MINERÍA DE DATOS

Maximiliano Ojeda

muojeda@uc.cl

IIC-2433

Minería de Datos

¿De qué trata este curso?

El curso tiene como objetivos principales:

- Comprender las teorías y prácticas fundamentales de la Minería de Datos.
- Dominar las técnicas clave para diseñar programas que extraigan conocimiento de diversas fuentes y tipos de datos.
- Adquirir sólidos fundamentos teóricos que permitan elegir la herramienta más adecuada, conociendo sus ventajas y limitaciones.
- Experimentar de forma práctica en un entorno realista la aplicación de dichas técnicas y herramientas.

Contenido

Preprocesamiento y Transformación de Datos

- Librerías para trabajo con datos
- Limpieza de datos
- Técnicas de Reducción de Dimensionalidad

Modelos de predicción supervisados

- Vecinos cercanos (KNN)
- Árboles
- Inferencia Causal

Clustering

- K-means
- DBSCAN / HDBSCAN
- Mixture of Gaussians

Otros

- Reglas de Asociación
- Redes Bayesianas
- Información Semi-estructurada

Metodología

Clase Expositiva

En el primer bloque de la clase se verán contenidos teóricos

Clase Práctica

Laboratorio práctico para realizar **durante la clase**

Ayudantías

El bloque de los días jueves se reservará únicamente para días de presentaciones. (2 o 3 días del semestre)

Evaluación

Ponderación

- Tarea 1: 20%
- Tarea 2: 20%
- Tarea 3: 20%
- Proyecto: 40%

Ponderación Proyecto

- Propuesta: 10%
- Avance: 30%
- Entrega Final: 60%

Importante

- Restricciones Aprobación: **Tarea 3 ≥ 4.0** y **Proyecto ≥ 4.0**
- No hay Interrogaciones ni Examen
- Actividades Formativas en clases: Una décima al promedio final por actividad, son 9 o 10 actividades en el semestre

Herramientas del Curso

- **Jupyter Notebook**

- **Google Colab**

- **GitHub**

Repositorio del Curso

En el GitHub del curso encontraran cada semana los archivos necesarios para realizar la actividad formativa, además de las presentaciones y material adicional.

https://github.com/MaxOjeda/IIC2433

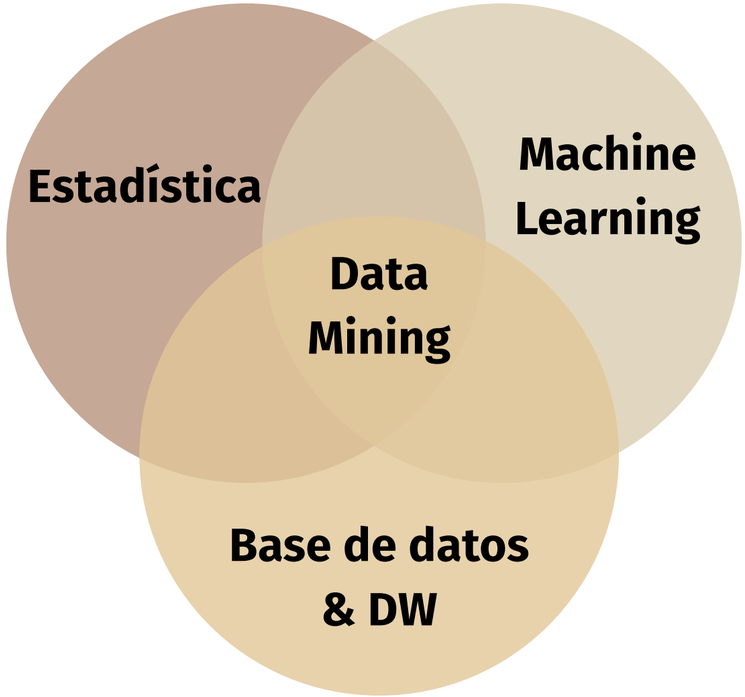
Introducción

Minería de Datos

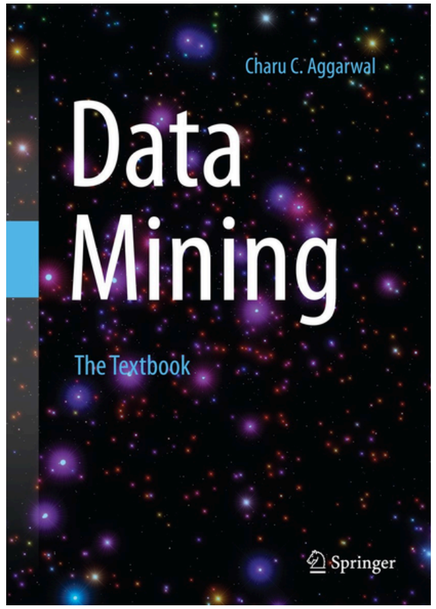
“La minería de datos es la extracción de información implícita, previamente desconocida y potencialmente útil de los datos. Se centra en construir programas informáticos que examinan bases de datos automáticamente en busca de regularidades o patrones”

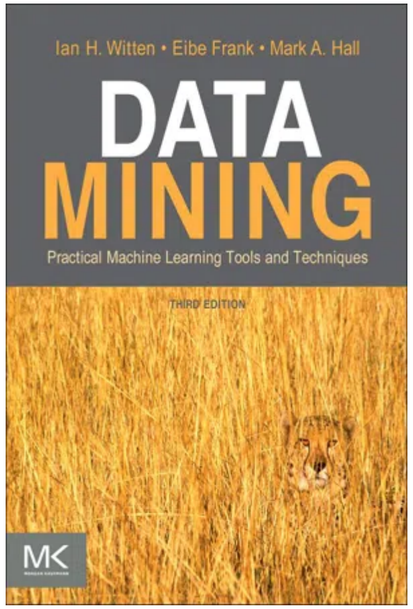
Data Mining: Practical Machine Learning Tools and Techniques
Ian H. Witten & Eibe Frank

Minería de Datos



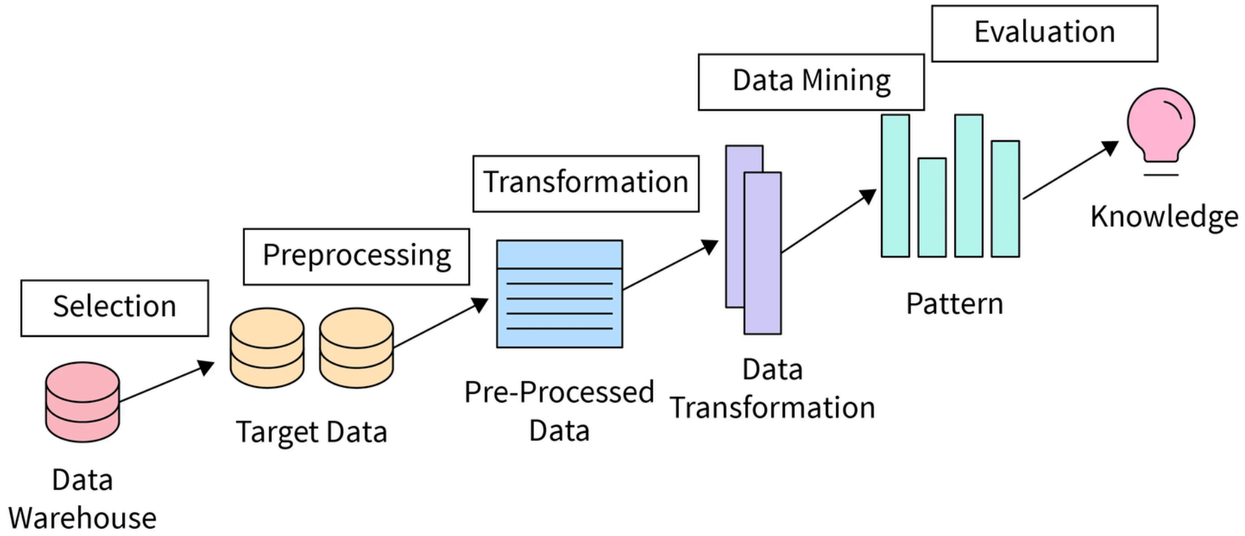
¿Por qué “minería”?





Así como un minero extrae oro oculto en una roca aparentemente inútil, la minería de datos extrae conocimiento valioso de “materia prima” (datos brutos)

Knowledge Discovery in Databases (KDD)



Importancia de Data Mining

Radica en su capacidad para transformar grandes volúmenes de datos brutos en conocimiento valioso

Toma de decisiones basada en datos

Extrae patrones que respaldan decisiones estratégicas en marketing, finanzas, operaciones o innovación

Optimización de procesos y eficiencia

Permite detectar cuellos de botella y planificar mantenimiento predictivo

Ventaja competitiva sostenible

Permite descubrir oportunidades de mercado y reaccionar a nuevos comportamientos del cliente

Casos Prácticos

Comercio minorista y comercio electrónico 🛒

Se utilizan reglas de asociación para sugerir productos complementarios o anticipar demanda en periodos especiales. “Dippers & Beers”

Astronomía 🪐

Se descubrieron más de 5000 exoplanetas con técnicas de minería que permitieron identificaron diminutas caídas de brillo

Finanzas 💳

Bancos y emisores procesan millones de transacciones en tiempo real; algoritmos de árboles y redes neuronales reducen pérdidas por fraude hasta 40 % al bloquear patrones atípicos

Preprocesamiento de Datos

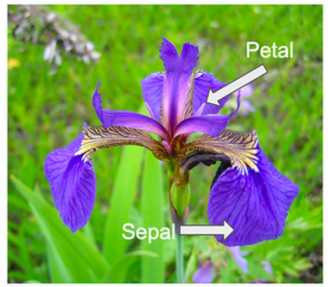
“Success depends upon previous preparation, and without such preparation there is sure to be failure” - Confucius

Datos estructurados

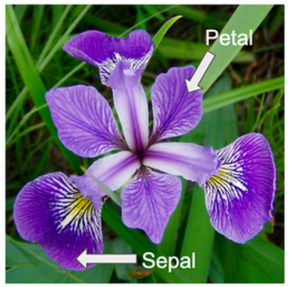
Datos organizados en un formato predefinido, lo que los hace fácil de comprender tanto para las personas como para las máquinas. Normalmente se almacena en tablas con filas y columnas.

Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa
13	4.8	3	1.4	0.1	Iris-setosa
14	4.3	3	1.1	0.1	Iris-setosa
15	5.8	4	1.2	0.2	Iris-setosa
16	5.7	4.4	1.5	0.4	Iris-setosa
17	5.4	3.9	1.3	0.4	Iris-setosa
18	5.1	3.5	1.4	0.3	Iris-setosa
19	5.7	3.8	1.7	0.3	Iris-setosa

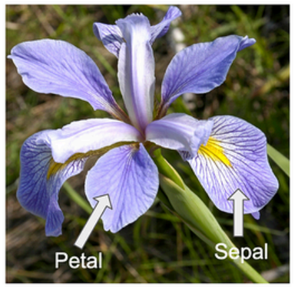
Iris setosa



Iris versicolor



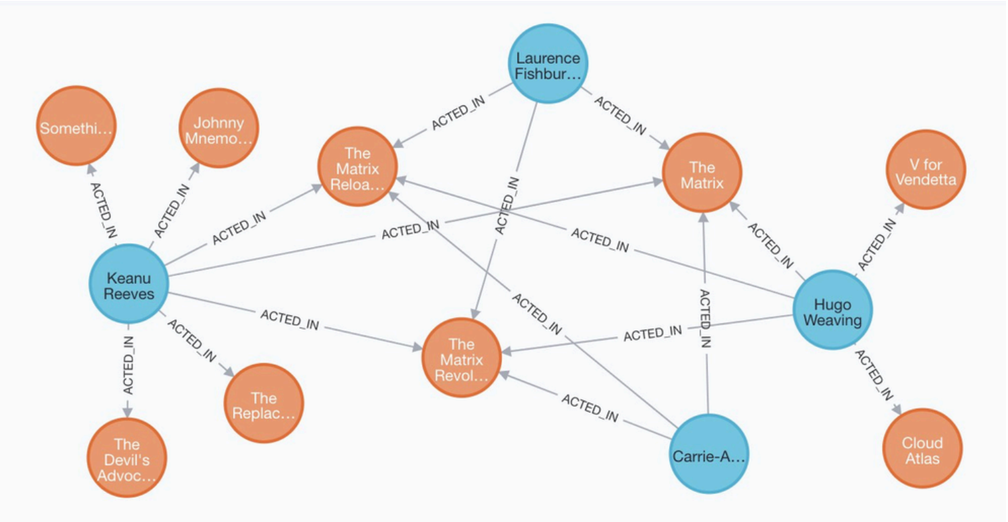
Iris virginica



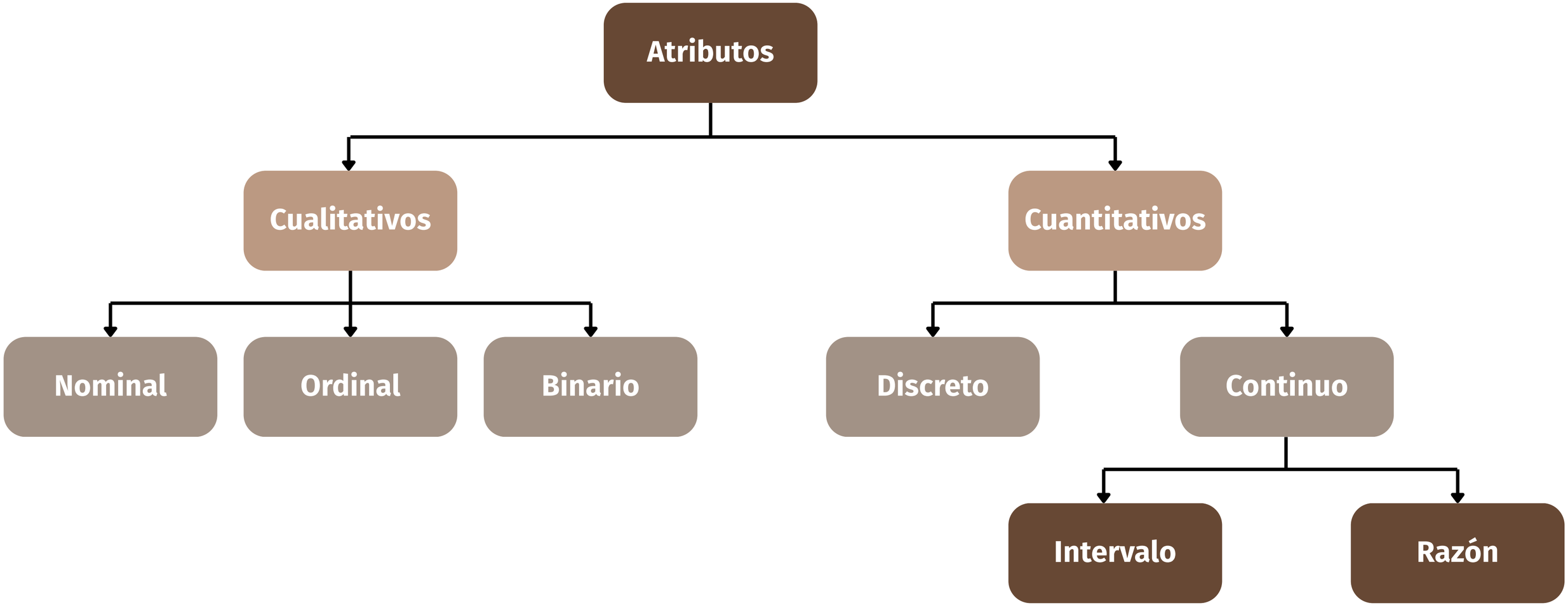
Datos no estructurados

Datos que no se ajusta a un formato predefinido ni presenta una estructura específica que permita organizarla fácilmente en filas y columnas. A menudo se almacenan en su formato nativo e incluyen una gran variedad de formatos, como texto, audio, imágenes, grafos.

	text	polarity
0	Bromwell High is a cartoon comedy. It ran at t...	1
1	Homelessness (or Houselessness as George Carli...	1
2	Brilliant over-acting by Lesley Ann Warren. Be...	1
3	This is easily the most underrated film inn th...	1
4	This is not the typical Mel Brooks film. It wa...	1



Tipos de Atributos



Data Portability

Numérico a Categórico

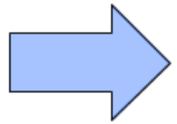
Esta conversión es de las más comunes, discretización. Consiste en dividir los valores de un atributo número en ϕ rangos

Edad	Rango	Edad_cat
15	[10, 20)	1
24	[20, 30)	2
22	[20, 30)	2
10	[10, 20)	1

Categórico a Numérico

En muchos algoritmos de minería de datos es deseable convertir datos categóricos a numéricos

Color
Red
Red
Yellow
Green
Yellow



Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

Data Cleaning

Datos faltantes

Entradas no especificadas durante la recolección o por la naturaleza de los datos

Datos erróneos

Información de viene de distintas fuentes puede causar una inconsistenacia

Scaling

Los datos están expresados en escalas muy distintas (ej: edad vs salario)

Datos Faltantes

	School ID	Name	Address	City	Subject	Marks	Rank	Grade
0	101.0	Alice	123 Main St	Los Angeles	Math	85.0	2	B
1	102.0	Bob	456 Oak Ave	New York	English	92.0	1	A
2	103.0	Charlie	789 Pine Ln	Houston	Science	78.0	4	C
3	NaN	David	101 Elm St	Los Angeles	Math	89.0	3	B
4	105.0	Eva	NaN	Miami	History	NaN	8	D
5	106.0	Frank	222 Maple Rd	NaN	Math	95.0	1	A
6	107.0	Grace	444 Cedar Blvd	Houston	Science	80.0	5	C
7	108.0	Henry	555 Birch Dr	New York	English	88.0	3	B

- Cualquier registro de datos que contenga un valor faltante puede eliminarse por completo.
- Los valores faltantes pueden imputarse.

¿Qué desventajas tienen estas posibles soluciones?

Datos Inconsistentes

Edad	Color_ojos	Fecha
15	Café	16-06-1998
24	Verde	18 de Abril 1992
22	café	27-01-2003
10	Cafe	24/12/1990

Scaling & Normalization

En muchos escenarios, las distintas características representan escalas de referencia diferentes y, por lo tanto, pueden no ser comparables entre sí.

Standarization

$$z_i^j = \frac{x_i^j - \mu_j}{\sigma_j}$$

Min-Max Scaling

$$y_i^j = \frac{x_i^j - min_j}{max_j - min_j}$$

MINERÍA DE DATOS

Maximiliano Ojeda

muojeda@uc.cl

IIC-2433