UNIVERSIDAD POLITÉCNICA DE TAPACHULA

# ANÁLISIS DE DATOS

SECRETARÍA DE EDUCACIÓN PÚBLICA

# INTRODUCCIÓN AL ANÁLISIS DE DATOS

La razón por la que cada vez más empresas están invirtiendo en personal, infraestructura y equipo para procesar grandes volúmenes de datos, es por el gran valor e impacto que genera para su proceso estratégico de toma de decisiones, el poder transformar los datos sobre productos o servicios en información clave, identificar patrones, tendencias, desarrollar nuevos nichos de mercado, realizar predicciones de preferencias y necesidades de los clientes, aprender más sobre su negocio para ser capaces de innovar, entre otros innumerables beneficios.

Las empresas actuales emplean el análisis de datos de forma diaria mediante diferentes plataformas de Business Intelligence y Big Data para diseñar procesos automatizados que les faciliten procesar, visualizar y monitorear sus datos relevantes (Cupas, 2021).

# ¿QUÉ ES EL ANÁLISIS DE DATOS?

Partiendo de la definición del término análisis que se refiere al estudio detallado de algo, separándolo en elementos para descubrir su composición, podemos decir que:

**Análisis de datos** es un proceso que busca examinar un conjunto de variables aleatorias para conocerlo en detalle y transformarlo en información concreta a través de una serie de procesos y técnicas ejecutadas en pasos secuenciales.

Estos pasos, en su mayoría automatizados, permiten:

- ❑ Recopilar y extraer datos de múltiples fuentes.
- ❑ Clasificarlos, almacenarlos y depurarlos dentro un repositorio.
- ❑ Medirlos y modelarlos con una serie de modelos matemáticos.
- ❑ Evaluar, interpretar y visualizar los resultados obtenidos.

# ¿CUÁL ES EL PROPÓSITO DEL ANÁLISIS DE DATOS?

El propósito principal del análisis de datos es descubrir información significativa que se oculta tras los datos en bruto de un data lake de cifras, etiquetas, imágenes, etc. a fin de dar respuestas coherentes a preguntas que teníamos en mente, pero también a preguntas totalmente desconocidas.

**Pasos previos al análisis de datos**

Antes de dar paso al análisis de datos, independientemente de la metodología que elijamos, es recomendable que:

- ❑ Establezcamos un objetivo claro y específico del análisis, es decir, cuál es la información que se pretende obtener.
- ❑ Seleccionemos los conjuntos de datos con los cuales trabajar, tanto de fuentes de datos internas como externas.
- ❑ Aseguremos la recopilación de datos relevantes para que el resultado del análisis sea efectivo.
- ❑ Validemos la calidad de los datos.

# IDENTIFICACIÓN DE PATRONES EN CONJUNTO DE DATOS

A menudo recopilamos datos para poder encontrar patrones en los datos, tales como números que tienen una tendencia hacia arriba o correlaciones entre dos conjuntos de números.

Dependiendo de los datos y los patrones, a veces podemos ver ese patrón en una presentación tabular sencilla de los datos. Otras veces, ayuda visualizar los datos en una gráfica, como una serie de tiempo, una gráfica de líneas o una gráfica de dispersión (Fox, 2020).

**Detectar tendencias**

Una cantidad de tendencia es un número que por lo general está aumentando o disminuyendo.

# IDENTIFICACIÓN DE PATRONES EN CONJUNTO DE DATOS

**Hacer predicciones**

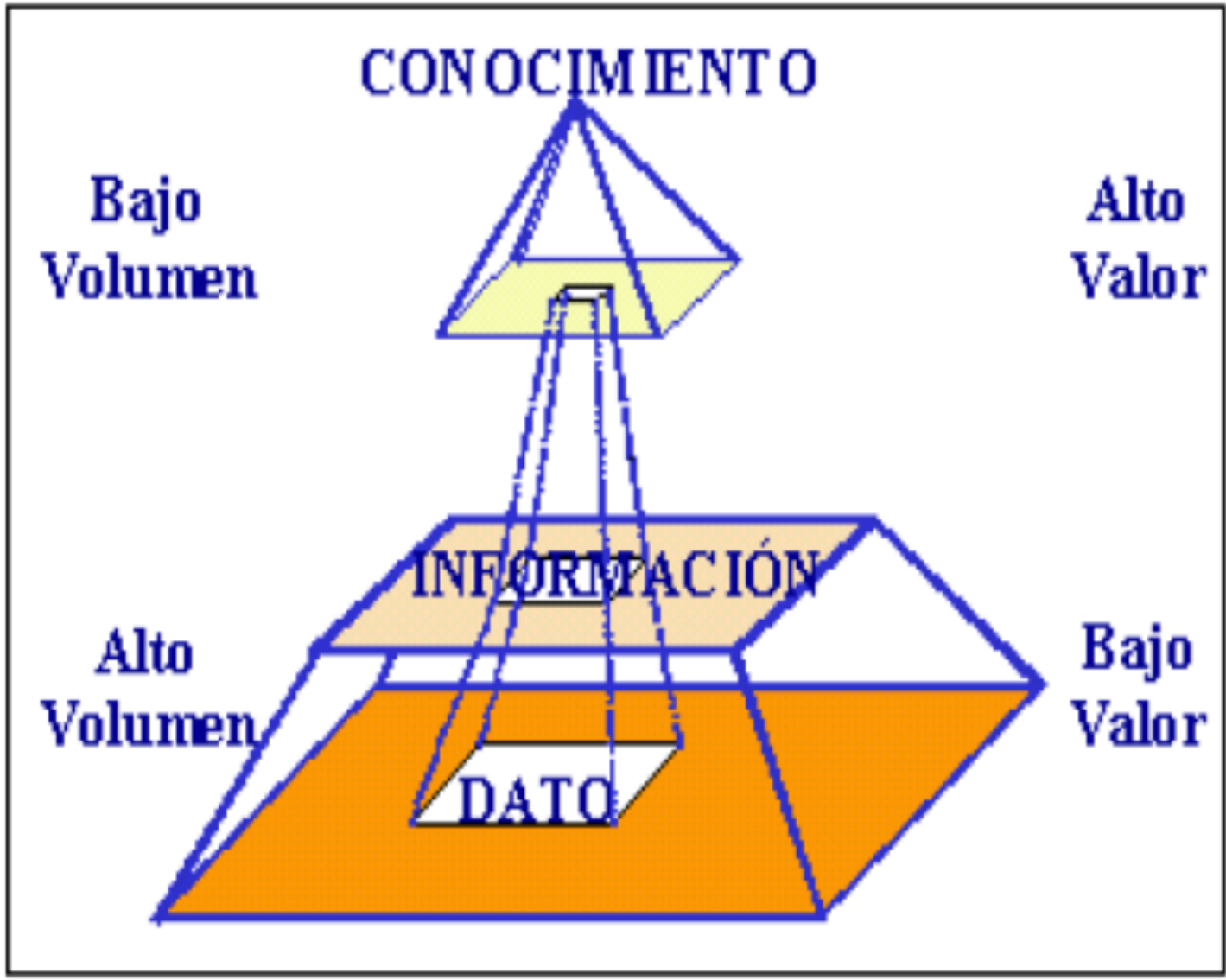
Una razón por la que analizamos datos es para obtener predicciones.

**Encontrar correlaciones**

Otro objetivo de analizar datos es calcular la correlación, la relación estadística entre dos conjuntos de números. Una correlación puede ser positiva, negativa o no existir en absoluto. Una gráfica de dispersión es una manera común de visualizar la correlación entre dos conjuntos de números.

# MINERÍA DE DATOS

- La minería de datos es el proceso de detectar la información procesable de los conjuntos grandes de datos.
- Utiliza el análisis matemático para deducir los patrones y tendencias que existen en los datos.
- Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos porque las relaciones son demasiado complejas o porque hay demasiado datos.
- Estos patrones y tendencias se pueden recopilar y definir como un modelo de minería de datos.



# MINERÍA DE DATOS

Los modelos de minería de datos se pueden aplicar en escenarios como los siguientes:

•**Pronóstico**: cálculo de las ventas y predicción de las cargas del servidor o del tiempo de inactividad del servidor.

•**Riesgo y probabilidad**: elección de los mejores clientes para la distribución de correo directo, determinación del punto de equilibrio probable para los escenarios de riesgo, y asignación de probabilidades a diagnósticos y otros resultados.

# MINERÍA DE DATOS

•**Recomendaciones**: determinación de los productos que se pueden vender juntos y generación de recomendaciones.

•**Búsqueda de secuencias**: análisis de los artículos que los clientes han introducido en el carrito de la compra y predicción de posibles eventos.

•**Agrupación**: distribución de clientes o eventos en grupos de elementos relacionados, y análisis y predicción de afinidades.

# MINERÍA DE DATOS

La generación de un modelo de minería de datos forma parte de un proceso mayor que incluye desde la formulación de preguntas acerca de los datos y la creación de un modelo para responderlas, hasta la implementación del modelo en un entorno de trabajo. Este proceso se puede definir mediante los seis pasos básicos siguientes:

1. Definición del problema
2. Preparación de datos
3. Explorar los datos
4. Creación de modelos
5. Exploración y validación de modelos
6. Implementar y actualizar los modelos

# TAREAS DE MINERÍA DE DATOS

1. **Tareas descriptivas:** Orientadas a describir un conjunto de datos.
   - **Clasificación:** Se asigna una categoría a cada caso. Cada caso tiene un conjunto de atributos, donde uno de ellos es el atributo clase. Se busca un modelo que describa el atributo clase como una función de los atributos de salida. Existen principalmente dos tipos de clasificación: Clasificación basada en árboles de decisión y Clasificación neuronal.

   - **Segmentación (agrupación):** Esta tarea también es conocida como segmentación, y se encarga de identificar grupos naturales basándose en un conjunto de atributos. Existen diversas técnicas: clustering y segmentación neuronal.

# TAREAS DE MINERÍA DE DATOS

1. **Tareas descriptivas:** Orientadas a describir un conjunto de datos.
   - **Asociación:** Organizar según relaciones entre atributos (Análisis de la cesta de la compra). Expresa las afinidades entre elementos siguiendo el modelo de las reglas de asociación X->Y, facilitando una serie de métricas como el soporte y confianza.

   - **Regresión:** Tarea muy similar a la de clasificación pero con el objetivo de buscar patrones para determinar su valor único.

# TAREAS DE MINERÍA DE DATOS

2. **Tareas Predictivas:** Orientadas a estimar valores de salida.
   - **Previsión:** A partir de la entrada, conjunto de valores obtenidos a lo largo de un tiempo determinado de los que se extrae un comportamiento futuro. Para la estimación de variables cuantitativas, los métodos más usados son: funciones de base radial y predicción neuronal.

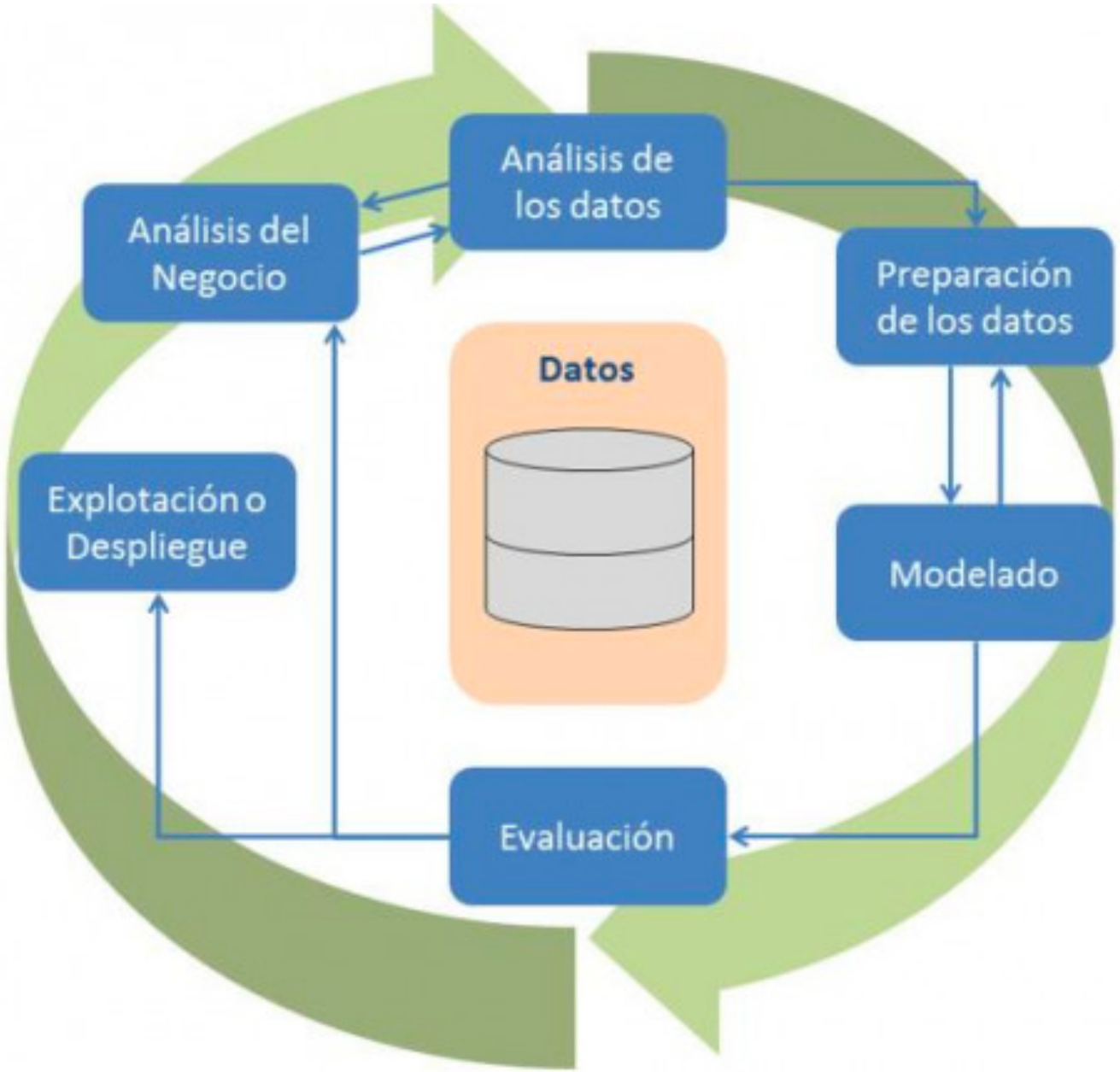
   - **Análisis de secuencia:** Se encarga de la búsqueda de patrones en una serie de eventos denominados secuencias o transacciones, lo que permite optimizar las ventas a lo largo del tiempo

# TAREAS DE MINERÍA DE DATOS

2. **Tareas Predictivas:** Orientadas a estimar valores de salida.
   - **Análisis de desviaciones:** Busca datos distintos, raros, diferentes en comparación con el resto de los datos obtenidos.

   - **Análisis de similitud en series temporales:** Detecta todas las ocurrencias de secuencias similares en una colección de series temporales.

# CICLO DE UN PROYECTO DE MINERÍA DE DATOS



# CICLO DE UN PROYECTO DE MINERÍA DE DATOS

1. **Entendimiento del negocio:**
   - Formulación del problema de negocio (uno de los ya antes mencionados: previsión, gestión de riesgos, segmentación de clientes, etc).

2. **Entendimiento de los datos:**
   - Recolección de datos.

3. **Preparación de los datos:**
   - Transformación de datos: Generalmente, el formato de los datos contenidos en las fuentes de datos no es el idóneo, y la mayoría de las veces no es posible aplicar algún algoritmo de minería sobre los datos iniciales sin que requieran algún cambio (Por ejemplo, transformaciones numéricas).

# CICLO DE UN PROYECTO DE MINERÍA DE DATOS

**3. Preparación de los datos:**

- Limpieza o filtrado de datos: En esta fase se filtran los datos con el objetivo de eliminar valores erróneos o desconocidos, según las necesidades y el algoritmo a utilizar.

- Preprocesado: Se analizan las propiedades de los datos, en especial los histogramas, diagramas de dispersión, presencia de valores atípicos y ausencia de datos (valores nulos) y se obtienen muestras de los datos en busca de mayor velocidad y eficiencia de los algoritmos o se reduce el número de valores posibles, mediante tareas como: Redondeo, Agrupación y Agregación

# CICLO DE UN PROYECTO DE MINERÍA DE DATOS

4. **Modelado:** Creación del modelo.
   - Selección de variables: Después de haber sido preprocesados y realizar la limpieza de datos, se sigue teniendo una cantidad enorme de variables o atributos. La selección de características reduce el tamaño de los datos, eligiendo las variables más influyentes del problema.

   - Extracción de Conocimiento: La extracción del conocimiento es la esencia de la Minería de Datos donde mediante una técnica, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. Estos modelos se representan mediante: reglas, árboles y redes neuronales.

# CICLO DE UN PROYECTO DE MINERÍA DE DATOS

**5. Evaluación:**
Evaluación de la integridad del modelo en el negocio. Una vez obtenido el modelo, se procede a su validación; comprobando que las conclusiones obtenidas son válidas y satisfactorias.

**6. Implantación:**
Integración en aplicaciones para solucionar el problema de negocio expuesto.

# ALGORITMOS DE MINERÍA DE DATOS

- **Forecasting (Predicción) :** Dada una tendencia de los datos se busca cuál será su previsión.

- **Supervisados (predictivos):** Predicen un dato (o un conjunto de ellos) desconocido a priori, a partir de otros conocidos.

- **No supervisados (o del descubrimiento del conocimiento):** Se descubren patrones y tendencias en los datos.

# INTELIGENCIA DE NEGOCIOS (BI)

- Se define como la habilidad para transformar los datos en información, y la información en conocimiento, de forma que se pueda optimizar el proceso de toma de decisiones en los negocios.

- Conjunto de procesos, aplicaciones y tecnologías que facilitan la obtención rápida y sencilla de datos provenientes de los sistemas de gestión empresarial para su análisis e interpretación, de manera que puedan ser aprovechados para la toma de decisiones y se conviertan en conocimiento para los responsables del negocio.

- Una solución BI completa permite:
  - Observar ¿qué está ocurriendo?
  - Comprender ¿por qué ocurre?
  - Predecir ¿qué ocurriría?
  - Colaborar ¿qué debería hacer el equipo?
  - Decidir ¿qué camino se debe seguir?

# MINERÍA DE DATOS Y SUS RELACIONES

Trabajan todos juntos para cumplir el propósito de **proporcionar información basada en datos**.

| MINERÍA DATOS | ESTADISTICA | TIC'S | BI |
|---|---|---|---|
| • Se buscan patrones ocultos en los datos mediante métodos algorítmicos y utilizando un sistema automatizado.<br>• Se utiliza para modelos predictivos<br>• Es un campo de la Estadística y de las Ciencias de la Computación | Trabaja con muestras de la información | Hace uso de tecnologías de la información para el análisis de datos | Requiere un enfoque más analítico por parte del usuario, sin necesidad de recurrir a modelos matemáticos predictivos<br><br>Se consideraría la categoría general en la que habitan la estadñistica y la minería. |

# MINERÍA DE DATOS Y SUS CAMPOS DE APLICACIÓN

Los campos de aplicación de estas nuevas técnicas dentro de la industria son numerosos:

- control de calidad,
- identificación de sistemas,
- determinación de causas en fallos del proceso,
- detección de anomalías,
- prevención de fallos,
- modelización de sistemas,
- obtención de reglas y patrones de comportamiento,
- búsqueda de causas y relaciones entre variables, etc.