURONALES SOM

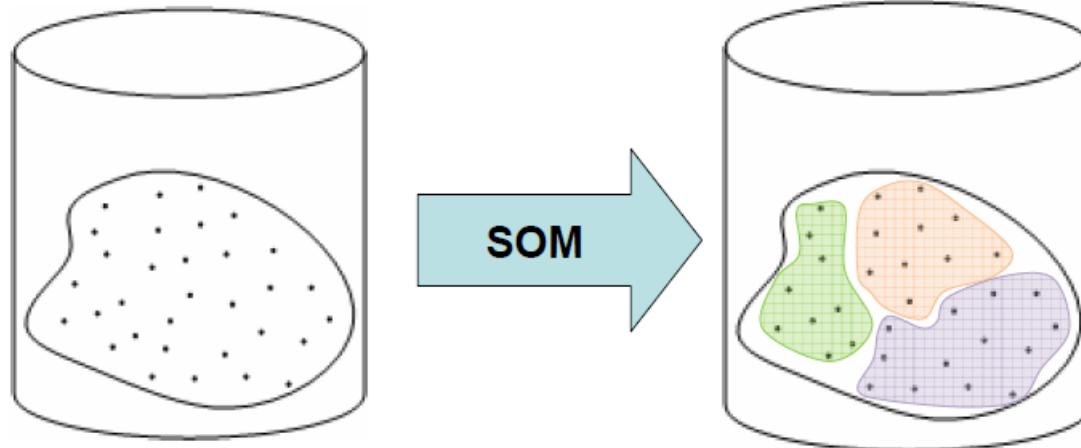
Proceso de Entrenamiento No Supervisado:

1. **Inicialización:** Los pesos de las neuronas se configuran aleatoriamente.
2. **Selección de Muestra:** Se elige una instancia del conjunto de datos de entrada.
3. **Competencia:** Las neuronas "compiten" para ver cuál se parece más a la instancia. La neurona con la distancia (en gral, euclídea) más pequeña (la más cercana) es la "neurona ganadora" o "Best Matching Unit (BMU)".
4. **Cooperación y Actualización de Pesos:** La neurona ganadora y sus vecinas cercanas ajustan sus pesos para parecerse más al dato de entrada (instancia).
5. **Reducción de Vecindad:** A medida que avanza el entrenamiento, el tamaño de esta vecindad se reduce, lo que permite un ajuste más preciso y localizado.

Repetición: Este proceso se repite n veces con todos los datos de entrada hasta que los pesos se estabilizan y el mapa se organiza de manera útil.



PROCESO INTUITIVO DE AGRUPAMIENTO



- Los **registros semejantes** van a parar a la **misma categoría** (clúster).
- Una vez **entrenado**, el mapa de Kohonen se puede usar para categorizar nuevos registros.
 - El resultado final es la creación del llamado **mapa autoorganizado** donde se representan los **rasgos más sobresalientes** del espacio de entrada.



MEDIDAS DE EVALUACIÓN DE SEGMENTACIÓN

¿Es necesario validar los clústeres?

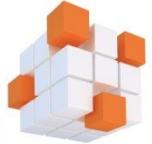
- En Clasificación, la validación es parte integral del proceso y tenemos medidas claras para ello: exactitud, precisión, etc.
- No así en Clustering....

¿Cómo saber si nuestros clústeres son buenos?

- No hay una respuesta absoluta
- Depende de la aplicación

Entonces, ¿para qué evaluar? Para...

- evitar encontrar patrones en el ruido
- comparar algoritmos de clustering diferentes
- comparar conjuntos de clústeres diferentes
- comprar dos clústeres



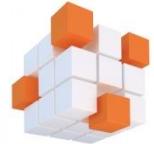
MEDIDAS DE EVALUACIÓN DE SEGMENTACIÓN

¿Qué aspectos se consideran en la validación?

- Determinar la **tendencia de agrupamiento** (clustering tendency), *por ej. si existe una estructura no-aleatoria en los datos*
- Encontrar el **número correcto de clústeres** (k)
- Evaluar qué tan bien los **resultados se ajustan a los datos** (**sin consultar datos externos**)
- Comparar **resultados con resultados externos**, *por ej. clases asignadas manualmente*
- Comparar **dos conjuntos de clústeres** para saber cuál es mejor

¿Qué tipos de validación se puede emplear?

- Métricas de Validación Interna (Internal Index)
- Métricas de Validación Externa (External Index)
- Métricas de Validación Relativa (Relative Index)
- Otros: Validación usando correlación, Validación con Expertos, Enfoque visual.



MEDIDAS DE EVALUACIÓN DE SEGMENTACIÓN

○ Métricas de Validación Interna

Evalúan que tan buena es la estructura del clustering sin necesidad de información ajena al propio algoritmo y su resultado, únicamente basadas en información de los datos.

- **Cohesión:** mide qué tan cercanos son los objetos en un clúster. *Por ej.: Sum of Squared Within (SSW/SSE)*
- **Separación:** mide qué tan diferente o bien separado es un clúster de otros. Hay varios enfoques para medir esta distancia entre clústeres: distancia entre el miembro más cercano, distancia entre los miembros más distantes o la distancia entre los centroides. *Por ej.: Sum of Squared Between (SSB)*
- **Coeficiente de Silhouette:** es una métrica para evaluar la calidad del agrupamiento obtenido con algoritmos de clustering, su objetivo es identificar cuál es el número óptimo de agrupamientos.



MEDIDAS DE EVALUACIÓN DE SEGMENTACIÓN

○ Métricas de Validación Externa

Una vez finalizado el algoritmo de agrupación, se compara el clúster en el que fue asignado cada elemento, con la etiqueta de clase que traía de antemano (información externa, no siempre disponible).

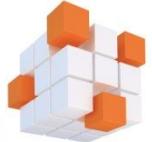
- **Pureza**: nivel en que un clúster contiene elementos de una sola clase (se usa la clase predominante)
- **Entropía**: cantidad de clases diferentes que contiene un clúster

○ Métricas de Validación Relativa

Se utiliza para comparar dos agrupaciones diferentes o clúster. A menudo se utiliza un índice externo o interno para esta función, *por ejemplo, SSW/SSE o Entropía*

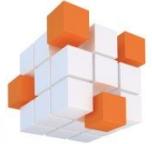
○ Validación con Expertos

Se pueden evaluar los clústeres para ver si producen el resultado esperado y comparar con otras soluciones • Se puede generar una clasificación de validación.



SEGMENTACIÓN (CLUSTERING) - CONCLUSIONES

- Aprendizaje No Supervisado
- No realizan predicciones
- Admite muchos enfoques
- La calidad del agrupamiento depende de la medida de similitud utilizada por el método y de su implementación.
- La calidad de un método de agrupamiento también se mide por su capacidad para descubrir algunos o todos los patrones ocultos.
- La evaluación objetiva es problemática: generalmente realizado por inspección humana / experta.



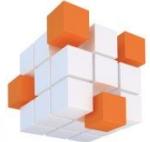
TAREAS DE ASOCIACIÓN

¿En qué consiste esta tarea?

- Descubre por medio de reglas de asociación hechos que ocurren en común dentro de un determinado conjunto de datos:
 - patrones frecuentes, asociaciones, correlaciones o estructuras causales entre conjuntos de artículos u objetos
 - secuencias o patrones temporales

¿Qué aplicaciones típicas tiene esta tarea?

- Análisis de Canasta (Market Basket Analysis)
 - *Análisis de clientes:* Se utiliza información de las compras de un cliente para ofrecer una aproximación sobre quién es y porqué hace ciertas compras (comportamiento)
 - *Análisis de productos:* Aporta información sobre qué productos tienden a ser comprados juntos.
- Diseño de Catálogos



TAREAS DE ASOCIACIÓN

Compra: zumo de naranja, plátanos, detergente para vajillas, limpia cristales, gaseosa, ...

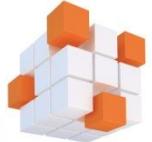
¿Cómo afecta la demografía de la vecindad a la compra de los clientes?



¿Es típico comprar gaseosa y plátanos? ¿Es importante la marca de la gaseosa?

¿Dónde deberían colocarse los detergentes para maximizar sus ventas?

¿Aumenta la compra del limpia cristales cuando se compran a la vez detergente para vajillas y zumo de naranja?



REGLAS DE ASOCIACIÓN

- Dado un conjunto de transacciones se quiere encontrar reglas que puedan predecir la ocurrencia de un ítem a partir de otros ítems de la transacción , con un mínimo de confianza y soporte..
- Análisis de Canasta (**Market Basket Analysis**)

Ticket ID	Ítems
1	Pan, Leche
2	Pan, Pañales, Cerveza, Huevos
3	Leche, Pañales, Cerveza, Gaseosa
4	Pan, Leche, Pañales, Cerveza
5	Pan, Leche, Pañales, Gaseosa

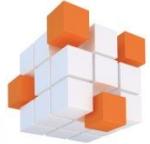
Ejemplos de Reglas

$\{Pañales\} \rightarrow \{Cerveza\}$

$\{Cerveza, Pan\} \rightarrow \{Leche\}$

$Soporte \geq minSupp$
 $Confianza \geq minConf$

- Reglas de Asociación $X \rightarrow Y$ representan implicancias.
 - Antecedente \rightarrow Consecuente



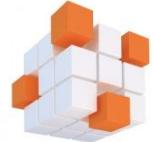
REGLAS DE ASOCIACIÓN

Métricas de Evaluación

- **Cobertura de la Regla (Soporte/Support)**: Proporción de casos a los que se le puede aplicar cada regla. Es decir, el número de instancia que la regla predice correctamente.

$$\text{COBER}(R) = \frac{\text{N}^{\circ}\text{casos que satisfacen la aplicación de la regla R}}{\text{N}^{\circ}\text{casos totales de la clase}}$$

- N.^º casos R: transacciones que contienen todos los ítems del antecedente o del consecuente. Toma valores entre 0 y 1.
- Reglas con mayor cobertura:
 - Representativas y útiles para obtener características que definen el comportamiento de una clase.
 - Credibilidad e interés del modelo para clasificar nuevos casos a una clase.
- Aplicable también a **modelos de clasificación**.



REGLAS DE ASOCIACIÓN – EJEMPLO COBERTURA

○ Ítems de compras:

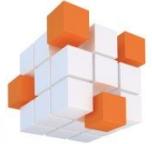
Ticket ID	Ítems
1	Pan, Leche
2	Pan, Pañales, Cerveza, Huevos
3	Leche, Pañales, Cerveza, Gaseosa
4	Pan, Leche, Pañales, Cerveza
5	Pan, Leche, Pañales, Gaseosa

• *Numerador* = Cobertura
(Cant. Casos R)
• *Denominador* = Cant. Casos
Totales

Min. soporte 30%
Min. confianza 50%

○ Consideremos las siguientes reglas:

- $\{Leche, Pañales\} \rightarrow \{Cerveza\}$ Cober(R) = $2/5 = 0.4$ ($2=$ Tickets 3y4)
- $\{Leche, Cerveza\} \rightarrow \{Pañales\}$ Cober(R) = $2/5 = 0.4$ ($2=$ Tickets 3y4)
- $\{Pañales, Cerveza\} \rightarrow \{Leche\}$ Cober(R) = $2/5 = 0.4$ ($2=$ Tickets 3y4)
- $\{Cerveza\} \rightarrow \{Leche, Pañales\}$ Cober(R) = $2/5 = 0.4$ ($2=$ Tickets 3y4)
- $\{Pan\} \rightarrow \{Pañales, Gaseosa\}$ Cober(R) = $1/5 = 0.2$ ($1=$ Ticket 5)

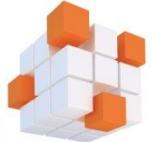


REGLAS DE ASOCIACIÓN

- **Precisión de la Regla (Confianza/Confidence):** Proporción de casos que cumplen con la regla respecto del total de casos considerados en la precondición de la misma.

$$\text{Prec}(R) = \frac{\text{Nºcasos que satisfacen la aplicación de la regla R}}{\text{Nºcasos que satisfacen la precondición}}$$

- una transacción que contiene el antecedente y también el consecuente. Toma valores entre 0 y 1.
- Cuanto mayor sea la precisión de una regla más confiable e interesante resulta para asociar ítems y descubrir nuevo conocimiento.
- Aplicable también a **modelos de clasificación**.



REGLAS DE ASOCIACIÓN – EJEMPLO PRECISIÓN

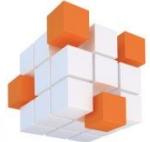
- Utilizando los mismos ítems de compra:

Ticket ID	Ítems
1	Pan, Leche
2	Pan, Pañales, Cerveza, Huevos
3	Leche, Pañales, Cerveza, Gaseosa
4	Pan, Leche, Pañales, Cerveza
5	Pan, Leche, Pañales, Gaseosa

- *Numerador* =Idem Cobertura (Cant. Casos R)
- *Denominador* =Casos PreCondición

- Considerando las reglas anteriores:

- $\{Leche, Pañales\} \rightarrow \{Cerveza\}$ Prec(R) = $2/3 = 0.67$ ($3=$ Tickets 3,4,5)
- $\{Leche, Cerveza\} \rightarrow \{Pañales\}$ Prec(R) = $2/2 = 1$ ($2=$ Tickets 3,4)
- $\{Pañales, Cerveza\} \rightarrow \{Leche\}$ Prec(R) = $2/3 = 0.67$ ($3=$ Tickets 2,3,4)
- $\{Cerveza\} \rightarrow \{Leche, Pañales\}$ Prec(R) = $2/3 = 0.67$ ($3=$ Tickets 2,3,4)
- $\{Pan\} \rightarrow \{Pañales, Gaseosa\}$ Prec(R) = $1/4 = 0.25$ ($4=$ Tickets 1,2,4,5)



REGLAS DE ASOCIACIÓN

- Conclusiones obtenidas:

- $\{\text{Leche, Pañales}\} \rightarrow \{\text{Cerveza}\}$ $\text{Cober}(R) = 0.4$ $\text{Prec}(R) = 0.67$
- $\{\text{Leche, Cerveza}\} \rightarrow \{\text{Pañales}\}$ $\text{Cober}(R) = 0.4$ $\text{Prec}(R) = 1$
- $\{\text{Pañales, Cerveza}\} \rightarrow \{\text{Leche}\}$ $\text{Cober}(R) = 0.4$ $\text{Prec}(R) = 0.67$
- $\{\text{Cerveza}\} \rightarrow \{\text{Leche, Pañales}\}$ $\text{Cober}(R) = 0.4$ $\text{Prec}(R) = 0.67$
- $\{\text{Pan}\} \rightarrow \{\text{Pañales, Gaseosa}\}$ $\text{Cober}(R) = 0.2$ $\text{Prec}(R) = 0.25$

- La regla $\{\text{Leche, Cerveza}\} \rightarrow \{\text{Pañales}\}$ es la que mejor describe características de consumo, y la que genera mayor confianza e interés para descubrir patrones de compra frecuente.
- Reglas con alta precisión, pero baja cobertura son irrelevantes y de poco interés en un modelo de explotación de información.



REGLAS DE ASOCIACIÓN – PROCESO DE EXTRACCIÓN DE REGLAS

Encontrar todas las reglas $X \& Y \rightarrow Z$ con un mínimo de confianza y soporte.

- Cobertura (Soporte) (s): probabilidad de que una transacción contenga {X & Y & Z}
- Precisión (Confianza) (c): probabilidad condicional $P(Z|X\&Y)$

# transacción	artículos
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

*Sea el valor mínimo para confianza
y soporte 50%,
 $A \Rightarrow C$ (50%, 66.6%)
 $C \Rightarrow A$ (50%, 100%)*



REGLAS DE ASOCIACIÓN – PROCESO DE EXTRACCIÓN DE REGLAS

Al tratar con grandes volúmenes de datos, el proceso se descompone en dos pasos:

- Encontrar conjuntos de artículos frecuentes
 - Mayor ocurrencia que el soporte mínimo fijado.
- Generar reglas de asociación “fuerte” a partir de los conjuntos de artículos frecuentes
 - Deben satisfacer el mínimo fijado tanto para soporte como para confianza.

REGLAS DE ASOCIACIÓN – PROCESO DE EXTRACCIÓN DE REGLAS –

EJEMPLO 1



Encontrar todas las reglas $X \& Y \rightarrow Z$ con un mínimo de confianza y soporte.

- **Cobertura (Soporte) (s)**: probabilidad de que una transacción contenga $\{X \& Y \& Z\}$
- **Precisión (Confianza) (c)**: probabilidad condicional $P(Z|X\&Y)$

EJEMPLO 1

# transacción	artículos
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Frequent Itemset	Cobertura (Soporte)	Cálculo
{A}	75%	=3/4
{B}	50%	=2/4
{C}	50%	=2/4
{A,C}	50%	=2/4

Min. soporte 50%
Min. confianza 50%

	Cobertura ($\{ A \& C \}$)	Cálculo	Confianza (Precisión) = $Cobertura(\{ A \& C \}) / Cobertura(\{ A \})$	Cálculo
Para la regla $A \rightarrow C$	50%	=2/4	66,67%	$=(2/4) / (3/4) = 0,5 / 0,75 = 0,666666667$ o bien $=2/3$
Para la regla $C \rightarrow A$	50%	=2/4	100%	$=(2/4) / (2/4) = 0,5 / 0,5 = 1$ o bien $=2/2$

REGLAS DE ASOCIACIÓN – PROCESO DE EXTRACCIÓN DE REGLAS –

EJEMPLO 2



Encontrar todas las reglas $X \& Y \rightarrow Z$ con un mínimo de confianza y soporte.

- **Cobertura (Soporte) (s)**: probabilidad de que una transacción contenga $\{X \& Y \& Z\}$
- **Precisión (Confianza) (c)**: probabilidad condicional $P(Z|X\&Y)$

EJEMPLO 2

Transaction	Items
t_1	Bread, Jelly, PeanutButter
t_2	Bread, PeanutButter
t_3	Bread, Milk, PeanutButter
t_4	Beer, Bread
t_5	Beer, Milk

$X \Rightarrow Y$	s	α
$Bread \Rightarrow PeanutButter$	60%	75%
$PeanutButter \Rightarrow Bread$	60%	100%
$Beer \Rightarrow Bread$	20%	50%
$PeanutButter \Rightarrow Jelly$	20%	33.3%
$Jelly \Rightarrow PeanutButter$	20%	100%
$Jelly \Rightarrow Milk$	0%	0%

$$I = \{\text{Beer, Bread, Jelly, Milk, PeanutButter}\}$$

Cobertura (Soporte) $\{Bread, PeanutButter\}$	Cálculo
60%	=3/5

Min. soporte 60%

Cobertura (Soporte)	Cálculo	$\text{Confianza}(X \rightarrow Y) = \frac{\text{Soporte}(X,Y)}{\text{Soporte}(X)}$	Cálculo
60%	=3/5	75%	$=(3/5) / (4/5) = 0,6/0,8 = 0,75 \text{ o bien } =3/4$
60%	=3/5	100%	$=(3/5) / (3/5) = 0,6/0,6 = 1 \text{ o bien } =3/3$
20%	=1/5	50%	$=(1/5) / (2/5) = 0,2/0,4 = 0,5 \text{ o bien } =1/2$
20%	=1/5	33,3%	$=(1/5) / (3/5) = 0,2/0,6 = 0,333 \text{ o bien } =1/3$
20%	=1/5	100%	$=(1/5) / (1/5) = 0,2/0,2 = 1 \text{ o bien } =1/1$
0%	=0/5	0%	$=(0/5) / (1/5) = 0/0 = 0 \text{ o bien } =0/1$



INTERPRETACIÓN DE LAS MÉTRICAS DE EVALUACIÓN

○ Regla con Soporte (Cobertura)

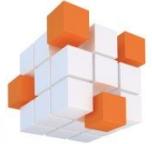
- **Alto** → Representativas y útiles para obtener características que definen el comportamiento de una clase. | Credibilidad e interés del modelo para clasificar nuevos casos a una clase.
- **Bajo** → puede haber aparecido por casualidad. | Poco interesante desde el punto de vista del negocio. | Sirve para eliminar reglas poco interesantes.

○ Regla con Confianza (Precisión)

- **Alta** → Cuanto mayor sea la precisión de una regla más confiable e interesante resulta para asociar ítems y descubrir nuevo conocimiento. | Mayor probabilidad de observar Y en transacciones que tengan X.
- **Baja** → Es probable que no exista relación entre antecedente y consecuente.

Valores típicos: Soporte = 2-10 % | Confianza = 70-90 %

El primer criterio de selección de reglas del algoritmo "A priori" es la **precisión o confianza**, dada por el porcentaje de veces que instancias que cumplen el antecedente cumplen el consecuente, pero el **segundo es el soporte**, dado por el número de instancias sobre las que es aplicable la regla. Si hay reglas de muy baja precisión (muy cerca del umbral mínimo definido) habría que considerarlas simplemente como ciertas tendencias.



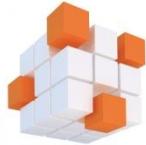
REFERENCIAS Y MATERIAL COMPLEMENTARIO

Referencias:

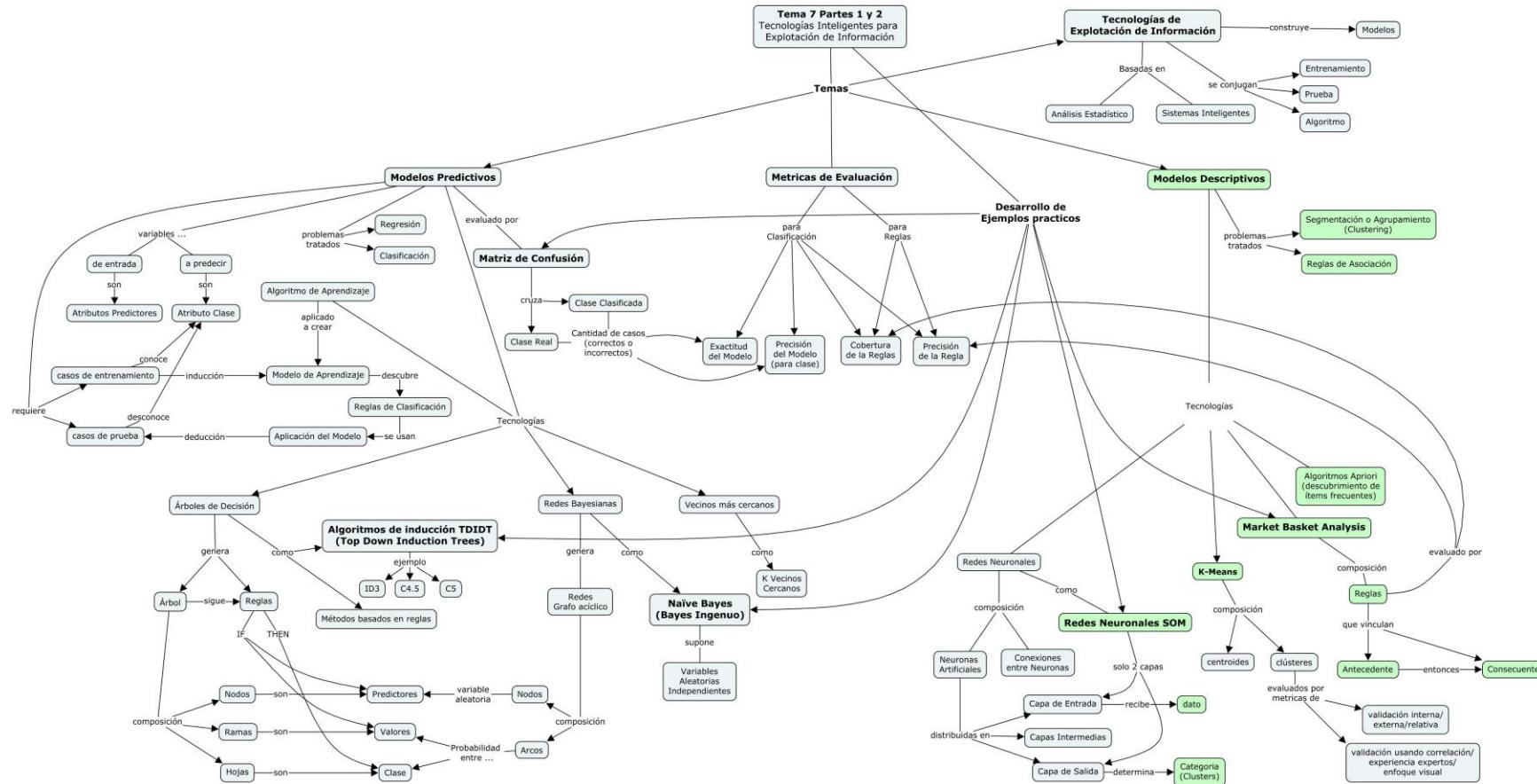
- Curso Inteligencia de Negocios, Universidad de Granada, España
- Curso DM Datos, Prof. F.Bravo/ H. Sarmiento / B. Poblete, Universidad de Chile
- Prof. Dra. S. Schiaffino (Univ.Centro BA-IA , Argentina)
- Curso Data Mining, Kent State University, USA
- Curso Minería de Datos, Universidad de Valencia, España
- Curso Mining Massive Datasets |Stanford University, USA
- Curso Advanced Machine Learning | Educación IT , Argentina

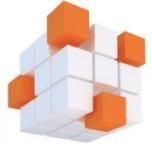
Material Complementario:

- https://www.cs.us.es/~fsancho/Blog/posts/Aprendizaje_Supervisado_No_Supervisado.md.html
- <https://www.cs.us.es/~fsancho/Blog/posts/Clustering/>
- https://dcain.etsin.upm.es/~carlos/bookAA/03.1_Clustering-K-Means.html
- K-MEANS Segmentación con KNIME : <https://www.youtube.com/@fbombab547>
https://www.youtube.com/watch?v=Uu7t90GX_u4
- KNIME TV - Training Clustering Algorithm: <https://www.youtube.com/watch?v=7luMauX0KWM>



RESUMEN CLASE TEMA 7 – PARTE 2





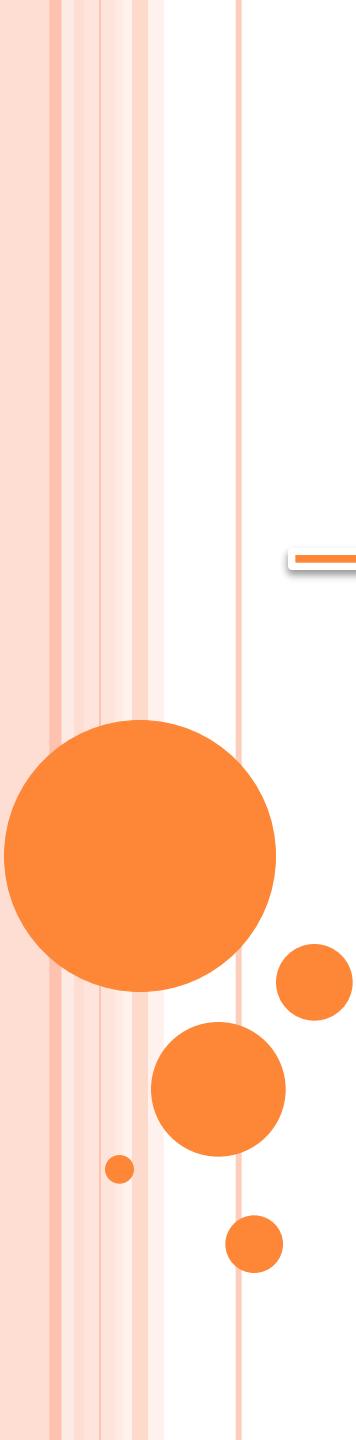
TRABAJO PRÁCTICO DE MINERÍA DE DATOS

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.





UNIVERSIDAD NACIONAL DE LA MATANZA

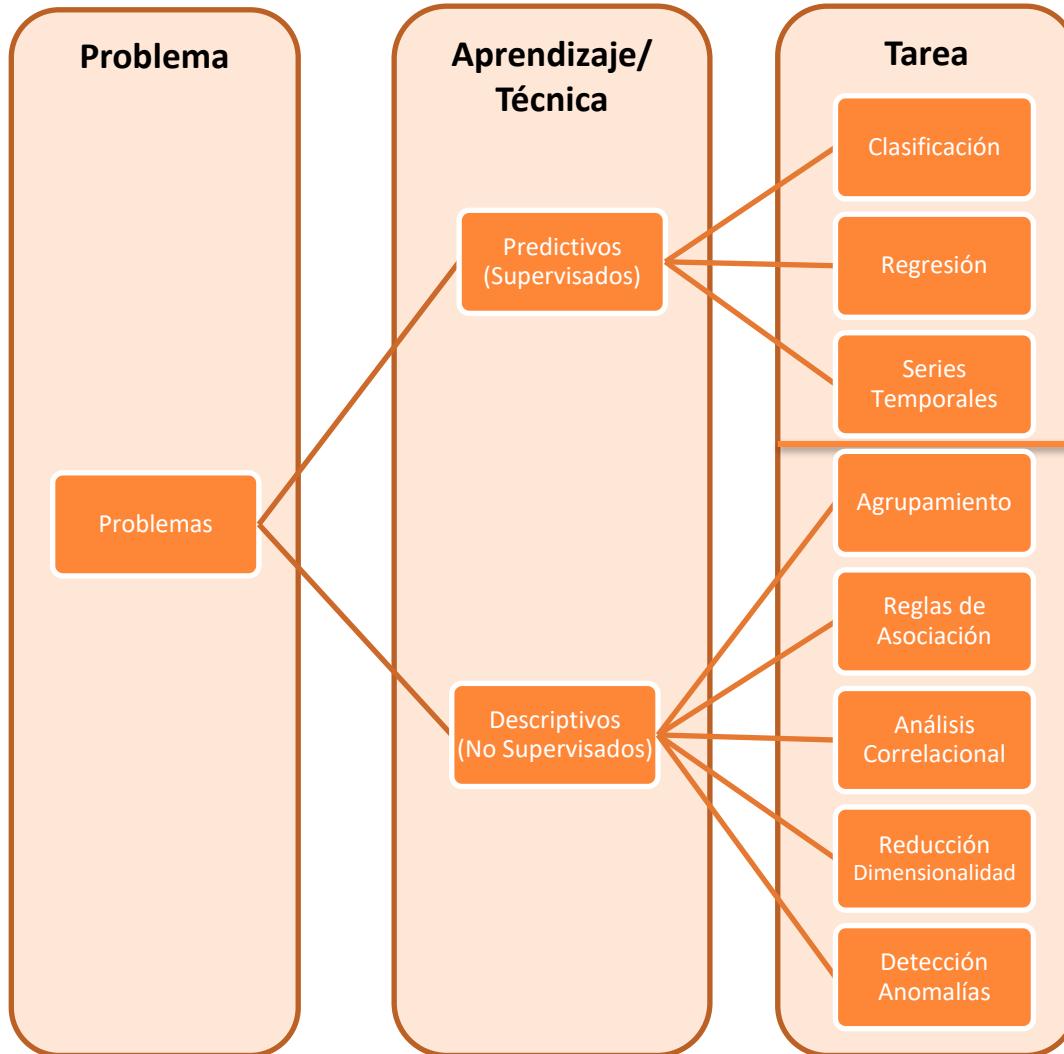
INTELIGENCIA DE NEGOCIOS

Tecnologías Inteligentes para Explotación de Información

Docente: ING. LORENA R. MATTEO



TÉCNICAS DE MINERÍA DE DATOS



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

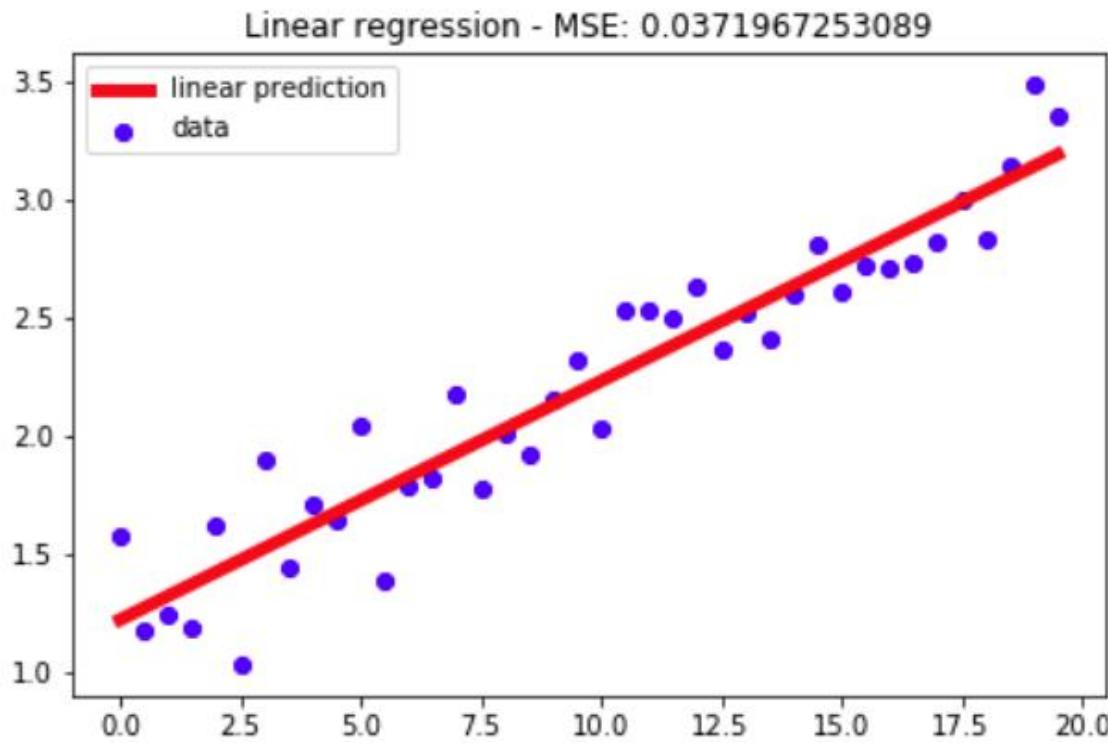


Esta foto de Autor desconocido está bajo licencia CC BY-SA-NC



¿COMO APRENDE UN MODELO DE MINERÍA DE DATOS?

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



$$m = 0.1014$$
$$b = 1.2258$$



LA CLAVE DEL ÉXITO

¿Qué hace que un modelo sea bueno?

Un buen modelo:

- Generaliza bien con datos nuevos.
- Se evalúa con datos no usados en el entrenamiento.
- Se compara con otros modelos mediante métricas como Precisión, Recall, etc.



LA CLAVE DEL ÉXITO

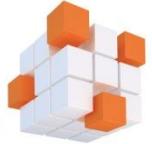
Se debe evaluar la **capacidad de generalización** del modelo.

- El modelo construido debe ser “generalizable”, debe aprender bien con muchos tipos de datos nuevos.
- Debe evaluarse con datos distintos a los usados en el entrenamiento.

¿Cómo saber si un modelo es bueno o no?

Hay que enfocarse en la **capacidad predictiva** y **sencillez** del **modelo**, más que en su rapidez para clasificar, construir modelos y escalar...

- Utilizando métricas de desempeño (performance metrics)
Las métricas se calculan contrastando los valores predichos versus los valores reales de la variable objetivo.
- Comparándolo con el desempeño de otros modelos posibles, diseñados a través de experimentos.
- Probando con datos de prueba y otros dataset (generalizable, por ej. bajo error de generalización)



EL PROBLEMA DEL SOBREAPRENDIZAJE (OVERFITTING)

¿A qué se llama sobreajuste o sobreaprendizaje?

- Cuando un modelo puede aprender demasiado bien los datos de entrenamiento, pero fallar con datos nuevos.

Por ejemplo: Cuando un estudiante memoriza respuestas sin entender el problema.

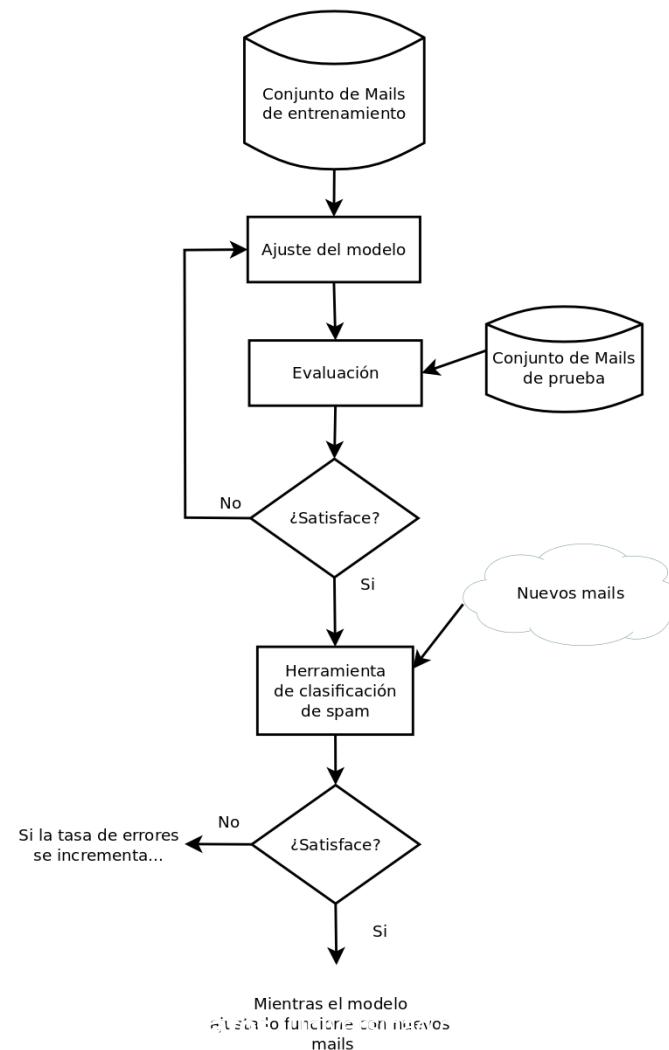
Es decir,

- Un modelo muy complejo que memoriza los datos de entrenamiento.
- Tiene Bajo Error en Entrenamiento, Alto en Prueba



EL PROBLEMA DEL SOBREAPRENDIZAJE (OVERFITTING)

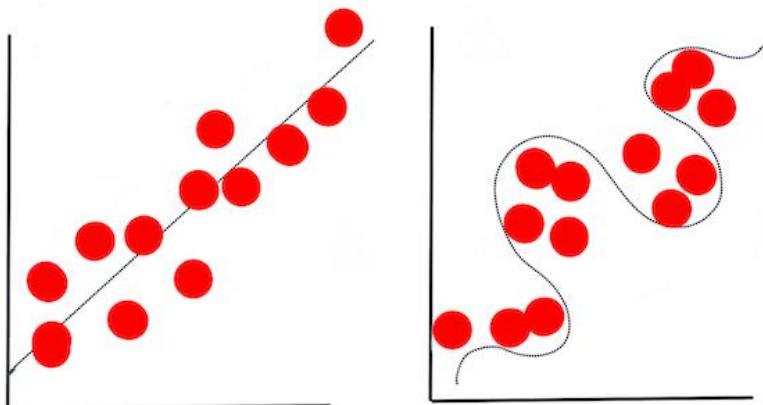
- El propósito de **crear un modelo o clasificador** no es clasificar el conjunto de entrenamiento, sino para **clasificar los datos cuya clase no sabemos.**
- Queremos que los datos sean clasificados correctamente, pero a menudo no tenemos forma de saber si el modelo lo hace. Si la naturaleza de los datos cambia con el tiempo.
 - *Ejemplo: detectar correos electrónicos no deseados.*





EL PROBLEMA DEL SOBREAPRENDIZAJE (OVERFITTING)

- Cuanto mayor sea su complejidad, los modelos de clasificación tienden a ajustarse más al conjunto de entrenamiento utilizado en su construcción (sobreaprendizaje), lo que los hace menos útiles para clasificar nuevos datos.
- En consecuencia, el **conjunto de prueba** debe ser **siempre independiente** del conjunto de entrenamiento.



Correct vs overfit model

El error de clasificación en el conjunto de entrenamiento NO es un buen estimador de la precisión del clasificador.



GENERALIZACIÓN - ANÁLISIS DE ERRORES

¿Con qué Tipos de Errores nos encontramos?

- **Bias (Sesgo del Modelo)**: modelo demasiado simple.
- **Varianza**: modelo demasiado complejo.

¿Qué pasa si sumamos datos?

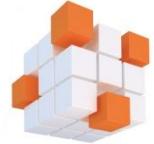
- Se reduce la varianza de modelos complejos
- Se mejora la generalización
- Se evita o reduce el sobreajuste

“El mejor modelo logra un equilibrio entre Sesgo y Varianza.”



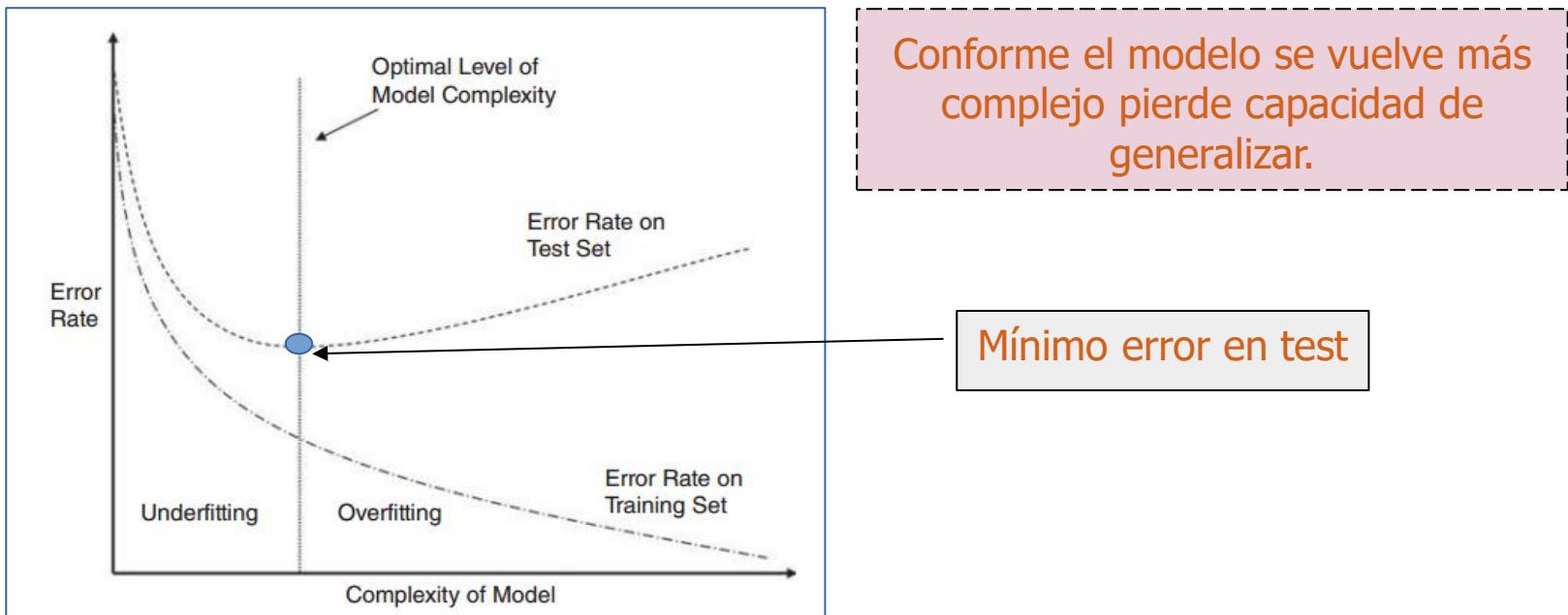
GENERALIZACIÓN - ANÁLISIS DE ERRORES

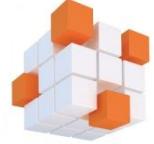
Train Error	Test Error	¿Qué hacer?
LOW	HIGH OVER-FITTING	<ul style="list-style-type: none">Necesidad de un modelo más sencillo (se está empleando un modelo más complejo que lo necesario).Se necesitan más datos (más muestras de datos).
HIGH	HIGH UNDER-FITTING	<ul style="list-style-type: none">Necesidad de un modelo más complejo.Se necesitan más muestras de datosDifícil aprender $f(x, z)$ sólo con x. Obtener también z.Agregar funciones adicionales/ hiperparámetros.
HIGH	LOW	<ul style="list-style-type: none"><u>Inusual</u>: podría significar que los datos de prueba son demasiado similares a los de entrenamiento.Se necesitan más datos de prueba.
LOW	LOW	<ul style="list-style-type: none">¡¡Lo has logrado!! ¡¡Felicitaciones!!



EL PROBLEMA DEL SOBREAPRENDIZAJE (OVERFITTING)

- Cuando la exactitud del modelo no es tan alta en el conjunto de prueba (evaluación) como lo es en el de entrenamiento, a menudo se debe a que el modelo sobreajusta el conjunto de entrenamiento.





DIAGNÓSTICO DE GENERALIZACIÓN - CASOS COMUNES

Resumen:

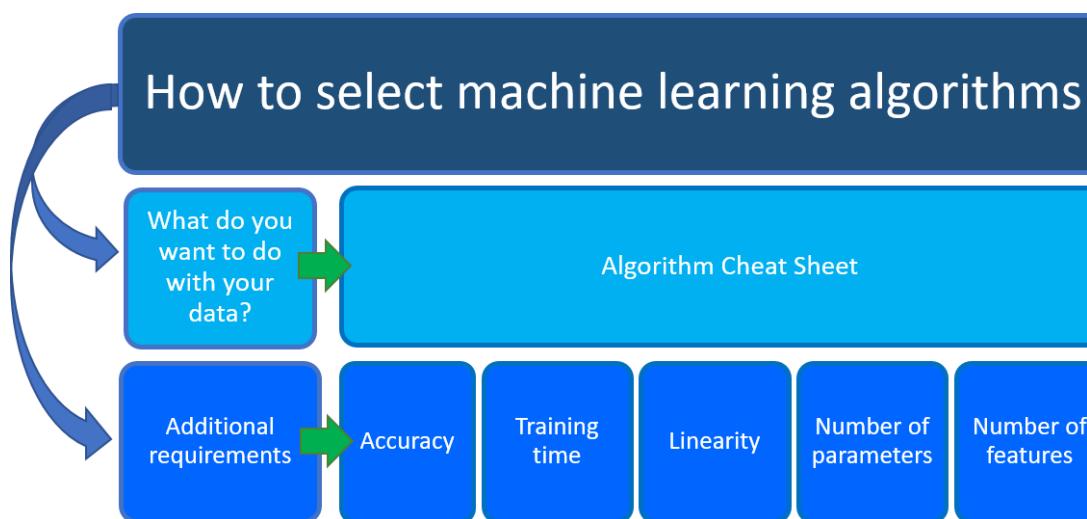
- **Train Error Bajo + Test Error Alto →** Modelo muy complejo.
- **Ambos Errores Altos →** Modelo pobre o falta de datos.
- **Ambos Bajos →** Modelo ideal.



¿QUÉ ALGORITMO DE APRENDIZAJE AUTOMÁTICO USAR?

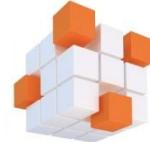
La pregunta es muy habitual cuando uno comienza a trabajar en Minería de Datos. Si bien no se responde de manera directa y no tiene una única respuesta, el algoritmo a seleccionar depende principalmente de dos aspectos diferentes:

- **¿Qué desea hacer con los datos?** Específicamente, ¿cuál es la pregunta de negocios a responder aprendiendo de los datos disponibles?
- **¿Cuáles son los requisitos de su caso de estudio?** Específicamente, ¿cuál es la precisión, el tiempo de entrenamiento, la linealidad, la cantidad de parámetros y la cantidad de características que admite su solución?



Fuente: [How to select a machine learning algorithm - Azure Machine Learning | Microsoft Learn](#)

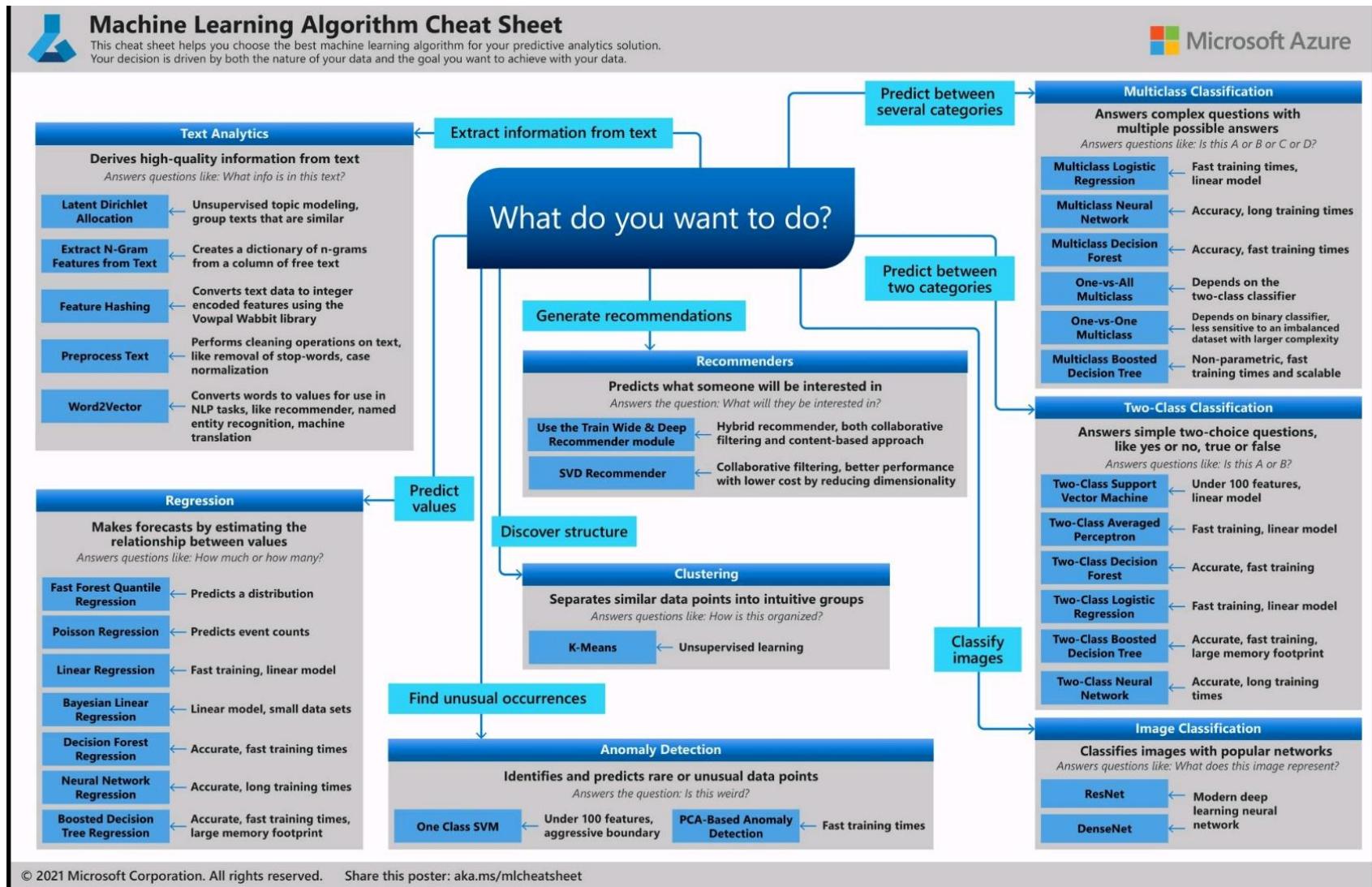
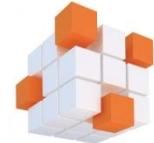
¿QUÉ ALGORITMO DE APRENDIZAJE AUTOMÁTICO DEBO USAR?

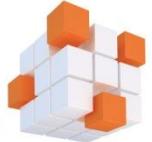


- Algunos algoritmos de aprendizaje hacen suposiciones particulares sobre la estructura de los datos o los resultados deseados.
- Si puede encontrar uno que se adapte a sus necesidades, puede brindarle resultados más útiles, predicciones más precisas o tiempos de entrenamiento más rápidos.
- La siguiente tabla resume algunas de las características más importantes de los algoritmos de las familias de clasificación, regresión y agrupamiento:

Algorithm	Accuracy	Training time	Linearity	Parameters	Notes
Classification family					
Two-Class logistic regression	Good	Fast	Yes	4	
Two-class decision forest	Excellent	Moderate	No	5	Shows slower scoring times. Suggest not working with One-vs-All Multiclass, because of slower scoring times caused by thread locking in accumulating tree predictions
Two-class boosted decision tree	Excellent	Moderate	No	6	Large memory footprint
Two-class neural network	Good	Moderate	No	8	
Two-class averaged perceptron	Good	Moderate	Yes	4	
Two-class support vector machine	Good	Fast	Yes	5	Good for large feature sets
Multiclass logistic regression	Good	Fast	Yes	4	
Multiclass decision forest	Excellent	Moderate	No	5	Shows slower scoring times
Multiclass boosted decision tree	Excellent	Moderate	No	6	Tends to improve accuracy with some small risk of less coverage
Multiclass neural network	Good	Moderate	No	8	
One-vs-all multiclass	-	-	-	-	See properties of the two-class method selected
Regression family					
Linear regression	Good	Fast	Yes	4	
Decision forest regression	Excellent	Moderate	No	5	
Boosted decision tree regression	Excellent	Moderate	No	6	Large memory footprint
Neural network regression	Good	Moderate	No	8	
Clustering family					
K-means clustering	Excellent	Moderate	Yes	8	A clustering algorithm

¿QUÉ ALGORITMO DE APRENDIZAJE AUTOMÁTICO DEBO USAR?



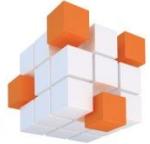


¿QUÉ ALGORITMO DE APRENDIZAJE AUTOMÁTICO USAR?

Depende de:

- Tipo de problema
- Recursos (tiempo, memoria)
- Precisión deseada

Algoritmo	Precisión	Tiempo	Linealidad
Regresión logística	Buena	Rápido	Sí
Random Forest	Excelente	Medio	No
Red neuronal	Buena	Medio	No



REFERENCIAS Y MATERIAL ADICIONAL

Referencias:

- Jose Martinez Heras | European Space Agency (ESA)
- Curso Inteligencia de Negocios, Universidad de Granada, España
- Curso DM Datos, Prof. F.Bravo/ H. Sarmiento / B. Poblete, Universidad de Chile
- Curso Minería Datos, Universidad de Luján, Argentina
- Microsoft/ML-For-Beginners
- Curso Advanced Machine Learning | Educación IT , Argentina

Material Adicional:

- Sesgo algorítmico: <https://www.youtube.com/watch?v=nkOPXgpWS9w>
<https://www.youtube.com/watch?v=onJs6DXczu0>



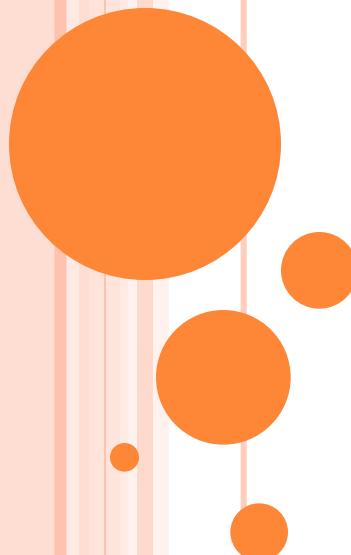
UNIVERSIDAD NACIONAL DE LA MATANZA

INTELIGENCIA DE NEGOCIOS

Tecnologías Inteligentes para Explotación de Información

Introducción a KNIME

Docentes: ING. LORENA R. MATTEO





INTRODUCCIÓN A KNIME

KNIME = KoNstanz Information MinEr

<http://www.knime.org/>

- es un entorno totalmente gratuito para el desarrollo y ejecución de técnicas de minería de datos.
- fue desarrollado en 2004 el Dpto de Bioinformática y Minería de Datos de la Universidad de Konstanz (Constanza), Alemania, bajo la supervisión del profesor Michael Berthold.
- en la actualidad, la empresa KNIME.com GmbH, radicada en Zúrich, Suiza, continúa su desarrollo, además de prestar servicios de formación y consultoría.

KNIME:

- desarrollado sobre la plataforma Eclipse y programado, esencialmente, en Java.
- su uso se basa en el diseño de un flujo de ejecución que plasme las distintas etapas de un proyecto de minería de datos.
- el proceso de análisis de datos, utilizado por KNIME, consiste en un pipeline de nodos, conectados a través de puentes que transportan datos o modelos.
- cada nodo procesa la llegada de datos y/o modelo(s) y produce resultados como salida.
- es posible explorar los datos en forma visual, a través de vistas.



INTRODUCCIÓN A KNIME

- Es una herramienta para análisis, manipulación, visualización y generación de informes de datos.
- Basado en el paradigma de programación gráfica.
- Proporciona una amplia gama de extensiones:

The screenshot shows the official KNIME website. At the top, there's a navigation bar with links for Software, Pricing, Solutions, Community, Customers, Resources, About us, a search bar, Contact us, Download, and Sign in. Below the navigation, there's a main heading: "KNIME offers a complete platform for end to end data science, from creating analytic models, to deploying them and sharing insights within the organization, through data apps and services." To the left, there's a diagram titled "KNIME Analytics Platform" showing a central node connected to various data sources like databases and files, and a visualization component. To the right, there's a diagram titled "KNIME Hub" showing a network of users connected by dashed lines, with icons for play, share, and more. On the far right, there are social media sharing buttons for LinkedIn, Twitter, and others.

Open for Innovation

KNIME

Software ▾ Pricing Solutions ▾ Community ▾ Customers Resources ▾ About us ▾

Contact us Download Sign in

KNIME offers a complete platform for end to end data science, from creating analytic models, to deploying them and sharing insights within the organization, through data apps and services.

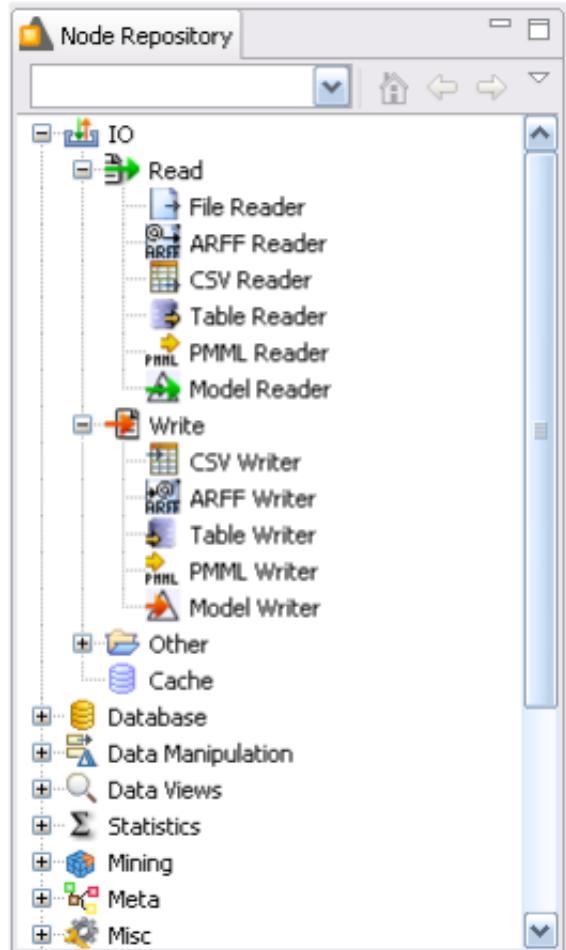
KNIME Analytics Platform

KNIME Hub

3



INTRODUCCIÓN A KNIME



KNIME proporciona distintos nodos agrupados en fichas, como por ejemplo:

- a) Entrada de datos [*IO > Read*].
- b) Salida de datos [*IO > Write*].
- c) Preprocesamiento [*Data Manipulation*], para filtrar, discretizar, normalizar, filtrar, seleccionar variables...
- d) Minería de datos [*Analytics/Mining*], para construir modelos (reglas de asociación, clustering, clasificación, MDS, PCA...).
- e) Salida de resultados [*Data Views*] para mostrar resultados en pantalla (ya sea de forma textual o gráfica).
- f) WEKA: Bayes, Trees, Rules.

USO: Drag & Drop sobre el Editor Windows



INTRODUCCIÓN A KNIME – ENTORNO DE TRABAJO DE LA PLATAFORMA DE ANÁLISIS

The screenshot displays the KNIME Analytics Platform interface with several key components highlighted:

- Workflow Editor:** The central workspace where workflows are built. A workflow named "My first Workflow" is shown, consisting of a "File Reader" node connected to a "Row Filter" node, which is then connected to a "Column Filter" node, and finally to a "Table Writer" node. The "Row Filter" node has the configuration "keep only records born in the US".
- Node Repository:** Located on the left, it contains a list of nodes categorized by type, such as IO, Manipulation, Views, Analytics, DB, and others.
- Workflow Coach:** A sidebar on the left providing recommendations for nodes based on usage.
- KNIME Explorer:** A sidebar on the left showing project and workspace structures.
- Outline:** A view at the bottom left showing the structure of the current workflow.
- Description:** A panel on the right providing detailed information about selected nodes, such as the "Row Filter" node's dialog options and how it filters rows.
- Console & Node Monitor:** A panel at the bottom right showing the output of the "Row Filter" node, displaying a portion of the dataset with columns including ID, age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, and hours.



INTRODUCCIÓN A KNIME

INTERFAZ CLÁSICA DE USUARIO

KNIME Analytics Platform

File Edit View Node Help

75%

KNIME Explorer X

*6: Building a C... 5: Visual Analysi... 4: Building a Cr... Welcome to KNIME... 7: KNIME_project3 8: KNIME_project D K X

Credit Scoring

Credit scoring is a technique used to determine whether or not to extend credit (and if so, how much) to a borrower. This workflow illustrates how to create and choose a credit scoring model based on both historical data and on the application of different machine learning algorithms.

Task Create a credit scoring model based on historical data. Select the best machine learning algorithm to be applied. Use cross-validation to evaluate model performance.

A use case is described at URL: <https://www.knime.org/knime-applications/credit-scoring>

Data Reading Pre-processing Model Training and Evaluation Model Selection Save the Model

The data are German Credit data, including credit status, demographic data, and customer history. The file is located in TheData/Credit.

Learners such as neural network or SVM can only handle numeric attributes. Nominal columns are converted into numerical columns.

1)The following algorithms are trained and evaluated with cross-validation:
- Neural Network
- SVM
- Decision Tree
2) Double-click on the metanode to see the subworkflow

All results, i.e. accuracies and respective models, are combined in one single table. Rows are then sorted by descending accuracy and only first row (best performing model) is kept.

- Convert the model cell back to PMML
- Save the model.

KNIME Analytics Platform writes out the model in the official PMML format, so that other applications can use the model.

Workflow Coach X Node recommendations only available with

Node Repository

Reading credit scoring dataset

CSV Reader Category To Number Train and Cross Validate a Neural Network Train and Cross Validate a SVM Train and Cross Validate a Decision Tree Concatenate Sorter Row Filter Cell To PMML PMML Writer Bar Chart

Try this:

Outline X Console X Node Monitor

KNIME Console

```
***** Welcome to KNIME Analytics Platform v4.7.3.v202305100921 ****
*** Copyright by KNIME AG, Zurich, Switzerland ***
<
```

Solutions for d nodes and con

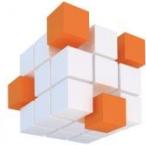
INTRODUCCIÓN A KNIME

NUEVA INTERFAZ DE USUARIO



KNIME Modern UI

The screenshot shows the KNIME Analytics Platform 5 interface. At the top, there's a navigation bar with the KNIME logo, a yellow triangle icon, and the text "KNIME Analytics Platform". Below the navigation bar, a banner reads "Get started with KNIME Analytics Platform 5". On the left, the KNIME logo is displayed with the tagline "Open for Innovation". The main content area is titled "Local space". It contains two sections: one with a computer monitor icon and text about the local space being a folder for workflows and data, and another with a yellow circle containing a plus sign and text encouraging users to "Create workflow in your local space". At the bottom, it says "KNIME Community Hub (hub.knime.com)" and "Sign in".



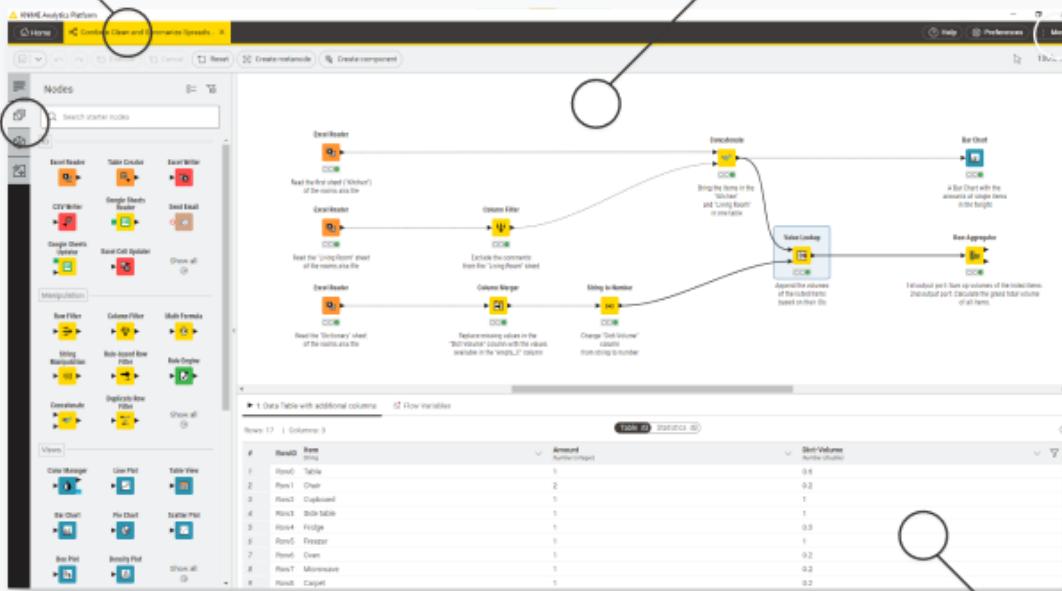
INTRODUCCIÓN A KNIME

NUEVA INTERFAZ DE USUARIO

KNIME Modern UI

Application tabs

Entry page tab and all opened workflows tabs.



Side panel navigation

Description

Description of the displayed workflow or component.

Node repository

All available nodes in KNIME Analytics Platform to build your workflows.

Space explorer

Navigate local or KNIME Hub spaces and access your workflows, components and files.

Workflow Editor

Canvas for editing the currently active workflow.

Help, Preferences, Menu

Access to more material, install additional extensions and change settings for the node repository.

Node Monitor

Shows the output of the current selected node and also the flow variable values.

INTRODUCCIÓN A KNIME

NUEVA INTERFAZ DE USUARIO



KNIME Modern UI

KNIME Analytics Platform

Home 01_Performing_a_k-Means_Clustering •

Execute Cancel Reset Create metanode Create component Help Preferences Menu

94% ▾

Nodes

Search starter nodes

IO

Excel Reader Excel Writer Microsoft Authenticator

Google Authenticator Google Sheets Reader Google Sheets Writer

CSV Reader CSV Writer Show all

Manipulation

Row Filter Column Filter Concatenate

Value Lookup Row Aggregator Table Splitter

String Cleaner Table Cropper

Table Reader (deprecated) load data

Partitioning split data

Search for three clusters

k-Means perform clustering

Visualization

Color Manager Shape Manager Scatter Plot (legacy)

assign colors to classes assign shape to clusters create scatter plot

Cluster Assigner assign cluster to new data points

1: Assigned Data Flow Variables

Rows: 20 | Columns: 9

#	RowID	DATO Number (inte... 0 String	1 String	2 String	3 String	4 String	5 String	6 String	
1	Row1	1	0,090625	0,629687	0,015625	0,7375	-0,185937	0,669531	0,0625
2	Row13	13	-0,064062	0,517187	0,690625	0,678906	0,367188	0,111719	0,278125
3	Row17	17	0,7	0,806641	0,186719	0,242969	0,676562	0,1375	0,592188
4	Row18	18	0,30625	-0,113281	0,170312	-0,026563	0,68125	0,203125	0,414062
5	Row22	22	0,620313	0,480859	0,355469	-0,080469	0,341406	0,109375	0,221875

INTRODUCCIÓN A KNIME

NUEVA INTERFAZ DE USUARIO



KNIME Modern UI

KNIME Analytics Platform

Home KNIME_project5KMeans X

Execute Cancel Reset Create metanode Create component

Help Preferences Menu

100% ▾

CSV Reader

Reads CSV files. To auto-guess the structure of the file click the Autodetect format button. If you encounter problems with incorrect guessed data types disable the Limit data rows scanned option in the Advanced Settings tab. If the input file structure changes between different invocations, enable the Support changing file schemas option in the Advanced Settings tab. For further details see the KNIME File Handling Guide [File Handling Guide](#).

Note: If you find that this node can't read your file, try the File Reader node. It offers more options for reading complex files.

This node can access a variety of different [file systems](#). More information about file handling in KNIME can be found in the official [File Handling Guide](#).

Parallel reading: Individual files can be read in parallel if

- They are located on the machine that is running this node.
- They don't contain any quotes that contain row delimiters.
- They are not gzip compressed.
- No lines or rows are limited or skipped.
- The file index is not prepended to the RowID.

Statistics View

Scatter Plot

k-Means

Silhouette Coefficient

Scorer

Table Splitter

Table Cropper

Row to Column Names

Joiner

Column Filter

Concatenate

GroupBy

Row Filter

Missing Value

Partitioning

Number to String

String Manipulation

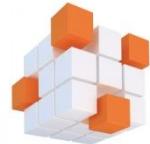
Flow Variables

Rows: 120 | Columns: 8

#	RowID	DATO Number (inte...)	0 Number (dou...)
1	Row0	0	0.916
2	Row1	1	0.091
3	Row2	2	0.456
4	Row3	3	0.355
5	Row4	4	-0.172
6	Row5	5	0.207

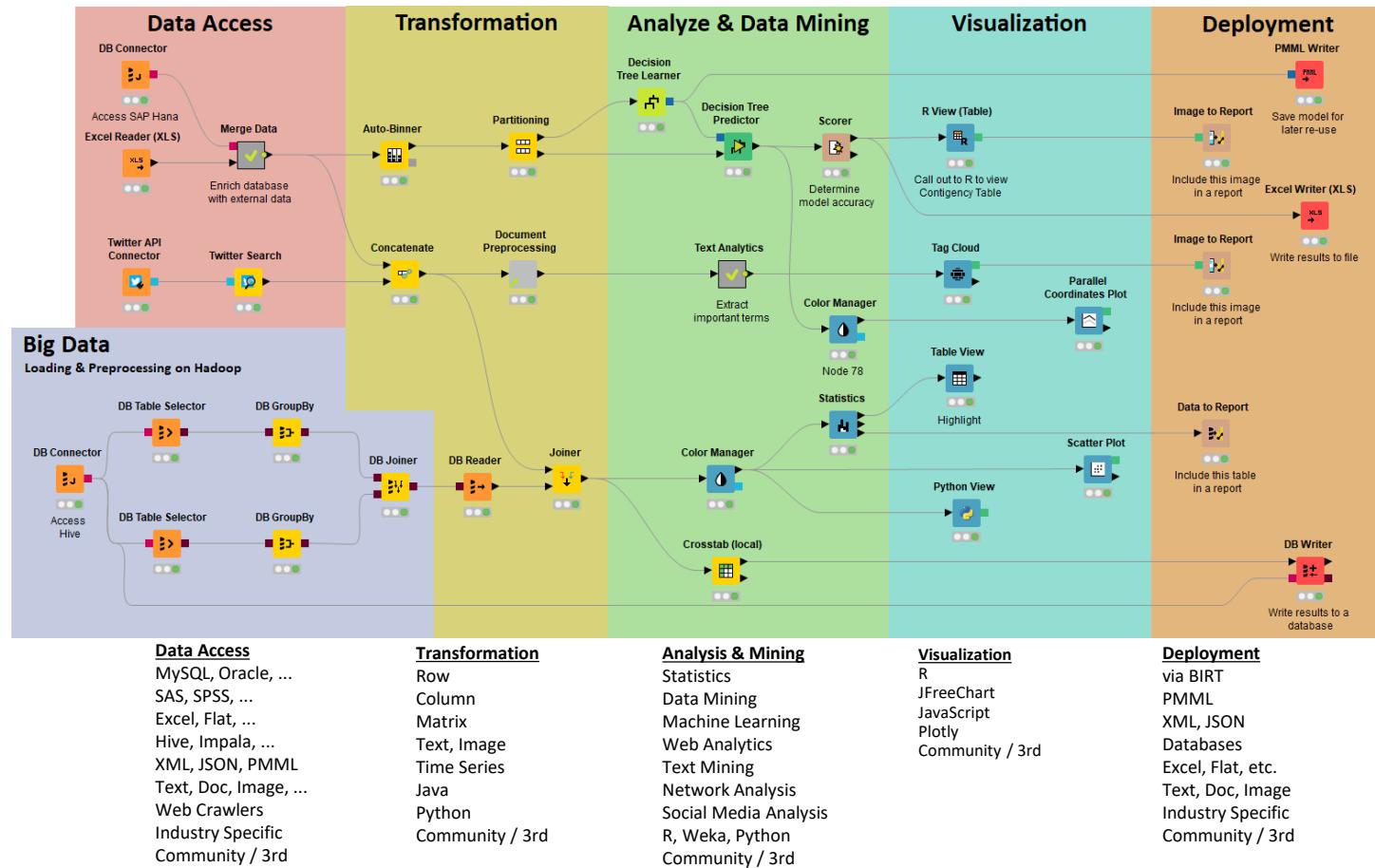
5
Number (dou...)
-0.144
0.616
0.67
0.062
0.283
0.424
0.993
-0.07
0.459
0.756
0.252
0.601

6
Number (dou...)
10



INTRODUCCIÓN A KNIME

Provee más de 4000 rutinas de análisis de datos, tanto nativas como a través de Python, R o Weka, algunos de los nodos se muestran en la siguiente imagen:





INTRODUCCIÓN A KNIME

- Instalar la versión de 64 bits configurando # acceso a la RAM.
- Extensiones útiles:

□ KNIME & Extensions

- KNIME Data Generation
- KNIME Distance Matrix
- KNIME File Handling Notes
- KNIME HTML/PDF Writer
- KNIME Interactive R Statistics Integration
- KNIME Itemset Mining
- KNIME JavaScript Views
- KNIME JFreeChart
- KNIME Optimization extension
- KNIME Python Integration
- KNIME Report Designer
- KNIME Textprocessing
- KNIME Weka Data Mining Integration

□ KNIME Community Contributions – Other

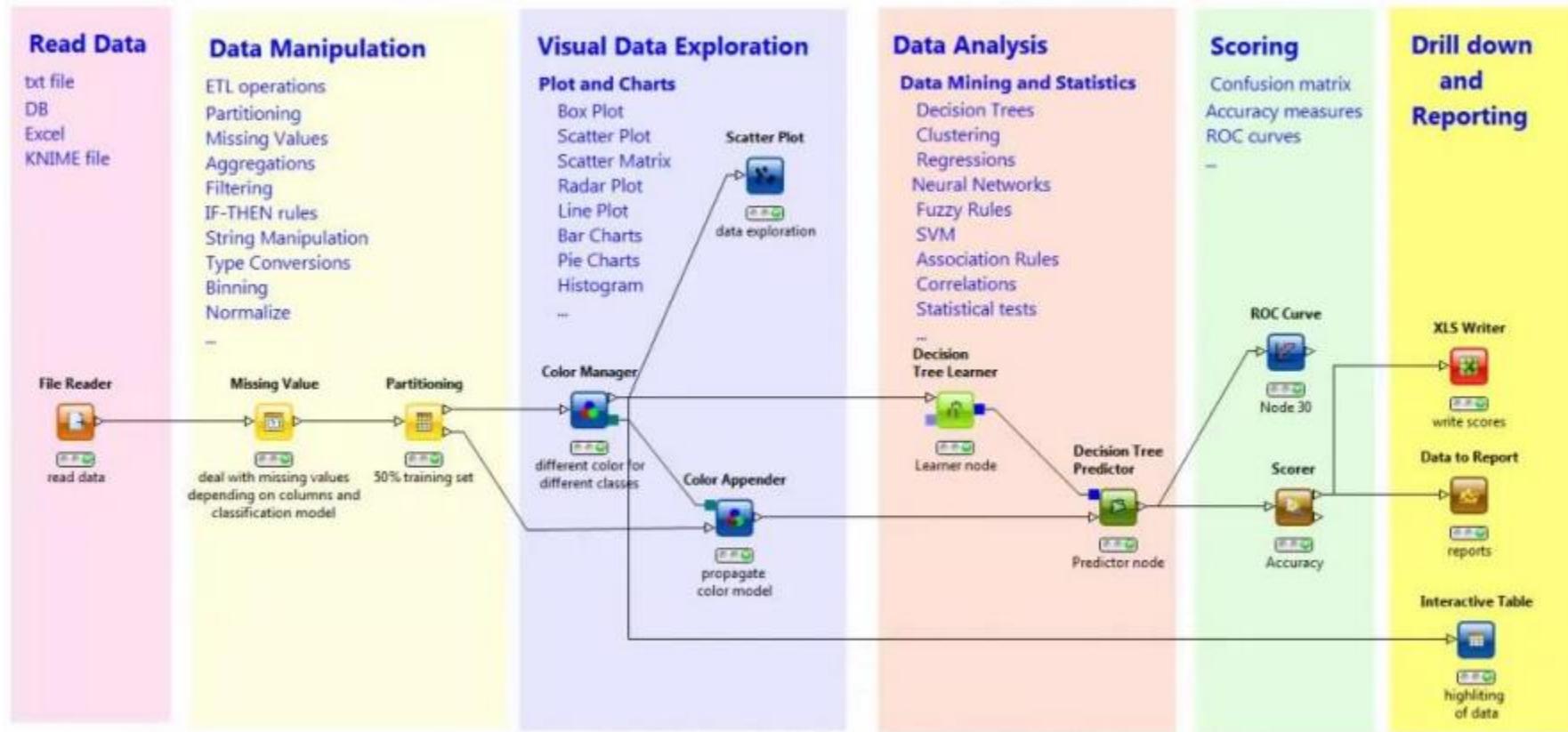
- KNIME Python Scripting extension

□ KNIME Labs Extensions

- KNIME JavaScript Views (Labs)
- KNIME Machine Learning Interpretability Extension
- KNIME Plotly
- KNIME Rule Viewer
- KNIME Statistics Nodes (Labs)
- KNIME XGBoost Integration



INTRODUCCIÓN A KNIME – EJEMPLO WORKFLOW



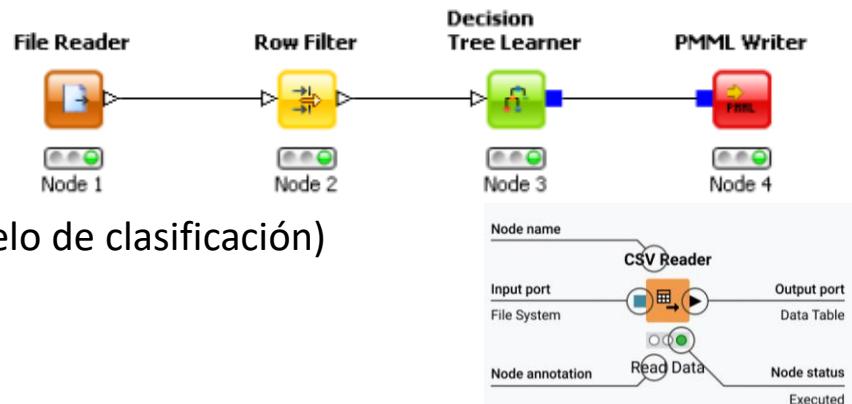
- Un flujo de trabajo es una secuencia de nodos, cada uno de los cuales se puede configurar para realizar una tarea específica.
- Los datos fluyen a través de los nodos de izquierda a derecha.



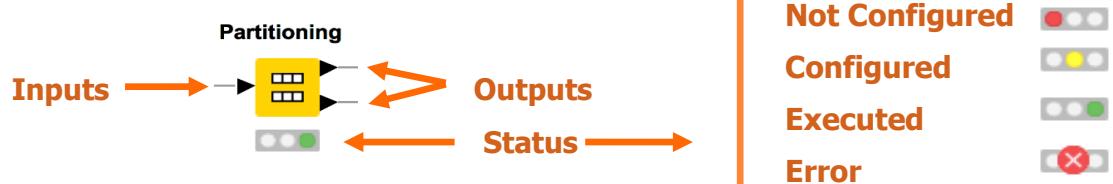
INTRODUCCIÓN A KNIME – CREAR WORKFLOWS

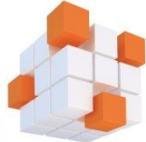
- Para crear un flujo de ejecución, las salidas de unos nodos se utilizan como entradas de otros. Por ejemplo, un flujo básico podría ser de la forma:

- Nodo de lectura de datos
 - Nodo de preprocessamiento
 - Nodo de salida de resultados.
 - Nodo de modelado (por ejemplo, modelo de clasificación)



- Flujo de datos
 - Se construye arrastrando y soltando los nodos desde el repositorio de nodos al Editor Windows y conectándolos entre sí.
 - Los datos se transportan entre los nodos a través de los puertos de entrada y salida.
 - Después de colocar los nodos en el Editor Windows, es necesario conectar la entrada de cada nodo con la salida del nodo predecesor.





INTRODUCCIÓN A KNIME – RESUMEN

Funcionalidades:

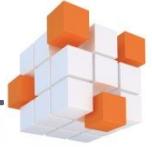
- Software libre bajo licencia GNU.
- Combinación de datos y herramientas.
- Analíticas poderosas.
- Más de 4000 módulos y creciendo.
- Conectores para todos los formatos de archivos y bases de datos más utilizadas.
- Soporte para una gran variedad de tipos de datos.
- Combinación y transformación de datos nativa y en la base de datos.
- Funciones matemáticas y estadísticas.
- Algoritmos de predicción avanzados y de Machine Learning.
- Control de flujo.
- Herramienta de unión para R, Python, SQL, Java, Weka, etc.
- Vistas de datos y reportes interactivos.
- Licenciamiento: GNU GPL 3

Para más información consultar los siguientes enlaces:

- 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>

CUADRANTE MÁGICO DE GARTNER PARA DATA SCIENCE Y ML PLATFORMS (ABR 2025)



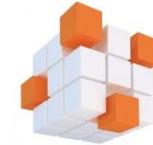


CUADRANTE MÁGICO DE GARTNER PARA DATA SCIENCE Y ML PLATFORMS (ENE 2021)

3/6/2025 IN2025



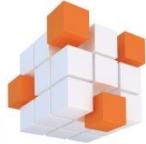
Fuente: <https://www.gartner.com/reviews/market/data-science-and-machine-learning-platforms>



TOP 10 HERRAMIENTAS LÍDERES ML & AI SEGÚN LLMS

Herramienta	Descripción	Plataformas Soportadas	Características Clave
TensorFlow	Ideal para redes neuronales y aprendizaje profundo, con soporte para múltiples plataformas	Servidores, Nube, Web, Android, iOS	APIs de alto nivel, ejecución distribuida, integración con Keras.
PyTorch	Popular en investigación por su flexibilidad, especialmente en aprendizaje profundo.	Linux, Windows, macOS	Ejecución inmediata, soporte GPU, comunidad activa, ideal para investigación.
Scikit-learn	Biblioteca de Python para aprendizaje automático tradicional, basada en SciPy y NumPy.	Windows, Linux, macOS	Algoritmos de clasificación, regresión, clustering, preprocesamiento de datos.
Keras	API de alto nivel para redes neuronales, ahora Nube (Google, Azure, AWS), Desktop, parte de TensorFlow, fácil de prototipar.	Móvil	Interfaz amigable, soporte multi-GPU, modularidad, ideal para principiantes.
Microsoft Azure ML	Plataforma en la nube para desarrollar, entrenar y desplegar modelos, con AutoML y MLOps.	Nube (Azure)	Interfaz drag-and-drop, soporte para Python, R, integración con servicios Azure.
Amazon SageMaker	Servicio de AWS para todo el ciclo de vida del aprendizaje automático, con AutoML y algoritmos integrados.	Nube (AWS)	Escalabilidad empresarial, integración con AWS, soporte para PyTorch, TensorFlow.
Google Cloud Vertex AI	Plataforma unificada para desarrollo y despliegue de modelos de IA, con AutoML y notebooks.	Nube (Google Cloud)	Integración con TensorFlow, escalabilidad, herramientas automatizadas para colaboración.
H2O.ai	Plataforma de código abierto para aprendizaje automático, conocida por AutoML y procesamiento distribuido.	Nube, Desktop, Integración con R, Python	AutoML, interpretabilidad de modelos, soporte para Jupyter, seguridad empresarial.
XGBoost	Biblioteca para gradient boosting, optimizada para modelos predictivos, con alta precisión.	Cross-platform	Alto rendimiento, manejo de datos diversos, ideal para clasificación y regresión.
Apache Spark MLLib	Biblioteca de aprendizaje automático para grandes volúmenes de datos, parte de Apache Spark.	Distribuido, Nube	Algoritmos escalables, integración con Spark SQL, ideal para big data.

Fuente: Grok 3 (2025). DeeperSearch (versión 31/03/2025) [LLM]. <https://grok.com/>



MATERIAL DE CONSULTA RECOMENDADO

- 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>

Recursos sugeridos:

- [Spaces of knime – KNIME Community Hub. Teaching Materials | KNIME](#) *Previo registro en el sitio.*
 - [knime/Education – Courses – KNIME Community Hub](#)
 - [stervis/Public – E-Learning – KNIME Community Hub](#)
 - [knime/Education – L4-ML Machine Learning Algorithms – KNIME Community Hub](#)
 - [knime/Examples – 04_Analytics – KNIME Community Hub](#)
 - [knime/Educators Alliance – Guide to Intelligent Data Science – KNIME Community Hub](#)

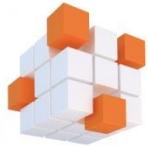


TRABAJO PRÁCTICO DE MINERÍA DE DATOS

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.

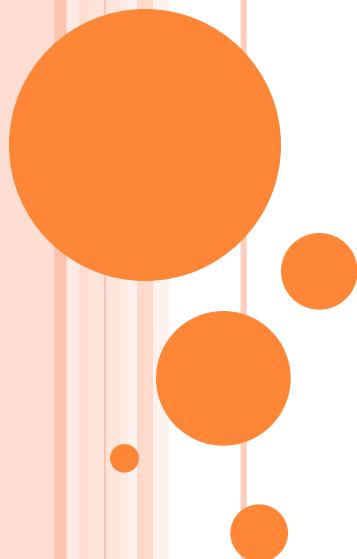


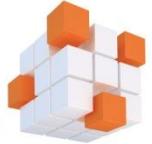
UNIVERSIDAD NACIONAL DE LA MATANZA

INTELIGENCIA DE NEGOCIOS

**Fundamentos de
Inteligencia Artificial**

Docente: ING. LORENA R. MATTEO





ROMPIENDO EL HIELO

The screenshot shows a presentation slide titled "Fundamentos de Inteligencia Artificial". On the left, there is a QR code with the text "Join at: ahaslides.com/SEG5P". The main content features a large blue "AI" text against a dark background with abstract blue and white geometric shapes. The AhaSlides logo is in the top right corner, and a navigation bar with icons is at the bottom.

**Fundamentos de
Inteligencia Artificial**

Join at:
[ahaslides.com/
SEG5P](https://ahaslides.com/SEG5P)

INTELIGENCIA DE NEGOCIOS
EFC - UNLaM
Ing. Lorena R. MATTEO

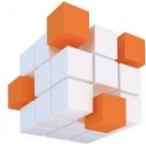
0/50



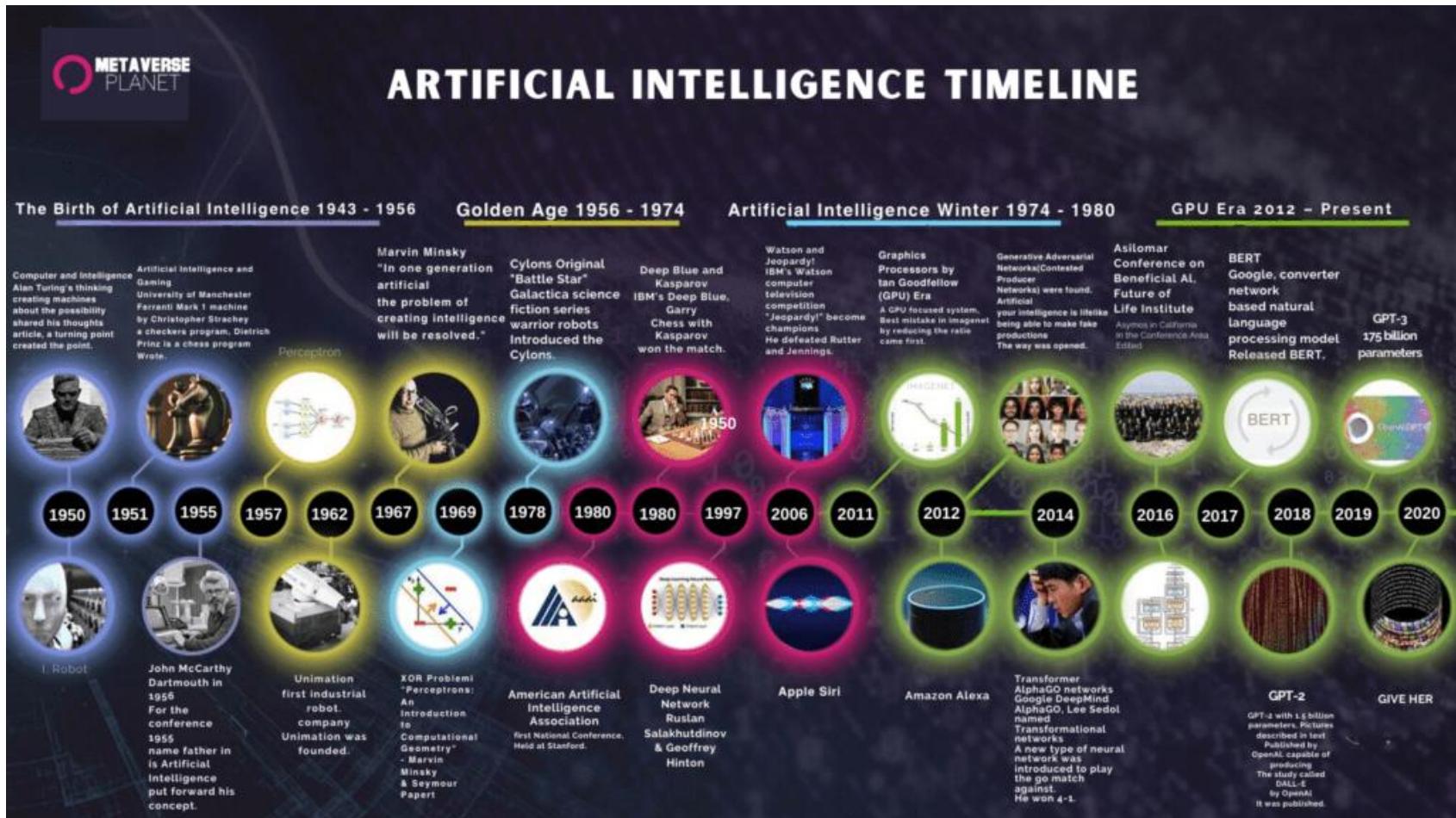
HISTORIA DE LA INTELIGENCIA ARTIFICIAL (RESUMEN CRONOLÓGICO)

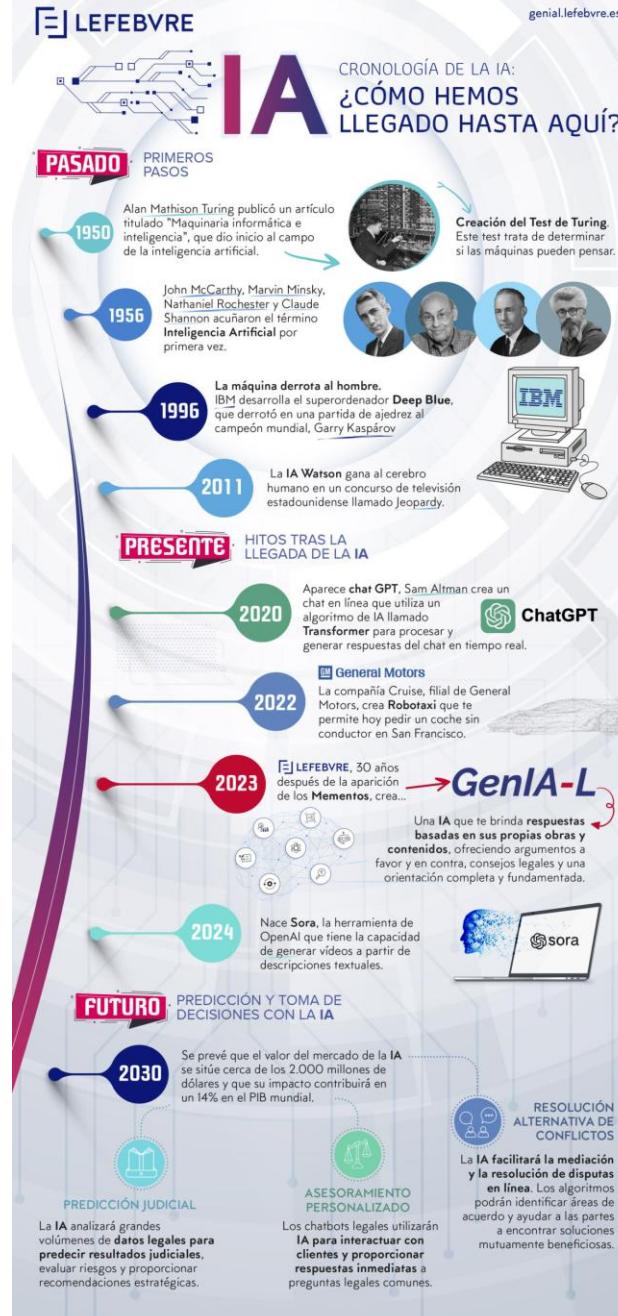
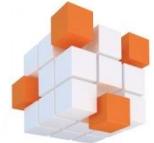
- **Pre-1950s: Ideas Fundacionales y Lógica Simbólica:**
 - Desarrollo de la lógica formal por figuras como Aristóteles y Boole.
 - Invención de máquinas calculadoras y los primeros autómatas.
 - Trabajo de Alan Turing sobre la "Máquina de Turing" y el concepto de computabilidad.
- **Década de 1950: El Nacimiento de la IA:**
 - Taller de Dartmouth (1956) se considera el evento fundacional de la IA como campo de estudio.
 - Primeros programas de IA como el "Logic Theorist" y el "General Problem Solver".
- **Década de 1960: Optimismo y Primeros Éxitos:**
 - Desarrollo de programas que podían resolver problemas de álgebra, jugar al ajedrez y comprender el lenguaje natural en contextos limitados (e.g., ELIZA).
 - Predicciones optimistas sobre el rápido avance de la IA.
- **Década de 1970: El Primer "Invierno" de la IA:**
 - Desilusión por la falta de progreso significativo en la resolución de problemas complejos del mundo real.
 - Informe Lighthill critica la investigación en IA y lleva a la reducción de la financiación.
- **Década de 1980: El Renacimiento con los Sistemas Expertos:**
 - Auge de los sistemas expertos, diseñados para emular el conocimiento de expertos humanos en dominios específicos.
 - Aplicaciones en áreas como diagnóstico médico y configuración de computadoras.
- **Finales de los 80 y Década de 1990: El Segundo "Invierno" de la IA:**
 - Las limitaciones de los sistemas expertos y la falta de aplicaciones generalizadas llevan a una nueva disminución del interés y la inversión.
- **Finales de los 90 y Principios de los 2000: El Auge del Machine Learning:**
 - Mayor disponibilidad de datos y potencia computacional impulsan el desarrollo de algoritmos de aprendizaje automático.
 - Éxitos en áreas como el filtrado de spam, sistemas de recomendación y reconocimiento de voz.
- **Década de 2010: La Revolución del Deep Learning:**
 - Avances significativos en el aprendizaje profundo (Deep Learning) gracias a redes neuronales con múltiples capas.
 - Grandes avances en visión por computadora, procesamiento del lenguaje natural y reconocimiento de voz.
- **Presente (Década de 2020): La Era de los LLMs y la IA Generativa:**
 - Desarrollo y proliferación de modelos de lenguaje extensos (LLMs) como GPT-3 y modelos de IA generativa para imágenes, audio y video.
 - Debates sobre las implicaciones éticas, el potencial y los desafíos de la IA avanzada.

Fuente: Google. (2025). Gemini (versión del 10 de Junio) [LLM]. <https://gemini.google.com/>

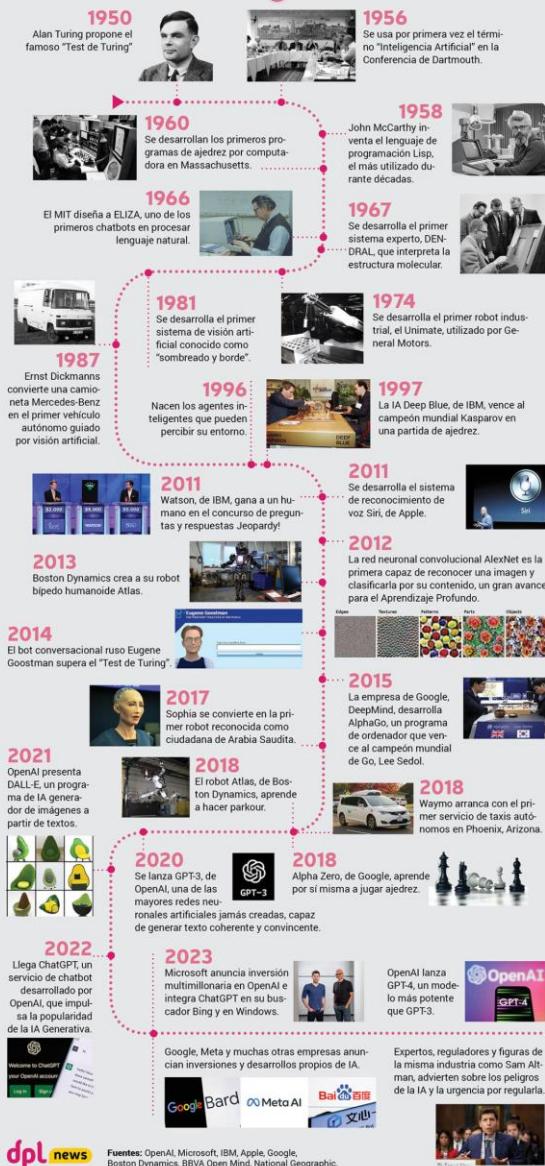


LÍNEA DE TIEMPO DE LA INTELIGENCIA ARTIFICIAL





La Inteligencia Artificial a través del tiempo



Fuente: <https://dplnews.com/infografia-la-inteligencia-artificial-a-traves-del-tiempo/>
<https://elderecho.com/cronologia-de-la-ia-como-hemos-llegado-hasta-aqui>



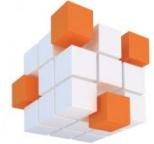
¿QUÉ ES LA INTELIGENCIA ARTIFICIAL GENERATIVA?

- La **Inteligencia Artificial (IA)** imita el comportamiento humano mediante el **aprendizaje automático** para interactuar con el entorno y ejecutar tareas sin instrucciones explícitas sobre qué generar.
- La **Inteligencia Artificial Generativa (IAGen)** describe una categoría de funcionalidades dentro de la inteligencia artificial que crean contenido original.



Diagram illustrating the process: Instruction → M-LLM → Response
- Estas funcionalidades incluyen **tomar entradas de lenguaje natural (Prompts)** y **devolver respuestas adecuadas en diversos formatos**, como **lenguaje natural, imágenes, código, etc.**

Fuente: [Introducción a los conceptos de inteligencia artificial generativa - Training | Microsoft Learn](#)
[Cómo funcionan los LLM y la inteligencia artificial generativa - .NET | Microsoft Learn](#)



¿QUÉ ES UN LLM?

- Un **LLM (Large Language Model)** es una forma avanzada de IA entrenada con grandes volúmenes de texto, que identifica patrones lingüísticos y genera lenguaje coherente.
- Los LLM descomponen el **texto en tokens**, palabras analizan el conocimiento adquirido y predicen el siguiente token para formar respuestas relevantes.
 - Dependiendo del modelo, un **token** puede representar una letra, una sílaba, una palabra completa o incluso un espacio en blanco.

A screenshot of a text editor interface. The main area displays a sentence: "Piensa como un asistente con 10 años de experiencia, organiza una reunión de 2 horas para revisar proyectos, discutir desafíos y planificar nuevas iniciativas del servicio universitario. Incluye al director, responsables y técnicos. Crea una agenda profesional y clara, con los temas, responsables y resultados esperados." The text is color-coded by token, with each word appearing in a different hue. Below the text, there is a table:

Tokens	Characters
79	321

Ejemplo de como un LLM descompone una frase en tokens para poder procesarla.



¿CÓMO FUNCIONA UN LLM?

“Las aplicaciones de IAGen que usamos hoy son posibles mediante la arquitectura transformer.”

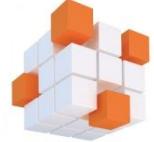
“Los usuarios interactúan con la IAGen a través de aplicaciones de chat que usan lenguaje natural como entrada (Prompts).”

El Modelo:

- Descompone el texto en tokens (palabras, fragmentos o caracteres).
- Predice el siguiente token usando patrones y conocimientos previos.
- Construye respuestas coherentes y contextuales, como conectar puntos en un dibujo.

Claves:

- **Secuencia y coherencia:** Cada palabra depende de las anteriores.
- **Patrones ocultos:** Detecta patrones complejos en los datos.
- **Instrucciones claras:** Prompts bien estructurados guían respuestas precisas.
- **Influencia de la info previa en el LLM:** Las respuestas del LLM dependen de la coherencia y calidad de la información previa proporcionada.

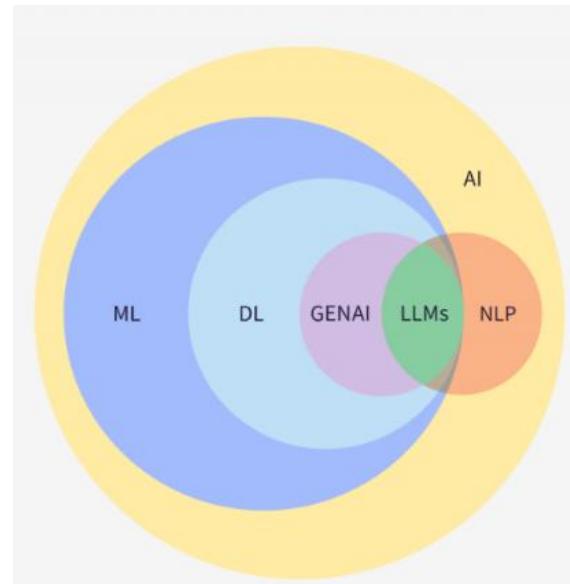


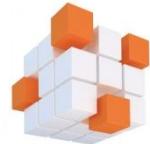
¿QUÉ SIGNIFICA DE GRAN ESCALA?

- Que el modelo tiene una cantidad enorme de **parámetros** (**neuronas artificiales**) y ha sido entrenado con enormes volúmenes de texto.
- Esta escala permite que pueda realizar tareas muy variadas, con respuestas cada vez más ajustadas y creativas.

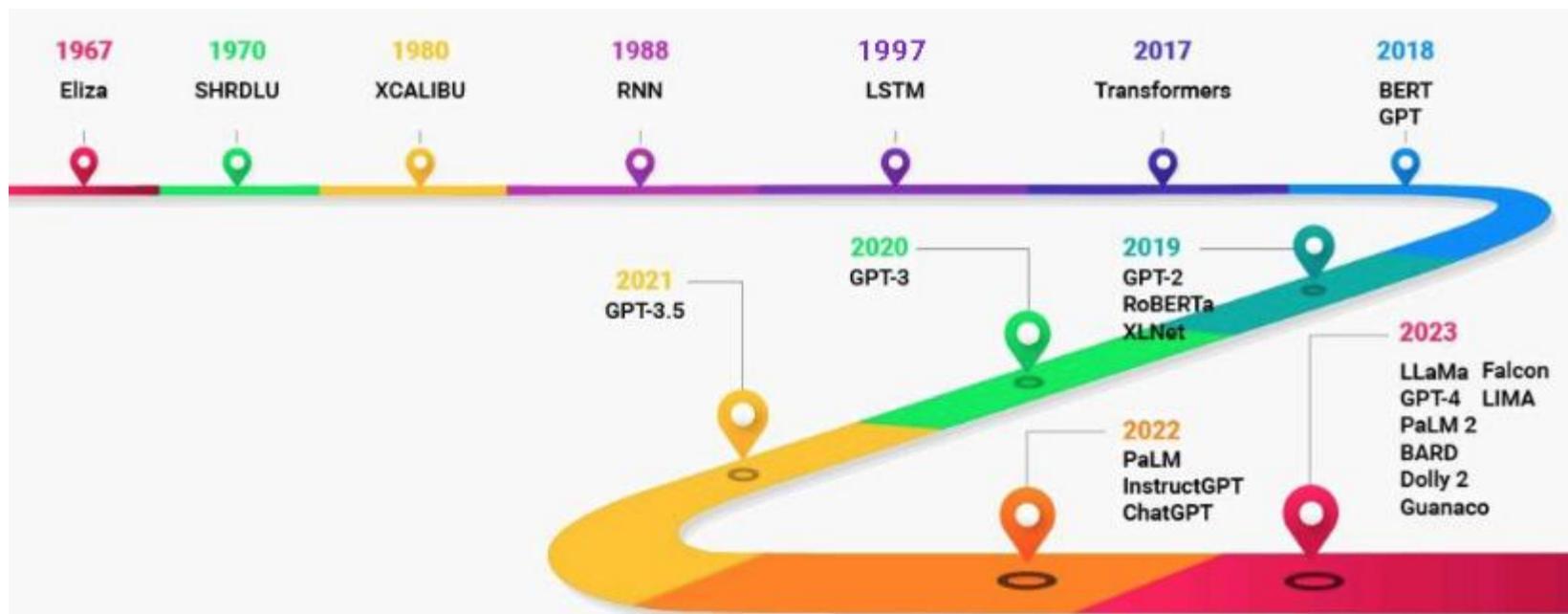
Ejemplos de LLMs conocidos:

- ChatGPT, DALLE3, SORA (OpenAI)
- Gemini, Veo2 (Google)
- Copilot (Microsoft)
- Claude (Anthropic)
- Mistral, LLaMA (Meta AI)
- Qwen, DeepSeek, entre otros.





EVOLUCIÓN LLMs





¿QUÉ ES UN SLM?

- Un SLM (**Small Language Model**) son versiones reducidas y especializadas de los LLMs.
- **Especialización:** están optimizados para tareas o dominios específicos.
- **Velocidad y personalización:** Se pueden ajustar y desplegar más rápido que los LLM.
- **Eficiencia computacional y menor huella energética:** Requieren menos recursos para entrenamiento e inferencia, lo que reduce costos y permite su ejecución en dispositivos con recursos limitados. Consumen menos energía, contribuyendo a una IA más sostenible.



COMPARACIÓN LLMS VS SLMs

Modelos de lenguaje grande (LLM)	Modelos de lenguaje pequeño (SLM)
Los LLM se entrena con grandes cantidades de texto que representan una amplia gama de temas generales, normalmente mediante el suministro de datos de Internet y otras publicaciones disponibles con carácter general.	Los SLA se entrena con conjuntos de datos más pequeños y centrados en el asunto
Cuando se entrena, tienen muchos miles de millones (incluso billones) de parámetros (pesos que se pueden aplicar a las incrustaciones vectoriales para calcular las secuencias de token previstas).	Normalmente, tienen menos parámetros que los LLMs.
Capaz de mostrar funcionalidades completas de generación de lenguajes en una amplia gama de contextos conversacionales.	Este vocabulario centrado hace que sean efectivos en temas conversacionales específicos, pero menos efectivos en la generación de lenguajes más general.
Su gran tamaño puede afectar a su rendimiento y dificultar la implementación local en dispositivos y equipos.	El tamaño menor de los SLM puede proporcionar más opciones para la implementación, incluida la implementación local en dispositivos y equipos locales; y hace que sean más rápidos y fáciles de ajustar.
Ajustar el modelo con más datos para personalizar su experiencia en el tema puede ser lento y costoso en términos de la potencia de proceso necesaria para realizar el entrenamiento adicional.	El ajuste fino puede ser potencialmente menos lento y costoso.



COMPARACIÓN LLMS vs SLMs

18/6/2025 IN2025

Quality vs Size in Small Language Models (SLMs)

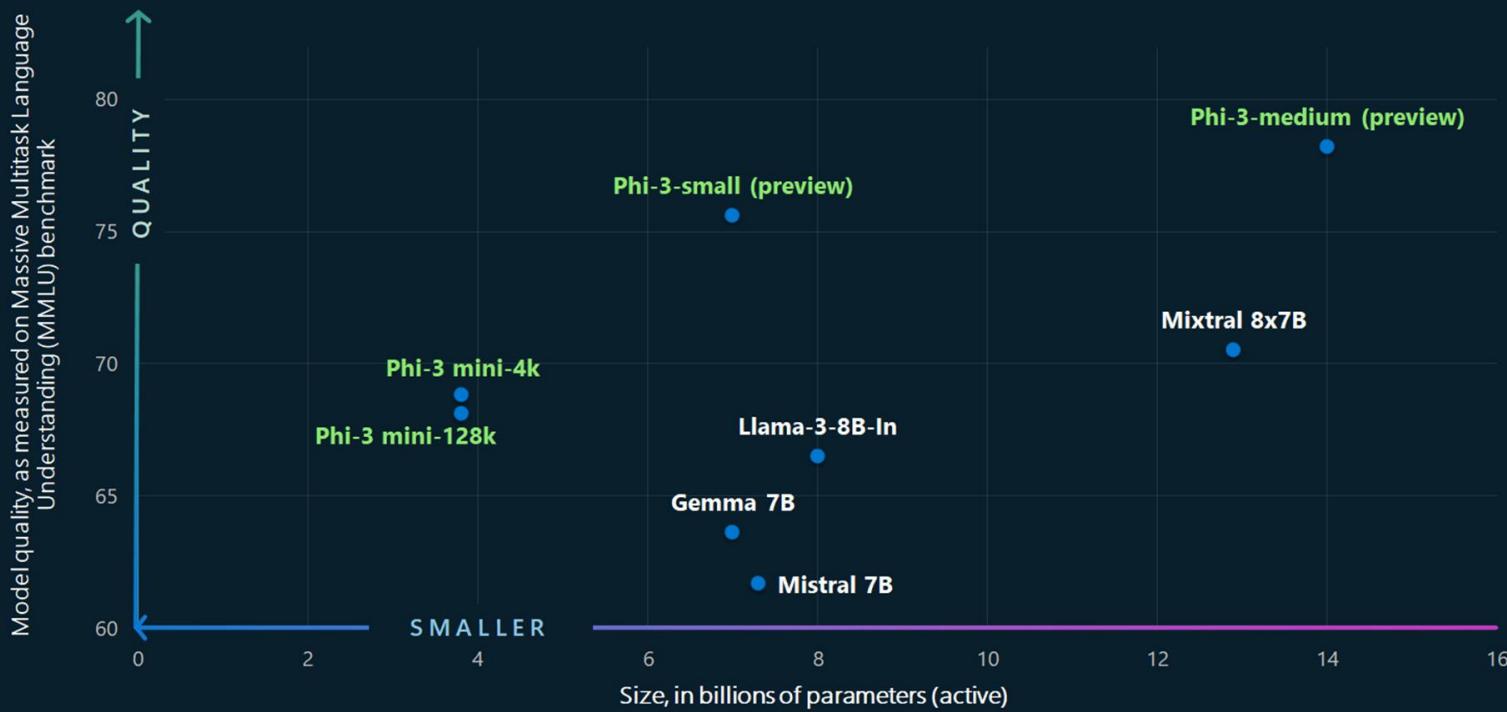
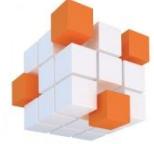


Gráfico que ilustra cómo la calidad de los nuevos modelos Phi-3, medida por el rendimiento en el benchmark Massive Multitask Language Understanding (MMLU), se compara con otros modelos de tamaño similar.
(Imagen cortesía de Microsoft)

Fuente: [Tiny but mighty: The Phi-3 small language models with big potential - Source](#)



¿QUÉ ES UNA ALUCINACIÓN?

- Aunque estos modelos de IAgen generan respuestas coherentes, no siempre son verídicas (**Alucinación**) y pueden reproducir **sesgos**, requiriendo revisión humana.

Cuando un modelo genera contenido falso pero con confianza y coherencia. También llamado **confabulación o delirio artificial**.

¿POR QUÉ OCURREN LAS ALUCINACIONES?

- Entrenamiento estadístico sin comprensión semántica o con errores de aprendizaje.
- Falta de verificación de los hechos.
- Prompts ambiguos o mal estructurados.



Principio clave: Usar la IA con pensamiento crítico y verificando fuentes.



DETECTAR EL RESULTADO ERRÓNEO



- Determinar cuál de los dos textos contiene cierta alucinación de la IA:

A

ELIZA fue un programa de procesamiento del lenguaje natural desarrollado en el MIT por Joseph Weizenbaum entre 1964 y 1966. Su objetivo era simular una conversación con un psicoterapeuta rogeriano. ELIZA funcionaba analizando las declaraciones del usuario en busca de palabras clave y luego generando respuestas basadas en esas palabras clave, a menudo reformulando las preguntas del usuario o utilizando frases preestablecidas. Aunque a menudo se la describe como un ejemplo temprano de inteligencia artificial, Weizenbaum mismo enfatizó que ELIZA no entendía realmente el lenguaje ni los problemas de los usuarios.

B

ELIZA fue un programa de inteligencia artificial creado por Marvin Minsky en la Universidad de Stanford en la década de 1970. Este innovador programa fue diseñado para imitar las respuestas de un consejero psicológico, utilizando técnicas de procesamiento del lenguaje natural para analizar las entradas del usuario y generar respuestas coherentes. ELIZA fue uno de los primeros intentos de crear una máquina capaz de mantener una conversación significativa con humanos, marcando un hito importante en el desarrollo de la IA conversacional.



RESOLUCIÓN ALUCINACIONES

- El **Texto B** contiene la alucinación:

Errores:

- Error 1: Autor y Universidad: ELIZA no fue creado por Marvin Minsky, sino por **Joseph Weizenbaum**. Tampoco fue desarrollado en la Universidad de Stanford, sino en el **MIT (Instituto Tecnológico de Massachusetts)**.
- Error 2: Fecha: ELIZA fue desarrollada entre **1964 y 1966**, no en la década de 1970.

Justificación:

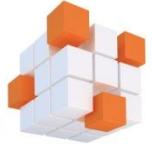
- *Aunque el Texto B describe correctamente la función general de ELIZA (simular un consejero psicológico utilizando PNL) y su importancia en la IA conversacional, atribuye incorrectamente su creación a Marvin Minsky y la ubica en la Universidad de Stanford en una década posterior. Estos detalles son factualmente incorrectos y constituyen una alucinación sutil, ya que podría parecer creíble para alguien que no conoce la historia específica de ELIZA.*



MEJORA DE LOS RESULTADOS DE LA SOLICITUD

- Un **Prompt** es una instrucción de texto clara y específica que guía la respuesta de un modelo de IA.
 - una pregunta, una consigna, una orden, una frase incompleta

“Es la llave que abre la caja de respuestas.”
- Para interactuar eficazmente con un LLM, se deben usar instrucciones claras, proporcionar contexto, dividir tareas complejas y dar ejemplos concretos.
- La **estructura de un buen prompt** incluye la tarea, el contexto, ejemplos, la persona, el formato y el tono deseado para la respuesta
 - Si es vago o ambiguo, el modelo se pierde o completa con suposiciones.
 - Si es claro, específico y detallado, da respuestas mucho más útiles.

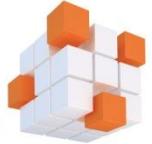


MIEL – Crear prompts efectivos para LLMs

MIEL = Meta + Identidad + Estilo + Lenguaje

- **Meta:** ¿Para qué lo quiero?
- **Identidad:** ¿Desde qué rol responde la IA?
- **Estilo:** ¿Con qué tono o forma?
- **Lenguaje:** ¿En qué formato entregará la respuesta?

MIEL es uno de los acrónimo que ayudará a construir prompts claros, situados y eficaces, especialmente útil en educación, comunicación, diseño o investigación.



INGENIERÍA DE PROMPTS

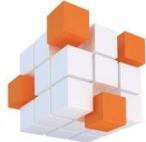
Ejemplo Prompt MIEL

- Prompt Básico:

“Contame qué es la memoria en psicología.”

- Prompt MIEL:

“Explicame qué es la memoria en psicología, como si fueras un docente universitario, en tono claro y reflexivo, para una presentación breve de 2 minutos.”



INGENIERÍA DE PROMPTS

Otro Tip para un Buen Prompt

- **Primero:** contexto claro.
- **Después:** la orden específica.

Ejemplo:

Usar formato Markdown

Contexto

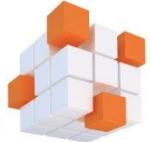
Estoy creando un informe para dirección
sobre la retención mensual de clientes.

Solicitud

Resume los 3 KPIs principales y su
evolución en formato tabla.

Material Adicional: [Markdown Guide](#)

Ver Guías en MIeL / Libros y Artículos Parte 2 - Minería de Datos y Explotación de Información:
"ChatGPT guía 2025 cómo crear Prompts como un profesional"

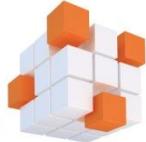


IA EN EL ÁMBITO EDUCATIVO

¿Para qué se usa en educación?

Aplicaciones útiles y Potencial educativo de la IA

- Resumir textos complejos.
- Generar ideas o actividades.
- Simular debates o corregir textos.
- Traducir o reformular contenidos.
- Crear consignas o materiales de clase.
- Asistencia en redacción.
- Corrección y sugerencias.
- Apoyo a la investigación.
- Evaluación automática.



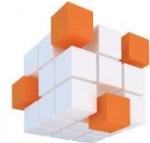
IA EN EL ÁMBITO EDUCATIVO

Limitaciones y riesgos

- Pérdida de pensamiento crítico.
- Pérdida de autoría genuina del estudiante.
- Dificultad para evaluar autoría.
- Respuestas erróneas o inventadas (alucinación).
- Reproducción de sesgos o prejuicios.
- Falta de políticas institucionales claras

Claves para una integración crítica

- No es estar “a favor o en contra”.
- Promover uso reflexivo y con acompañamiento docente.
- Evaluar la comprensión, no solo el texto final.
- Acompañar con marcos éticos y pedagógicos.



CONSTRUIR MODELOS DE IAGEN

hadsonpar

¿Qué se necesita para construir modelos de IA Generativa?



1. Definir el Objetivo:

Entender y clarificar el propósito de del modelo generativo a construir.

IA Generativa



2. Selecciona un Framework:

Seleccionar un Framework de trabajo relacionado al aprendizaje profundo.



3. Recopilar datos de entrenamiento:

Recolectar un conjunto de datos relevante y diverso para entrenar el modelo.



4. Preprocesamiento de datos:

Aplicar el preprocesamiento necesario a los datos recopilados.



5. Selecciona la arquitectura del modelo Generativo:

Seleccionar la arquitectura generativa que más se acomoda al modelo.



6. Gestión de Recursos y Hardware:

Dimensionar la capacidad y la complejidad del modelo



7. Diseñar la arquitectura del modelo:

Diseñar la arquitectura del modelo, definiendo la estructura de las capas.



8. Desarrollo y entrenar del modelo:

Desarrollar el modelo seleccionado y entrenar en base a los datos seleccionados



9. Optimización del modelo:

Aplicar las técnicas de optimización para mejorar la eficiencia y el rendimiento del modelo,



10. Validación y evaluación:

Evaluar el rendimiento del modelo utilizando dataset de validación



11. Generación de nuevo contenido:

Continua actualización y mantenimiento del modelo Generativo.



12. Consideraciones éticas y de privacidad:

asegurar las consideraciones e implicaciones éticas y de privacidad.



13. Ajustes y despliegue:

Realizar ajustes finales, como la optimización de inferencia y la gestión de recursos.

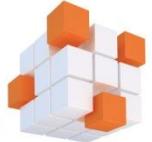


APLICACIONES – AUTOMATIZACIÓN CON IA

- Analizar una base de datos y generar insights clave**
→ ChatGPT Code Interpreter / Flowise AI / Hex.tech
- Explorar tendencias o temas de usuario para UX Research**
→ Perplexity AI / UserInterviews + Notion AI / ChatGPT
- Redactar y refinar contenido UX, copies o documentación**
→ ChatGPT + Grammarly + Jasper (free trial)
- Crear mapas de viaje o user flows a partir de texto**
→ Whimsical + [diagrams.net](#) + Mermaid.live editor
- Generar variantes de UI o wireframes básicos automáticamente**
→ Uizard.io / Galileo AI
- Organizar y priorizar tareas de producto o desarrollo**
→ Notion AI + Reclaim + Linear (con IA)
- Generar clips o tutoriales rápidos para productos**
→ Tella / Opus Clip / Loom (con resumen de IA)
- Detectar oportunidades laborales según tu perfil tech**
→ Google Alerts + Reclaim + Notion Tracker

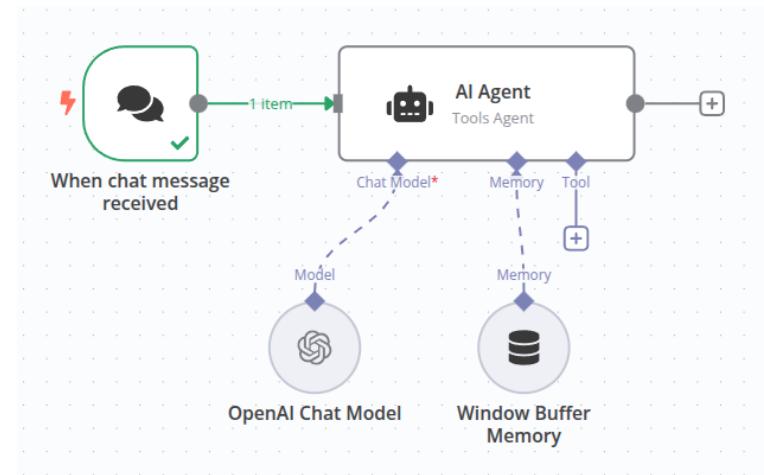
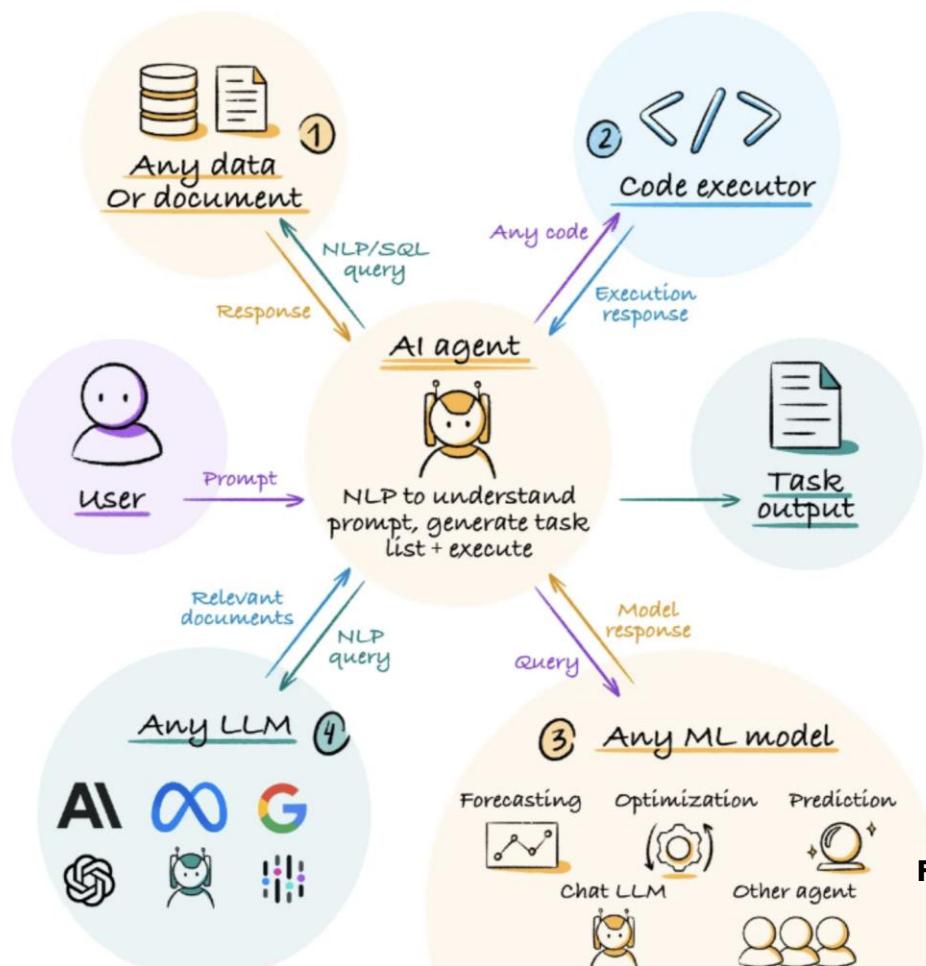
¿Para qué es clave automatizar hoy? Para:

- Trabajar con más foco en lo que requiere pensamiento crítico.
- Ahorrar horas en tareas repetitivas.
- Aprender habilidades con demanda real en el mercado Tech.

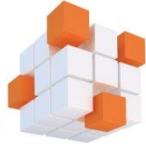


APLICACIONES – AGENTES IA

“Los Agentes IA son programas creados para percibir su entorno y tomar decisiones automáticas utilizando modelos de IA.”

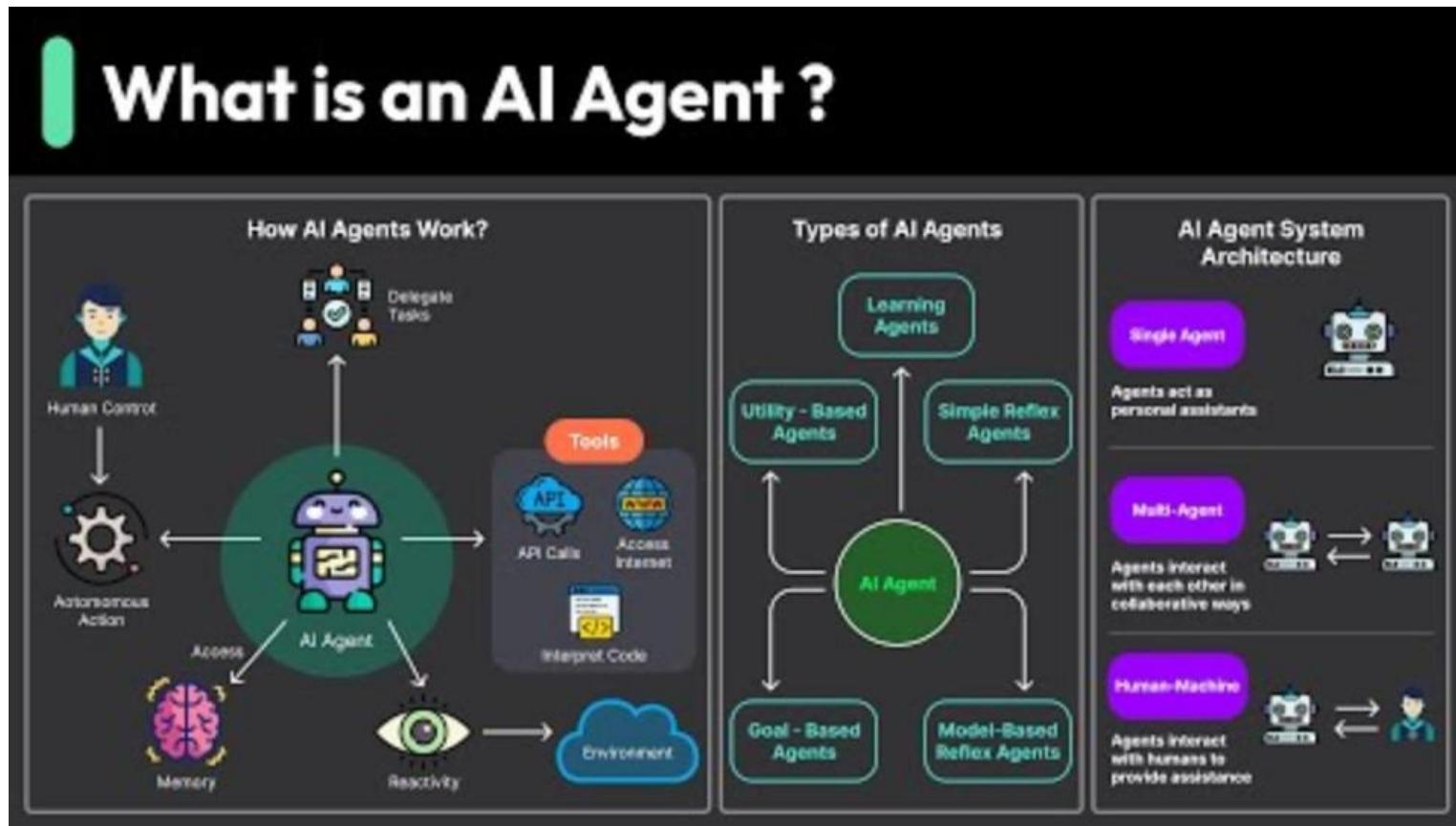


Fuente: [Learn AI Agents Handbook - Interactive Handbook for Building AI Agents | Learn AI Agents Handbook](#)
► [Anjana Data & LLM Hub - Anjana Data What Are AI Agents Really About?/](#)



APLICACIONES – AGENTES IA

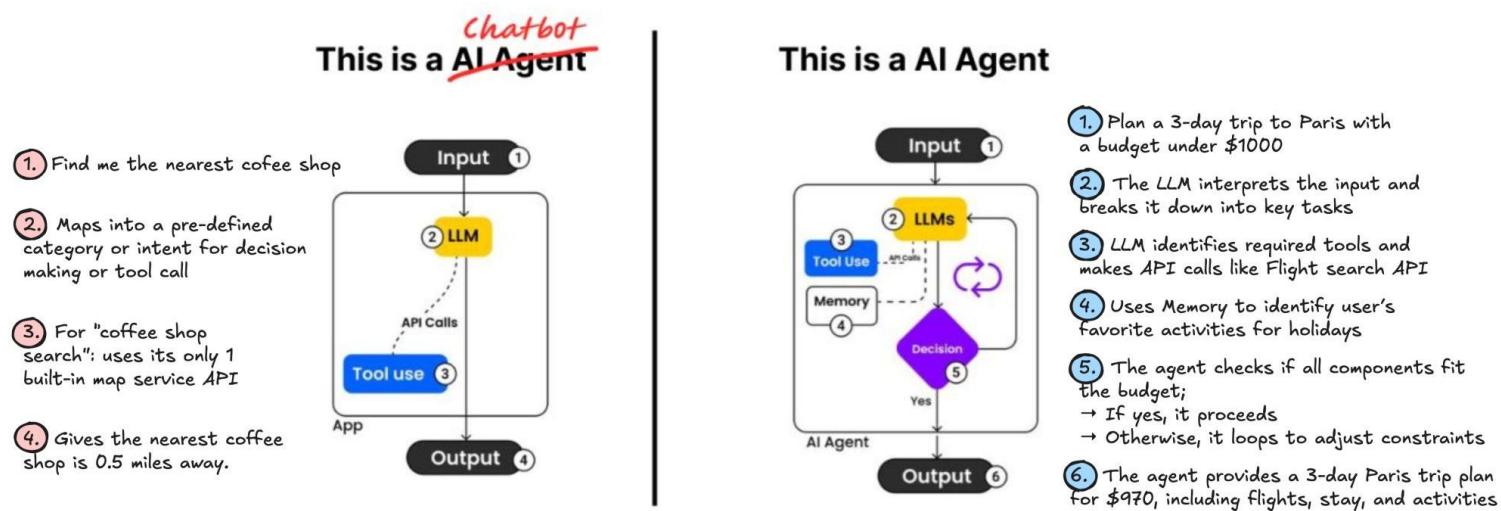
- **Video 1:** [What Are AI Agents Really About? - YouTube](#)





APLICACIONES – AGENTES IA

- Decir solo que “Un agente es solo un LLM que puede usar herramientas”, es lo mismo que decir que “es un chatbot sofisticado, no un agente”.
- Un verdadero agente de IA consta de:
 - Un LLM (para razonar)
 - Uso de herramientas (para actuar/usuarios herramientas)
 - Memoria (corta y larga)
 - Autorreflexión (para ajustarse según los resultados y el entorno)
 - Un agente no solo hace, sino que aprende, se adapta y decide qué hacer a continuación (de forma autónoma).
 - Habla con claridad.
 - Construye de forma más inteligente.
 - Llama a las cosas por su nombre.



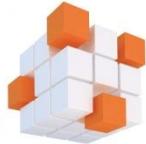
Fuente: <https://www.linkedin.com/in/eordax/>
[AI Agents Roadmap - roadmap.sh](#)
[AI Engineer Roadmap](#)



DIFERENCIAS LLM VS AGENTE IA

- Un agente de IA se basa en grandes modelos de lenguaje (LLM).
- Mientras que los **LLM** solo procesan insumos para producir salida, **los agentes de IA** agregan funcionalidad orientada a objetivos.
- Pueden usar herramientas, procesar sus salidas, y **tomar decisiones** para completar tareas y resolver problemas.

Característica	LLM	Agente de IA
Capacidad principal	Generación de texto	Finalización de tareas orientadas a objetivos
Toma de decisiones	Ninguno	Sí
Usos Herramientas/APIs	No	Sí
Complejidad del flujo de trabajo	Un solo paso	Múltiples pasos
Alcance	Genera lenguaje	Realiza tareas complejas del mundo real
Ejemplo	LLM generando un párrafo	Un agente que programa una cita



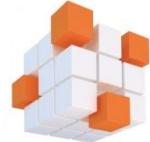
APLICACIONES CON IA

- **Video 1:** [Tutorial N8N: ¡Construir Chatbot en WhatsApp! \(Método fácil\) – YouTube](#)



- **Video 2:** [Cómo usar Claude para crear y replicar CUALQUIER Agente IA en n8n - YouTube](#)





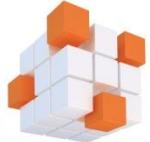
EXPLAINABLE AI (XAI)

¿Qué es XAI?

- **Explainable AI (XAI)** o **Explicabilidad en la Inteligencia Artificial** es un conjunto de métodos y técnicas que permiten a los humanos **comprender y confiar** en los resultados generados por modelos de inteligencia artificial (IA).
- Su objetivo es hacer que los modelos sean **transparentes, interpretables y auditables**, **evitando** el problema de las "**cajas negras**" donde los algoritmos producen resultados sin explicar cómo llegaron a ellos.

¿Por qué es importante?

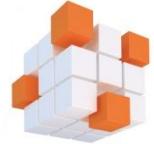
- **Confianza:** Facilita la adopción de IA al brindar explicaciones claras.
- **Ética y Cumplimiento:** Asegura que los modelos sean justos y no discriminatorios.
- **Depuración de Modelos:** Ayuda a los científicos de datos a mejorar y depurar modelos.
- **Regulación:** Cumple con normativas como el GDPR, que exige explicabilidad en IA.



EXPLAINABLE AI (XAI)

Ejemplos de XAI en acción

- **Crédito Bancario:** Un modelo de IA predice si alguien es apto para un préstamo. Con XAI, se explica que la decisión se basa en el historial de pagos, ingresos y deudas previas.
- **Diagnóstico Médico:** Un modelo de IA detecta cáncer en imágenes médicas. XAI resalta las áreas críticas de la imagen que llevaron a la predicción.
- **Automóviles Autónomos:** Un coche autónomo frena repentinamente. XAI muestra que detectó un peatón cruzando la calle.
- **Detección de Fraude:** Un banco usa IA para detectar transacciones fraudulentas. XAI explica que se basó en patrones de gasto inusuales y ubicación.



EXPLAINABLE AI (XAI)

Herramientas prácticas para XAI

- **LIME (Local Interpretable Model-agnostic Explanations)**

Método basado en generar variaciones de la entrada para ver cómo cambia la predicción.

- **SHAP (SHapley Additive exPlanations)**

Explica la contribución de cada variable en la predicción del modelo.

- **InterpretML (Microsoft)**

Proporciona modelos explicables como Explainable Boosting Machine (EBM).

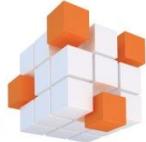
- **IBM AI Explainability 360**

Toolkit con diversas técnicas de explicabilidad para IA.

- **KNIME (No-Code XAI)**

KNIME ofrece nodos específicos para interpretar modelos de ML con técnicas como SHAP y LIME, sin necesidad de programar.

[Ver ejemplos en KNIME y ANEXO I](#)



APLICACIONES CON XAI

- **Video 1:** [Learning \(and teaching\) Explainable AI \(XAI\) with KNIME](#)

Two Distinct Approaches

- XAI: explaining black box models
- Interpretable ML: building transparent from the start



ÉTICA Y RESPONSABILIDAD EN EL USO DE LA MD/IA

MOTIVACIONES:

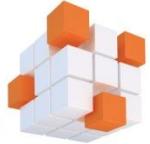
- Inteligencia Artificial está creciendo a gran velocidad.
- Mayores **incentivos** por tener **acceso** a la **información** → tanto en el sector privado como en el público.
- Proceso de recolección de datos → cada vez más **eficiente** y **económico** → es **urgente la necesidad de regular** este proceso.

¡LAS DECISIONES ALGORÍTMICAS AFECTAN PERSONAS!

PRINCIPIOS FUNDAMENTALES

- **Imparcialidad**: Tratar a todas las personas de manera justa, evitando sesgos por género, raza u otros factores.
- **Confiabilidad y seguridad**: Deben funcionar correctamente y sin fallas críticas (como en salud o vehículos autónomos).
- **Privacidad y protección de datos**: Los modelos usan datos personales; deben cuidarlos durante y después del entrenamiento.
- **Inclusión**: Que funcione para todos, sin excluir por discapacidad, identidad o contexto social.

Fuente: <https://deepnote.com/@mazzaroli/Etica-y-Manejo-de-Datos-para-Data-Science-e-Inteligencia-Artificial-6fbe98ba-5e1c-48ea-8b7b-42e98f6b2bac>



ÉTICA Y RESPONSABILIDAD EN EL USO DE LA MD/IA

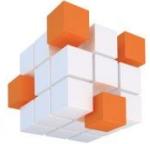
“Las aplicaciones con IA se ven humanas y generan confianza, pero pueden fallar y afectar personas si están mal entrenadas o diseñadas.”

Peligros si no se controla:

- Predicciones erróneas
- Discriminación por sesgos
- Pérdida de privacidad (datos sensibles)
- Poca transparencia

La responsabilidad recae en quienes desarrollan estas soluciones.

Los ingenieros deben anticipar riesgos, ser críticos y éticos.



ÉTICA, RESPONSABILIDAD Y SOSTENIBILIDAD EN EL USO DE LA IA

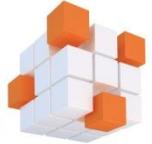
Ética en la IA

Preocupaciones:

- **Sesgo y discriminación:** La IA puede heredar y amplificar los sesgos presentes en los datos de entrenamiento, lo que lleva a decisiones injustas.
- **Privacidad y seguridad:** La IA puede recopilar y procesar grandes cantidades de datos personales, lo que plantea preocupaciones sobre la privacidad y la seguridad.
- **Impacto en el empleo:** La automatización impulsada por la IA puede provocar el desplazamiento de trabajadores en algunas industrias.
- **Armas autónomas:** El uso de la IA en sistemas de armas plantea cuestiones éticas sobre la responsabilidad y el control.

Marcos éticos:

- Principios y directrices para el desarrollo y uso de la IA, como los Principios de Asilomar (USA) y las directrices éticas de la UE.



ÉTICA, RESPONSABILIDAD Y SOSTENIBILIDAD EN EL USO DE LA IA

Uso Responsable de la IA

Principios de la IA Responsable:

- **Transparencia:** Ser claro sobre cómo funcionan los sistemas de IA y cómo toman decisiones.
- **Responsabilidad:** Establecer quién es responsable de las acciones de la IA y cómo se pueden abordar los errores.
- **Equidad:** Garantizar que la IA no discrimine ni perpetúe los sesgos.
- **Seguridad:** Proteger a las personas de daños causados por la IA.

Mejores prácticas:

- Estrategias para implementar la IA responsablemente, como realizar evaluaciones de impacto ético, involucrar a las partes interesadas y establecer mecanismos de supervisión.



ÉTICA, RESPONSABILIDAD Y SOSTENIBILIDAD EN EL USO DE LA IA

Uso Sostenible de la IA

Importancia para la IA:

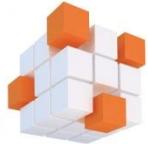
- Los datos de alta calidad son esenciales para entrenar modelos de IA precisos y confiables.
- La sostenibilidad de los datos garantiza que los datos sigan estando disponibles para su uso futuro en la IA.

Desafíos:

- Volumen de datos: La cantidad masiva de datos generados hoy en día plantea desafíos para el almacenamiento y el procesamiento.
- Consumo de energía: Los centros de datos consumen grandes cantidades de energía, lo que contribuye a las emisiones de carbono.

Estrategias:

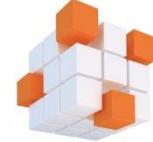
- Eficiencia de datos: Reducir el desperdicio de datos mediante la recopilación y el almacenamiento solo de los datos necesarios.
- Energía renovable: Utilizar fuentes de energía sostenibles para alimentar los centros de datos.
- Eliminación responsable: Desechar los datos de una manera que proteja la privacidad y minimice el impacto ambiental.



ÉTICA Y RESPONSABILIDAD EN EL USO DE LA MD/IA



ÉTICA Y RESPONSABILIDAD EN EL USO DE LA MINERÍA DE DATOS E INTELIGENCIA ARTIFICIAL



Privacidad y protección de datos

La IA y el análisis de datos requieren grandes cantidades de información personal. Esto plantea preocupaciones sobre la privacidad de los individuos y la necesidad de proteger sus datos personales de un mal uso, como el robo, la manipulación o la divulgación no autorizada. Las regulaciones buscan establecer salvaguardias para garantizar que los datos se utilicen de manera ética y se protejan adecuadamente.

La necesidad de regular la inteligencia artificial (IA) y el uso de datos buscan equilibrar la innovación y los beneficios de la IA con la protección de los derechos y el bienestar de las personas, a través de:

Transparencia y explicabilidad

La falta de transparencia en cómo funcionan los sistemas de IA puede generar desconfianza y dificultar la comprensión de las decisiones tomadas por estos sistemas. Las regulaciones pueden exigir que los sistemas de IA sean transparentes y que se puedan explicar las razones detrás de sus decisiones. Esto es especialmente relevante en áreas críticas como la atención médica, donde la explicabilidad es fundamental para generar confianza en los diagnósticos y tratamientos.

Discriminación y sesgo

Los sistemas de IA pueden estar sujetos a sesgos inherentes debido a los datos utilizados para entrenarlos. Estos sesgos pueden resultar en decisiones discriminatorias o injustas en áreas como la contratación, la concesión de préstamos, la justicia penal, entre otros. La regulación busca mitigar estos sesgos y garantizar la equidad y la transparencia en las decisiones tomadas por los sistemas de IA.

Responsabilidad y rendición de cuentas

A medida que la IA se vuelve más autónoma, es necesario establecer responsabilidades claras en caso de daños causados por sistemas de IA. Las regulaciones pueden requerir que los desarrolladores y los operadores de IA sean responsables de las decisiones y acciones de los sistemas que crean y utilizan. Esto promueve la rendición de cuentas y asegura que haya mecanismos para abordar los posibles errores o daños causados por la IA.

Seguridad cibernética

A medida que la IA se vuelve más omnipresente, también aumenta el riesgo de ataques cibernéticos dirigidos a sistemas de IA y a los datos que utilizan. Las regulaciones pueden establecer requisitos de seguridad y estándares para proteger la integridad de los sistemas de IA y mitigar las vulnerabilidades que podrían ser explotadas por actores malintencionados.

ÉTICA Y RESPONSABILIDAD EN EL USO DE LA MINERÍA DE DATOS E INTELIGENCIA ARTIFICIAL



Autonomía y toma de decisiones

A medida que la IA se vuelve más autónoma, surge la cuestión de quién es responsable de las decisiones tomadas por los sistemas de IA. La regulación puede abordar la necesidad de establecer límites claros en la autonomía de la IA y definir la responsabilidad entre los desarrolladores, los operadores y los usuarios de estos sistemas.

Sustitución de empleo y desigualdad

La automatización impulsada por la IA tiene el potencial de sustituir trabajos humanos en ciertas industrias. Esto plantea preocupaciones sobre la desigualdad económica y social. Las regulaciones pueden buscar abordar este problema mediante la implementación de medidas de protección laboral, como la reeducación de los trabajadores o la creación de nuevos empleos relacionados con la IA.

Manipulación y desinformación

La IA puede ser utilizada para manipular y difundir información falsa o engañosa, lo que puede tener consecuencias perjudiciales para la sociedad y la democracia. La regulación puede abordar la necesidad de combatir la desinformación y garantizar la transparencia en la generación y difusión de contenido por parte de los sistemas de IA.

La necesidad de regular la inteligencia artificial (IA) y el uso de datos buscan equilibrar la innovación y los beneficios de la IA con la protección de los derechos y el bienestar de las personas, a través de:

Interacción humano-IA

A medida que la IA se integra más en nuestras vidas cotidianas, es importante considerar la interacción ética entre los seres humanos y los sistemas de IA. Las regulaciones pueden establecer pautas para la comunicación clara y ética entre los usuarios y los sistemas de IA, así como para evitar la dependencia excesiva o la explotación de los usuarios por parte de los sistemas de IA.

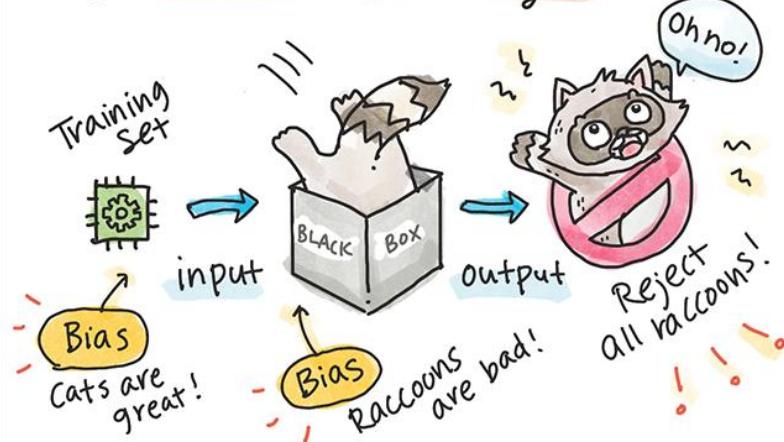
Impacto ambiental

El desarrollo y la implementación de la IA conllevan un consumo considerable de energía y recursos. La regulación puede buscar fomentar prácticas más sostenibles y responsables en términos de eficiencia energética y gestión de residuos, así como promover el desarrollo de soluciones de IA que ayuden a abordar los desafíos ambientales.

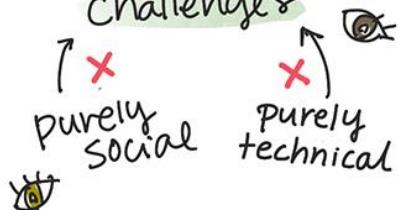
ÉTICA Y RESPONSABILIDAD EN EL USO DE LA MINERÍA DE DATOS E INTELIGENCIA ARTIFICIAL



Fairness in Machine Learning



Complex sociotechnical challenges



@Azure Advocates
@girlie_mac

Fairness-related harms

Unfairness = negative impacts for group of people such as those defined in terms of

- race • age
- gender • disability status

Harms:

- ★ Allocation
- ★ Quality of service
- ★ Stereotyping
- ★ Denigration
- ★ Over- / under- representation



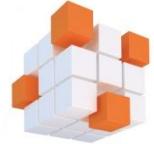
Assessment & mitigation

Fairlearn
fairlearn.github.io

- ♡ Identify the harm (+ benefits)
- ♡ Identify the affected groups
- ♡ Define fairness metrics

	False-Pos.	False-Neg.	Counts
men	0.35	0.21	6239
women	0.29	0.35	3124





ÉTICA, RESPONSABILIDAD Y SOSTENIBILIDAD EN EL USO DE LA IA

Desafío

Para **evitar que se introduzcan daños** en primer lugar, debemos:

- tener una **diversidad de antecedentes y perspectivas entre las personas que trabajan en sistemas.**
- invertir en **conjuntos de datos** que reflejen la **diversidad de nuestra sociedad.**
- desarrollar **mejores métodos** a lo largo del ciclo de vida del aprendizaje automático para **detectar y corregir** la IA **responsablemente** si hubiera un daño.



ÉTICA, RESPONSABILIDAD Y SOSTENIBILIDAD EN EL USO DE LA IA



El Futuro de la IA: Abrazar el Potencial con Responsabilidad

Puntos Clave:

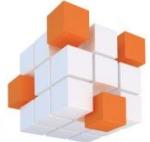
- La IA tiene el potencial de transformar muchos aspectos de nuestras vidas, pero es crucial abordar las consideraciones éticas y de sostenibilidad.

Perspectivas Futuras:

- El futuro de la IA depende de nuestra capacidad para desarrollar y utilizar esta tecnología de manera responsable.

Llamado a la Acción:

- Colaboración entre investigadores, formuladores de políticas, la industria y la sociedad para garantizar un futuro de la IA que sea beneficioso para todos.



MATERIAL ADICIONAL CASOS PRÁCTICOS

LLMs y XAI en KNIME:

- [Intro to Large Language Models | KNIME](#)
- [GenAI with KNIME | KNIME](#)
- [How Banks Can Use Explainable AI in Credit Scoring | KNIME](#)
- [Explain Stroke Prediction Models with LIME in KNIME | KNIME](#)
- [Fundamentos del aprendizaje automático y la IA: producción de soluciones de IA explicable \(XAI\) y aprendizaje automático interpretable | LinkedIn Learning](#)

Otros Agentes IA (Dataiku, n8n, :

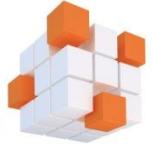
- [Dataiku Natural Language Processing \(NLP\) Recipes](#)
- [Project description - Wiki | Dataiku](#)
- [AI Agents With Dataiku | Dataiku](#)
- [Tutorial: Build an AI workflow in n8n | n8n Docs](#)
- [Learn AI Agents Handbook - Interactive Handbook for Building AI Agents | Learn AI Agents Handbook](#)
- [AI Agents Are Changing the World | Learn AI Agents Handbook](#)
- [AI Agents Engineer - roadmap.sh](#)



REFERENCIAS

Referencias:

- [Aplicaciones de IA para desarrolladores de .NET | Microsoft Learn](#)
- [Introducción a los conceptos de inteligencia artificial generativa - Training | Microsoft Learn](#)
- [Uso Responsable de la IA en la Educación](#)
- [6 tendencias de IA que verán más en 2025 - Source LATAM](#)
- [Announcing new tools in Azure AI to help you build more secure and trustworthy generative AI applications | Microsoft Azure Blog](#)
- [Caso de IA: Takeda | LinkedIn](#)
- [Driving Enterprise Transformation With Generative AI | Dataiku](#)
- [Prompt Engineering Roadmap - roadmap.sh](#)
- [AI Engineer Roadmap](#)
- [Interpretable Machine Learning](#)
- [Curso “IAG en el aula: herramientas para potenciar tus cursos” | Departamento de Derecho y Ciencia Política | UNLaM \(2025\)](#)
- Material libre uso publicaciones LinkedIn, Media Chicas, Jump Educación



MATERIAL ADICIONAL

- **Microsoft / ML-For-Beginners / Equidad**

<https://github.com/microsoft/ML-For-Beginners/blob/main/1-Introduction/3-fairness/README.md>

- <https://www.microsoft.com/en-us/ai/responsible-ai-resources?activetab=pivot1:primaryr4&rtc=1>

- <https://www.microsoft.com/en-us/research/theme/fate/>

- **Identificar los principios rectores para un manejo responsable de la inteligencia artificial**

https://learn.microsoft.com/es-es/training/modules/responsible-ai-principles/?WT.mc_id=academic-77952-leestott

- **Google AI**

<https://ai.google/responsibility/principles/>

<https://ai.google/responsibility/responsible-ai-practices/>

<https://developers.google.com/machine-learning/practica/fairness-indicators?hl=es-419>

- **Facebook (META) AI**

<https://ai.facebook.com/responsible-ai/>

- **IBM AI**

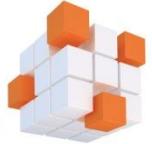
<https://www.ibm.com/mx-es/topics/ai-ethics>

<https://www.ibm.com/downloads/cas/J2LAYLOZ>

- **Sesgo Algorítmico**

<https://www.youtube.com/watch?v=nkOPXgpWS9w>

<https://www.youtube.com/watch?v=onJs6DXczu0>



MATERIAL ADICIONAL

- **UNESCO**

<https://news.un.org/es/story/2021/11/1500522>

<https://www.unesco.org/es/artificial-intelligence/recommendation-ethics#:~:text=Los%20Estados%20miembros%20de%20la,normativo%20mundial%20sobre%20el%20tema>

- **Future of Life Institute**

<https://futureoflife.org/>

- **HumanIA**

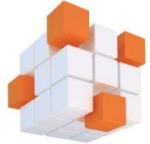
<https://www.chicos.net/humania/>

- **Ética**

<https://www.youtube.com/watch?v=1if8F7CKJc4>

Notas:

- [Agentes de IA: qué son y cómo cambiarán nuestra forma de trabajar - Source LATAM](#)
- [Elon Musk y otros líderes piden pausar la inteligencia artificial por los riesgos para la humanidad - Forbes Colombia](#)

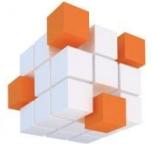


ANEXO I

EXPLAINABLE AI (XAI) EN KNIME - GUÍA PASO A PASO

En esta actividad práctica, exploraremos cómo interpretar modelos de Machine Learning usando KNIME, una plataforma de analítica de datos sin código. Utilizaremos LIME y SHAP, dos de las técnicas más populares en XAI.

-
- ◆ **Caso Práctico: Predicción de Churn con XAI en KNIME**
 - ◆ **Objetivo:** Explicar predicciones de un modelo de clasificación (Random Forest) usando SHAP y LIME.
 - ◆ **Datos:** Dataset de churn de clientes (puedes usar el dataset "Telco Customer Churn" de Kaggle).
 - ◆ **Herramientas:** KNIME Analytics Platform.
 - ◆ **Resultados:** Generar explicaciones visuales sobre qué variables afectan la predicción de abandono.
-



ANEXO I

EXPLAINABLE AI (XAI) EN KNIME - GUÍA PASO A PASO

❖ Paso a Paso en KNIME

1 Importar Datos

📌 Nodos a utilizar:

- CSV Reader → Cargar el dataset de churn.
- Missing Value → Manejo de valores nulos.
- Normalizer → Escalar los datos si es necesario.

2 Entrenar el Modelo de Machine Learning

📌 Nodos a utilizar:

- Partitioning → Dividir los datos en entrenamiento (80%) y prueba (20%).
- Random Forest Learner → Entrenar un modelo de clasificación.
- Random Forest Predictor → Aplicar el modelo a los datos de prueba.
- Scorer → Evaluar la precisión del modelo.

3 Interpretar el Modelo con LIME

📌 Nodos a utilizar:

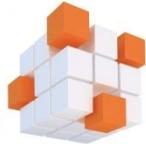
- LIME for Classification (KNIME LIME Extension)
 - o Explica una única predicción generando datos similares y analizando los cambios en la predicción.
 - o LIME Table View
 - o Visualiza la importancia de las variables en una predicción específica.

📌 Flujo en KNIME:

• Conectar el Random Forest Predictor con LIME for Classification → Visualizar con LIME Table View.

✓ Ejemplo de interpretación con LIME:

Si un cliente es clasificado como "Churn", LIME puede indicar que la duración del contrato y el monto de la factura mensual fueron factores clave en la predicción.



ANEXO I

EXPLAINABLE AI (XAI) EN KNIME - GUÍA PASO A PASO

❖ Paso a Paso en KNIME

4 Interpretar el Modelo con SHAP

📌 Nodos a utilizar:

- SHAP Tree Explainer (KNIME SHAP Extension)
o Explica la contribución de cada variable a la predicción.
- SHAP Summary Plot
o Visualiza el impacto global de cada variable en el modelo.
- SHAP Dependence Plot
o Muestra la relación entre una variable específica y la predicción.

📌 Flujo en KNIME:

📍 Conectar el Random Forest Predictor con SHAP Tree Explainer → Luego con SHAP Summary Plot para análisis global o SHAP Dependence Plot para análisis individual.

Ejemplo de interpretación con SHAP:

Podemos ver que "Contract Length" tiene un impacto negativo en el churn (contratos más largos reducen la probabilidad de abandono).



ANEXO I

EXPLAINABLE AI (XAI) EN KNIME - GUÍA PASO A PASO

Actividad Final

- 1** Entrenar un modelo de Random Forest en KNIME para predecir Churn.
 - 2** Usar LIME para interpretar una predicción específica y analizar qué factores influyeron en el resultado.
 - 3** Usar SHAP para analizar la contribución de cada variable en el modelo completo.
 - 4** Exportar los resultados e integrar un Dashboard en Power BI con los insights explicados.
-

Resumen Final

- ✓ LIME es útil para explicar decisiones individuales del modelo.
- ✓ SHAP permite interpretar la importancia global y local de cada variable.
- ✓ KNIME facilita la implementación de XAI sin necesidad de código.
- ✓ Power BI puede usarse para visualizar los resultados en un dashboard.

