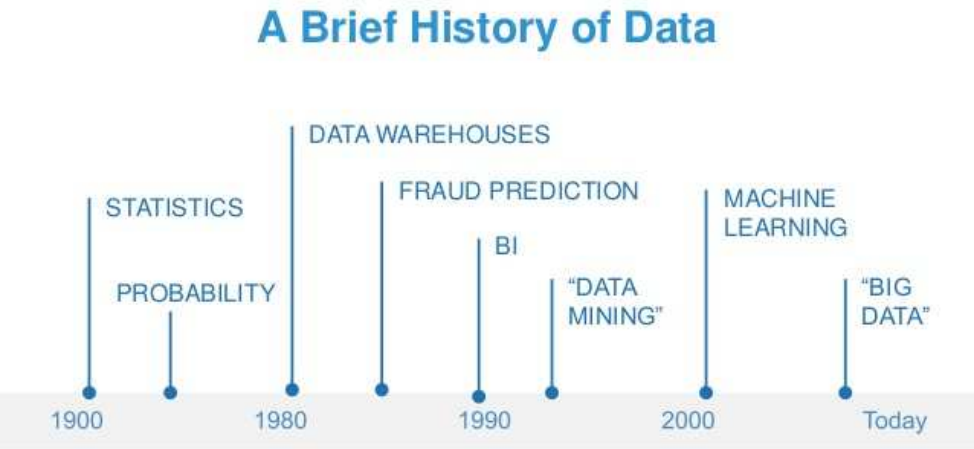
Minería de Datos

¿Qué es la Minería de Datos?

La minería de datos es el proceso de hallar **anomalías, patrones y correlaciones en grandes conjuntos de datos para predecir resultados**. Empleando una amplia **variedad de técnicas**, puede utilizar esta información para incrementar sus ingresos, recortar costos, mejorar sus relaciones con clientes, reducir riesgos y más.

Historia de la Minería de Datos

La minería de datos es el proceso de hallar **anomalías, patrones y correlaciones en grandes conjuntos de datos para predecir resultados**. Empleando una amplia **variedad de técnicas**, puede utilizar esta información para incrementar sus ingresos, recortar costos, mejorar sus relaciones con clientes, reducir riesgos y más.



Importancia de la Minería de Datos

¿Entonces por qué es importante la minería de datos? Ha podido apreciar los números asombrosos **– el volumen de datos producidos se duplica cada dos años**. Los **datos no estructurados por sí solos conforman el 90% del universo digital**. Pero más información no significa necesariamente más conocimientos.

La minería de datos le permite:

- **Filtrar** todos los datos no significativos y repetitivos.
- Entender qué es relevante y luego **hacer un buen uso de esa información** para evaluar resultados probables.
- **Acelerar el ritmo de la toma de decisiones informadas.**

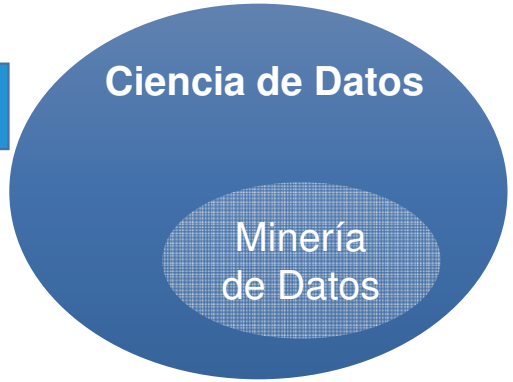
Minería de Datos vs Ciencia de Datos

Minería de Datos

- Existe desde los años 90’s
- Es una técnica
- Extraer información relevante

Ciencia de Datos

- Existe desde los años 60’s
- Es un área
- Desarrolla productos de datos para la industria



Casos de Éxito

amazon

Adoptó tecnología de vanguardia para **recolectar, analizar y utilizar la cantidad masiva de datos a la que tienen acceso a partir del historial de búsqueda y de compra de una persona**. Por eso, son los mejores en temas como **optimización de la cadena de suministro, optimización de precios y detección de fraudes.**

NETFLIX

Netflix no solo es capaz de predecir qué quiere ver una persona sino qué tipo de series o películas debe producir y qué actores tienen mejor acogida dependiendo del público. De hecho, desde que era un servicio de DVD por correo físico, una de las prioridades de la plataforma era recolectar datos para construir un **sistema de recomendaciones.**

Gracias al análisis de datos, Apple ha logrado posicionarse no solo como la mejor compañía de tecnología, sino como una de las que más clientes FIELES tiene alrededor del mundo. Así, las apps conocen a sus usuarios y la experiencia es cada vez más personalizada, al punto de que no puedan vivir sin sus productos Apple.

Casos de Éxito

El análisis de datos se ha utilizado en áreas como desarrollo de producto. Por ejemplo, el lanzamiento del sabor “Cherry Sprite” en 2017 nació en los datos recolectados de las máquinas dispensadoras de gaseosa que permiten que los consumidores mezclen sus propias bebidas. Así, Coca-Cola pudo identificar la mezcla más popular y convertirla en una bebida lista para ser consumida

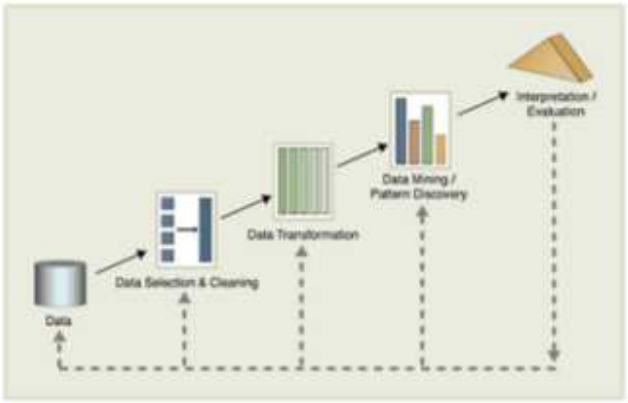
¿Alguna vez se ha preguntado cómo Starbucks puede abrir 5 tiendas en un radio de 3 kilómetros y aun así todos están llenos? Esta compañía cafetera utiliza el Big Data para determinar el éxito potencial de cada tienda nueva que piensan abrir. Recogen información sobre la ubicación, tráfico, área demográfica y comportamientos del consumidor.

Hacer este tipo de evaluación antes de abrir una tienda significa que Starbucks puede hacer una estimación bastante precisa de cuál será la tasa de éxito y elegir ubicaciones en función de la inclinación al crecimiento de los ingresos

Metodología de Minería de Datos

Las 3 metodologías dominantes para el proceso de Minería de Datos son:

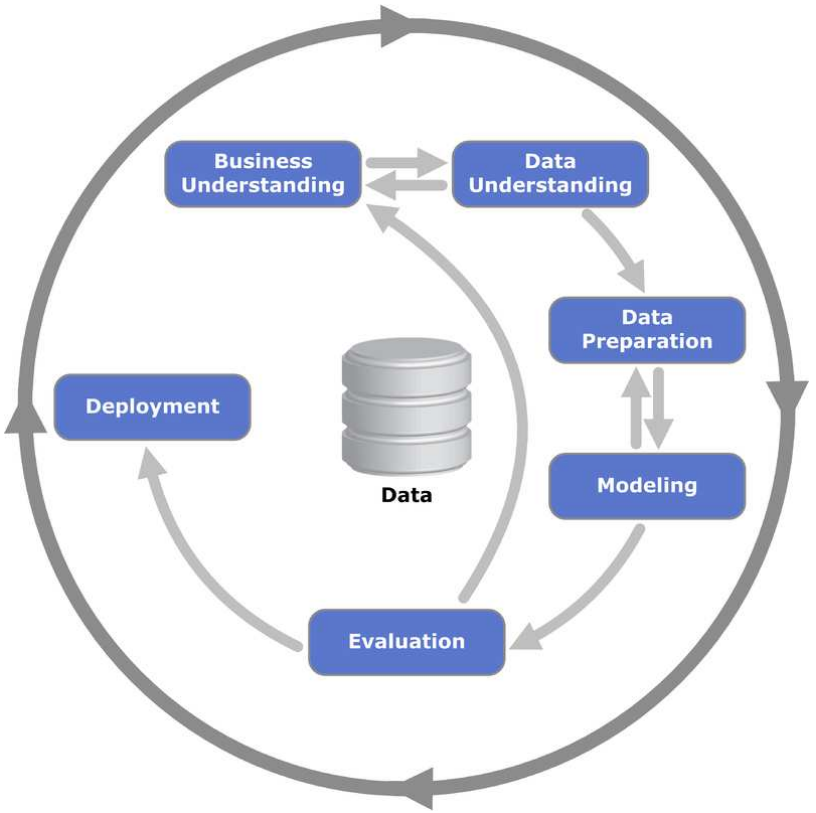
- **KDD**
- **CRISP-DM**
- **SEMMA**



¿Qué es CRISP-DM?

CRISP-DM (CRoss-Industry Standard Process for Data Mining) es una metodología de Minería de datos para desarrollo de proyectos analíticos. CRISP-DM **se explica como un proceso jerárquico**, que tiene **cuatro niveles de abstracción: Fase, tareas generales, tareas específicas e instancias de proceso.**

La metodología de referencia CRISP-DM contiene las fases de un proyecto, sus respectivas tareas y sus relaciones entre ellas. **El ciclo de vida de un proyecto de Minería de Datos consiste esencialmente en seis fases**. La **secuencia de las fases no es rígida**, **se puede regresar o adelantar a alguna de ellas siempre que se necesario**. Todo depende de los resultados de cada fase.



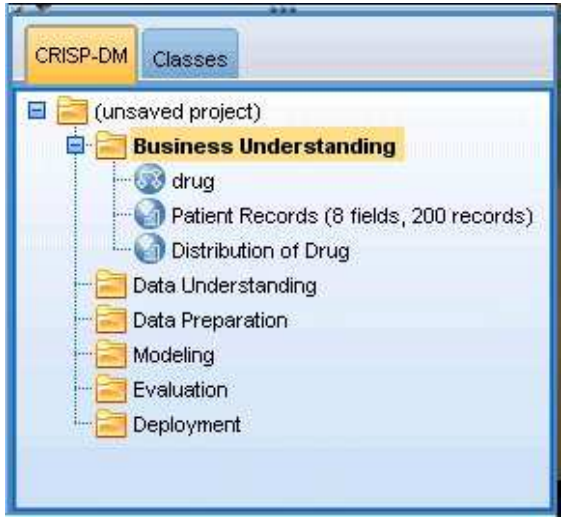
CRISP –DM: Fases

1. **Comprensión del negocio:**
   - ✓ Entendimiento de los objetivos y requerimientos del proyecto.
   - ✓ Definición del problema de Minería de Datos
2. **Comprensión de los datos**
   - ✓ Obtención conjunto inicial de datos.
   - ✓ Exploración del conjunto de datos.
   - ✓ Identificar las características de calidad de los datos
   - ✓ Identificar los resultados iniciales obvios.
3. **Preparación de Datos**
   - ✓ Selección de datos
   - ✓ Limpieza de datos
4. **Modelamiento**
   Implementación en herramientas de Minería de Datos
5. **Evaluación**
   - ✓ Determinar si los resultados coinciden con los objetivos del negocio
   - ✓ Identificar las temas de negocio que deberían haberse abordado
6. **Despliegue**
   - ✓ Instalar los modelos resultantes en la práctica
   - ✓ Configuración para minería de datos de forma repetida ó continua

CRISP Herramientas

La **herramienta de proyectos de CRISP-DM** proporciona un **método estructurado de minería de datos** que puede ayudarle a asegurar el rendimiento de su proyecto. Se trata esencialmente de una extensión de la **herramienta de proyectos estándar IBM® SPSS Modeler**. De hecho, puede alternar entre la vista de CRISP-DM y la vista Clases estándar para ver las rutas y los resultados organizados por el tipo o la fase de CRISP-DM.

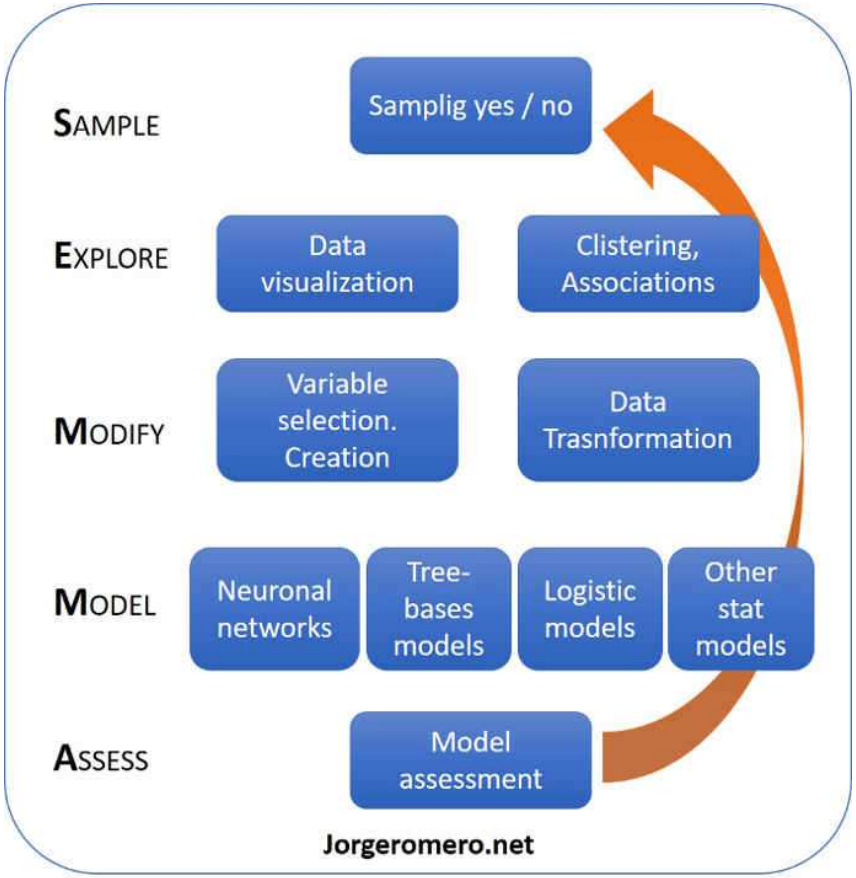
Figura 1. Herramienta de proyectos de CRISP-DM
Herramienta de proyectos CRISP-DM



Metodología SEMMA

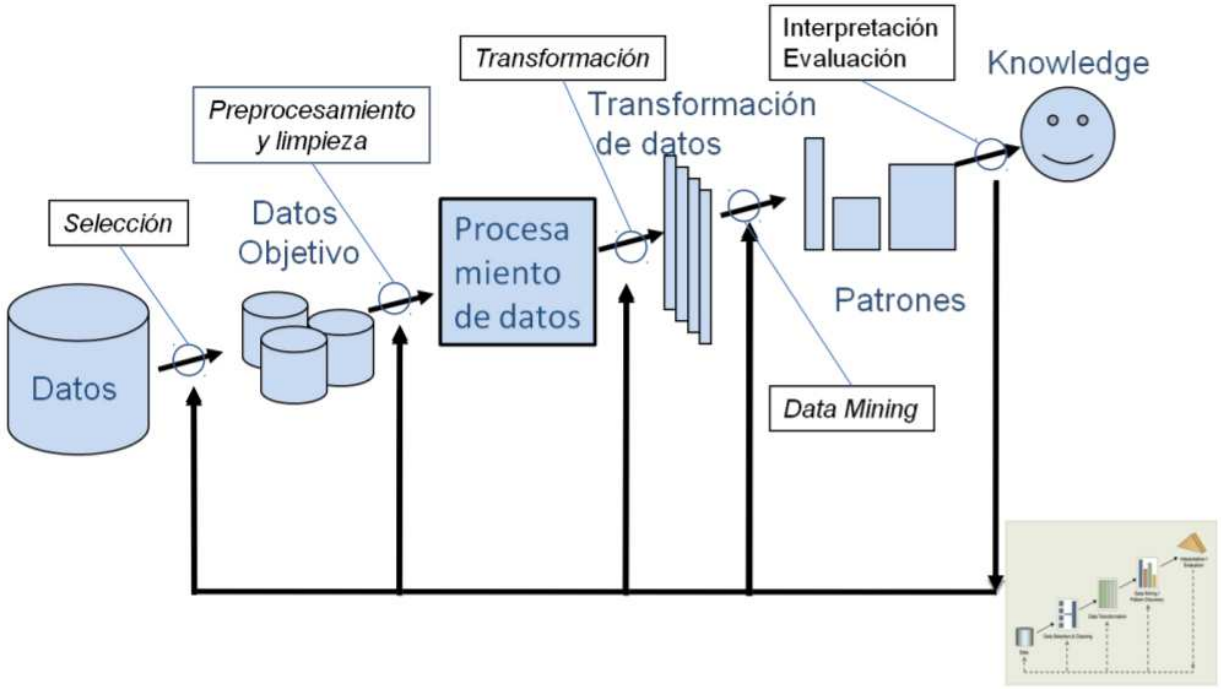
Por otra parte, la metodología llamada **SEMMA, desarrollada por el instituto SAS que es el acrónimo de SAMPLE, EXPLORE, MODIFY, MODEL, ASSESS**, que podemos interpretar como muestrea, explora, modifica, modela y evalúa, que se refiere al proceso básico para realizar Minería de Datos.

Esto es a partir de una muestra representativa de los datos, **se aplican técnicas estadísticas de exploración y visualización, se seleccionan y transforman variables, se modela con las variables para predecir los y se evalúa la exactitud del modelo.**



Metodología KDD

Es una metodología propuesta por Fayyad en 1996,propone 5 fases:
Selección, preprocesamiento, transformación, minería de datos y evaluación e implantación. Es un proceso iterativo e interactivo.



Técnicas de Minería de Datos