

Fundamentos de Análisis de Datos

Unidad 1

Fundamentos de ciencia de
datos

Agenda

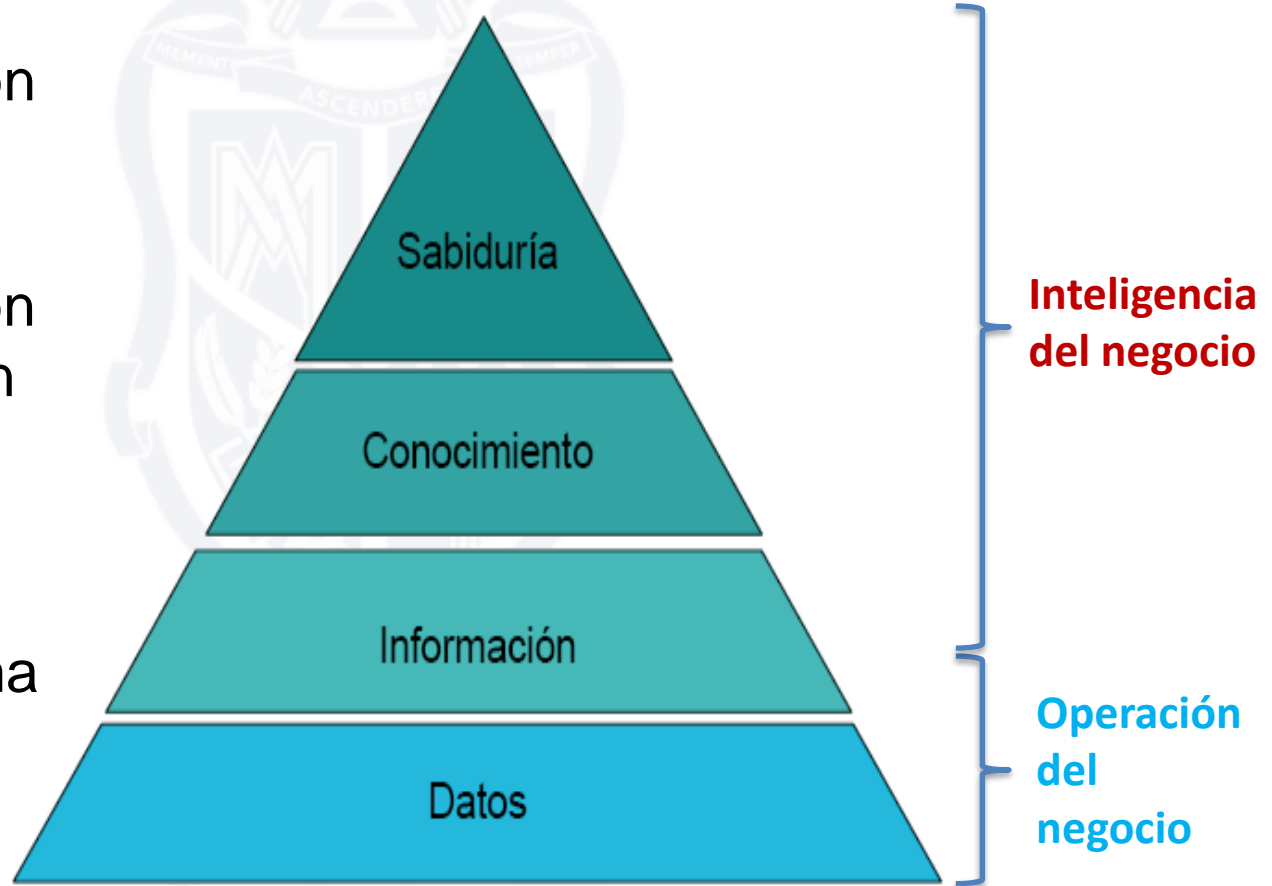
- Conceptos preliminares
 - Inteligencia de negocios
 - BigData
- Minería de Datos
- Ciencia de Datos



Conceptos preliminares

INTELIGENCIA DE NEGOCIOS

- Transformación de datos en información.
- Transformación de información en conocimiento.
- Orientación a una mejor toma de decisiones de negocio.



Conceptos preliminares

INTELIGENCIA DE NEGOCIOS

"Es el proceso de analizar y transformar los datos operacionales de una organización en un repositorio accesible con información de mucho valor, y una adecuada distribución de la información en la manera mas conveniente a las personas correctas, en el momento preciso y en forma oportuna para tomar mejores decisiones"

(Dan Pratte)

Conceptos preliminares

INTELIGENCIA DE NEGOCIOS



Brindar a los **empleados** de una organización la **información correcta** en el **momento** y **formato adecuado** para la **toma de decisiones** y cumplimiento de los objetivos empresariales.

Conceptos preliminares

INTELIGENCIA DE NEGOCIOS

Gestión del rendimiento del negocio (BPM)

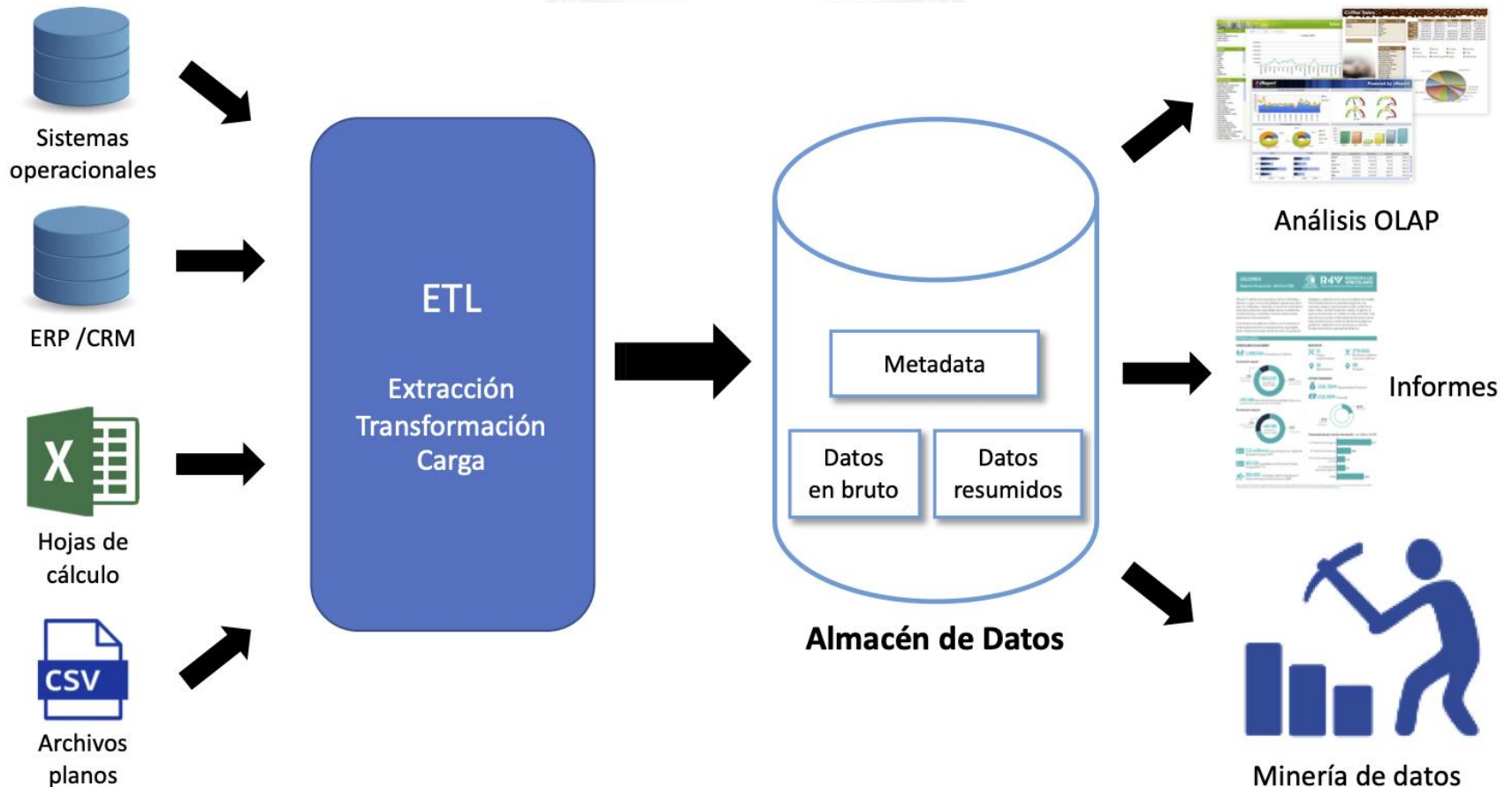


3

4

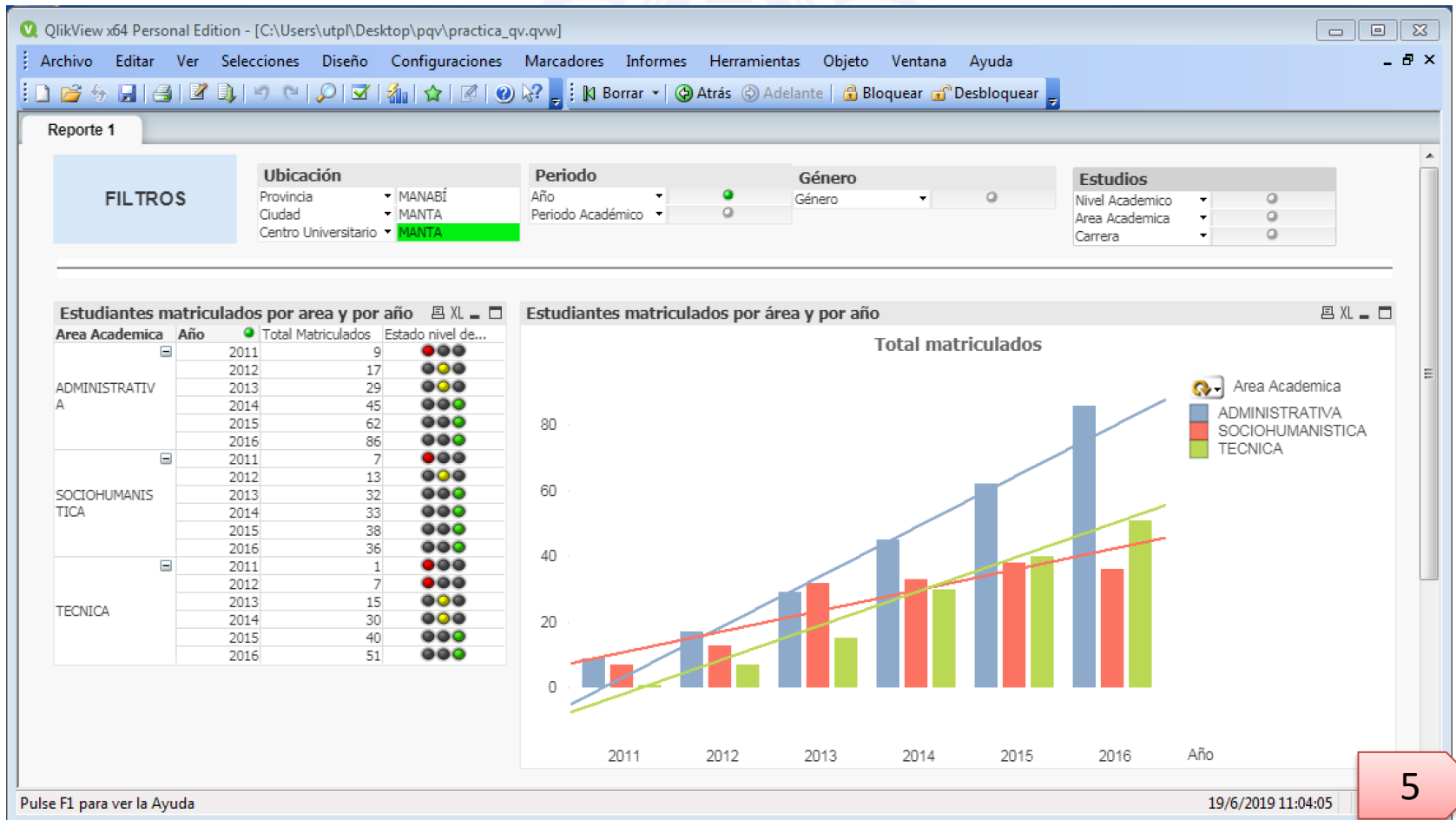
Conceptos preliminares

INTELIGENCIA DE NEGOCIOS



Conceptos preliminares

INTELIGENCIA DE NEGOCIOS



Conceptos preliminares

BIG DATA

Esto sucede en Internet en un minuto

Estimación de la cantidad de actividades y datos generados online en un minuto en 2020



Terabytes o petabytes de datos se producen en nuestras redes en el Internet y en sectores como negocios, educación, ciencia e ingeniería, medicina y casi cualquier otro aspecto de la vida cotidiana

Conceptos preliminares

BIG DATA

“El concepto de big data se puede definir como activos de información caracterizados por su alto volumen, variedad, y velocidad, que demandan soluciones innovadoras y eficientes de procesamiento para la mejora del conocimiento y toma de decisiones en las organizaciones”



Gartner, 2001

Conceptos preliminares

BIG DATA

“Conjunto de arquitecturas y herramientas informáticas destinadas a la manipulación, gestión y análisis grandes volúmenes de datos desde todo tipo de fuentes, diseñadas para extraer valor y beneficio de los mismos, con una amplia variedad en su naturaleza, mediante procesos que permitan capturar, descubrir y analizar información a alta velocidad y con un coste reducido”

UTPL
UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA *García et al., 2018*

Conceptos preliminares

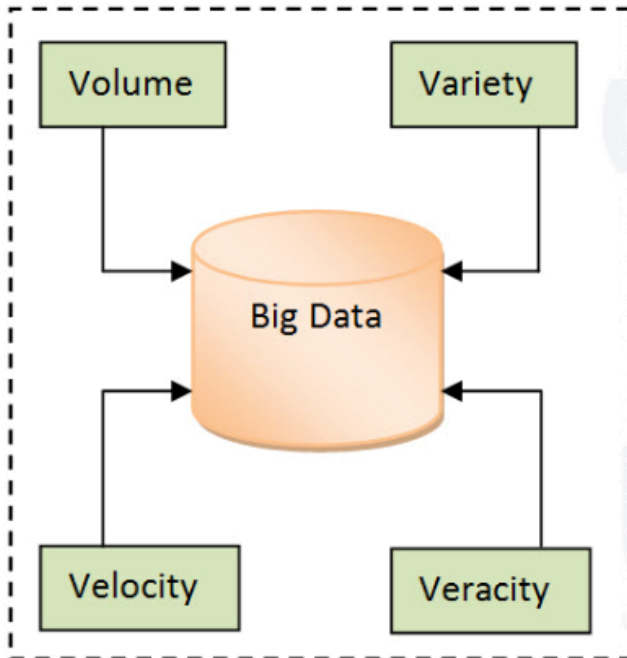
BIG DATA

- Corresponde a conjuntos de información enormes, que superan la capacidad del software de gestión de bases de datos convencionales para procesar datos en tiempo razonable.
- Los retos con este tipo de tecnologías se centran en la captura, almacenamiento, búsqueda, compartición, análisis y visualización de los datos

Conceptos preliminares

BIG DATA

En el marco de Big Data actualmente se viene imponiendo la terminología de las 4 V's para resumir las características fundamentales:



4 V's

- Volume** Grandes volúmenes de datos (petabytes, exabytes, etc.)
- Variety** No limitado a información relacional, puede incluir datos no estructurados
- Velocity** Los datos se deben procesar y analizar rápidamente
- Veracity** Veracidad, adhesión a la verdad, precisión.

Conceptos preliminares

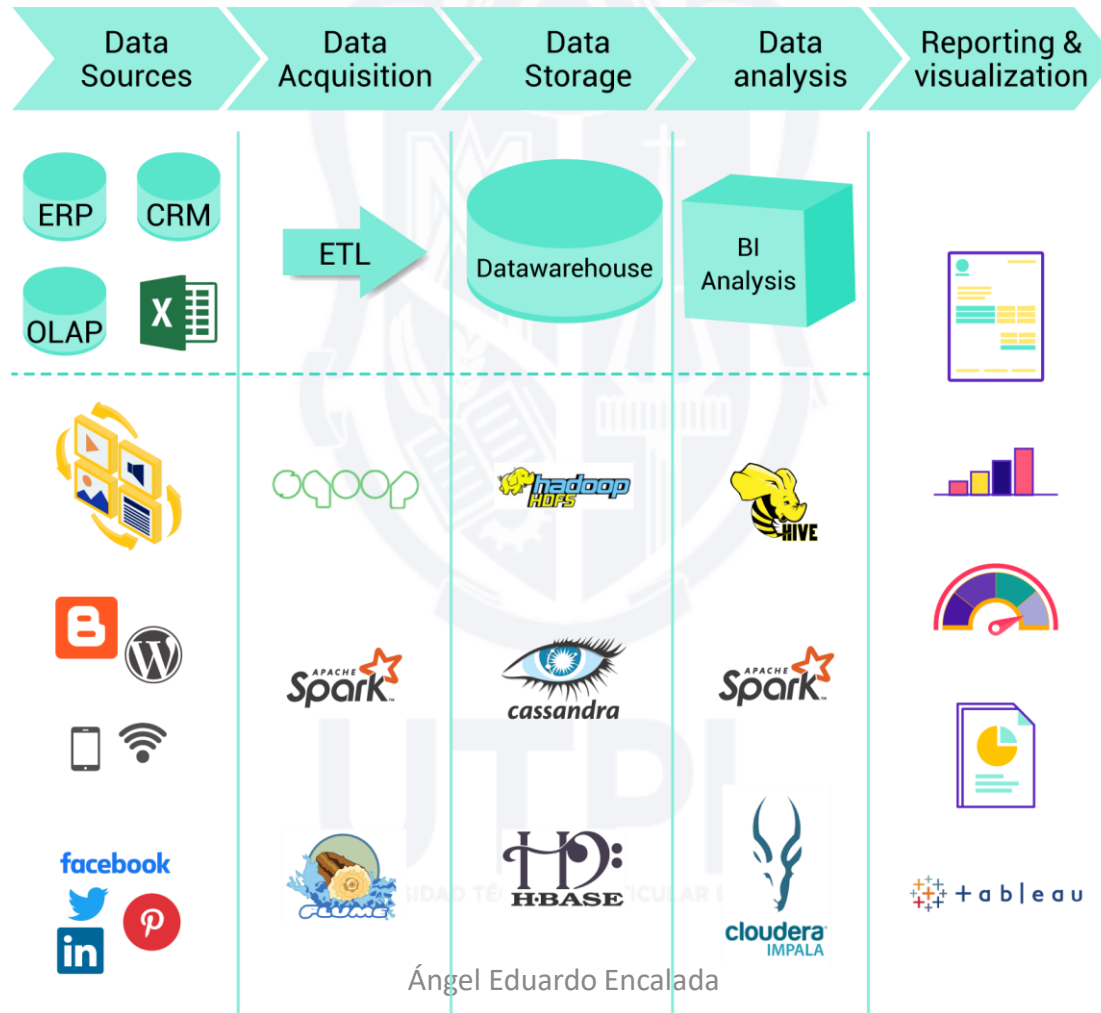
BIG DATA

7 V's

Volume	Grandes volúmenes de datos (petabytes, exabytes, etc.)
Variety	No limitado a información relacional, puede incluir datos no estructurados
Velocity	Los datos se deben procesar y analizar rápidamente
Veracity	Veracidad, adhesión a la verdad, precisión.
Variability	Datos cuyo significado cambia constantemente
Visualisation	Presentar los datos de forma comprensible
Value	Extraer información para la toma de decisiones

Conceptos preliminares

BIG DATA



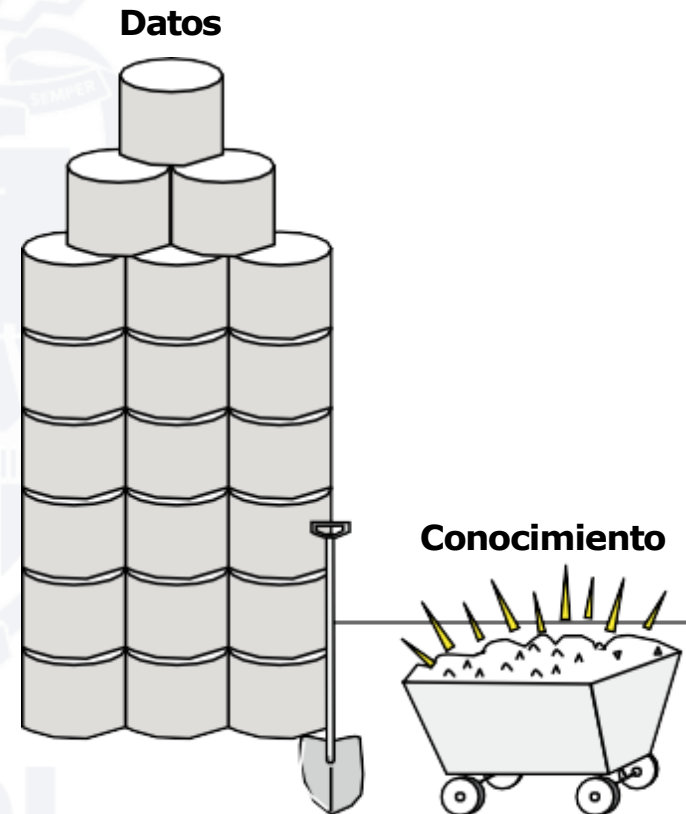
Minería de Datos

Minería de datos

La minería de datos es el conjunto de técnicas que permiten extraer información y descubrir patrones interesantes y conocimiento útil y comprensible desde grandes bases de datos.

Analogía

La minería de datos es similar al proceso de extracción de oro que requiere remover grandes cantidades de roca o arena (que equivaldría a los datos) para obtener el metal puro (que equivaldría al conocimiento).



Minería de Datos

TÉCNICAS

Clasificación

Agente bancario: ¿Debo ofrecer un préstamo a un cliente?

Regresión

Supervisor de fábrica: ¿Cuántas fallas para modelo X se esperan cada mes?

Reglas de asociación

Gerente retail: Clientes que compran pañales ¿Compran también cerveza?

Agrupamiento

Gerente de Talento Humano: ¿Qué tipo de empleados hay en la empresa?

El proceso KDD

Definición

KDD (*Knowledge Discovery in Databases*) o descubrimiento de conocimiento en bases de datos, es un proceso novedoso y no trivial de identificación de patrones de los datos para la obtención de nuevo conocimiento.

KDD es el motor de la Ciencia de Datos



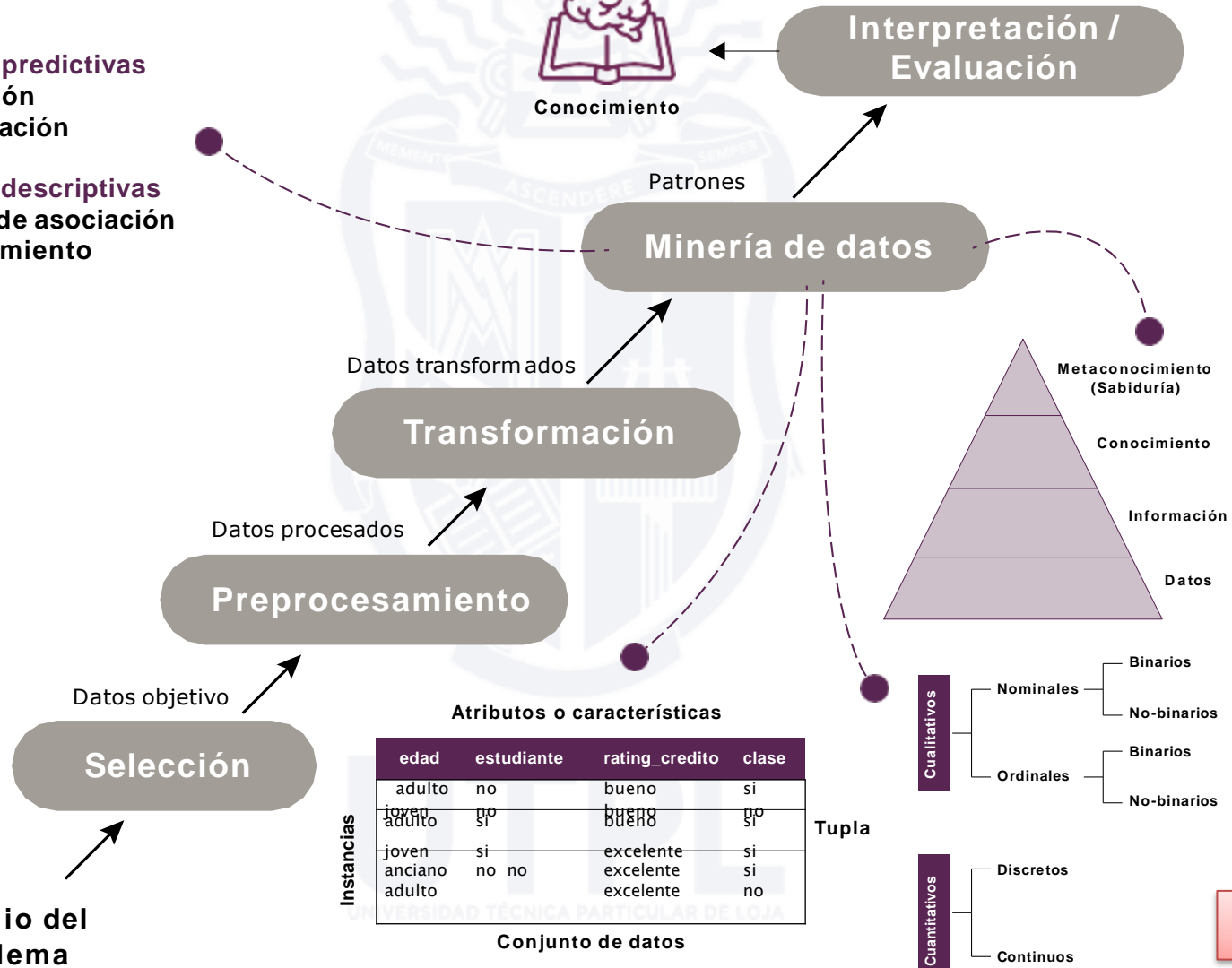
Etapas del proceso KDD

Técnicas predictivas

- Regresión
- Clasificación

Técnicas descriptivas

- Reglas de asociación
- Agrupamiento



Etapas del proceso KDD

Dominio del problema

Es un paso preliminar que desarrolla una comprensión del problema que se enfrenta. Establece las preguntas para entender lo que debería hacerse con los datos, y delimita claramente los objetivos del proceso a realizar.

También realiza la **integración y recopilación de datos** reconociendo las fuentes de información más importantes con las se trabajará en el futuro.

La salida de esta etapa es el **dominio del problema y los datos en “crudo”**.



Etapas del proceso KDD

1 Selección

Incluye la selección, extracción o muestreo de un **subconjunto de variables** desde diversas fuentes, y a partir de las cuales el descubrimiento de conocimiento debe ser realizado.

La salida de esta etapa es un **conjunto de datos objetivo**.

Etapas del proceso KDD

2 Preprocesamiento

Consiste eliminar imperfecciones propias de las fuentes originales de datos y, su objetivo es mejorar la fiabilidad de los datos.

También involucra tareas de **enriquecimiento** y el **análisis exploratorio de datos** (**EDA: *Exploratory data analysis***) que incluye métodos estadísticos para explorar la distribución e identificar características de los datos.

Tareas típicas: limpieza de datos (eliminar duplicados, corregir errores, tratar valores faltantes, filtrar outliers), integración de fuentes, resolver inconsistencias.

La salida de esta etapa es un **conjunto de datos pre-procesado**.

Etapas del proceso KDD

3 Transformación

Se mejoran los datos pre-procesados, con el fin de encontrar características útiles para representar los datos, dependiendo de la meta del proceso.

El objetivo es reformatear, reducir o enriquecer los datos para adaptarlos a los algoritmos de minería de datos.

Tareas típicas: Normalización o estandarización, Codificación de variables categóricas, Agregación y resumen, Generación de atributos derivados, Discretización, Reducción de dimensionalidad

La salida de esta etapa es un **conjunto de datos transformado**.

Etapas del proceso KDD

4 Minería de datos

El objetivo de la etapa es la búsqueda y descubrimiento de patrones insospechados y de interés.

a) Elección de la tarea de minería de datos

Tareas como regresión, clasificación, agrupamiento (*clustering*), asociaciones u otras.

b) Selección e implementación del algoritmo

Método/algoritmo para ser utilizado en la búsqueda de patrones en los datos. Cada algoritmo tiene parámetros y características que, suelen ser exitosas o no dependiendo del tipo de datos (categóricos, numéricos) a utilizar.

La salida de esta etapa son los **patrones reconocidos por los algoritmos**.

Etapas del proceso KDD

5 Interpretación/Evaluación

Se interpretan los patrones descubiertos y posiblemente se retorna a las anteriores etapas para posteriores iteraciones.

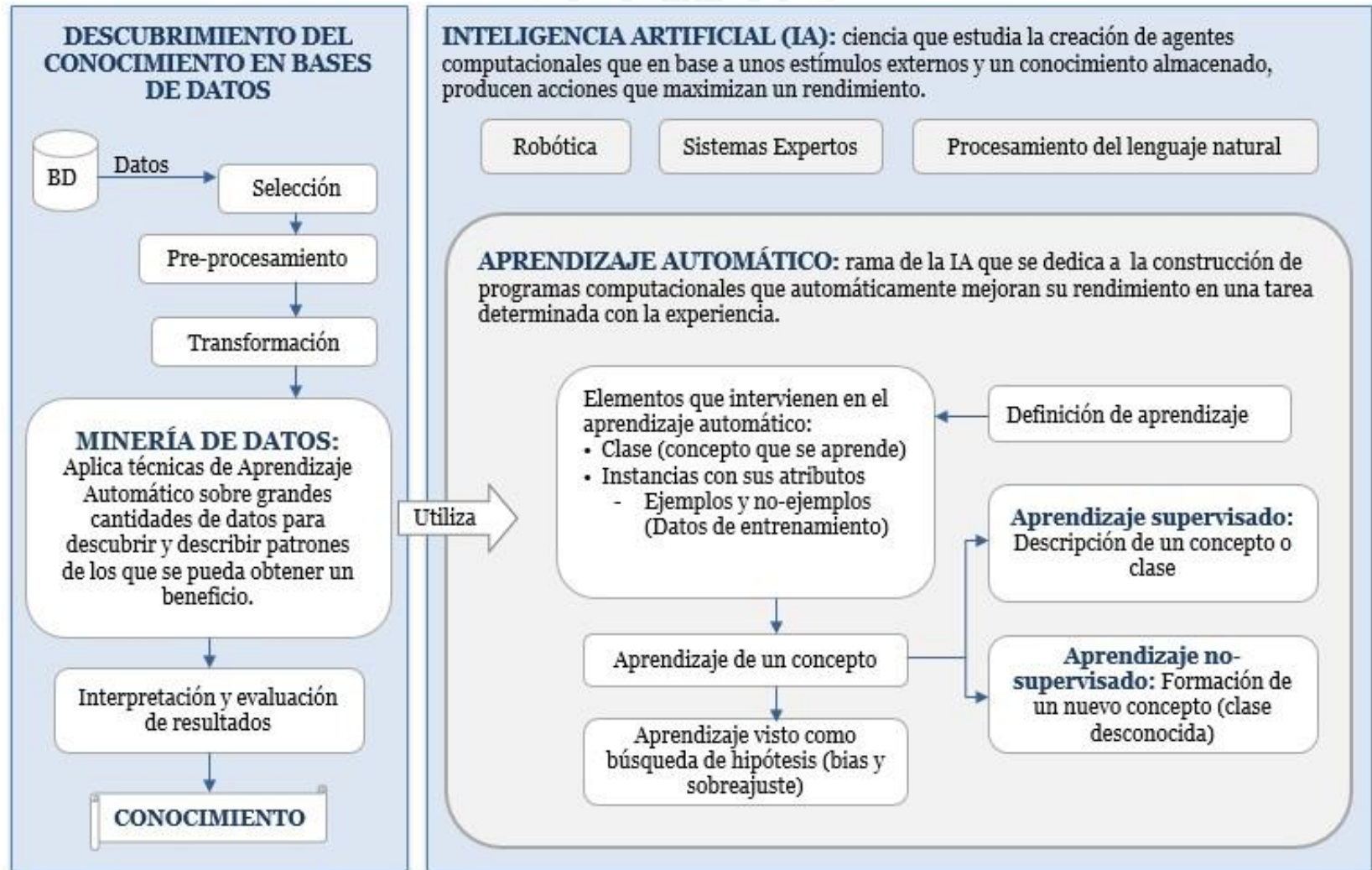
Se analiza si los patrones encontrados son útiles, comprensibles y aplicables.

Se puede incluir la visualización de los patrones extraídos, que sean entendibles para el usuario.

El resultado final debe ser **conocimiento accionable** para tomar decisiones.

La salida de esta etapa es el conocimiento descubierto en los datos.

Etapas del proceso KDD



Técnicas de minería de datos

Las técnicas pueden considerarse como un tipo de problema para ser resuelto por un algoritmo de minería de datos.



Técnicas de minería de datos

Técnicas predictivas (Supervisadas)

Regresión

Objetivo: Predecir un **valor numérico continuo**.

Ejemplo: Estimar el precio de una casa según su ubicación, tamaño y número de habitaciones.

Técnicas comunes: Regresión lineal, regresión polinómica, regresión logística (para clasificación).

Clasificación

Objetivo: Asignar una **categoría o clase** a una instancia.

Ejemplo: Determinar si un correo es *spam* o *no spam*, o si un cliente es *fiel* u *ocasional*.

Algoritmos comunes: Árboles de decisión, K-NN, Naive Bayes, Random Forest, SVM.

Análisis de series de tiempo

Objetivo: Predecir valores futuros **basados en el comportamiento temporal pasado**.

Ejemplo: Pronosticar ventas mensuales o la demanda de electricidad.

Algoritmos comunes : Modelos ARIMA, redes neuronales recurrentes (RNN), Prophet.

Técnicas de minería de datos

Técnicas descriptivas (No supervisadas)

Asociación

Objetivo: Encontrar relaciones frecuentes entre variables o eventos.

Ejemplo: En una tienda, descubrir que quienes compran *pan* también suelen comprar *mantequilla*. Usado en: Análisis de canasta de mercado.

Algoritmos comunes: Apriori, FP-Growth.

Agrupamiento (clustering)

Objetivo: Agrupar instancias similares **sin conocer previamente las categorías**.

Ejemplo: Segmentar clientes en grupos con comportamientos de compra similares. Muy usado también en: Marketing, segmentación, bioinformática.

Algoritmos comunes: K-means, DBSCAN, Hierarchical clustering.

Patrones secuenciales

Objetivo: Detectar **secuencias de eventos** que se repiten en el tiempo.

Ejemplo: Los clientes que compran una laptop, luego compran un mouse y después, en una siguiente visita, una mochila para laptop. Aplicaciones: análisis de rutas de navegación, comportamiento de clientes, fraude financiero.

Algoritmos comunes: PrefixSpan, GSP.

Técnicas de minería de datos

Clasificación

Agente bancario: ¿Debo ofrecer un préstamo a un cliente?

Histórico

ID	plazo	monto	salario	adeuda	...	ofrecer
101	18	60000	2967	2	...	no
102	12	20000	1676	0	...	si
103	36	45000	2967	1	...	no
104	12	8000	2034	0	...	si
105	9	700	2967	0	...	si
...
908	60	900	1484	1	...	no

Patrón/modelo

if adeuda > 0 **then** ofrecer =no
if adeuda = 0 **and** salario > 2500 **then** ofrecer =si

Técnicas de minería de datos

Regresión

Supervisor de fábrica: ¿Cuántas fallas para modelo X se esperan?

Histórico

modelo	mes-12	...	mes-3	mes-2	mes-1	fallas esperadas
X	50	...	14	33	21	?
Y	67	...	19	10	12	?
Z	90	...	21	19	17	?
W	21	...	11	8	20	?
R	12	...	13	4	19	?
...
P	40	...	25	5	3	?

Patrón/modelo

$$\text{fallas esperadas} = 0.62 + 0.33 \times \text{mes-1} + 0.12 \times \text{mes-2} + \dots + 0.03 \times \text{mes-12}$$

Técnicas de minería de datos

Reglas de asociación

Supervisor de fábrica: Clientes que compran pañales ¿Compran también cerveza?

Histórico

carrito	aceite	pan	huevos	leche	pañales	cerveza
001	si	no	si	si	no	si
002	si	si	no	no	si	no
003	no	si	no	si	si	no
004	no	no	si	no	si	si
005	no	si	no	si	si	si
...
878	no	si	si	no	si	si

Patrón/modelo

pañales \Rightarrow **cerveza**: Confianza=75% , Soporte =35%

Técnicas de minería de datos

Agrupamiento

Gerente de talento humano: ¿Que tipo de empleados hay en la empresa?

Histórico

salario	casado	carro	vivienda	sindicado	hijos	genero
2967	si	si	propia	no	1	M
980	no	si	propia	si	3	M
2034	no	no	alquilada	no	0	F
1676	no	no	propia	no	2	M
1484	no	si	propia	si	5	M
...
650	si	no	alquilada	si	0	F

Patrón/modelo

Grupo 1: Sin hijos/as y vivienda alquilada.

Grupo 2: Con hijos/as y de género masculino. Con carros.

Grupo 3: Solteros/as y sindicados/as.

Ejemplos KDD

Ejemplos de Aplicación del Proceso KDD y Técnicas de Minería de Datos

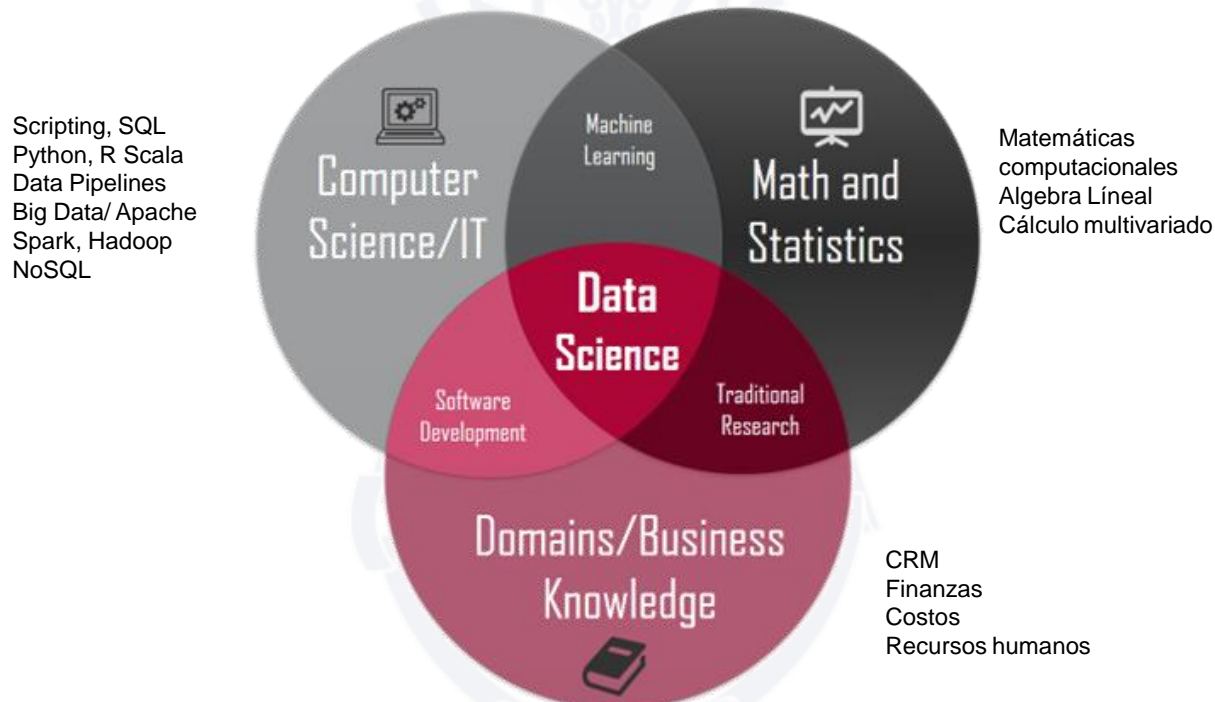


Actividad

Identificación de Caso, Datos y Técnicas de Minería de Datos



CIENCIA DE DATOS



Fuente https://www.researchgate.net/figure/Figura-1-Diagrama-de-Venn-de-la-ciencia-de-datos_fig1_366133299

La ciencia de datos es un paso evolutivo en campos interdisciplinarios que incorpora la informática, el modelado, la estadística, la analítica, y las matemáticas en uno solo proceso orientado al análisis de datos

CIENCIA DE DATOS

- Combinación de conocimiento de negocio, estadístico/matemático e informático
- Metodología estructurada para resolver problemas, basándose en datos
- Suele incluir Big Data, Minería de Datos, Aprendizaje Automático, Inteligencia Artificial, Analítica Empresarial.



Científico de datos



- Profesional con perfil multidisciplinario:
 - **Estadística**
 - Matemáticas
 - Informática (BBDD, Programación, BigData, IA, etc.)
 - Experiencia de dominio
 - Visualización de datos
- Es uno de los perfiles más demandados actualmente.

UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

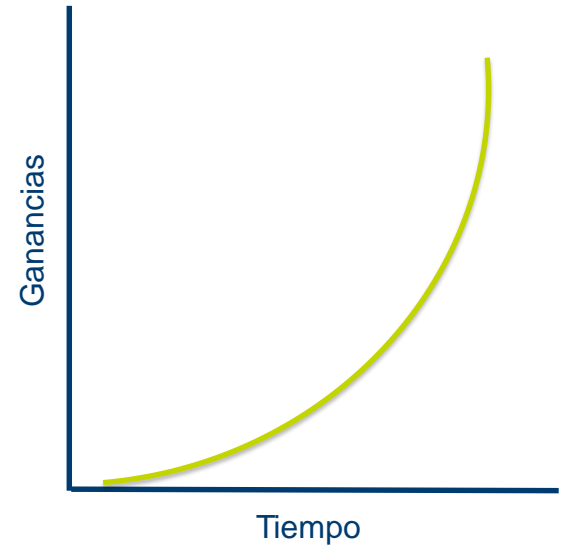
Científico de datos

MITO

Superhéroe (Científico(a) de Datos)



Inevitable y
Automáticamente



Fuente: <http://cio.com.mx/cientifico-datos-cuanto-gana-aporta/>

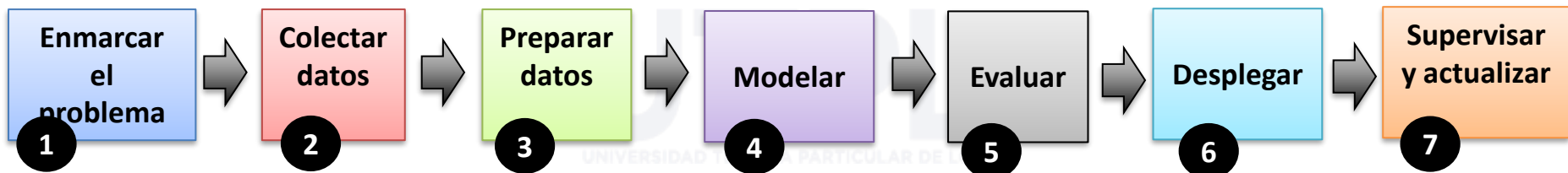
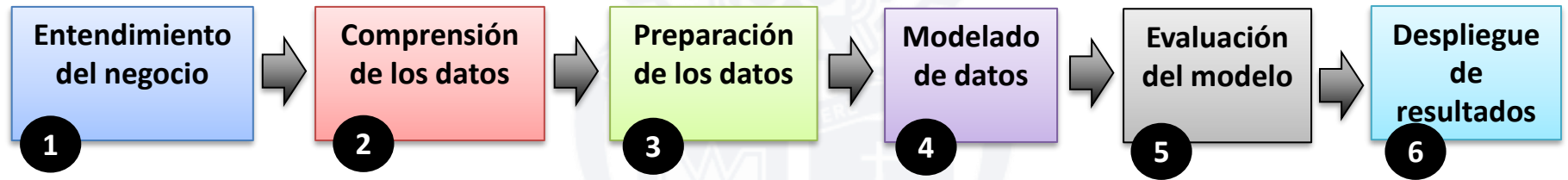
UTPL
el Confidencial
UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

Científico de datos

El científico de datos es un experto de ciencia de datos que resuelve problemas complejos de diferentes sectores haciendo uso de análisis de datos, y extracción de conocimiento de valor para una toma de decisiones más acertada y eficiente.



Ciclo de vida de proyectos de Ciencia de Datos



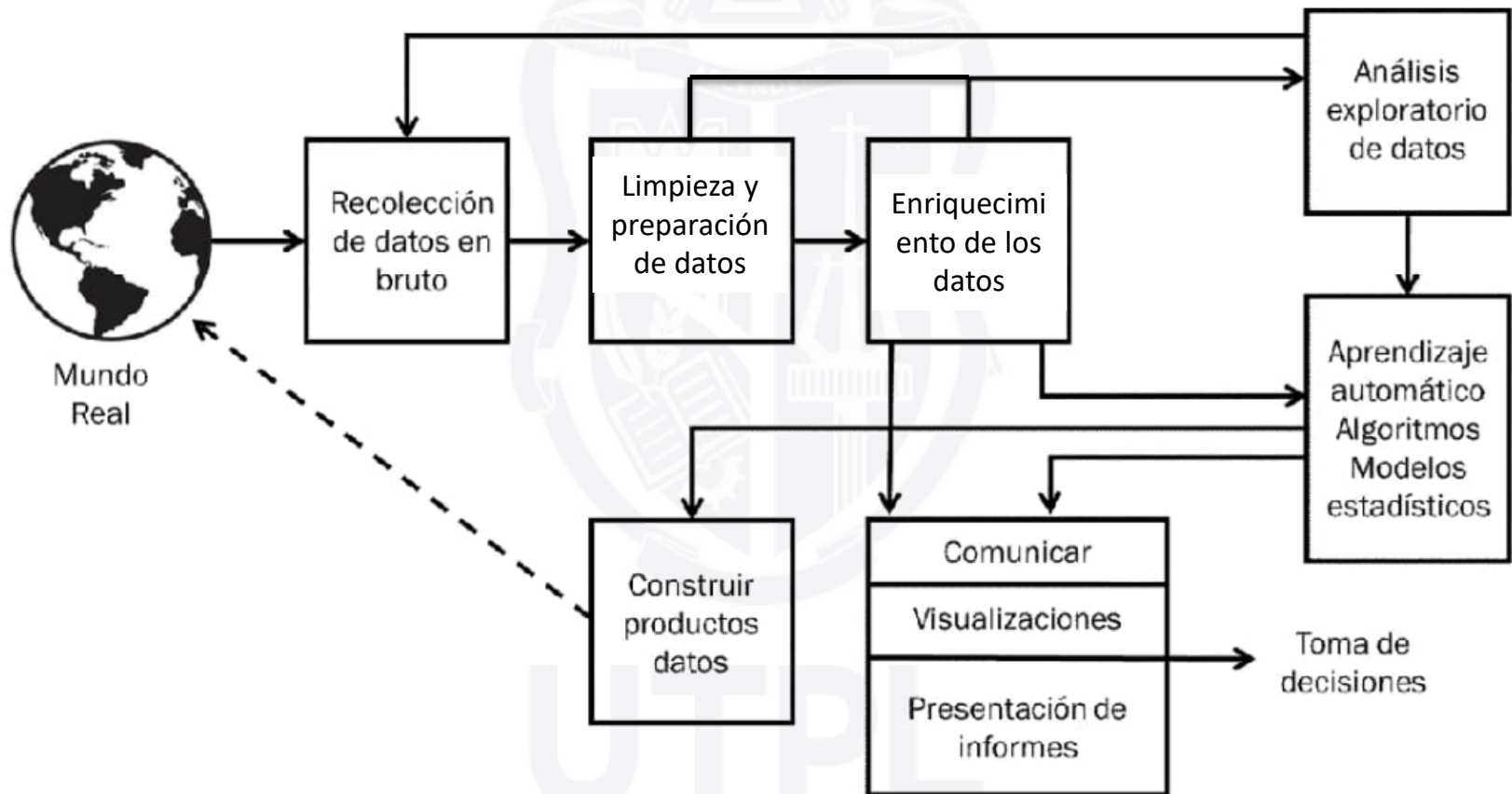
Donde entra el proceso KDD

- KDD (Knowledge Discovery in Databases) **entra en el núcleo técnico** del ciclo: entre la preparación de los datos y la evaluación del modelo. Es la **fase analítica intensiva**, donde se extrae conocimiento útil desde los datos ya disponibles.
- **El proceso KDD es el motor técnico del análisis de datos dentro del ciclo más amplio que incluye metas de negocio, comunicación y puesta en marcha de soluciones.**
- **KDD se enfoca en la extracción de conocimiento** desde los datos.
- El ciclo de vida de ciencia de datos incluye además **el entorno organizacional, el propósito y la entrega de valor real.**
- Un científico de datos no solo debe saber aplicar algoritmos (KDD), sino también traducir resultados en **impacto práctico.**

Comparando etapas — KDD dentro del ciclo de Ciencia de Datos

Etapa del Ciclo de Vida	Correspondencia con KDD	Explicación
1. Entendimiento del negocio	—	No es parte de KDD. Aquí se define el problema y los objetivos.
2. Comprensión de los datos	→ Selección	Se eligen las fuentes de datos más relevantes para el problema.
3. Preparación de los datos	→ Preprocesamiento y Transformación	Limpieza de datos, tratamiento de valores nulos, codificación, normalización, etc.
4. Modelado	→ Minería de datos	Aplicación de algoritmos para detectar patrones, hacer predicciones o clasificar datos.
5. Evaluación	→ Interpretación y evaluación	Se evalúa la calidad de los patrones descubiertos. ¿Son útiles? ¿Relevantes para el negocio?
6. Despliegue	—	KDD no abarca esta fase. Aquí se implementa la solución (por ejemplo, un dashboard o modelo en producción).

Ciclo de vida de proyectos de Ciencia de Datos



UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

Análisis exploratorio de datos (EDA)

Exploratory Data Analysis (EDA)



El EDA es una **fase preliminar** del ciclo de ciencia de datos que tiene como objetivo **explorar, conocer y entender** los datos disponibles, antes de aplicar técnicas analíticas o de minería de datos.

Se utiliza para explorar conjuntos de datos a fin de detectar problemas, patrones y particularidades en los datos que deben tratarse en las etapas de preprocesamiento y transformación, con la ayuda de estadística básica, tablas de frecuencia, análisis de correlación, visualizaciones básicas, etc.

“El EDA no es para responder preguntas de negocio, sino para preparar el terreno antes de sembrar. Su rol es técnico, no estratégico.”

Análisis exploratorio de datos (EDA)

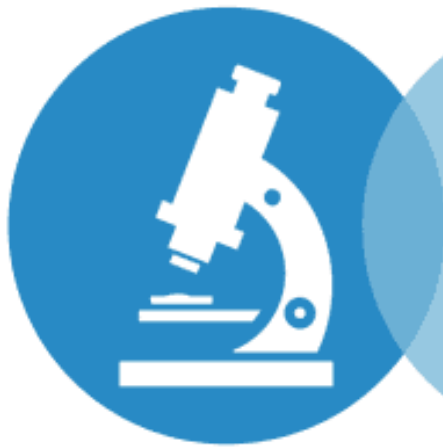
¿Qué se hace en el EDA?

- Identificación de **valores nulos o vacíos**
- Detección de **valores atípicos (outliers)**
- Verificación de **distribuciones** (¿es normal, sesgada?)
- Revisión de **formatos y consistencia de variables** (fechas, texto, numérico)
- Exploración de **correlaciones** entre variables (solo de forma preliminar)

El EDA no tiene como finalidad **generar conocimiento para la toma de decisiones finales**, sino **servir como diagnóstico inicial del estado de los datos**, para luego:

- Decidir qué técnicas de limpieza aplicar
- Transformar variables si es necesario
- Seleccionar variables relevantes
- Decidir si los datos están listos para el modelado

Tipos de análisis de valor



Descriptive

Explains what happened.



Diagnostic

Explains why it happened.



Predictive

Forecasts what might happen.



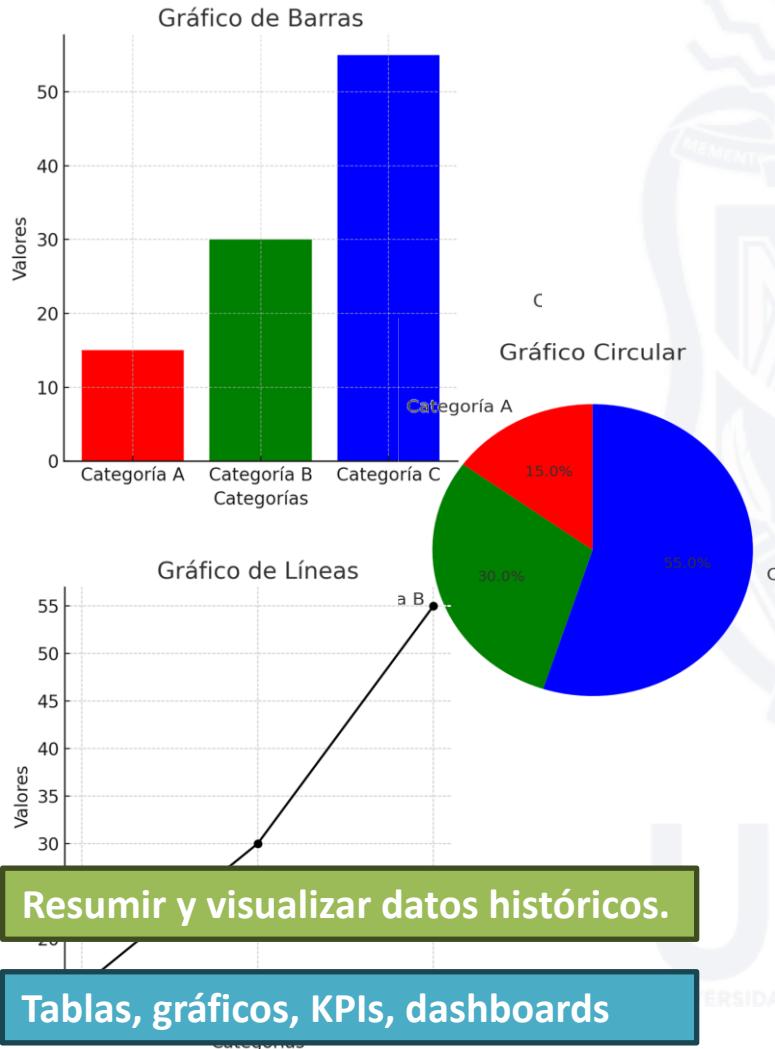
Prescriptive

Recommends an action based on the forecast.

Source: [enlace](#)

Análisis descriptivo

¿Qué pasó?



Es la forma más simple de análisis. Su propósito es resumir los hallazgos y entender lo que está sucediendo. Saber si está bien o mal, si buscar el por qué.

Utilizado para entender lo que ha ocurrido hasta un punto en el tiempo. Incluye la elaboración de reportes de ventas, métricas de rendimiento, resúmenes estadísticos, etc.

Ejemplo: Un supermercado analiza las ventas del último mes por categoría.

Descubre que los productos frescos fueron los más vendidos en Quito.

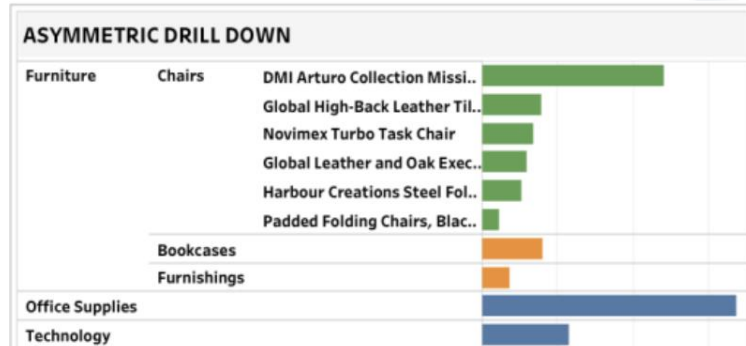
Análisis diagnóstico

¿Por qué pasó?

Permite determinar por qué sucedió algo en el pasado -> analiza los datos a profundidad para comprender las causas de los eventos.

Utilizado para determinar las razones detrás de ciertos resultados, como baja en la productividad de una empresa, o aumento en la tasa de morosidad de un banco.

Ejemplo: El mismo supermercado detecta que las ventas de productos frescos crecieron un 30% debido a una campaña de descuentos en redes sociales

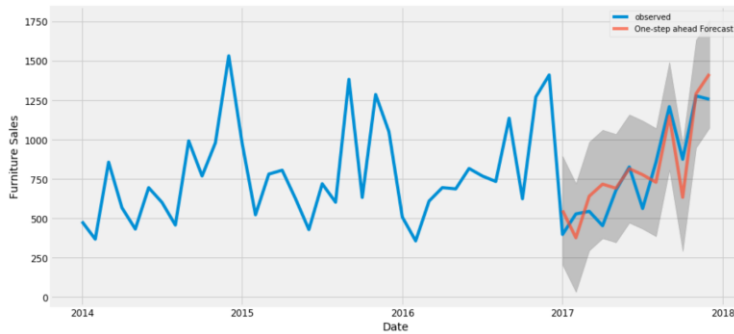


Identificar causas o correlaciones

Análisis de correlación, árboles de decisión, análisis de cohortes, regresión

Análisis predictivo

¿Qué podría pasar?



Se utiliza para predecir los resultados futuros -> pronostica cuáles son las posibilidades de la ocurrencia del evento.

Utilizado para hacer proyecciones sobre futuros eventos como la predicción de la demanda de productos, el riesgo de impago en un préstamo, o la probabilidad de éxito en una inversión

Ejemplo: Se entrena un modelo para predecir qué clientes probablemente abandonarán el programa de fidelización en los próximos 3 meses.



Estimar valores futuros o comportamientos probables

Modelos de regresión, series temporales, clasificación, machine learning

Análisis Prescriptivo

¿Qué deberíamos hacer?

Utiliza analítica predictiva, e intenta sugerir soluciones o comportamientos futuros. Puede sugerir resultados favorables de acuerdo con un curso de acción específico y también sugiere varios cursos de acción para llegar a un resultado en particular.

Utilizado para tomar decisiones informadas y estratégicas en la gestión de negocios, como optimizar rutas de entrega, o para optimizar procesos productivos.

Recomendar acciones óptimas para un resultado deseado

Simulación, optimización, algoritmos de decisión, sistemas de recomendación

Ejemplo: Según las predicciones de abandono, el sistema recomienda ofrecer un cupón del 15% solo a clientes de alto valor para retenerlos.

	A. Descriptivo What happened? ¿Qué pasó?	A. Diagnóstico Why did something happen? ¿Por qué ocurrió algo?	A. Predictivo What is LIKELY to happen? ¿Qué es PROBABLE que ocurra?	A. Prescriptivo What action should be taken? ¿Qué acción se debe tomar?
Entrada	Datos brutos de múltiples fuentes	Puntos de datos adicionales como el tráfico del sitio web, los presupuestos de marketing, la disponibilidad del inventario de productos y hacer una correlación para descubrir relaciones.	Utilizar datos históricos para la proyección	Datos históricos internos pero también información externa debido a la naturaleza de los algoritmos en los que se basa
Meta	Mirar si algo está mal o bien, sin explicar por qué	Buscar causa y efecto para ilustrar por qué sucedió algo. Comparar ocurrencias pasadas para determinar causas	Para determinar una tendencia, correlación, causalidad o probabilidad para el próximo evento	Prescribir qué acción tomar para eliminar un problema futuro o aprovechar al máximo una tendencia prometedora
Técnicas	Visualización (barras, gráficos circulares,...), estadísticas descriptivas, o cualquier formato de fácil lectura	Técnicas para identificar anomalías en los datos: valores atípicos, aislar patrones, descubrir relaciones. Teoría de la probabilidad, análisis de correlación, análisis de regresión.	Aprendizaje automático o aprendizaje profundo. Utiliza los hallazgos del análisis descriptivo y diagnóstico para detectar agrupaciones y excepciones y para predecir tendencias futuras, lo que lo convierte en una herramienta valiosa para la proyección	Optimización, simulación, análisis de decisión
Ejemplo	Identificar si las ventas aumentaron y disminuyeron en el último año, específicamente en qué región y en qué porcentaje.	Identificar la razón por la cual hay una caída repentina en las ventas.	Determinar cuál sería el volumen de ventas esperado en el próximo mes, trimestre o año, etc.	Si las ventas están cayendo, determinar las acción que se deben tomar de manera oportuna, como reducir precios, aumentar el marketing o discontinuar el producto.

Ejemplos Tipos de Análisis

Ejemplos adicionales sobre Tipos de Análisis



Relación de las técnicas de minería con los tipos de análisis de valor

Técnica	Tipo de análisis en el que se usa más frecuentemente
Regresión	Predictivo (cuando la variable a predecir es continua)
Clasificación	Predictivo (cuando la variable a predecir es categórica)
Asociación	Diagnóstico / Prescriptivo (para descubrir relaciones entre ítems)
Agrupamiento (clustering)	Descriptivo (para descubrir segmentos o patrones sin etiqueta previa)

Tipos de análisis de valor

"Describir, Diagnosticar, Predecir y Prescribir son los cuatro escalones del valor en la analítica de datos: desde mirar el pasado hasta decidir el futuro."

