



The data mining process

The illustration depicts a person with dark hair and glasses, wearing an orange shirt, seated at a desk and working on a computer. The desk features three monitors. The top monitor displays a network diagram with red nodes and connecting lines. The two bottom monitors show data visualizations consisting of horizontal bars and text. A keyboard and a mouse are on the desk. A small potted plant is positioned to the right of the desk. In the upper right corner, a network diagram shows three red circular nodes connected by lines, with ellipses indicating further connections. A large, light-orange semi-circle is in the background behind the person.

Sofia Barreto y Tomás Clavijo



Introducción a data mining process:

Implica **cuatro fases:**

01

Exploración

02

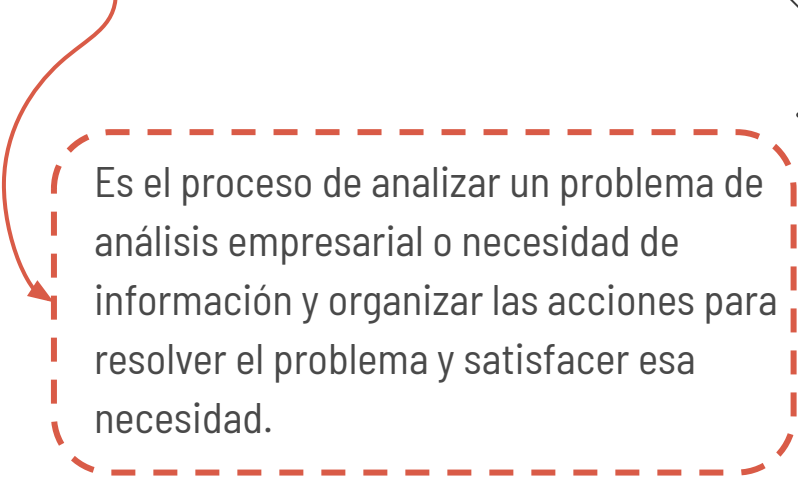
Analisis

03

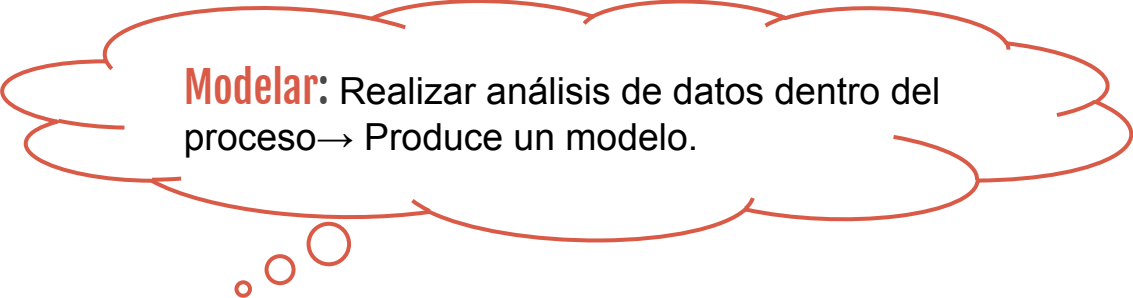
Interpretación

04

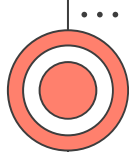
Explotación



Es el proceso de analizar un problema de análisis empresarial o necesidad de información y organizar las acciones para resolver el problema y satisfacer esa necesidad.



Modelar: Realizar análisis de datos dentro del proceso → Produce un modelo.



01 FASE de EXPLORACIÓN:

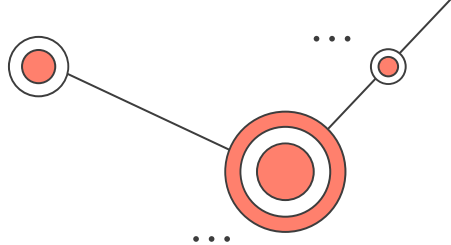
Implica entender el entorno empresarial y la necesidad de información del negocio.

Función del analista:

- Convierte las preguntas del negocio en preguntas analíticas bien formuladas y prepara los conjuntos de datos necesarios para ser analizados.
- Explora, adquiere y prepara los conjuntos de datos necesarios para ser analizados o minados en busca de respuestas durante esta etapa.



02 FASE de ANALISIS:



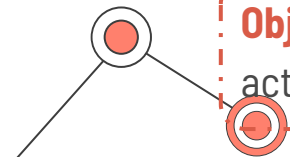
Tiene **2 etapas**

1. Comprensión profunda de los elementos de datos. → Implica analizar cada variable con métodos estadísticos básicos.

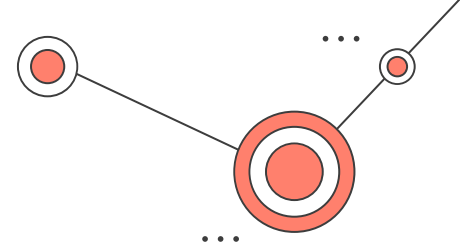
2. Data mining process: Busca una comprensión más profunda, saber por qué ocurrieron las cosas, va más allá de tabular lo sucedido.

Se **caracteriza** por la búsqueda de respuestas a preguntas bien formuladas a través de métodos analíticos.

Objetivo: Construir un modelo de datos sofisticado que se utilizará en el futuro a medida que se actualicen los datos y se necesite responder repetidamente a las preguntas.



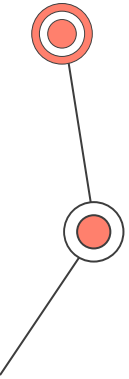
03 FASE de INTERPRETACIÓN:



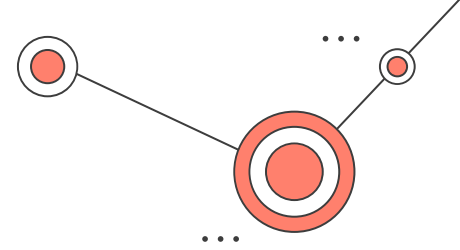
Es la base para los resultados de "toma de decisiones basada en datos"

- ★ **Implica** juicio empresarial y comprensión de las limitaciones de los modelos construidos.
- ★ **Busca** conclusiones significativas basadas en los hechos obtenidos del análisis como aportes imparciales al proceso de toma de decisiones empresariales.

Completa la actividad de recopilación de información deduciendo conclusiones derivadas de esos hechos.



04 FASE de EXPLOTACIÓN:



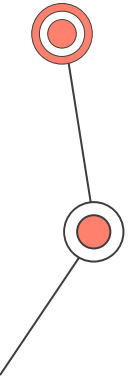
Implica la institucionalización de los esfuerzos de data mining process → Significa que cualquier respuesta obtenida y conclusión derivada se entregan a los gerentes y ejecutivos de negocios que necesitan tomar decisiones basadas en datos.

Los **modelos** a menudo requieren:

- Interfaces de usuario adecuadas para que las utilicen usuarios inexpertos.
- Recalibración para adaptarse a las preguntas planteadas y a las cambiantes condiciones económicas, políticas, legales y empresariales.



Esta fase garantiza que produzcamos y mantengamos una **herramienta empresarial utilizable por todos** los que la necesiten en esta etapa final.

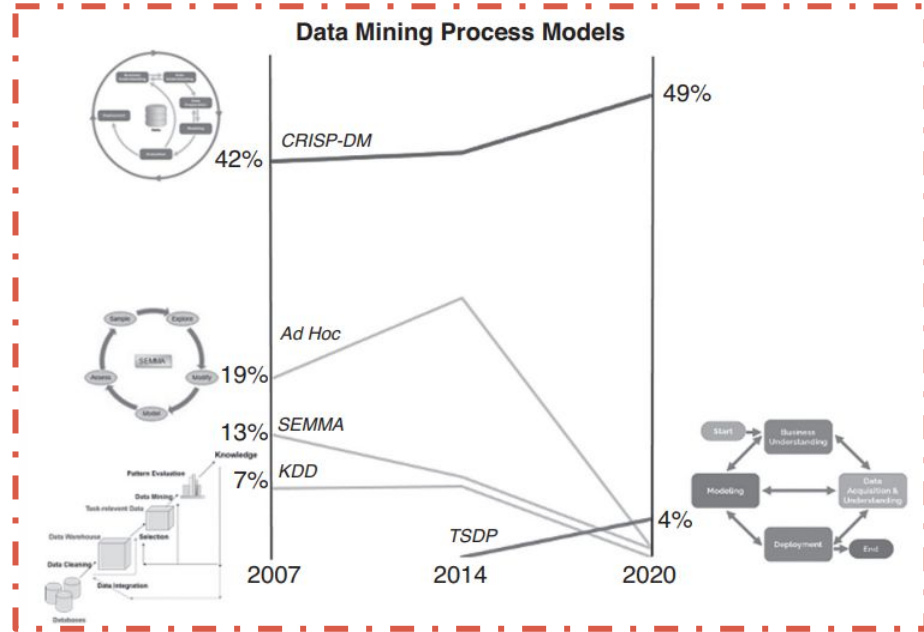


Selecting a data mining process:

Existen varias alternativas para adoptar un proceso de minería de datos, como:

- CRISP-DM → + utilizado
- KDD
- SEMMA
- TDSP, de microsoft

↓
Posible lider del futuro



CRISP-DM:

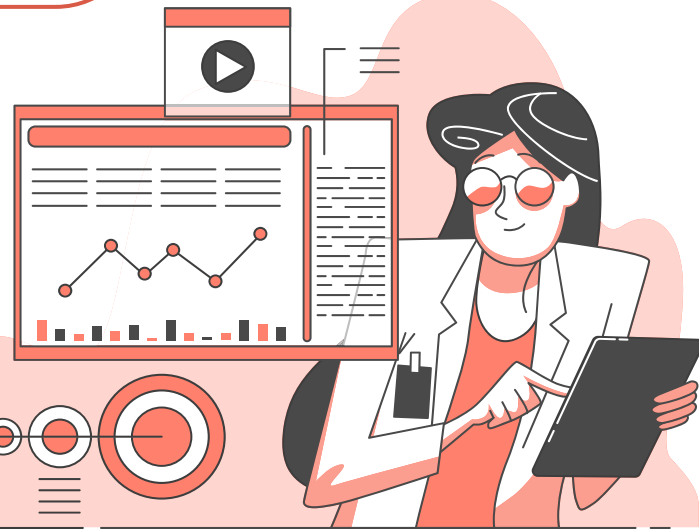
Es un modelo de 6 fases:

1. Comprensión del negocio
2. Comprensión de los datos
3. Preparación de los datos
4. Modelado
5. Evaluación
6. Implementación

Describe naturalmente el ciclo de vida de la ciencia de datos

KDD:

Proceso general de extraer información de patrones o de grandes conjuntos de datos utilizando aprendizaje automático, estadísticas y sistemas de bases de datos.

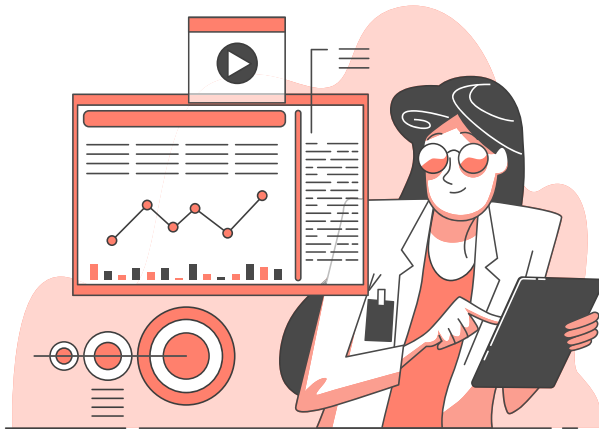


SEMMA:

Define 5 fases de un proyecto:

1. Muestra
2. Exploración
3. Modificación
4. Modelado
5. Evaluación

Diseñado originalmente para guiar a los usuarios a través de herramientas en SAS Enterprise Miner.



TDSP:

Define 5 etapas del ciclo de vida de la ciencia de datos:

1. Comprensión del negocio
2. Adquisición y Comprensión de los Datos
3. Modelado
4. Implementación
5. Aceptación del Cliente

Combina aspectos de Scrum y CRISP-DM, incorpora el aspecto de trabajo en equipo en la ejecución de proyectos de datos.



Profundizando en **CRISP-DM**

Repasando los pasos..

1. Comprensión del negocio: Comprender los objetivos y requisitos del proyecto desde una perspectiva empresarial.

2. Comprensión de los datos: Recopilación inicial de datos y familiarización con los mismos para identificar problemas de calidad de los datos y/o detectar subconjuntos interesantes para formular hipótesis sobre información oculta.

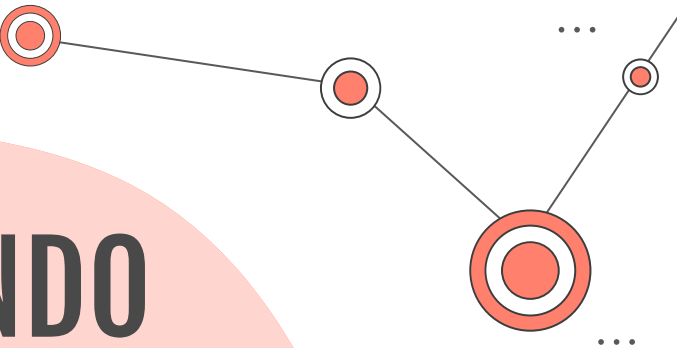
3. Preparación de los datos: Abarca todas las actividades necesarias para construir el conjunto final de datos, sin orden específico

4. Modelado: Se seleccionan y aplican técnicas de modelado, y se calibran sus parámetros para obtener los valores óptimos.

5. Evaluación: Revisión del modelo para asegurarse de que el mismo logre los objetivos empresariales.


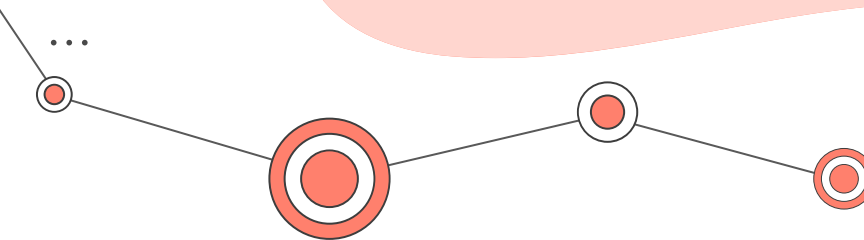
6. Implementación: Dependiendo de los requisitos, la fase de despliegue puede ser tan simple como generar un informe o tan compleja como implementar un proceso de data mining repetible en toda la empresa.

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives Background Business Objectives Business Success Criteria	Collect Initial Data Initial Data Collection Report	Select Data Rationale for Inclusion/Exclusion	Select Modeling Techniques Modeling Technique Modeling Assumptions	Evaluate Results Assessment of Data Mining Results w.r.t. Business Success Criteria	Plan Deployment Deployment Plan
Assess Situation Inventory of Resources Requirements, Assumptions, and Constraints Risks and Contingencies Terminology Costs and Benefits	Describe Data Data Description Report	Clean Data Data Cleaning Report	Generate Test Design Test Design	Review Process Review of Process	Plan Monitoring and Maintenance Monitoring and Maintenance Plan
Determine Data Mining Goals Data Mining Goals Data Mining Success Criteria	Explore Data Data Exploration Report	Construct Data Derived Attributes Generated Records	Build Model Parameter Settings Model Descriptions	Determine Next Steps List of Possible Actions Decision	Produce Final Report Final Report Final Presentation
Produce Project Plan Project Plan Initial Assessment of Tools and Techniques	Verify Data Quality Data Quality Report	Integrate Data Merged Data	Assess Model Model Assessment Revised Parameter Settings	Review Project Experience Documentation	
		Format Data Reformatted Data Dataset Dataset Description			



SELECCIONANDO LENGUAJES DE ANALÍTICA DE DATOS

Es esencial comprender qué decisiones debe tomar un analista al elegir plataformas de análisis o lenguajes de programación informática



En el paso 4 del modelo CRISP-DM, nos involucramos en el análisis concreto de datos → En algunos casos, implementamos modelos sofisticados de aprendizaje automático.

La selección del lenguaje de analítica de datos **depende del conjunto de datos que queramos procesar y/o analizar**

EXCEL:

- Elección básica.
- Para datos pequeños (hasta 100 MB).
- Práctica para preparación rápida de conjuntos de datos
- Posee tablas dinámicas, herramientas de gráficos y cálculo.

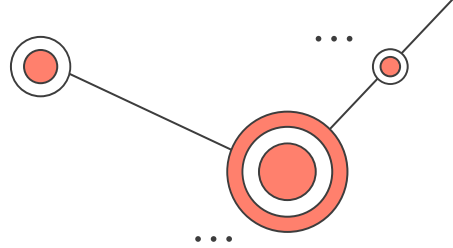
TABLEAU:

- Conjunto de datos que deben extenderse de almacenes de datos y programas de aplicación comerciales basados en un RDBMS.
- Conocido por su visualización de datos, análisis poderosas y paneles personalizados
- Necesita familiaridad con **SQL**.

SQL:

- ★ Es el más utilizado.
- ★ Solo para la extracción → NO recomendado para procesamiento ni análisis.
- ★ Usado para extraer datos de bases de datos estructuradas
- ★ No procedural (no requiere uso de lógica de programación tradicional)

Lenguajes de análisis más populares:

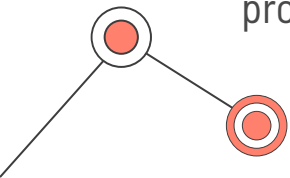


PYTHON:

- Lenguaje más utilizado.
- Proporciona todas las bibliotecas necesarias para las 4 etapas principales de trabajar con datos: Recolección y limpieza de datos, exploración de datos, modelado de datos y visualización de datos.

Lenguaje R:

- Excelente para cálculos estadísticos y gráficos.
- Puede manejar conjuntos de datos grandes y complejos.
- Código abierto.
- Tiene deficiencias en cuanto a seguridad, lo que dificulta la protección de los modelos escritos en él.





Gracias!