Introducción a la Ciencia de Datos

Héctor Alán de la Fuente Anaya

hector.delafuente@cinvestav.mx

19 de octubre del 2022

Contenido

Cantidad de datos

El mundo real gira entorno a los datos

- Ciencia
  - Bases de datos de astronomía, genómica, datos medioambientales, datos de transporte, ...
- Ciencias Sociales y Humanidades
  - Libros escaneados, documentos históricos, datos sociales, ...
- Negocios y Comercio
  - Ventas de corporaciones, transacciones de mercado, censos, tráfico de aerolíneas, ...
- Entretenimiento y Ocio
  - Imágenes en internet, películas, ficheros, MP3, ...
- Medicina
  - Datos de pacientes, datos de escáner, radiografías, ...
- Industria, Energía, ...
  - Sensores, ...

Cantidad de datos

Hoy en día estamos inundados de datos:

- Creación de herramientas para la recolección de información
- Avance en la tecnología de base de datos
- Reducción en costos del hardware
- Disponemos de cantidades gigantescas de datos almacenados en bases de datos, datawarehouses y otros tipos de almacenes de información
- De acuerdo con la Estrategia Europea de Datos, se prevé un incremento en el volumen global de datos. De los 33 zettabytes en 2018 a 175 zettabytes para 2025.

Riqueza en datos y pobreza en conocimiento

El progreso y la innovación ya no se ven obstaculizados por la capacidad de recopilar datos, sino por la capacidad de **descubrir conocimiento de los datos** recopilados, de manera oportuna y en una forma escalable.

Contenido

¿Qué es la ciencia de datos?

- Aún no existe una definición consensuada.
- Convergencia multidisciplinar de temas actuales
- Ciencia de Datos es el área de conocimiento que engloba todo lo relacionado con el análisis de datos masivos.

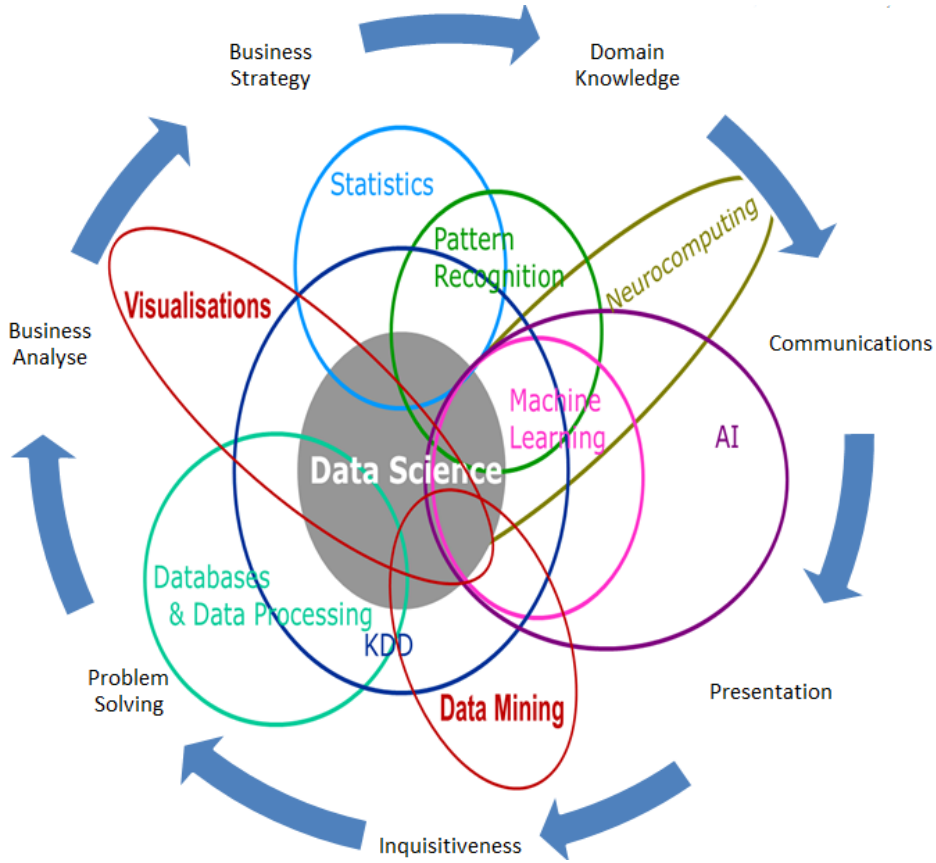
Ciencia de datos

Se refiere a un área de trabajo emergente relacionada con la **recopilación**, **preparación**, **análisis**, **visualización**, **gestión** y **preservación** de grandes colecciones de información, utilizando técnicas de diferentes campos.[a]

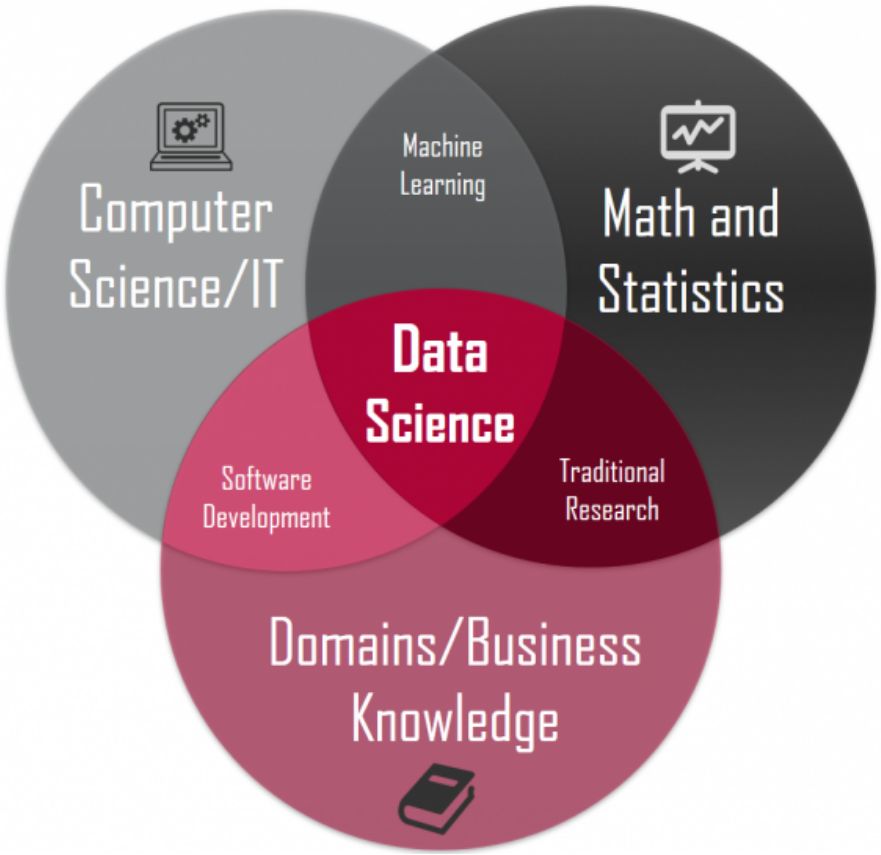
[a] 2012, Jeffrey Stanton, "An Introductión to Data Science"

La ciencia de datos es multidisciplinaria

La ciencia de datos es **multidisciplinaria**. Se basa en técnicas y tareas de muchos campos.



2012, Brendan Tierney



2018, Data Science Society

Objetivo de la Ciencia de datos

Objetivo:
Extraer conocimiento de datos y la creación de productos de información.

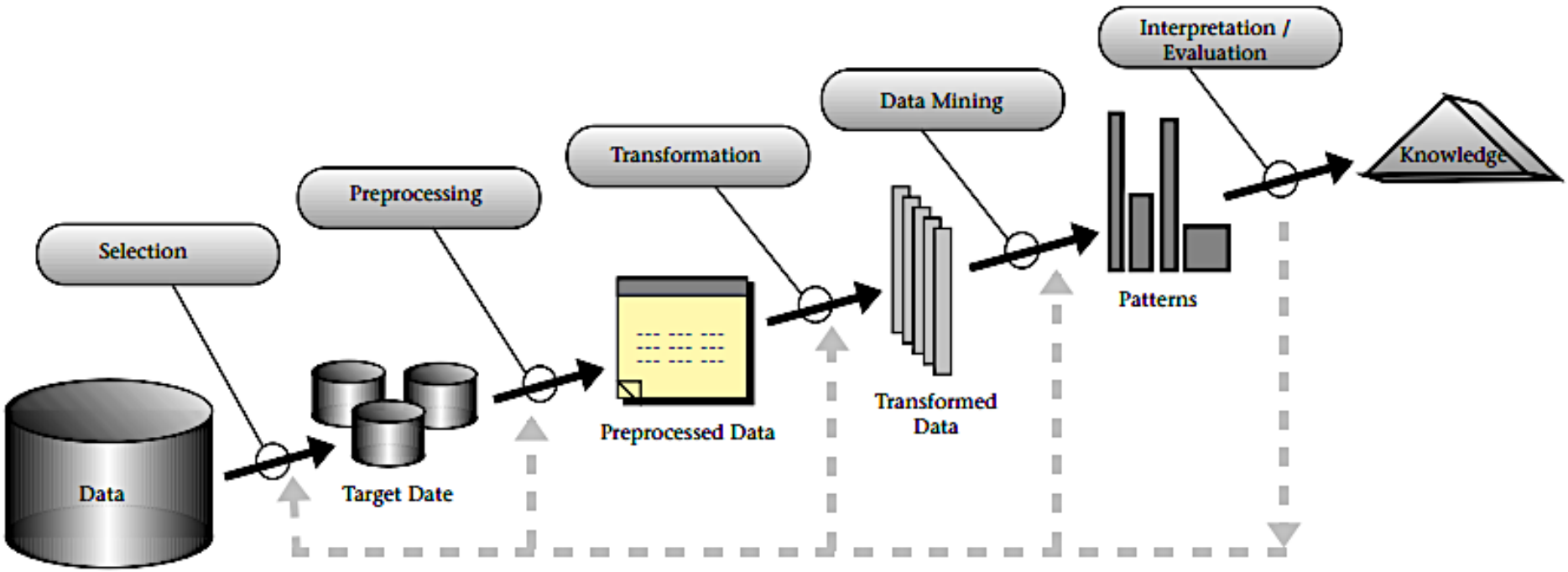
- La ciencia de datos busca utilizar todos los datos disponibles y relevantes para **extraer conocimiento** que pueda ser fácilmente comprendido por los expertos en el área de aplicación.

- Para extraer conocimiento se necesita que los datos sean:
  - **Almacenados**
  - **Gestionados**
  - **Analizados**

Contenido

Proceso KDD

El Descubrimiento de conocimiento en bases de datos (KDD, del inglés Knowledge Discovery in Databases) es un proceso automático que consiste en **descubrir patrones** en forma de reglas o funciones, a partir de los datos, para que el usuario los **analice**.



2014, Bernhard Hitpass

Proceso KDD

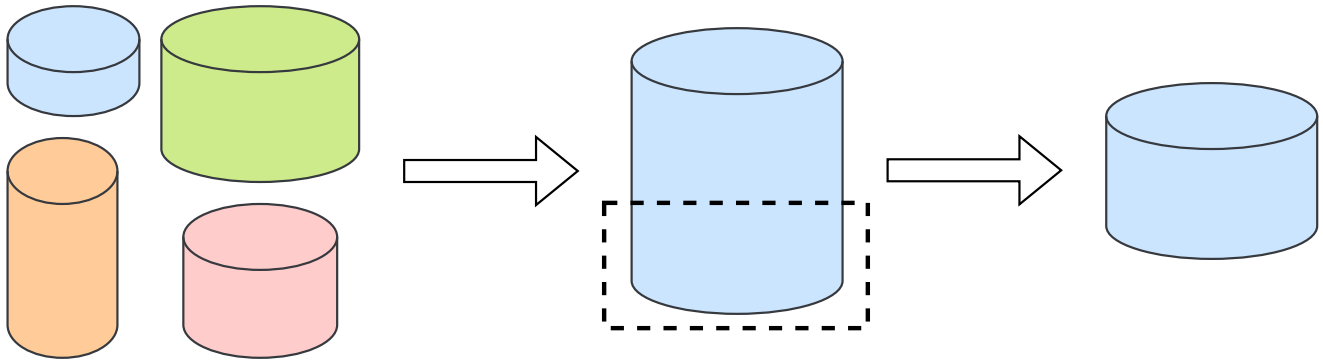
Comprensión del dominio del estudio y establecimiento de objetivos

- Desarrollo de un entendimiento sobre el dominio
- Descubrimiento de conocimiento previo que sea relevante
- Definición del objetivo del KDD
- Se identifica el conocimiento relevante y prioritario y se definen las metas del proceso KDD, desde el punto de vista del usuario final.

Proceso KDD

Etapa 1: Selección

- Selección e integración de los datos objetivo provenientes de fuentes **múltiples** y **heterogéneas**.
- Se crea un conjunto de datos objetivo, seleccionando todo el conjunto de datos o una muestra representativa de este, sobre el cual se realiza el proceso de descubrimiento.
- La selección de los datos varía de acuerdo con los objetivos del KDD.

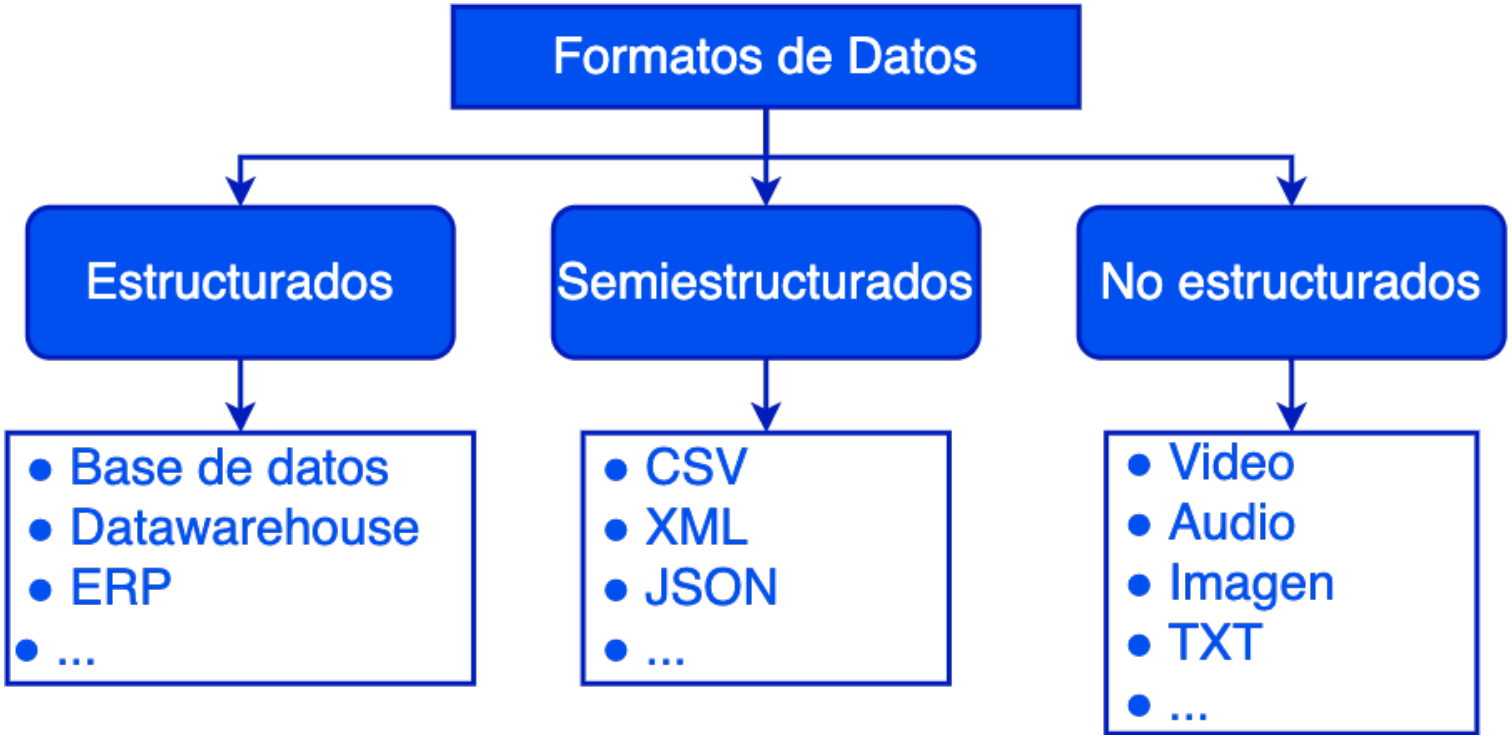


Etapa 1: Selección

Ejemplo:

En un hospital pueden encontrarse datos del personal médico, de pacientes, citas, farmacia, facturación, estudios de sangre, radiografías, etc., presentes en **diferentes formatos**.

Etapa 1: Selección



Proceso KDD

Etapa 2: Preprocesamiento
Datos sin calidad provocan resultados sin calidad.

Preprocesamiento de datos

El preprocesamiento de los datos mejorar la calidad de un conjunto de datos para que las técnicas de extracción de conocimiento puedan obtener mayor y mejor información.

Los datos en el mundo real son sucios:

- Incompletos: atributos con valores insuficientes, atributos de interés insuficientes, o que contienen sólo datos agregados.
- Ruidosos: contienen errores o outliers.
- Inconsistentes: contienen discrepancias en códigos, nombres, etc.

Etapa 2: Preprocesamiento

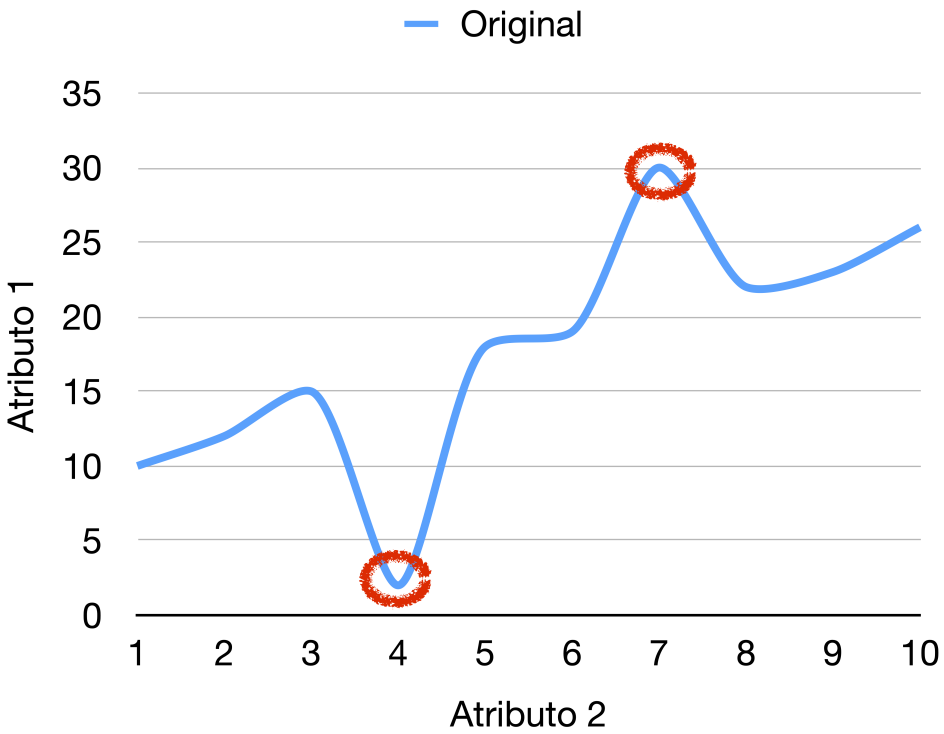
Aunque las técnicas de extracción de conocimiento sean correctas, las decisiones deben basarse en datos de calidad.

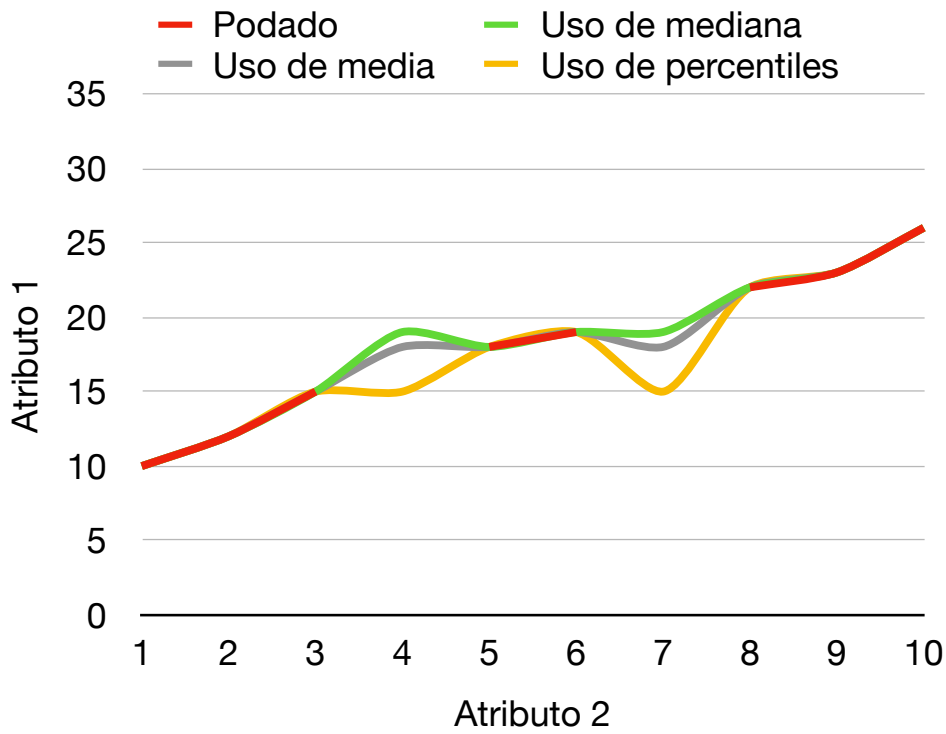
Tareas de preprocesamiento

1. Limpieza: Consiste en arreglar o eliminar los datos incorrectos, corruptos, mal formateados, duplicados o incompletos de un conjunto de datos.
2. Integración: Integra múltiples fuentes de datos atendiendo redundancia, incoherencia, duplicidad.

Etapa 2: Preprocesamiento

Limpieza de datos: Atendiendo outliers





Proceso KDD

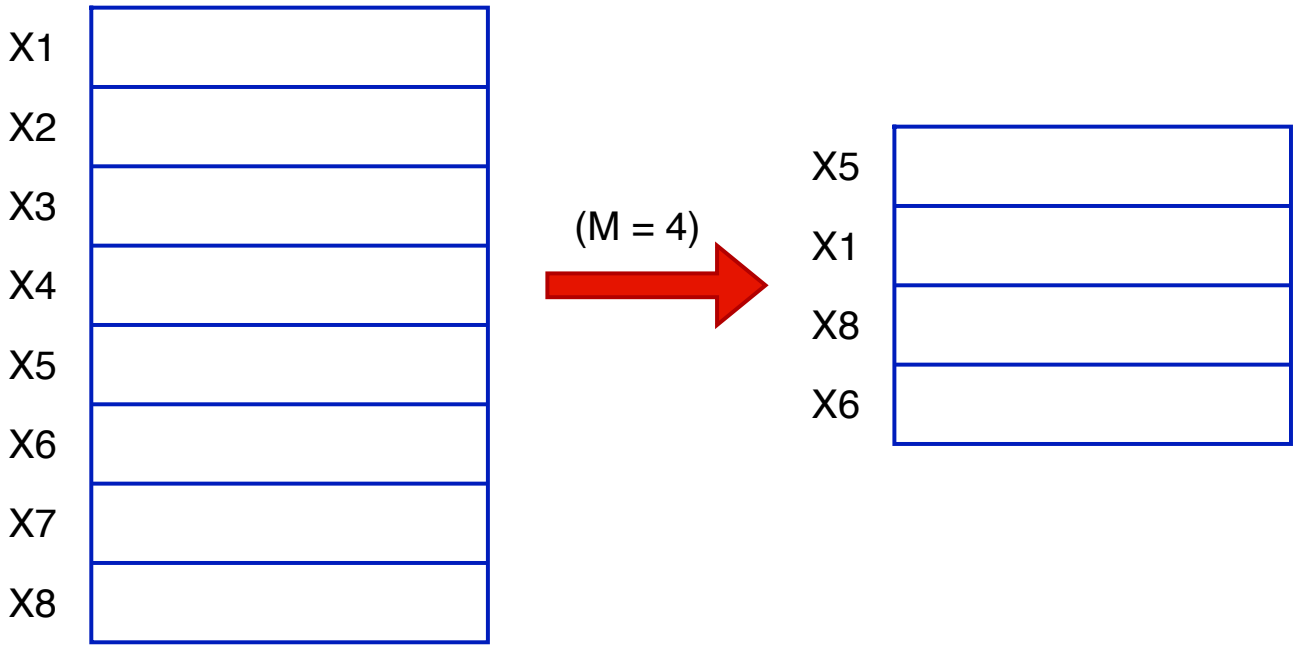
Etapa 3: Transformación

- En la etapa de transformación de datos, se buscan características útiles para representar los datos dependiendo de la meta del KDD.
- Cambia el formato, estructura o valores de los datos para ser utilizable de forma eficiente.
  1. Reducción: Produce una representación más pequeña de los datos dando resultados iguales o similares a los originales.
  2. Discretización: Convierte los valores de los atributos de los datos continuos en un conjunto finito de intervalos con la mínima pérdida de datos.
  3. Resumen: Presenta un informe para comprender las tendencias y los patrones del conjunto de datos de forma simplificada.
  4. Agregación: Utiliza la agregación en varios niveles de un cubo de datos para representar el conjunto de datos original.
  5. Normalización: Evita que una variable sea demasiado influyente, especialmente si se mide en diferentes unidades.

Etapa 3: Transformación

Reducción: Muestreo aleatorio

Del conjunto de datos X se extrae una muestra aleatoria de M = 4.



Etapa 3: Transformación

Discretización

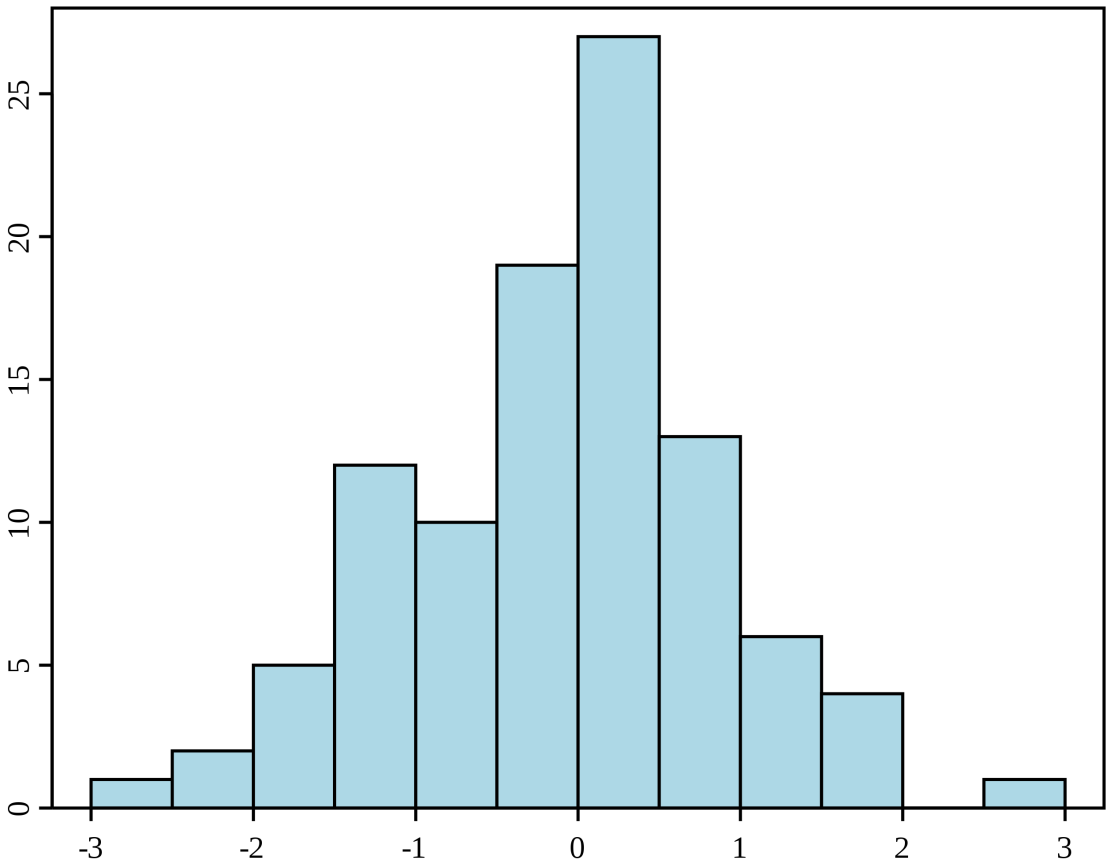
Edad
1
5
4
11
14
17
13
31
33
36
70
74

Edad	Discretización
1,5,4	Niñez
11,14,17,13	Juventud
31,33,36	Adultez
70,74	Vejez

Etapa 3: Transformación

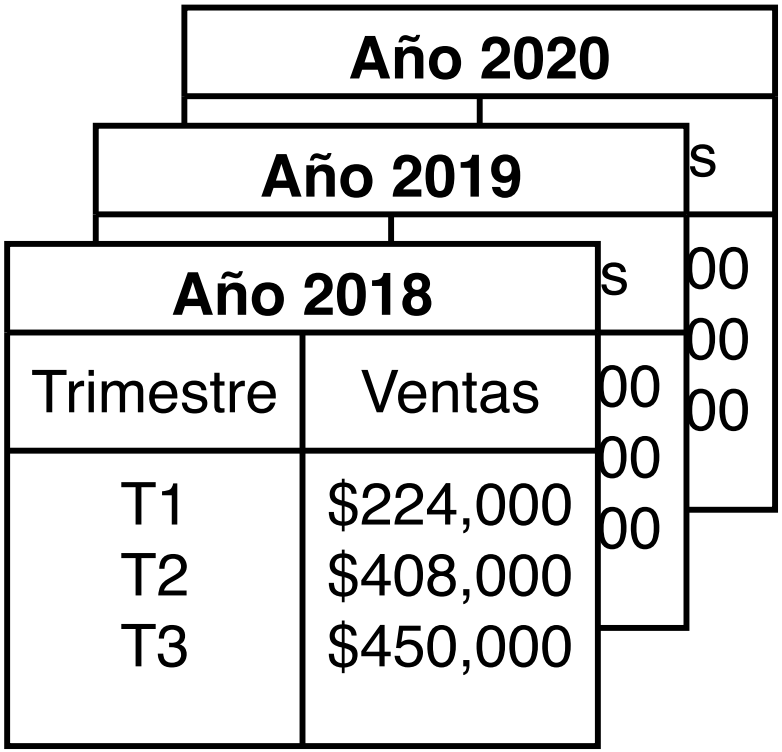
Resumen: Histograma de frecuencias

Representación gráfica de los datos en forma de barras, donde cada barra es proporcional a la frecuencia de los valores.



Etapa 3: Transformación

Agregación



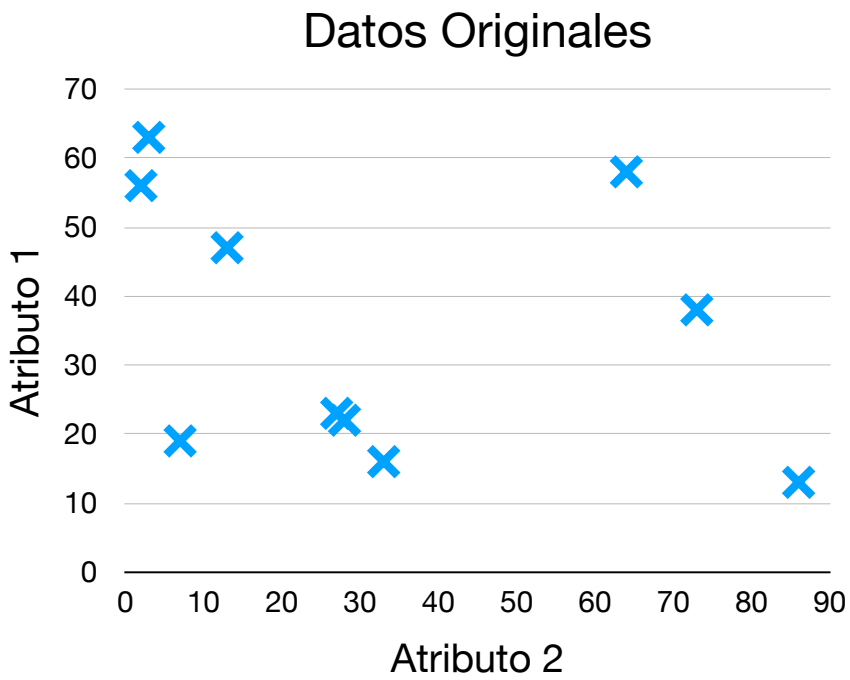
Año	Ventas
2018	$1,568,000
2019	$2,356,000
2020	$1,082,000

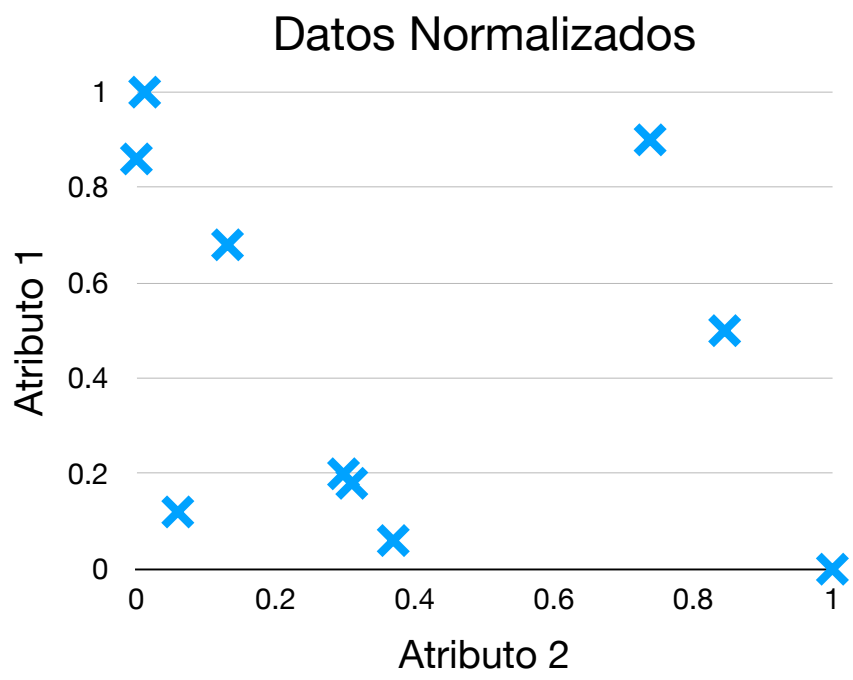
Etapa 3: Transformación

Normalización

Normalización min-max

$$x_{norm} = \frac{x - min(x)}{max(x) - min(x)}$$





Proceso KDD

Etapa 4: Minería de Datos

Tiene como objetivo la búsqueda y descubrimiento de patrones insospechados y de interés, aplicando tareas de descubrimiento.

- Las técnicas de minería de datos crean modelos que son predictivos.
- Los modelos predictivos pretenden estimar valores futuros o desconocidos.

Por ejemplo:

- Predecir si nuevos clientes son buenos o malos basados en su estado civil, edad, género y profesión.
- Determinar si nuevos estudiantes desertan o no en función de su zona de procedencia, facultad, estrato, género, edad y promedio de notas.

Etapa 4: Minería de Datos

Un algoritmo de minería de datos realiza una búsqueda de patrones en los datos, así como la decisión sobre los modelos y los parámetros más apropiados, dependiendo del tipo de datos (categóricos, numéricos) por utilizar.

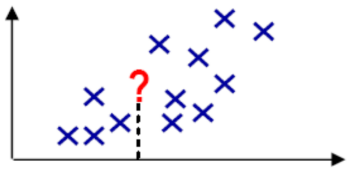
Métodos en minería de datos:

- Predictivos: Entrenan a un algoritmo por medio de datos para predecir una variable. Describe una instancia en relación con todas las demás.
- Descriptivos: Secciona los datos en grupos insospechables de antemano para mejorar la comprensión del conjunto total.

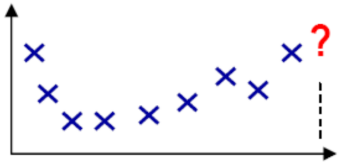
Etapa 4: Minería de Datos

Predictivos

- ***Interpolación***:



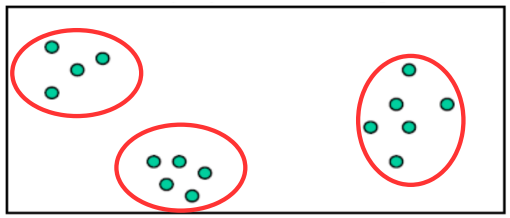
- ***Predicción secuencial***:
1,2,3,5,7,11,13,17,19, ...?



- ***Aprendizaje supervisado***:
(1,3) -> Sí
(3,5) -> Sí (4,2) -> ?
(7,2) -> No

Descriptivos

- ***Aprendizaje no supervisado:***



- ***Análisis exploratorio***:
  - -Correlaciones
  - -Asociaciones
  - -Dependencias

Proceso KDD

Etapa 5: Interpretación

- Se interpretan los patrones descubiertos
- Visualización de los patrones extraídos
- Remoción de los patrones redundantes o irrelevantes
- Traducción de los patrones útiles en términos entendibles para el usuario.
- Opcionalmente, se planea iteración futura.

Contenido

Aplicaciones

- Salud: Optimización de los diagnósticos médicos, análisis de las bases de datos clínicas, y detección temprana de enfermedades.
- Procesos productivos: Automatización de procesos, monitoreo y control de calidad y optimización de los sistemas de mantenimiento.
- Procesos comerciales: Determinación de patrones de consumo en los clientes, experiencias personalizadas, sistemas de precios dinámicos y atención al cliente con sistemas de inteligencia artificial.
- Comunicaciones: Interpretar patrones y conductas humanas. Se utiliza análisis de texto, análisis de emociones, analítica de imágenes y videos, y predicción de fake news.
- Recursos humanos: Estimar la adaptación y el aporte de un candidato para valorar el desempeño de los empleados o proyectar la probabilidad de abandono del puesto laboral.

Herramientas de Ciencia de Datos

Weka

Descarga:

`https://waikato.github.io/weka-wiki/downloading_weka`

Contacto:

M.C. Héctor Alán de la Fuente Anaya
hector.delafuente@cinvestav.mx