



Minería de Datos

PhD Jose Ricardo Zapata
< joser.zapata@upb.edu.co >

UNIVERSIDAD PONTIFICIA BOLIVARIANA
FACULTAD TIC
2018

AGENDA



1. Introducción
2. Metodologías
3. Herramientas
4. Análisis de Casos
5. Preparación de Datos
6. Técnicas y Algoritmos
7. Minería de Texto
8. Bibliografía

AGENDA



1. Introducción
2. Metodologías
3. Herramientas
4. Análisis de Casos
5. Preparación de Datos
6. Técnicas y Algoritmos
7. Minería de Texto
8. Bibliografía

PREGUNTA

¿Cuál es el tema de las siguientes imágenes?



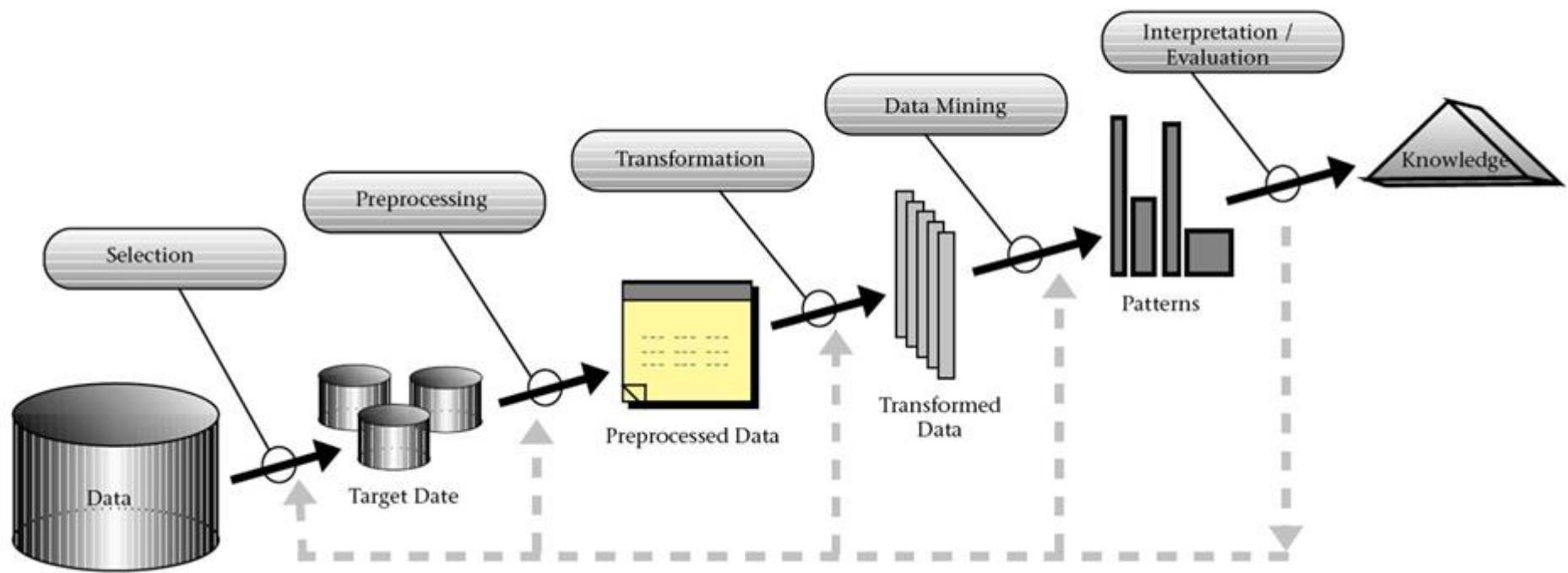


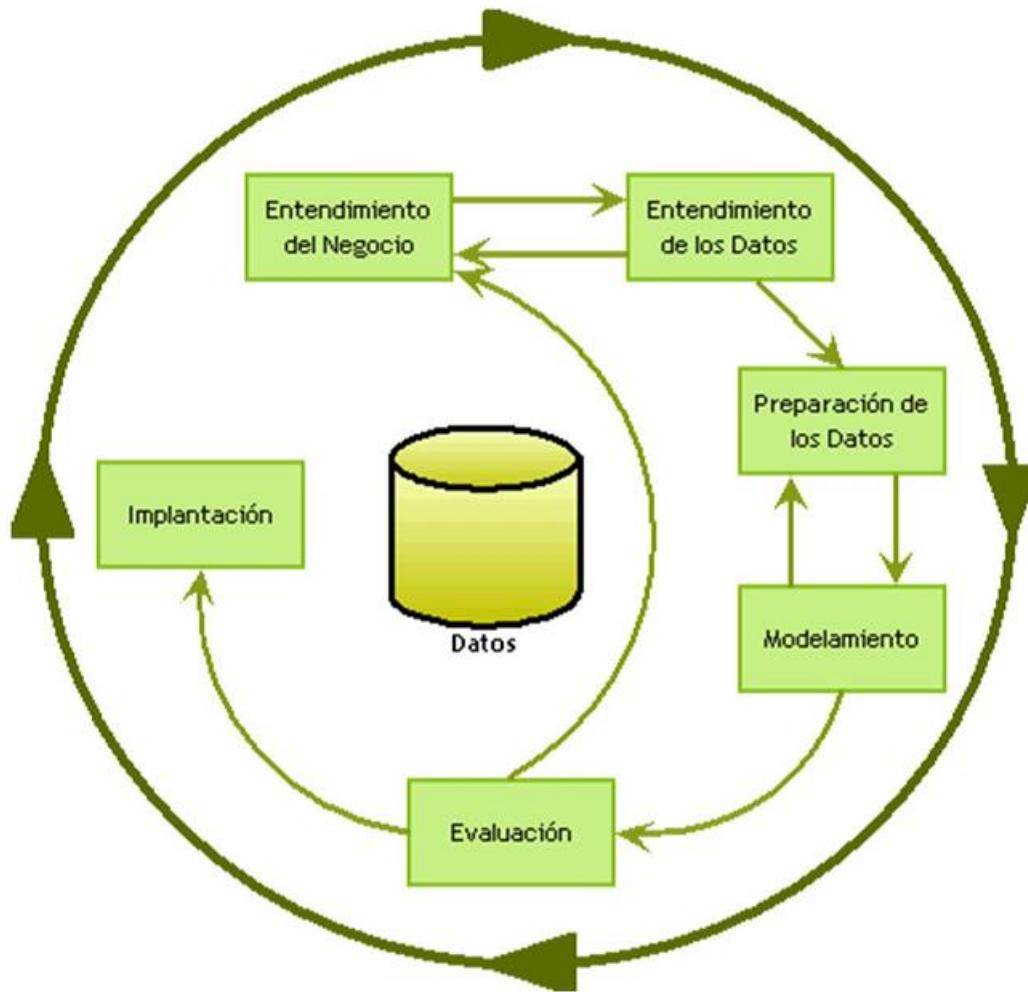


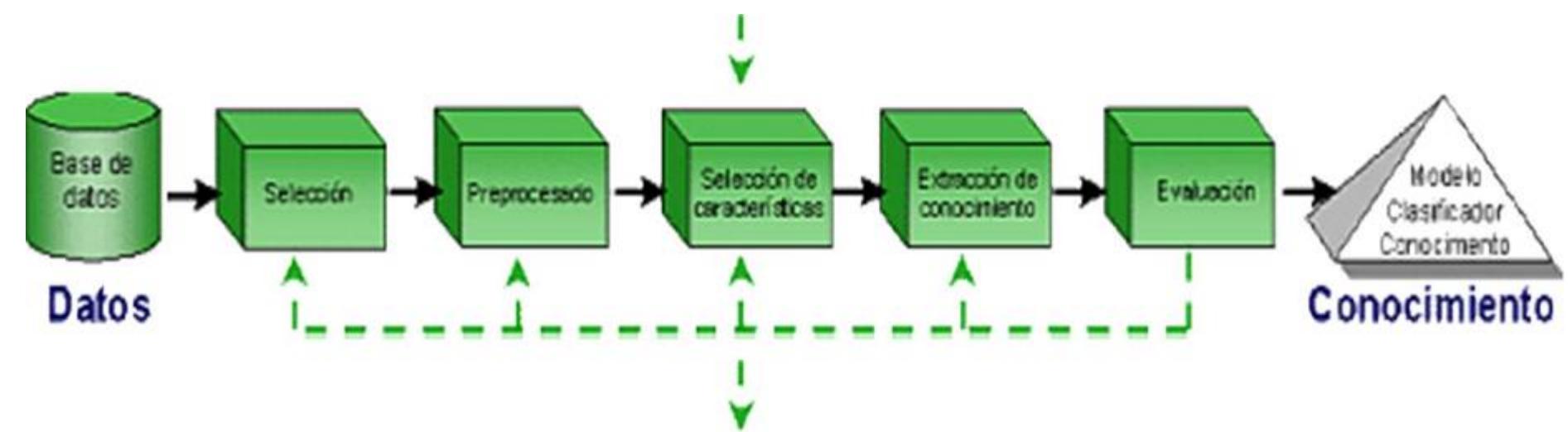


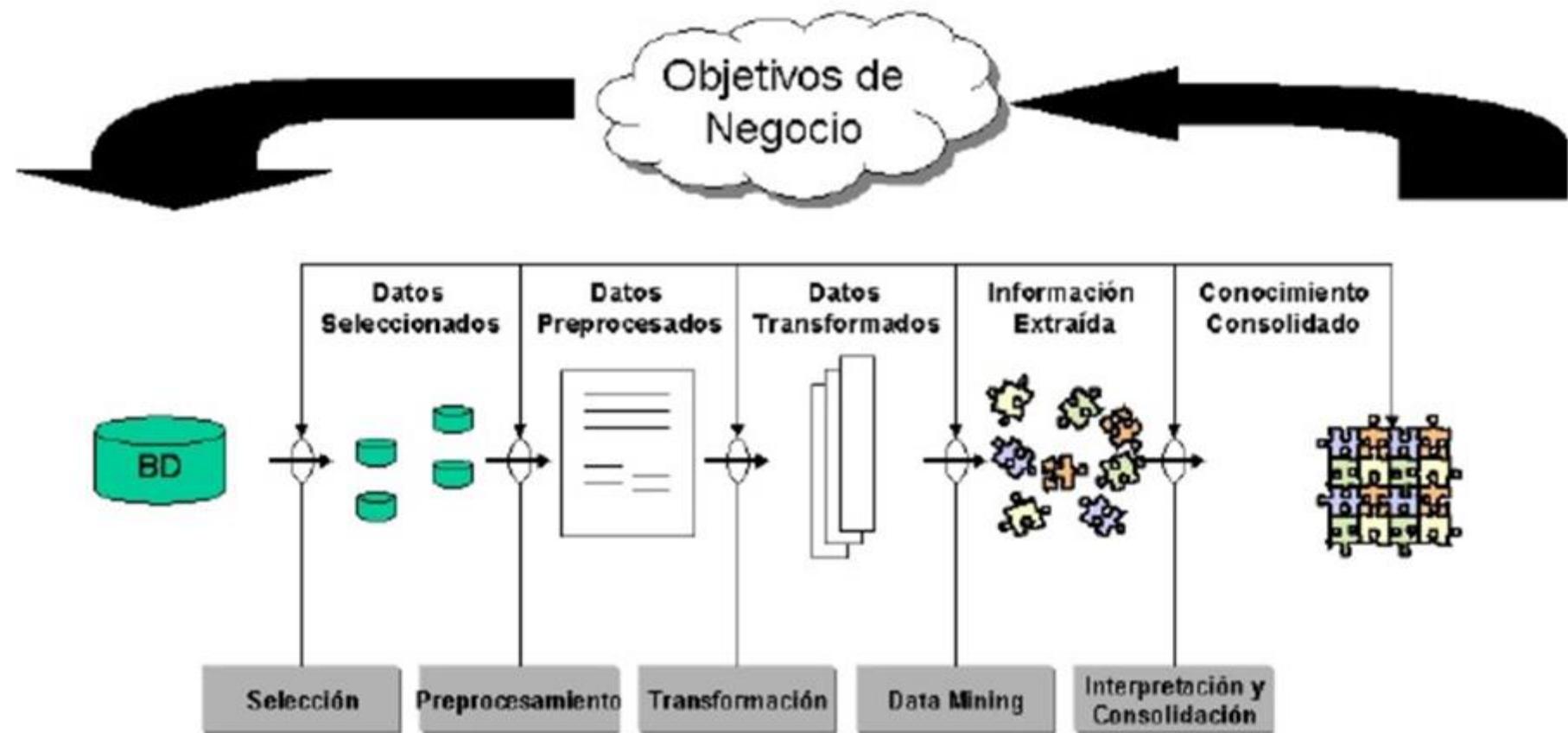
PREGUNTA

¿Cuál es el tema de las siguientes imágenes?









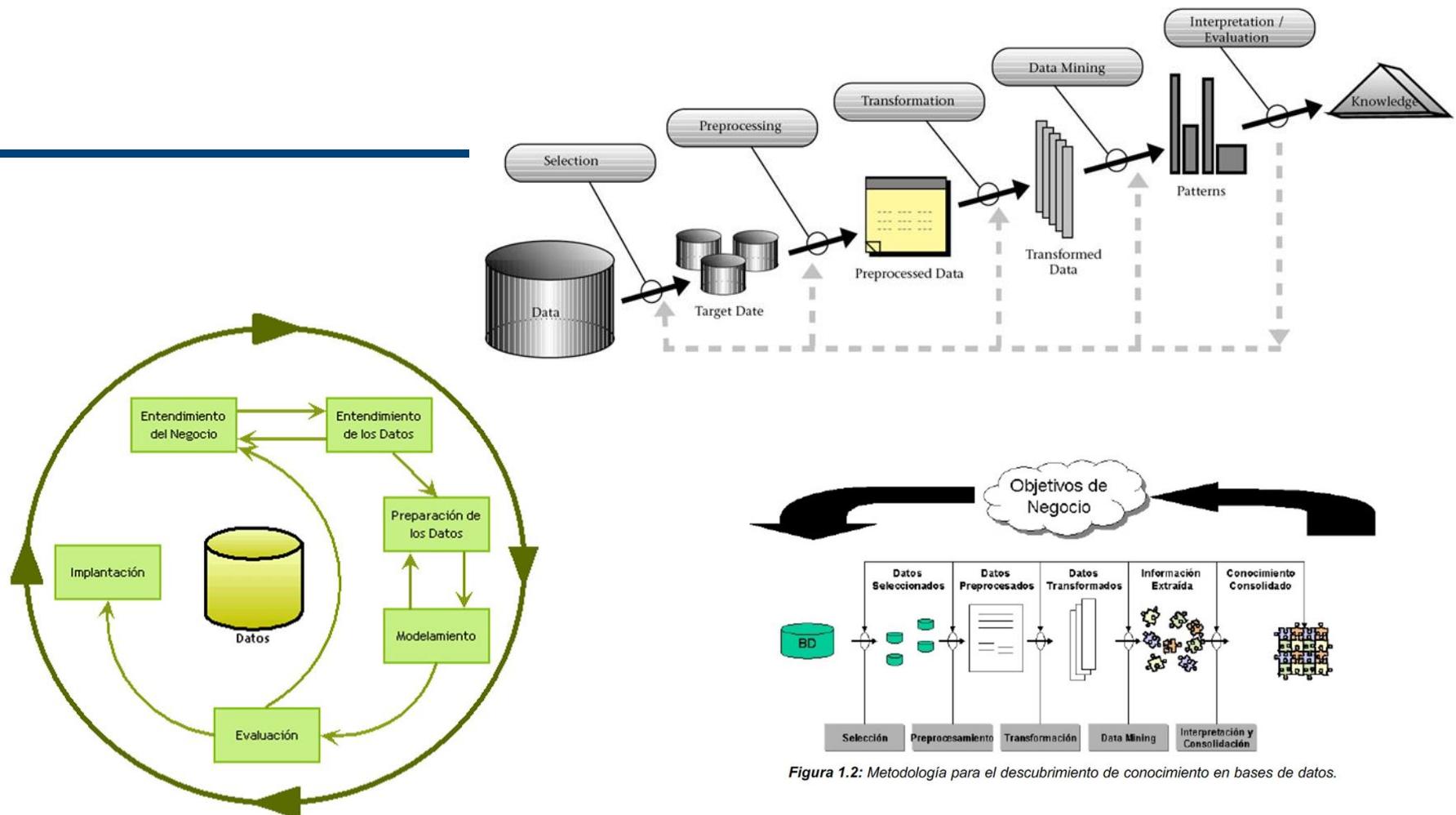
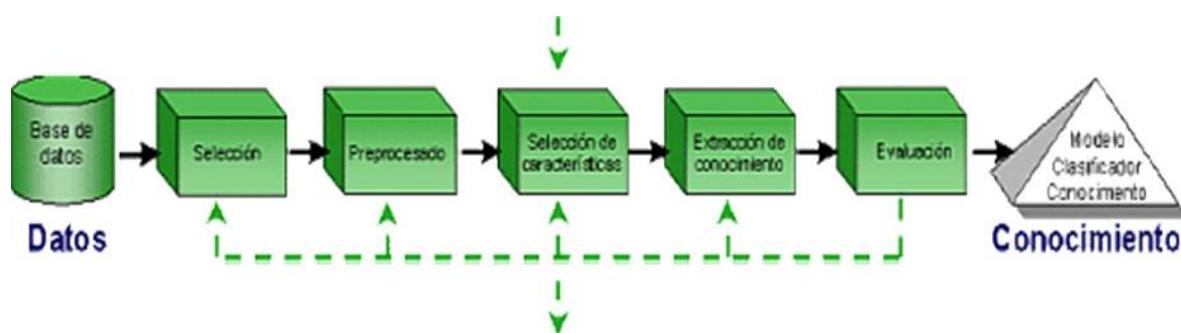


Figura 1.2: Metodología para el descubrimiento de conocimiento en bases de datos.



Minería de Datos

Extracción de conocimiento a partir de fuentes masivas de datos.



- Técnicas Estadísticas
- Técnicas de Aprendizaje Automático

Transformar datos en conocimiento!

Tipos de Análisis en Minería de Datos

Análisis Predictivo



- Predecir riesgos
- Predecir activación de nuevos clientes
- Series de tiempo
- Predecir inventario

Análisis Descriptivo



- Perfil de los clientes
- Selección de factores
- Detección de anomalías
- Canasta de mercado

Tipos de Análisis en Minería de Datos

Análisis Predictivo



- Predecir riesgos
- Predecir activación de nuevos clientes
- Series de tiempo
- Predecir inventario

- **Predicción Discreta o Clasificación**
- **Predicción Continua o Regresión**

Análisis Predictivo

Predicción Discreta o Clasificación

Estudio de categorías pre-definidas para catalogar nuevos elementos.



Ejemplo: Predecir el comportamiento de pago de clientes en una entidad financiera: BUENOS CLIENTES y MALOS CLIENTES.

ID	ATRIBUTO 1	ATRIBUTO 2	...	ATRIBUTO N	CLASE
1	10	alto		56	Cliente Oro
2	45	bajo		54	Cliente Plata
3	23	medio		34	Cliente Bronce
4	54	alto		24	Cliente Bronce
5	21	medio		43	Cliente Oro
6	54	medio		23	Cliente Oro
7	74	alto		65	Cliente Bronce
8	46	alto		47	Cliente Plata
9	43	bajo		83	Cliente Plata
10	34	bajo		59	Cliente Bronce

Histórico o Conjunto de Entrenamiento



Predicción de una clase

ID	ATRIBUTO 1	ATRIBUTO 2	...	ATRIBUTO N	CLASE
11	21	medio		43	?
12	54	medio		23	?
13	74	alto		65	?
14	46	alto		47	?
15	43	bajo		83	?
16	34	bajo		59	?

Datos futuros

Análisis Predictivo

Predicción Continua o Regresión

Estudio de datos con el objetivo de predecir un evento numérico futuro.



Ejemplos: Estimar la expectativa de vida de un cliente.

- Predecir ventas futuras (series de tiempo)

ID	ATRIBUTO 1	ATRIBUTO 2	...	ATRIBUTO N	PREDICCIÓN
1	10	alto		56	34
2	45	bajo		54	42
3	23	medio		34	15
4	54	alto		24	64
5	21	medio		43	36
6	54	medio		23	74
7	74	alto		65	34
8	46	alto		47	2
9	43	bajo		83	6
10	34	bajo		59	4

Histórico o Conjunto de Entrenamiento



Predicción de un número continuo

ID	ATRIBUTO 1	ATRIBUTO 2	...	ATRIBUTO N	PREDICCIÓN
11	21	medio		43	?
12	54	medio		23	?
13	74	alto		65	?
14	46	alto		47	?
15	43	bajo		83	?
16	34	bajo		59	?

Datos futuros

Tipos de Análisis en Minería de Datos

Análisis Descriptivo



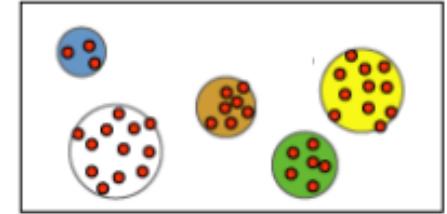
- Perfil de los clientes
- Selección de factores
- Detección de anomalías
- Canasta de mercado

- **Agrupamiento / Clustering**
- **Asociación**
- **Selección de Factores**

Análisis Descriptivo

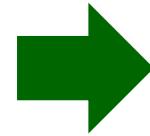
● Agrupamiento / Clustering

Organizar una población de datos heterogénea en un número de clúster homogéneos.

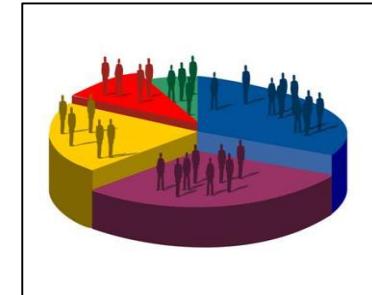


Ejemplos: Diseñar estrategias de mercadeo según el tipo de cliente. Detección de anomalías identificando datos que se alejen de los centroides de agrupación.

Id	Atributo 1	Atributo 2	...	Atributo n
1	10	alto		35
2	35	bajo		54
3	43	medio		28
4	26	bajo		65
5	87	alto		32
6	45	alto		29
7	76	bajo		55
8	5	medio		46
9	12	medio		43
10	54	bajo		27



Descripción en grupos



Análisis Descriptivo

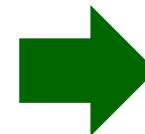
● Asociación

Identificar los elementos que tienen algún nivel de asociación a otros elementos por medio de reglas.



Ejemplo: Determinar los artículos que se pueden ofrecer juntos en promoción.

Id	Atributo 1	Atributo 2	...	Atributo n
1	10	alto		35
2	35	bajo		54
3	43	medio		28
4	26	bajo		65
5	87	alto		32
6	45	alto		29
7	76	bajo		55
8	5	medio		46
9	12	medio		43
10	54	bajo		27



Descripción en reglas

$P \rightarrow Q$
 $\{X, Y\} \rightarrow Z$
 $V \rightarrow \{W, U\}$

Análisis Descriptivo

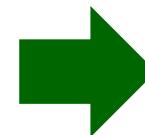
● Selección de Factores

Identificar los factores/variables que más influyen sobre algún evento.



Ejemplo: Determinar las variables que más influyen para la calidad del aire.

Id	Atributo 1	Atributo 2	...	Atributo n
1	10	alto		35
2	35	bajo		54
3	43	medio		28
4	26	bajo		65
5	87	alto		32
6	45	alto		29
7	76	bajo		55
8	5	medio		46
9	12	medio		43
10	54	bajo		27



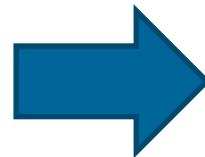
Factores seleccionados

**Atributo 1
Atributo 4
Atributo 6**

Técnicas en la Minería de Datos

Análisis Predictivo

- Clasificación
- Regresión

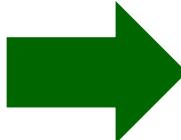


Técnicas Supervisadas

- Redes Neuronales
- Reglas de Decisión
- Árboles de Decisión
- Métodos Probabilísticos
- Máq. de Soporte Vectorial
- Métodos de Regresión
- Modelos Ocultos de Markov
- Métodos basados en Ejemplos

Análisis Descriptivo

- Clustering
- Asociación
- Selección de Factores



Técnicas NO Supervisadas

- Métodos Jerárquicos
- Métodos Particionales
- Redes Neuronales
- Métodos Probabilísticos
- Métodos Difusos
- Métodos Evolutivos
- Métodos basados en Kernel
- Métodos de reglas

Tareas Realizadas por Plataformas de Minería de Datos



Clasificación

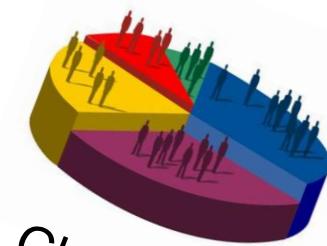


Regresión

	A	B	C	D	E
1	College Enrollment 2007 - 2008				
2	Student ID	Last Name	Initial	Age	Program
3	ST348-250	Graham	J.	20	Arts
4	ST348-248	James	L.	23	Nursing
5	ST348-252	Nash	S.	22	Arts
6	ST348-249	Peterson	M.	37	Science
7	ST348-254	Robitaille	L.	19	Drafting
8	ST348-253	Russell	W.	20	Nursing
9	ST348-251	Smith	F.	26	Business
10	ST348-247	Thompson	G.	18	Business
11	ST348-245	Walton	L.	21	Drafting
12	ST348-246	Wilson	R.	19	Science
13	ST348-255	Christopher	A.	22	Science
14	*				
15	*				



Asociación



Clustering



Selección de Factores

Tareas Realizadas por Plataformas de Minería de Datos

Análisis de Componentes Principales

Análisis de Correlaciones

Calidad de los Datos

Reducción de variables



Sistemas de Votación

Análisis Costo-Beneficio

Evaluación de modelos

Tendencias

- Minería de datos distribuida
- Minería multimedia
 - Minería de texto
 - Minería de imágenes
 - Minería de audio
 - Minería de videos
- Minería social
- Minería georeferenciada
- Minería de la web

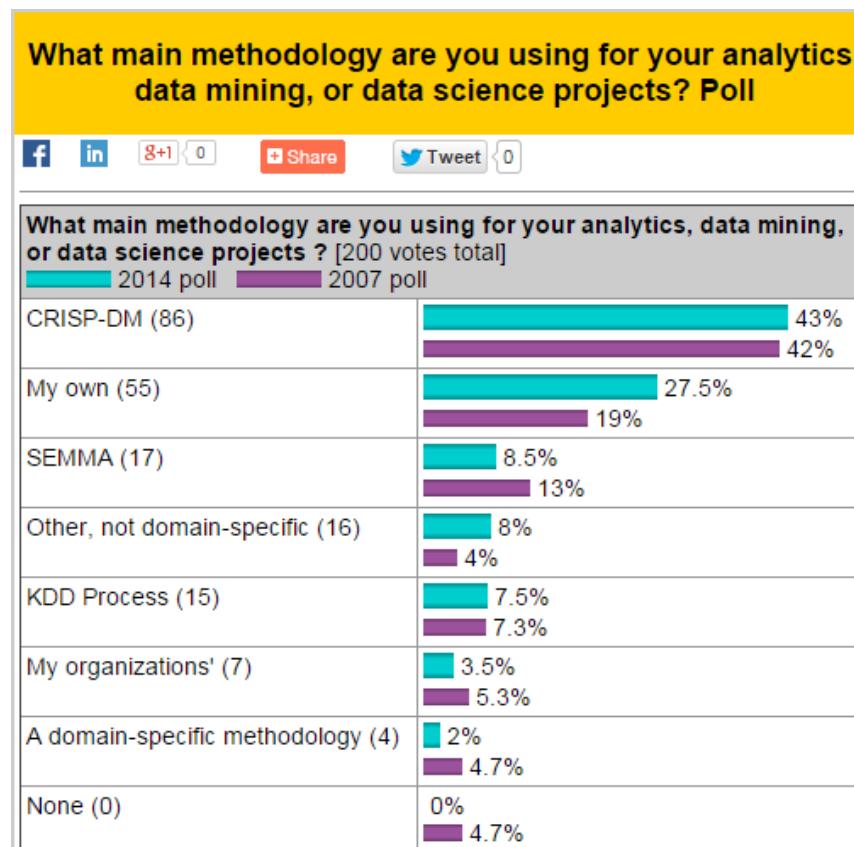
AGENDA



1. Introducción
- 2. Metodologías**
3. Herramientas
4. Análisis de Casos
5. Preparación de Datos
6. Técnicas y Algoritmos
7. Minería de Texto
8. Bibliografía

Metodologías para Minería de Datos

Conjunto de pasos sistematizados para guiar el proceso desde que se estudia el problema que se desea tratar hasta que se tienen las respuestas a los problemas formulados.

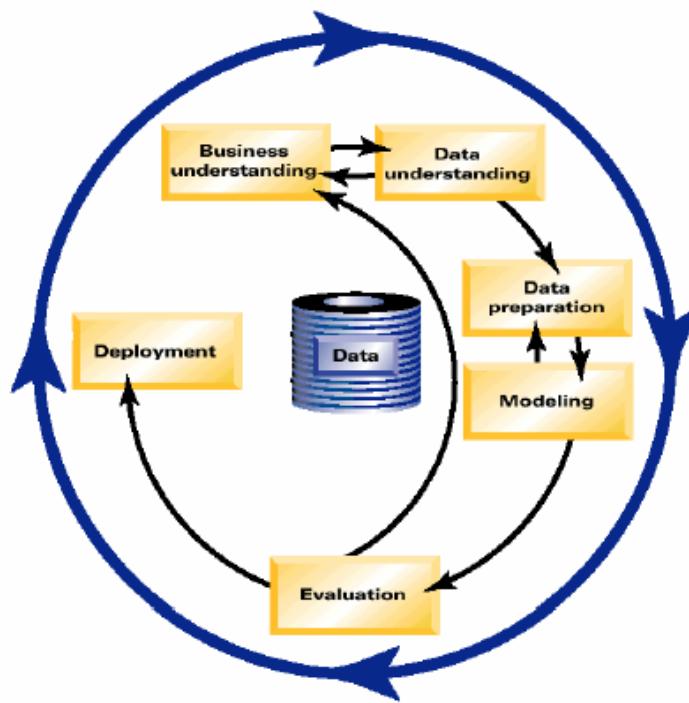


Metodologías para Minería de Datos

- CRISP-DM
- SEMMA
- KDD

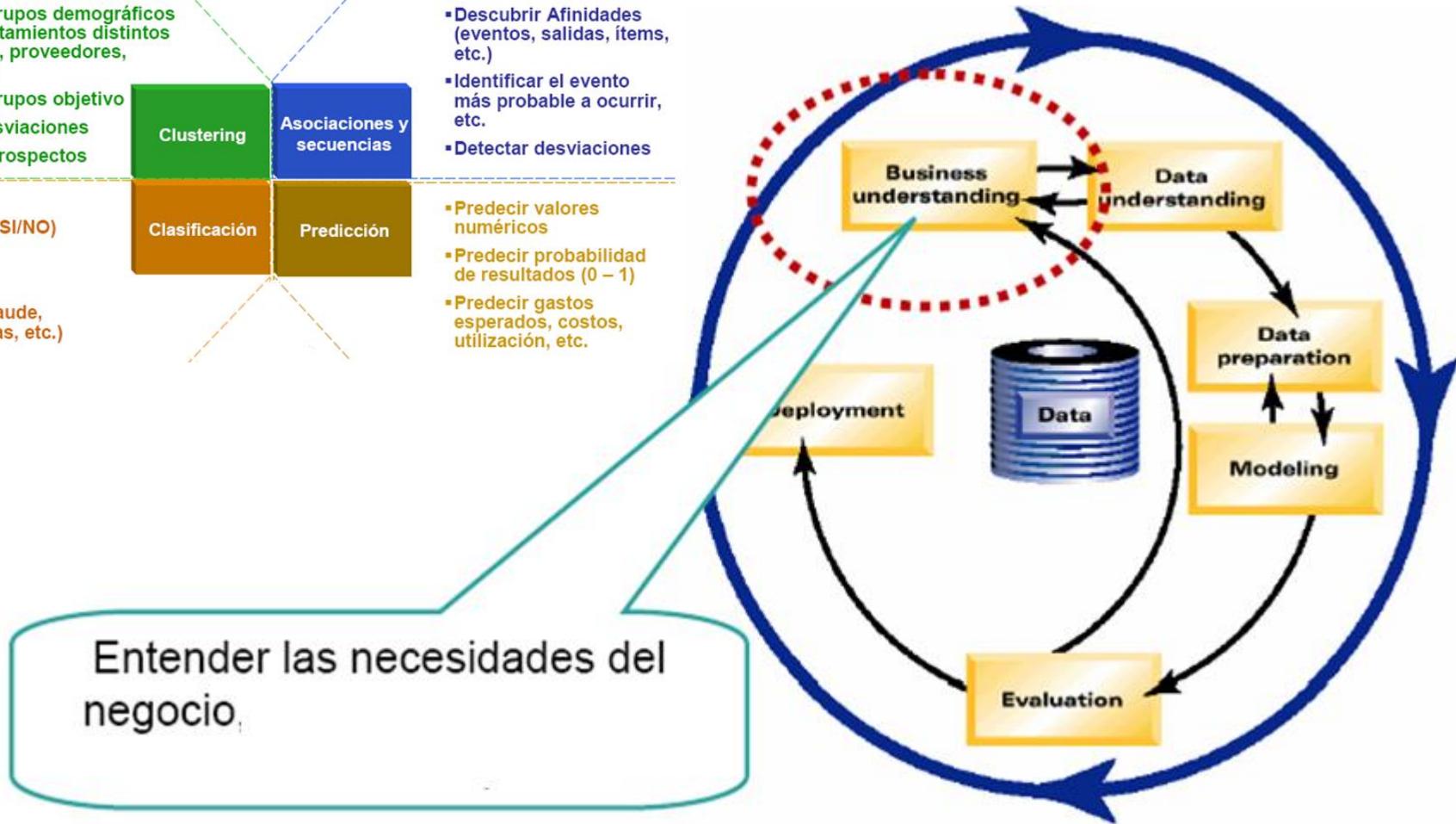
CRISP - DM

CRISP-DM (Cross- Industry Standard Process for Data Mining)
de NCR de Dinamarca, AG de Alemania, SPSS de Inglaterra y
OHRA de Holanda.



Fuente: <http://www.crisp-dm.org/>

CRISP - DM

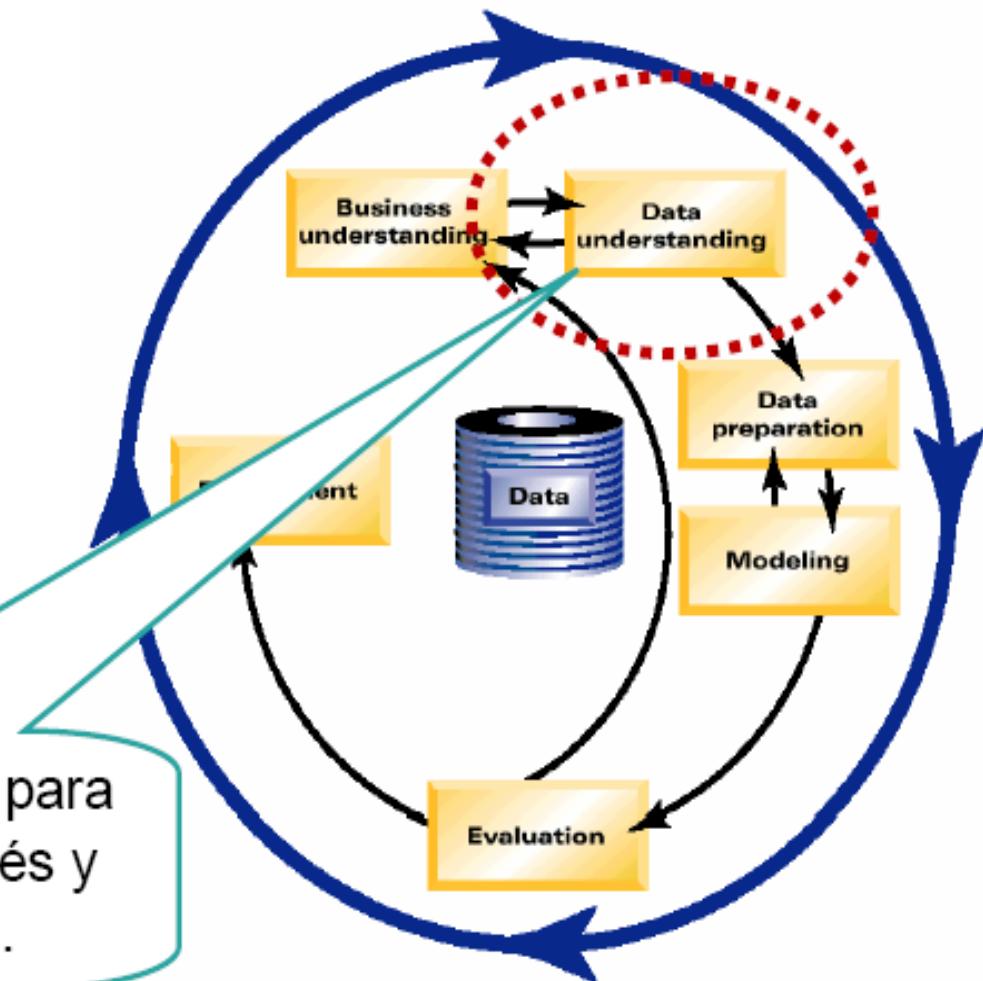


Basado parcialmente de material de Javier Rengifo, IBM de Colombia

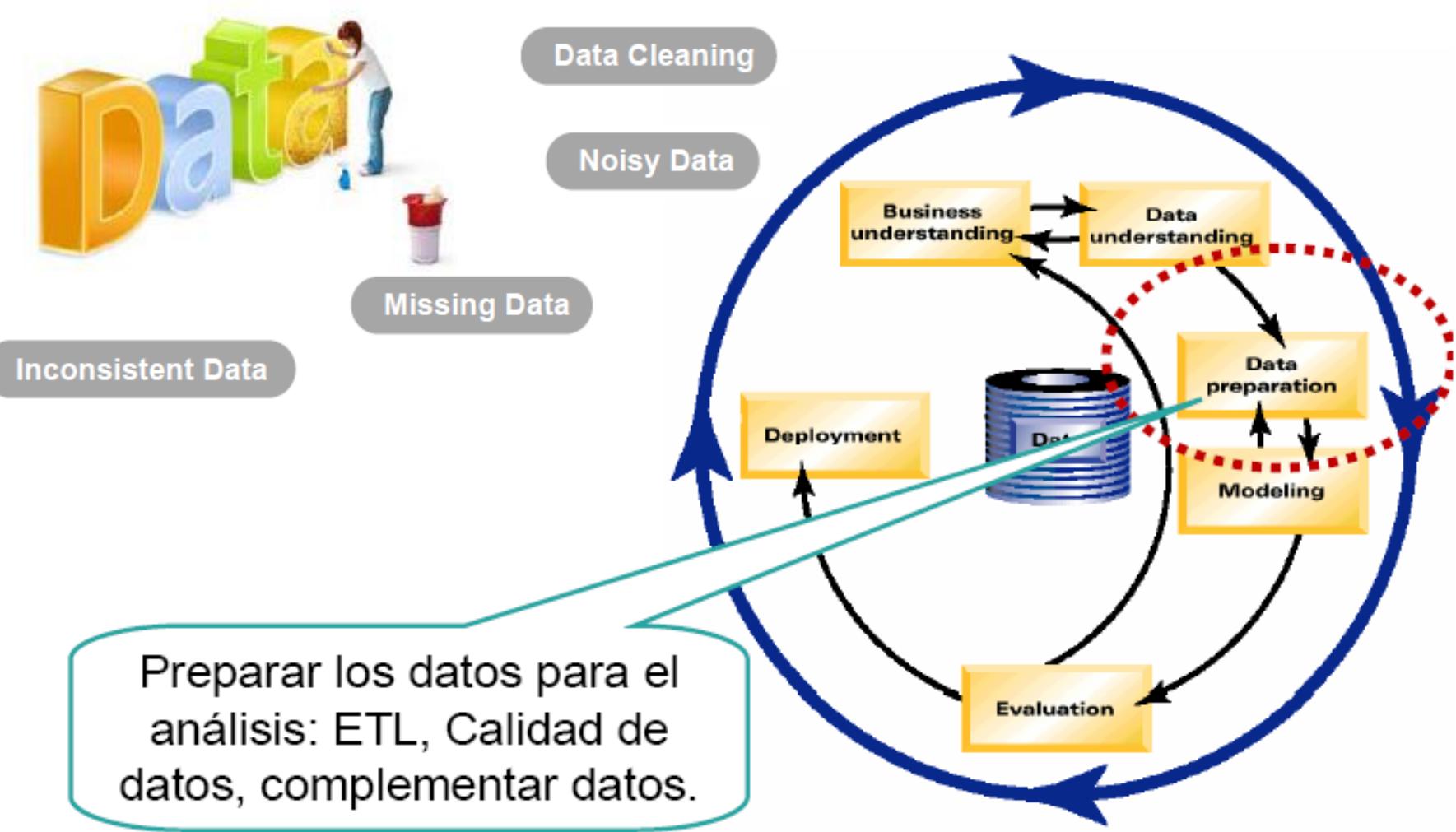
CRISP - DM

■ Estadísticas descriptivas:

- Visualización de datos
(Ej. Frecuencias, Distribuciones)
- Gráficas sobre datos numéricos
(Ej. histogramas)
- Medidas de tendencia central
(Ej. Media, mediana, moda)
- Estimaciones de varianza
(Ej. Desviación estandar)



CRISP - DM



Basado parcialmente de material de Javier Rengifo, IBM de Colombia

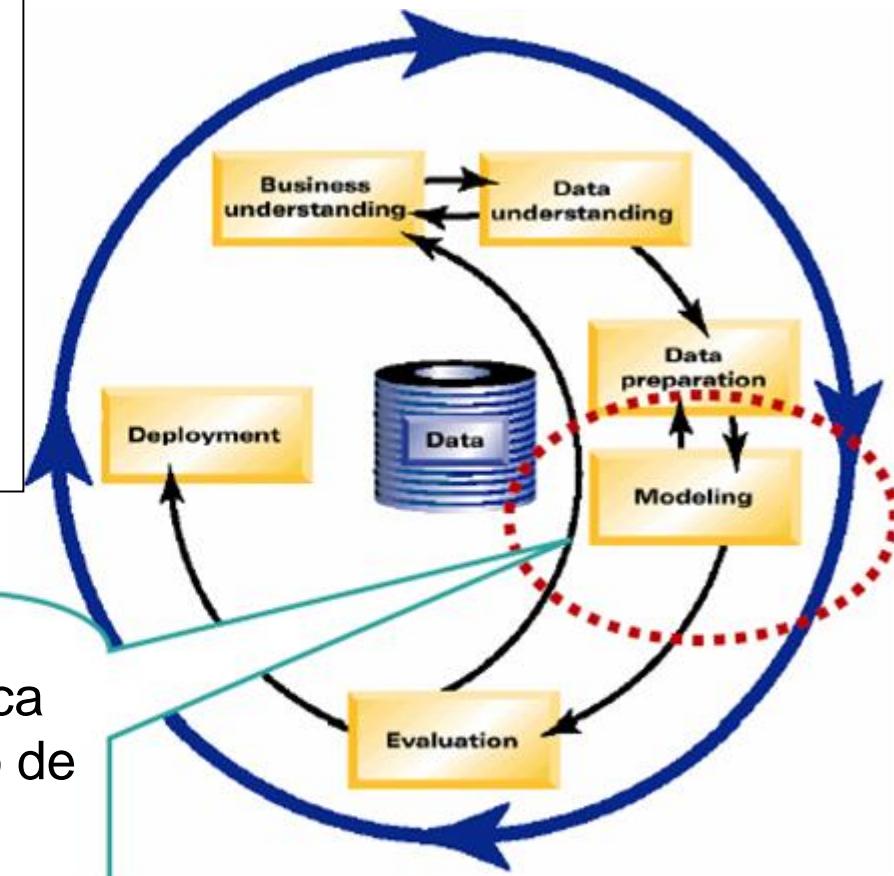
CRISP - DM

Técnicas Supervisadas

- Redes Neuronales
- Redes Semánticas
- Reglas de Decisión
- Árboles de Decisión
- Modelos Ocultos de Markov
- Métodos Probabilísticos
- Máq. De Soporte Vectorial
- Métodos de Regresión
- Métodos basados en Ejemplos
- Método Rocchio
- Métodos Evolutivos

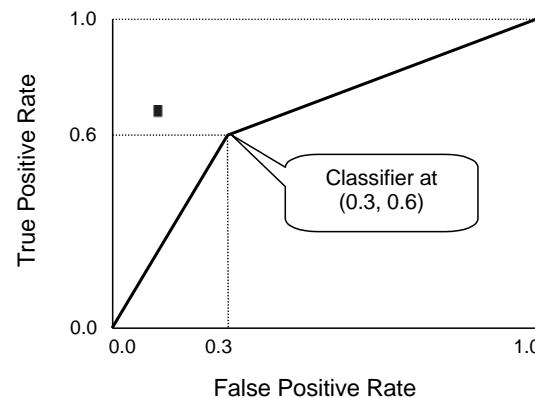
Técnicas NO Supervisadas

- Métodos Jerárquicos
- Métodos Particionales
- Redes Neuronales
- Métodos Probabilísticos
- Métodos Difusos
- Métodos basados en Grafos
- Métodos Evolutivos
- Métodos de Kernel
- Métodos Espectrales

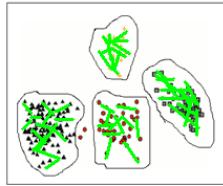


Descubrir de manera automática información de valor por medio de técnicas de minería de datos.

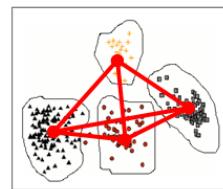
CRISP - DM



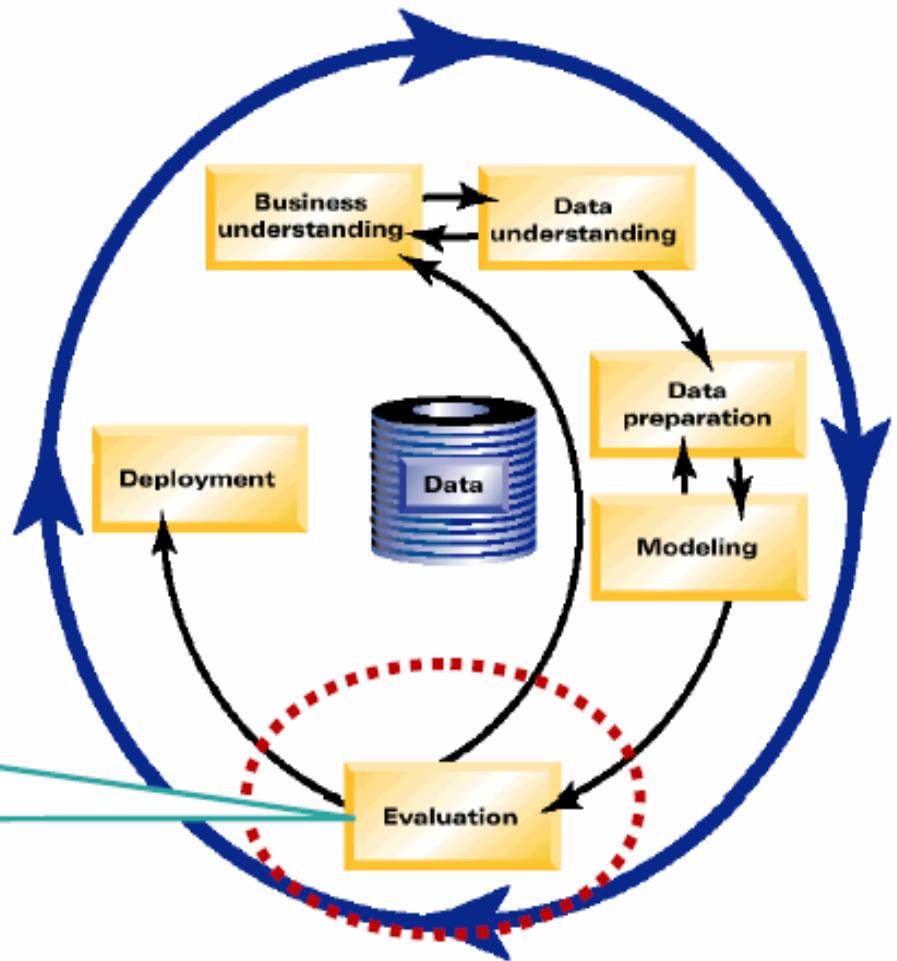
Compactness:



Separability:



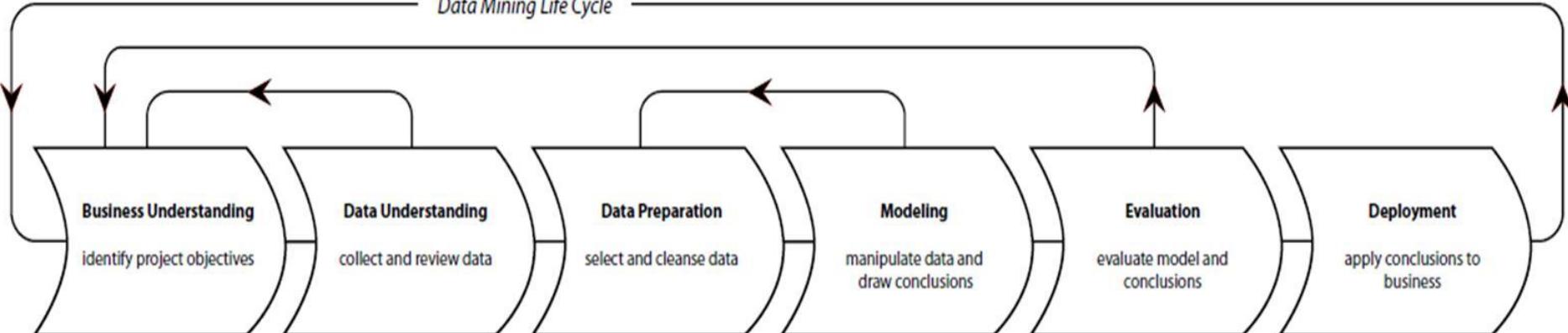
Evaluación de los resultados de acuerdo a la técnica utilizada



CRISP - DM



Data Mining Life Cycle



Determine Business Objectives

*Background
Business Objectives
Business Success Criteria
(Log and Report Process)*

Assess Situation

*Inventory of Resources,
Requirements, Assumptions,
and Constraints
Risks and Contingencies
Terminology
Costs and Benefits
(Log and Report Process)*

Determine Data Mining Goals

*Data Mining Goals
Data Mining Success Criteria
(Log and Report Process)*

Produce Project Plan

*Project Plan
Initial Assessment of Tools and
Techniques
(Log and Report Process)*

Collect Initial Data

*Initial Data Collection Report
(Log and Report Process)*

Describe Data

*Data Description Report
(Log and Report Process)*

Explore Data

*Data Exploration Report
(Log and Report Process)*

Verify Data Quality

*Data Quality Report
(Log and Report Process)*

Data Set

*Data Set Description
(Log and Report Process)*

Select Data

*Rationale for Inclusion/
Exclusion
(Log and Report Process)*

Clean Data

*Data Cleaning Report
(Log and Report Process)*

Construct Data

*Derived Attributes
Generated Records
(Log and Report Process)*

Integrate Data

*Merged Data
(Log and Report Process)*

Format Data

*Reformatted Data
(Log and Report Process)*

Select Modeling Technique

*Modeling Technique
Modeling Assumptions
(Log and Report Process)*

Generate Test Design

*Test Design
(Log and Report Process)*

Build Model Parameter Settings

*Models
Model Description
(Log and Report Process)*

Assess Model

*Model Assessment
Revised Parameter
(Log and Report Process)*

Evaluate Results

*Align Assessment of Data
Mining Results with
Business Success Criteria
(Log and Report Process)*

Approved Models

*Review Process
Review of Process
(Log and Report Process)*

Determine Next Steps

*List of Possible Actions
Decision
(Log and Report Process)*

Plan Deployment

*Deployment Plan
(Log and Report Process)*

Plan Monitoring and Maintenance

*Monitoring and
Maintenance Plan
(Log and Report Process)*

Produce Final Report

*Final Report
Final Presentation
(Log and Report Process)*

Review Project

*Experience
Documentation
(Log and Report Process)*

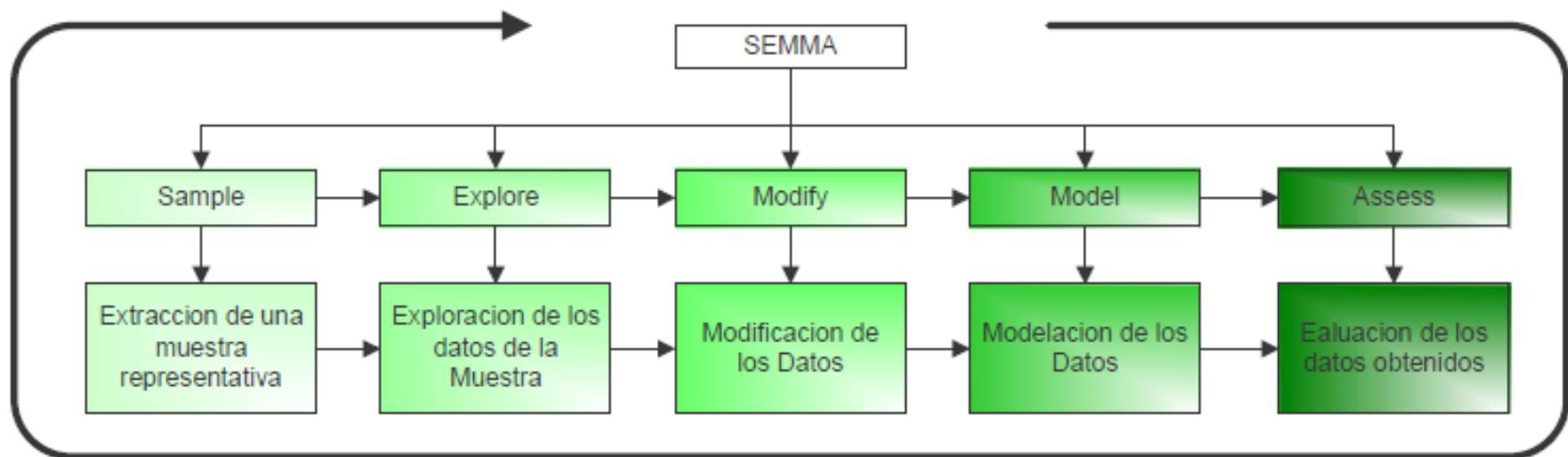
a visual guide to CRISP-DM methodology

SOURCE CRISP-DM 1.0

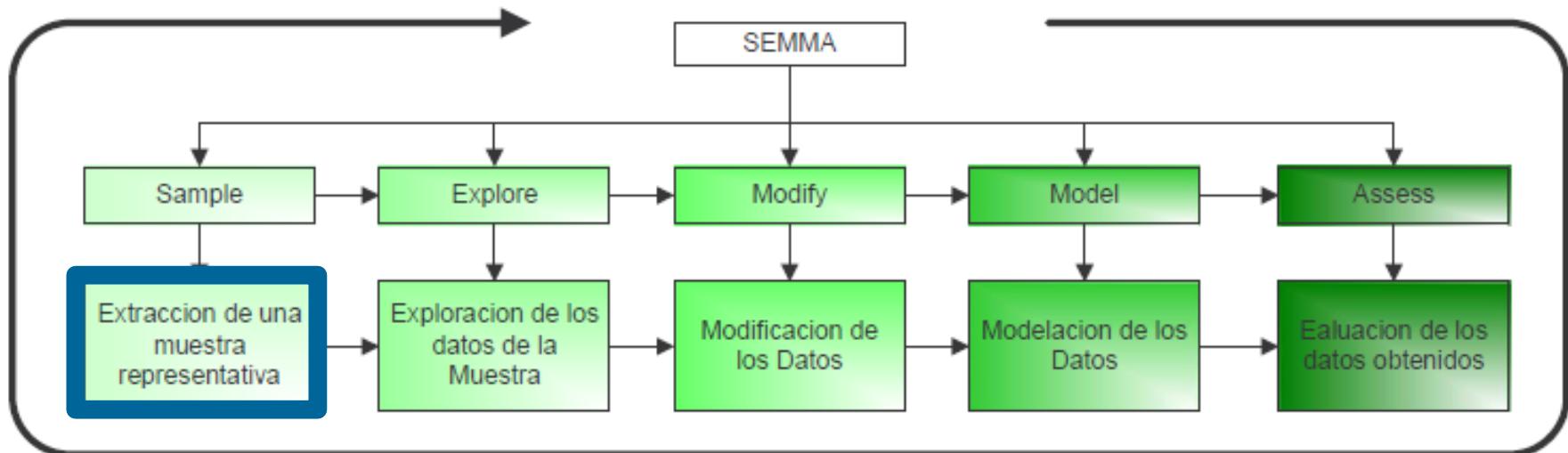
<http://www.crisp-dm.org/download.htm>

SEMMA

SEMMA (Simple, Explore, Modify, Model, Assess) de SAS Intelligent Miner.



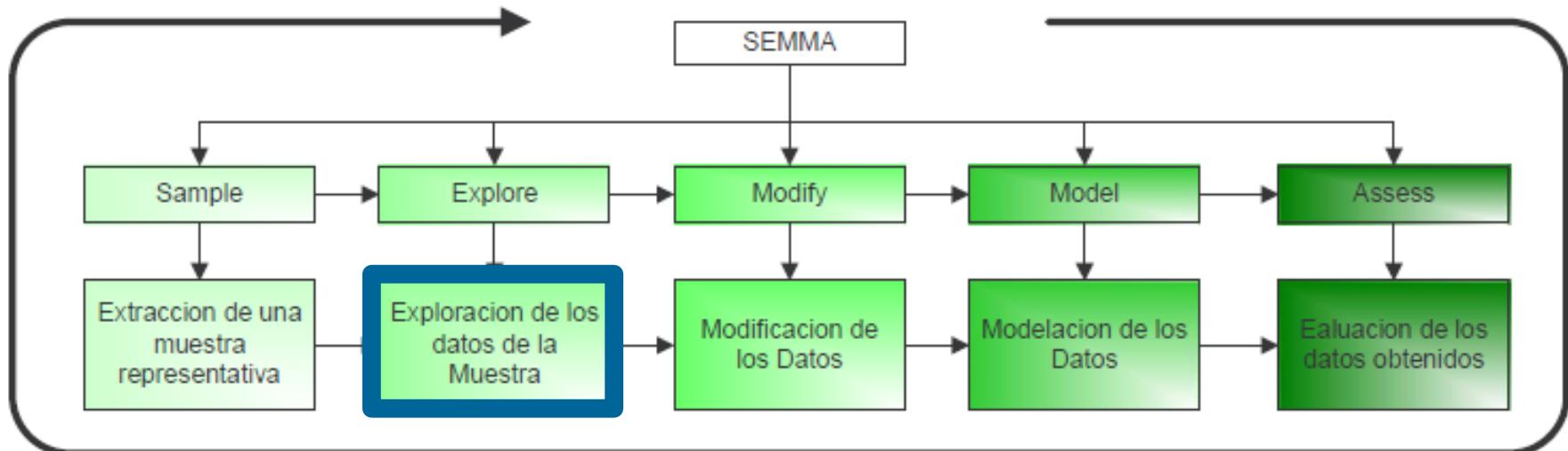
SEMMA



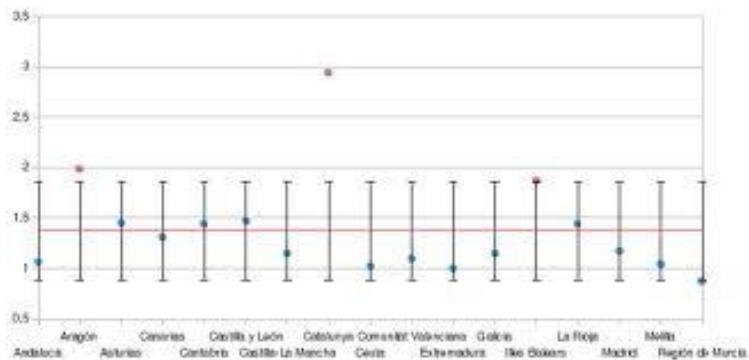
Recolección de una muestra estadística representativa de los datos, reduciendo el tiempo necesario para determinar el conocimiento nuevo para el negocio.



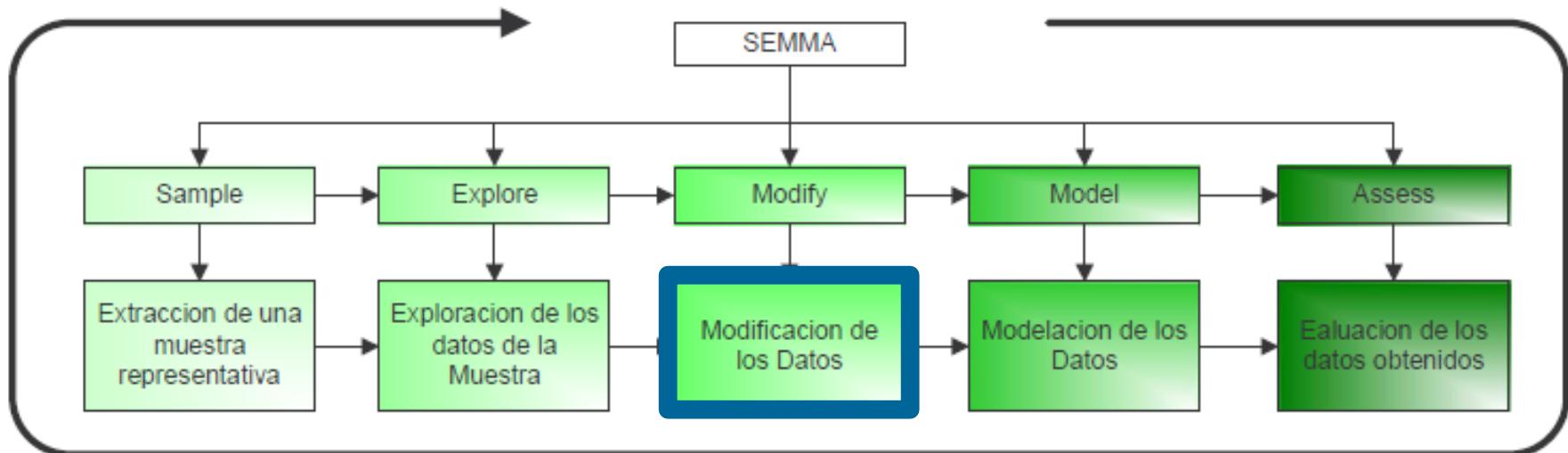
SEMMA



Detección, identificación y eliminación de datos anómalos, por medio de procesos de visualización y análisis estadístico.



SEMMA

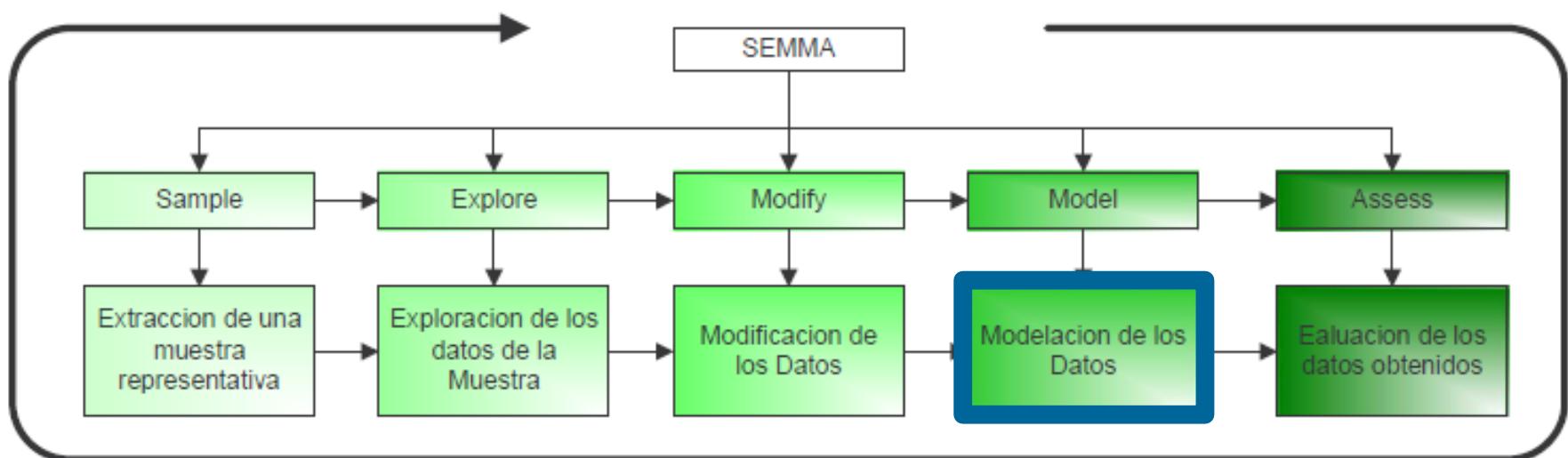


Selección y transformación de variables.

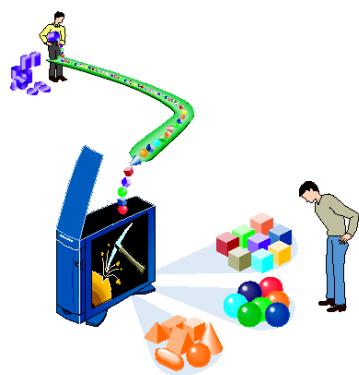
Unidades	Edad	Sexo	PAS	PAD	Peso	Talla
1	x	x	?	?	x	x
⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	x	x	?	?	x	x
101	x	x	x	x	?	?
⋮	⋮	⋮	⋮	⋮	⋮	⋮
200	x	x	x	x	?	?
201	x	x	?	?	?	?
⋮	⋮	⋮	⋮	⋮	⋮	⋮
300	x	x	?	?	?	?
301	x	x	x	x	x	x
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.000	x	x	x	x	x	x

Nota: El simbolo ? representa los valores ausentes
y x, los observados

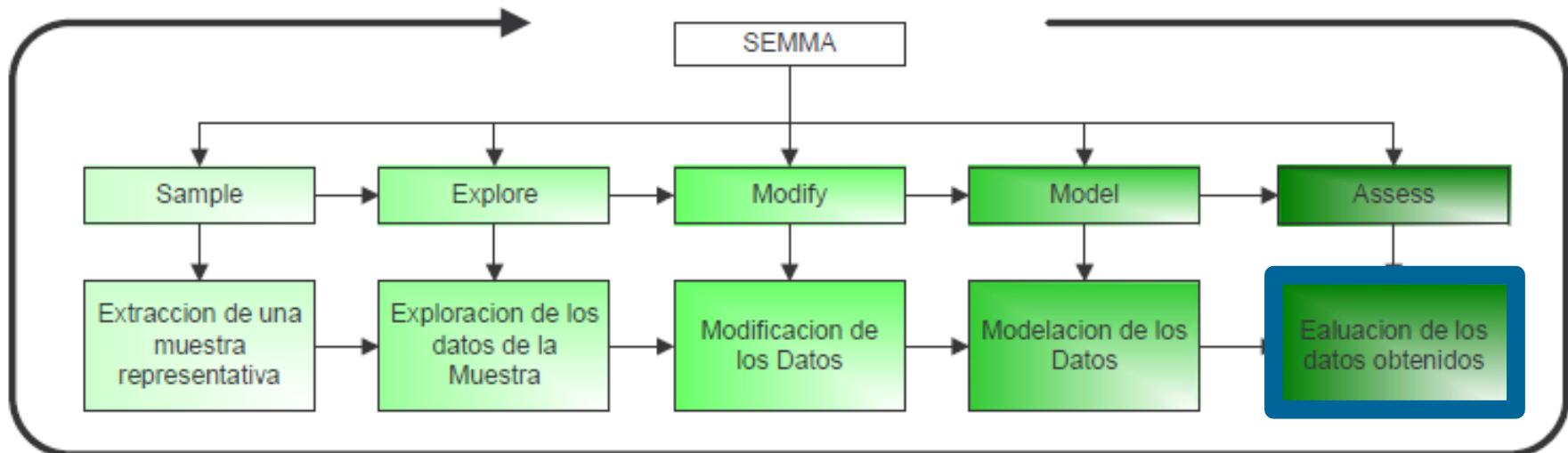
SEMMA



Aplicación de técnicas y métodos de minería de datos.



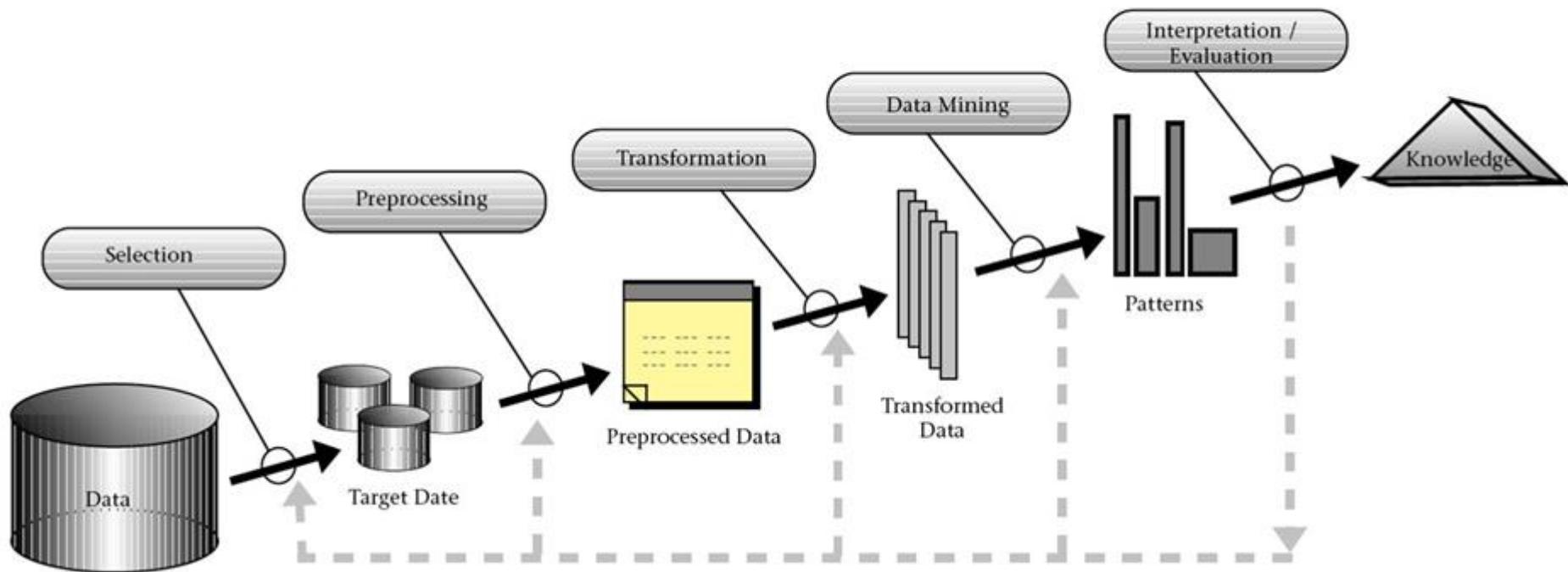
SEMMA



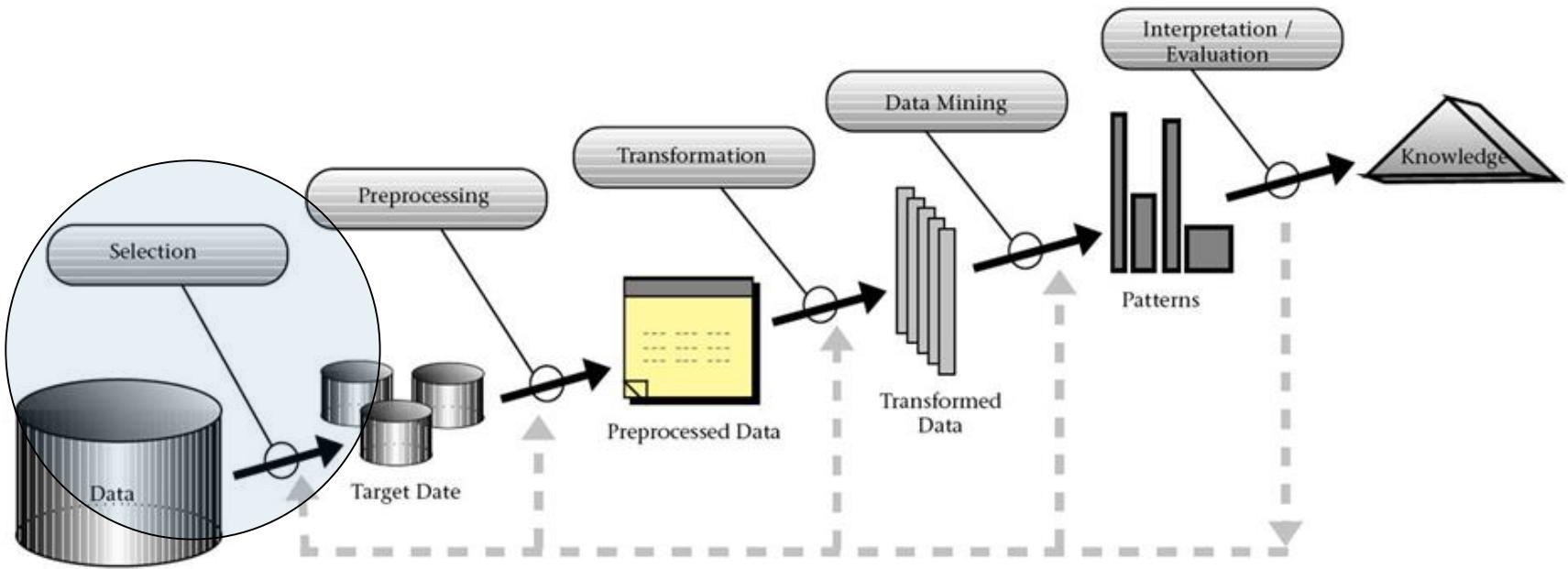
Determinar la calidad del modelo.



KDD: *Knowledge Discovery in Databases*

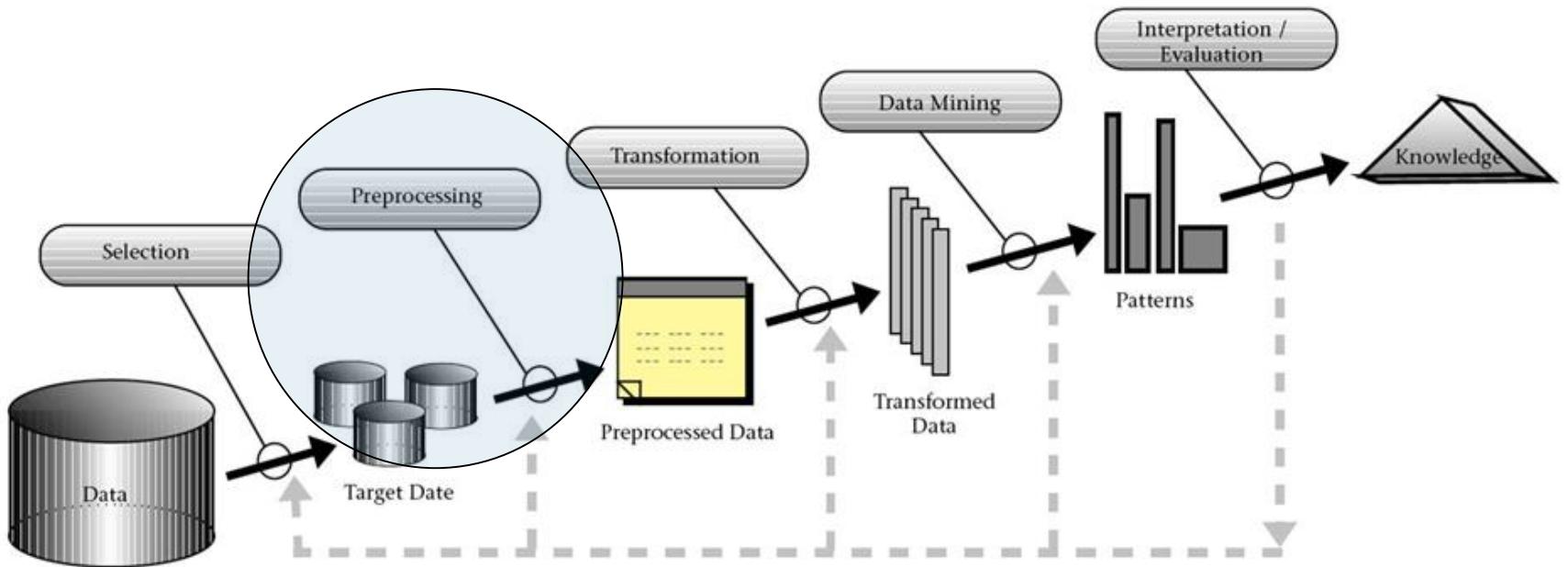


KDD: Knowledge Discovery in Databases



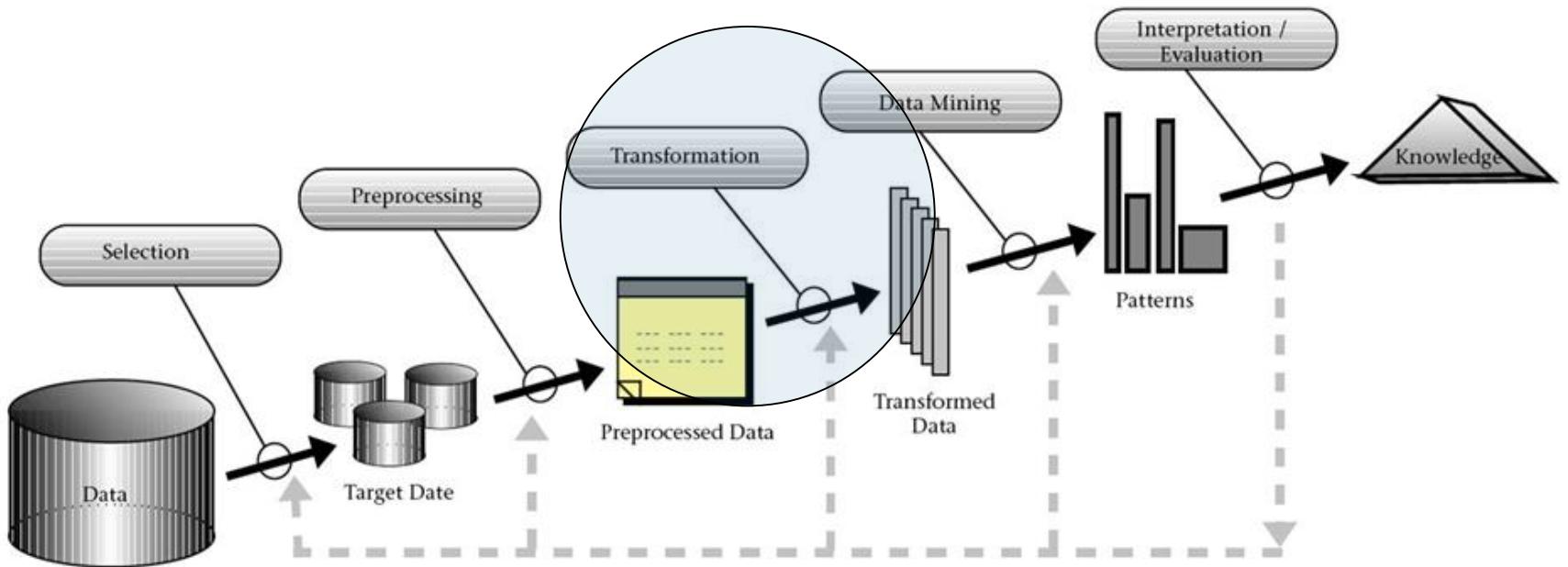
- Selección: identificación de datos relevantes para el proyecto.

KDD: Knowledge Discovery in Databases



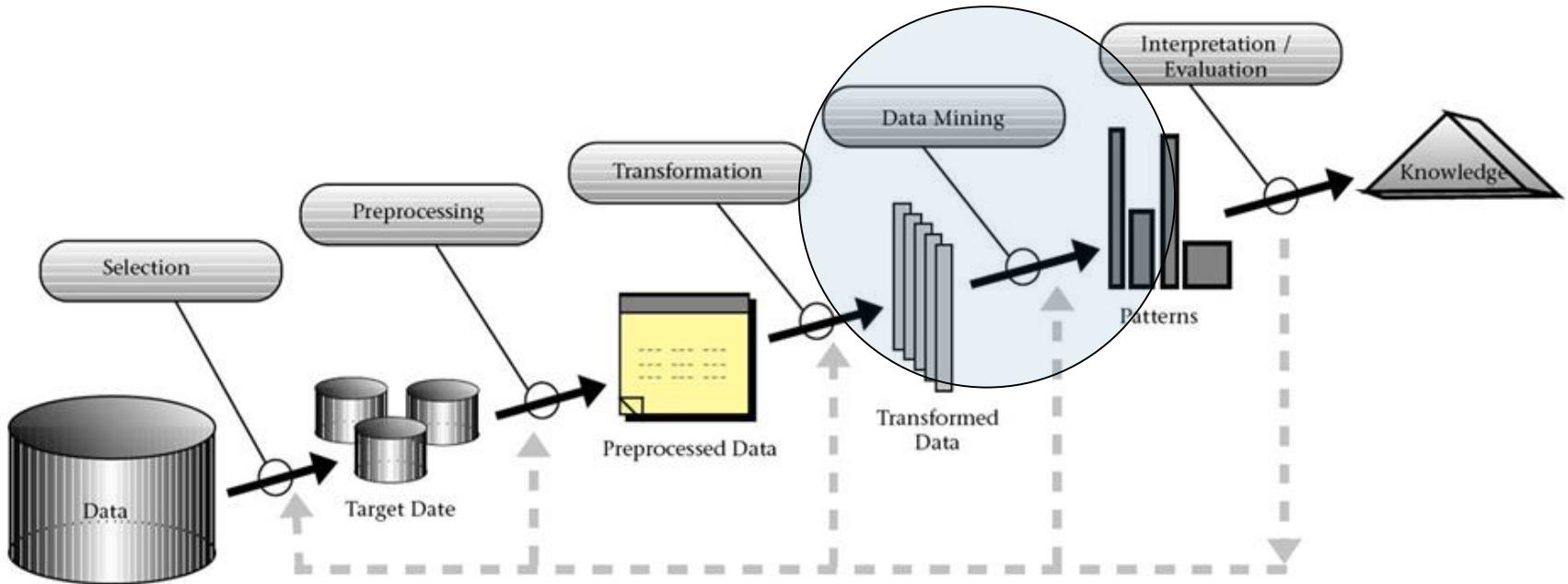
- Preprocesamiento: organización y limpieza de los datos, evaluando datos anómalos, faltantes y redundantes.

KDD: Knowledge Discovery in Databases



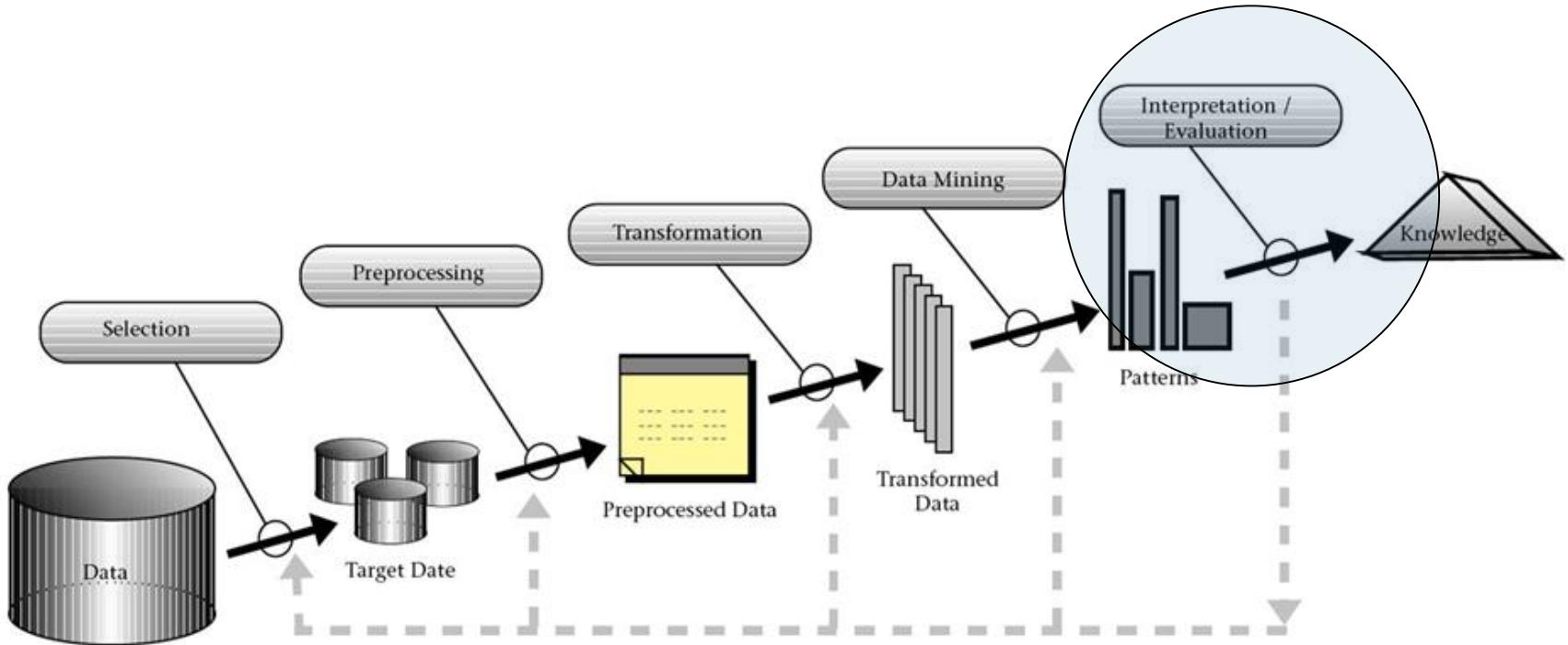
- Transformación: preparación de los datos para el análisis.

KDD: Knowledge Discovery in Databases



- Minería de datos: análisis predictivo y/o descriptivo.

KDD: Knowledge Discovery in Databases



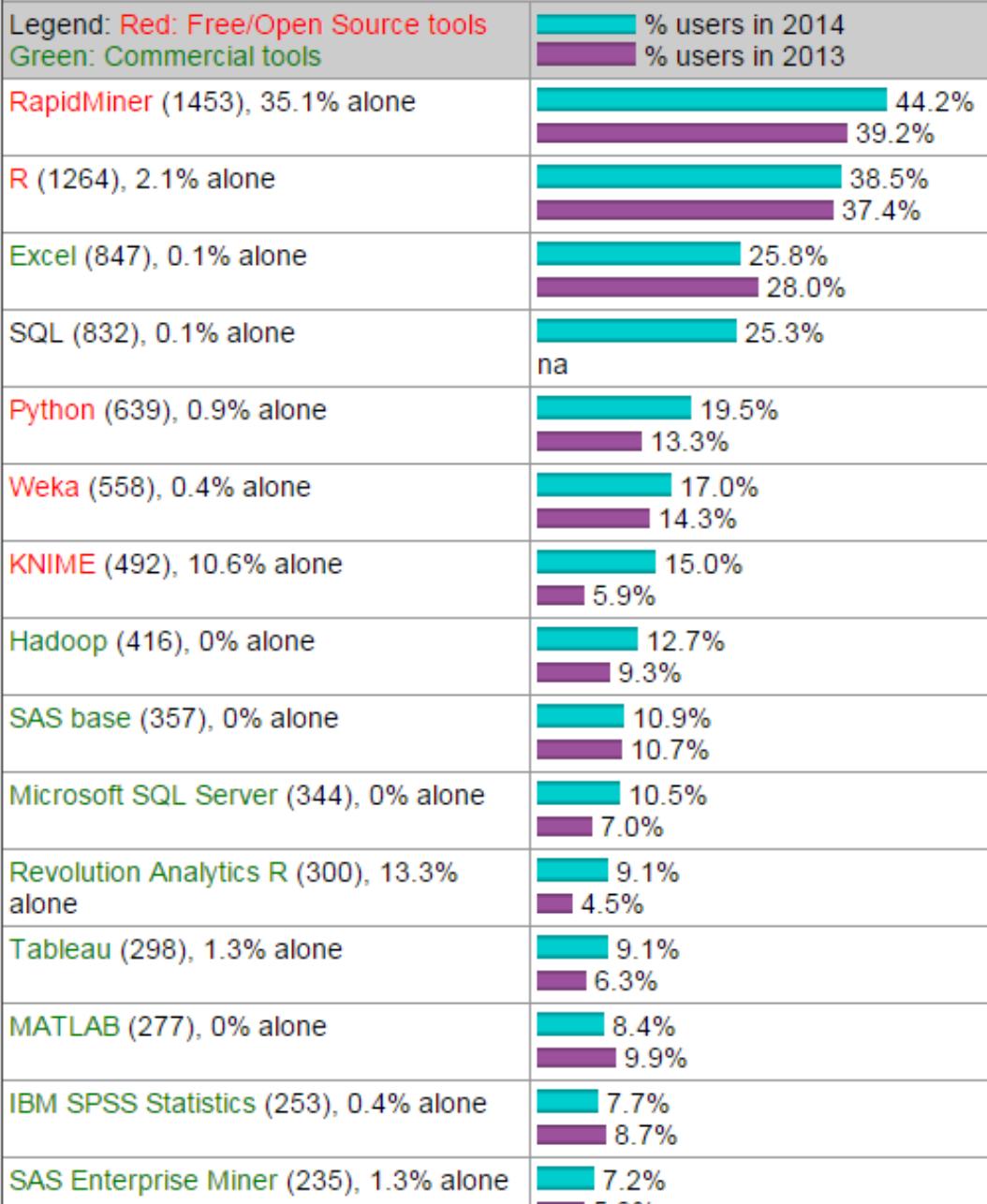
- Interpretación/Evaluación: medidas de calidad de resultados, visualización y representación del nuevo conocimiento.

AGENDA



1. Introducción
2. Metodologías
- 3. Herramientas**
4. Análisis de Casos
5. Preparación de Datos
6. Técnicas y Algoritmos
7. Minería de Texto
8. Bibliografía

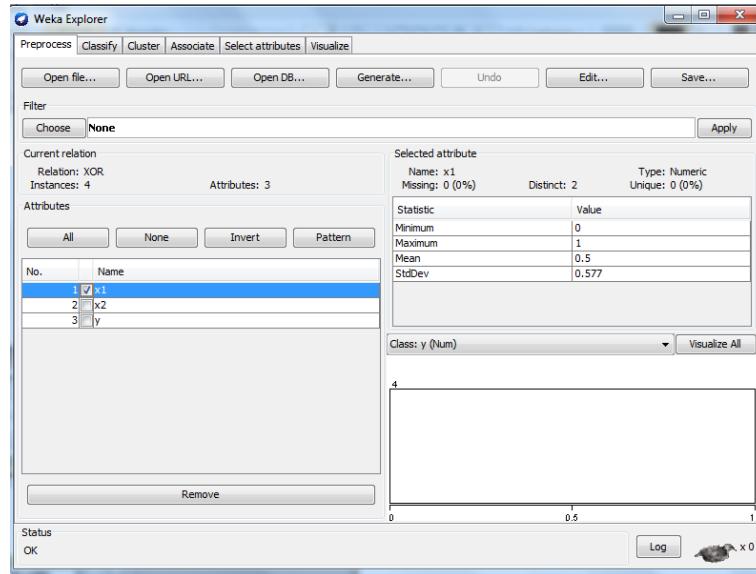
What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project? [3285 voters]



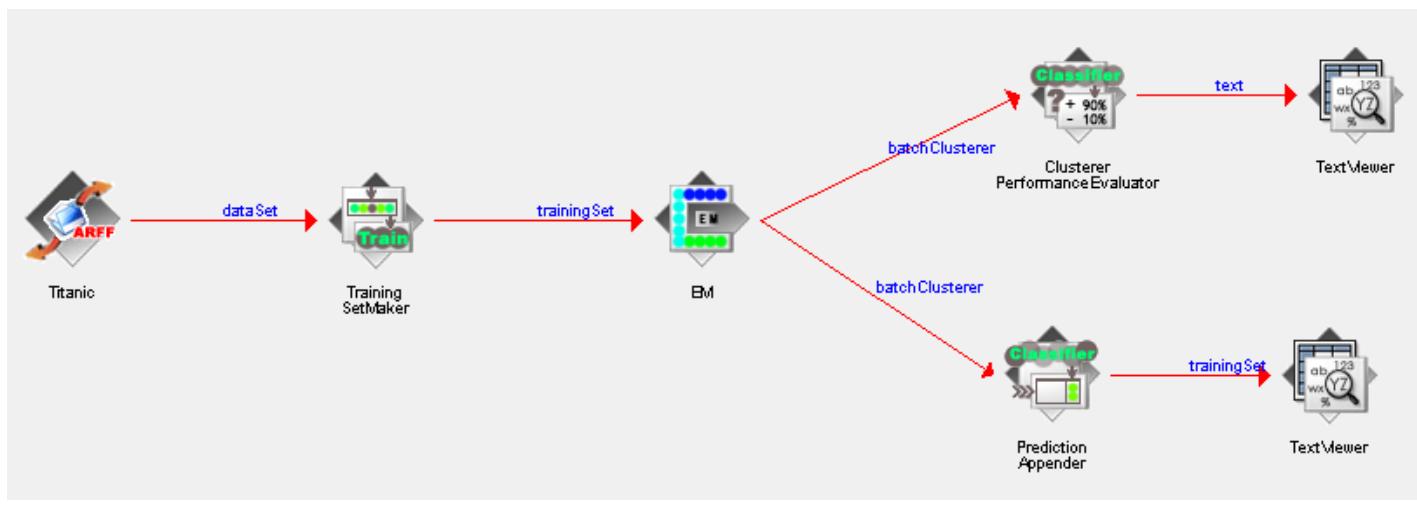
Herramientas

WEKA

Explorer



Knowledge Flow



RAPID MINER

GermanCredit - RapidMiner@AnalisaPC

File Edit Process Tools View Help

Overview Process XML

Main Process

```
graph LR; A[Read ARFF] --> B[Nominal to Num...]; B --> C[Clustering]; C --> D[Performance ...]
```

Parameters

Clustering (K-Means)

- add cluster attribute
- add as label
- remove unlabeled

k: 6
max runs: 10
 determine good start values
measure types: BregmanDiverg...

Help Comment

Problems Log

No problems found

Message Fixes Location

K-Means

Synopsis

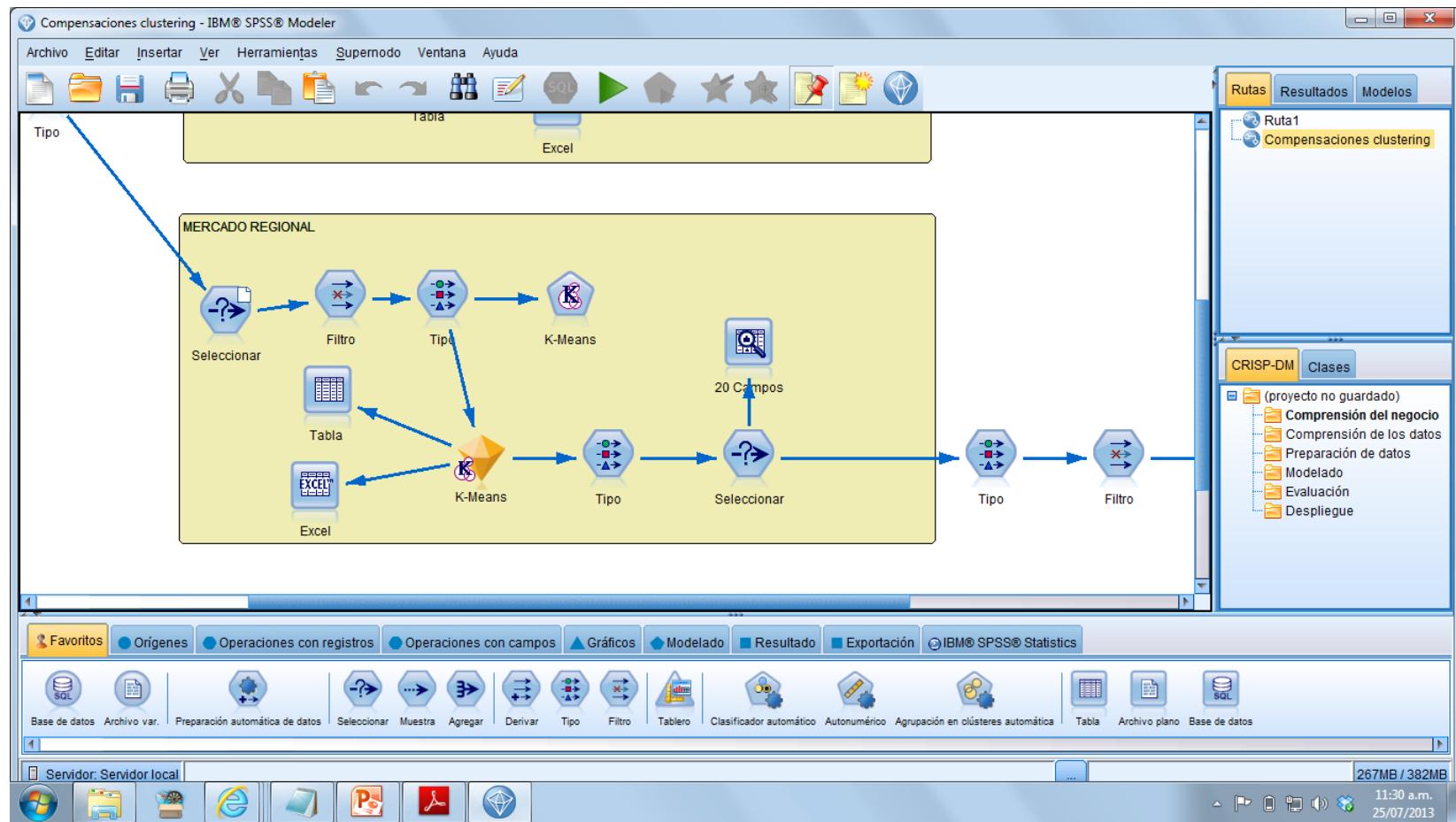
Clustering with k-means

Description

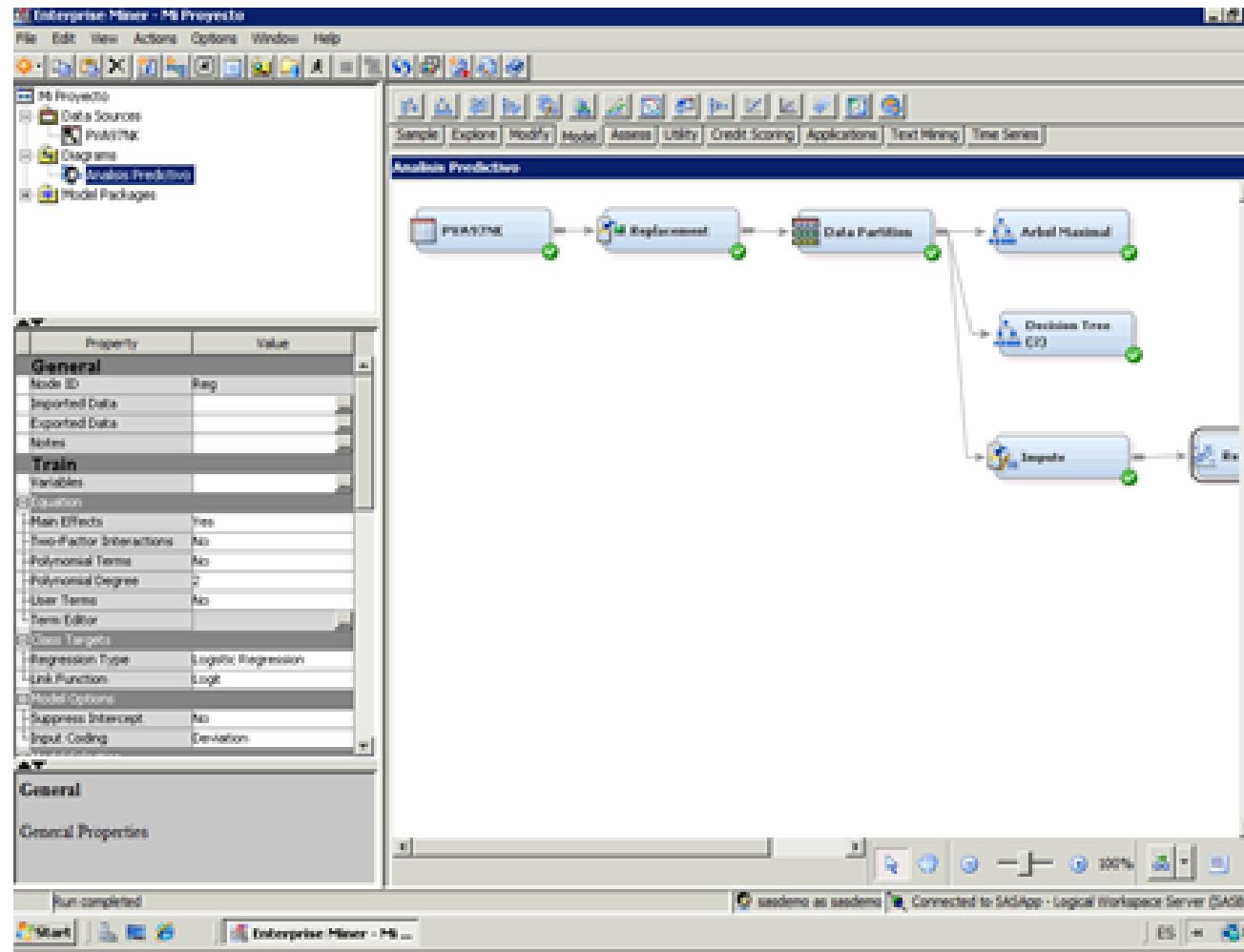
This operator represents an implementation of k-means. This operator will create a cluster attribute if not present yet.

The screenshot shows the RapidMiner graphical user interface. On the left, the 'Operators' palette is open, displaying a tree view of available operators categorized by type. The 'Clustering and Segmentation' category is expanded, showing various sub-operators including 'Weka (8)' and 'k-Means'. The 'k-Means' operator is highlighted with a yellow border. In the center, the 'Process' tab is active, showing a process flow titled 'Main Process'. The flow starts with a 'Read ARFF' operator, followed by a 'Nominal to Numerical' operator, then a 'Clustering' operator (which is the highlighted 'k-Means' operator), and finally a 'Performance' operator. The 'Parameters' tab on the right is also open, specifically for the 'Clustering (K-Means)' operator. It shows configuration options like 'add cluster attribute' (checked), 'add as label' (checked), and 'remove unlabeled' (unchecked). The 'k' value is set to 6, and 'max runs' is set to 10. Below the parameters, there is a 'Synopsis' section describing the operator as a clustering method and a 'Description' section providing more technical details. At the bottom, the Windows taskbar is visible with several application icons.

SPSS MODELER



SAS ENTERPRISE MINER



R

Screenshot of RStudio showing the following R code and a scatter plot:

```
View Project Workspace Plots Tools Help
File Go to file/function
Source on Save formatPlot.R diamonds Source
summary(diamonds)
summary(diamonds$price)
avsize <- levels(diamonds$clarity)
clarity <- round(mean(diamonds$clarity), 4)
p <- qplot(carat, price,
           data=diamonds, color=clarity,
           xlab="Carat", ylab="Price",
           main="Diamond Pricing")
format.plot(p, size=24)
```

The console output shows summary statistics for the diamonds dataset:

```
Min. : 0.000 1st Qu.: 4.710 Median : 5.731 Mean : 5.731 3rd Qu.: 6.540 Max. :10.740
1st Qu.: 4.710 Median : 5.710 Mean : 5.735 3rd Qu.: 6.540 Max. : 6.540
Min. 1st QU. Median 2nd QU. 3rd QU. Max.
326 950 2401 3933 5324 18820
```

The scatter plot is titled "Diamond Pricing" and shows Price on the Y-axis (0 to 15000) versus Carat on the X-axis (0.0 to 3.5). The data points are colored by Clarity level.

Python

Screenshot of a Jupyter Notebook cell showing Python code and a plot:

```
definitivo.py*
49 #!/usr/bin/python
50 # see crea_diccionario
51 # diccionario = tifid_vectorizer.get_feature_names()
52 # print 'Corroborar tamaño de la matriz Documentos vs Términos'
53 # print tfidf_matrix.shape
54 # print 'Obteniendo similitud de coseno entre 2 documentos (si son iguales el valor es
55 # 1)'
56 # print cosine_similitud(tifid_matrix[0], tifid_matrix[99])
57 # print 'Cálculo de distancia'
58 # dist = 1 - cosine
59 # print dist
60 # print 'Ángulo de separación de los documentos (grados)'
61 # print angle_in_radians = math.acos(cosine)
62 # print math.degrees(angle_in_radians)
63 # print 'Área de gráficos'
64 # print plt.figure(figsize=(10, 6))
65 # print plt.title('Diamond Pricing')
66 # print plt.xlabel('Carat')
67 # print plt.ylabel('Price')
68 # print plt.ylim(0, 15000)
69 # print plt.xlim(0, 3.5)
70 # print plt.scatter(carat, price, c=clarity, s=100)
71 # print plt.colorbar()
72 # print plt.show()
73 # print 'Clustering de distancia entre documentos'
74 # print 'VWS2'
75 # print 'VWS1'
76 # print 'IF'
77 # print 'Clustering de distancia entre documentos'
78 # print 'VWS2'
79 # print 'VWS1'
80 # print 'IF'
81 # print 'Clustering de distancia entre documentos'
82 # print 'VWS2'
83 # print 'VWS1'
84 # print 'IF'
85 # print 'Clustering de distancia entre documentos'
86 # print 'VWS2'
87 # print 'VWS1'
88 # print 'IF'
89 # print 'Clustering de distancia entre documentos'
90 # print 'VWS2'
91 # print 'VWS1'
92 # print 'IF'
93 # print 'Clustering de distancia entre documentos'
94 # print 'VWS2'
95 # print 'VWS1'
96 # print 'IF'
97 # print 'Clustering de distancia entre documentos'
98 # print 'VWS2'
99 # print 'VWS1'
100 # print 'IF'
```

The code performs clustering and generates a scatter plot titled "Diamond Pricing" with "Carat" on the X-axis and "Price" on the Y-axis. The data points are colored by Clarity level (VWS2, VWS1, IF).

AGENDA



1. Introducción
2. Metodologías
3. Herramientas
- 4. Análisis de Casos**
5. Preparación de Datos
6. Técnicas y Algoritmos
7. Minería de Texto
8. Bibliografía

CASO 1: Predicción del Fracaso Escolar

Unidad Académica Preparatoria de la Universidad Autónoma de Zacatecas (UAPUAZ) de México

- **Tarea:** Detectar cuáles son los factores que más influyen para que los estudiantes de enseñanza media o secundaria suspendan o abandonen para poder ofrecerles algún tipo de ayuda para tratar de evitar y/o disminuir el fracaso escolar.
- **Problemas:** Los datos suelen presentar una alta dimensionalidad (hay muchos factores que pueden influir) y suelen estar muy desbalanceados (la mayoría de los alumnos suelen aprobar y sólo una minoría suele fracasar).
- **Metodología:** Recopilación, Preprocesado, Minería de datos e Interpretación.
- **Herramienta:** Weka

CASO 1: Predicción del Fracaso Escolar

Unidad Académica Preparatoria de la Universidad Autónoma de Zacatecas (UAPUAZ) de México

● Desarrollo del Estudio

1. Recopilación: (1) encuestas a estudiantes, (2) examen de ingreso – estudio socioeconómico y (3) notas de los estudiantes.

2. Preprocesado:

- (1) Integración en único fichero
- (2) Limpieza eliminando registros incompletos
- (3) Transformación eliminando “ñ” y edad en años
- (4) Discretización de las notas
- (5) Validación cruzada
- (6) Selección de variables
- (7) Balanceo de datos

CASO 1: Predicción del Fracaso Escolar

Unidad Académica Preparatoria de la Universidad Autónoma de Zacatecas (UAPUAZ) de México

3. Minería de datos:

- Experimento 1: algoritmos de clasificación con todos las variables
- Experimento 2: algoritmos de clasificación con las variables seleccionadas
- Experimento 3: algoritmos de clasificación con las variables seleccionadas y los datos balanceados
- Experimento 4: evaluación de costos de la clasificación

Predicción del Clasificador	Clase Real		Clase Real
	Clase C_i	NO Clase C_i	
	Positivos para la clase C_i	Negativos para la clase C_i	
	Verdaderos Positivos (VP)	Falsos Positivos (FP)	
	Falsos Negativos (FN)	Verdaderos Negativos (VN)	

Predicción del Clasificador	Clase Real		Clase Real
	Clase C_i	NO Clase C_i	
	Positivos para la clase C_i	Negativos para la clase C_i	
	Beneficio de Verdaderos Positivos (B_{VP})	Costo de Falsos Positivos (C_{FP})	
	Costo de Falsos Negativos (C_{FN})	Beneficio de Verdaderos Negativos (B_{VN})	

$$\text{Beneficios: } (VP * B_{VP}) + (VN * B_{VN})$$
$$\text{Costos: } (FP * C_{FP}) + (FN * C_{FN})$$

4. Interpretación: Lectura de resultados

CASO 2: Predicción de Clientes que Responderán a Campaña de Mercadeo

Cadena de almacenes de ropa rk clothes

- **Tarea:** clasificación de clientes para una campaña de mercadeo a través de correo. Desean predecir los clientes que responderán a la compañía.
- **Metodología:** CRISP-DM
- **Preparación de datos:** Análisis costo/beneficio, correlación con la variable respuesta.
- **Método:** PCA, clustering, votación con redes neuronales, CART, C5.0 y regresión logística.
- **Herramienta:** Clementine de SPSS

Caso 3: Segmentación de Clientes

Banco Quebec - Norteamérica

- **Tarea:** segmentación de clientes para crear estrategias de mercadeo y ventas por segmentos.
- **Metodología:** CRISP-DM
- **Preparación de datos:** Se eliminaron variables redundantes o duplicadas por medio de un análisis de correlaciones.
- **Método:** Clustering demográfico (K vecinos más cercanos) y clustering neural.
- **Herramienta:** DB2 Intelligent Miner - IBM

CASO 4: Perfilamiento de Pacientes

Hospital en Israel

- **Tarea:** Perfiles de pacientes con trombosis de vena profunda (DVT) para determinar similitudes de comportamientos.
- **Metodología:** CRISP-DM
- **Preparación de datos:** PCA y arboles de clasificación para reducir cantidad de variables.
- **Método:** Clustering demográfico (K vecinos más cercanos)
- **Herramienta:** DB2 Intelligent Miner - IBM

Aspectos relevantes de los Casos de Estudio

- **Análisis de Correlaciones:**
 - ✓ Se eliminan variables que estén correlacionadas entre sí.
 - ✓ Cada variable debe tener correlación con la predicción.
- **Matriz de Costo-Beneficio:**
 - ✓ Se integra la matriz de confusión con costos y beneficios
 - Verdaderos Positivos – Verdaderos Negativos → Beneficios
 - Falsos Positivos – Falsos Negativos → Costos
- **Sistemas de Votación:**
 - ✓ Los clasificadores deben ser exactos y diversos
- **Análisis de Componentes Principales – PCA:**
 - ✓ Permite reducir la cantidad de variables
- **Árboles de Clasificación para reducir variables categóricas:**
 - ✓ Las variables con mayor información se encuentran en ramas altas del árbol.

AGENDA



1. Introducción
2. Metodologías
3. Herramientas
4. Análisis de Casos
- 5. Preparación de Datos**
6. Técnicas y Algoritmos
7. Minería de Texto
8. Bibliografía

Tipos de Variables

● Variables Numéricas (cuantitativas)

- Peso
- Años en la empresa
- Salario
- Edad
- Ventas
- Valor de deuda

● Variables Categóricas (cualitativas)

- Sexo = {Hombre, Mujer}
- Enfermedad= {Si, No}
- Estado civil= {Casado, Soltero}
- Estrato= {1,2,3,4,5}
- Religión= {Católica, Otra}
- Mayor de Edad={S, N}
- Nivel de formación= {Bachillerato, Profesional, Universitario}

● Cadenas de Carácteres (string)

● Fechas (date)

Preparación de los Datos

1. Integración de los Datos
2. Descripción de los Datos
3. Limpieza de Datos
4. Transformaciones
5. Selección de Variables
6. Análisis de Correlaciones
7. Reducción de Variables
8. Balanceo de Datos

1. Integración de Datos

- Se debe crear un registro por persona/producto/servicio/sede con todas las variables integradas:

Id	Nombre	Estrato	Sexo	Enfermedad
1	Ana Perez	3	F	SI
2	Mauricio Rios	2	M	NO
3	Samuel Ochoa	4	M	NO
4	Emilia Oviedo	3	F	NO
5	Carmen Reyes	4	F	SI
6	Alberto Arenas	4	M	NO
7	Maria Betancur	3	M	NO
8	Manuel Regino	2	M	NO
9	Sara Merino	2	F	NO
10	Pepe Corrales	2	M	SI
11	Raul Poveda	4	M	NO
12	Cristina Carrillo	3	F	NO
13	Olga Arias	2	F	NO
14	Yaned Cardona	4	F	SI
15	Paula Quintero	3	F	NO
16	Jeronimo Martinez	2	M	NO

Id	Colegio_U	Activo_Web	Asistencia	Entregas_Completas	Trabaja	Examen_Final
1	SI	BAJA	0,1	0,45	NO	APROBADO
2	NO	MEDIA	0,45	0,6	NO	APROBADO
3	NO	ALTA	0,5	0,75	SI	DESAPROBADO
4	NO	ALTA	0	0,5	SI	DESAPROBADO
5	NO	MEDIA	0,65	0,85	NO	APROBADO
6	NO	BAJA	0,1	0	NO	DESAPROBADO
7	NO	MEDIA	0,2	0,9	NO	APROBADO
8	NO	MEDIA	0,3	0,8	SI	DESAPROBADO
9	SI	BAJA	0,35	0,7	NO	APROBADO
10	NO	BAJA	0,75	0,5	SI	DESAPROBADO
11	SI	ALTA	0,7	0,6	NO	APROBADO
12	NO	MEDIA	0,9	0,8	NO	APROBADO
13	NO	ALTA	0,2	0,25	NO	DESAPROBADO
14	NO	ALTA	0,2	0,2	NO	DESAPROBADO
15	NO	BAJA	0,9	0,8	NO	APROBADO
16	NO	MEDIA	1	1	NO	APROBADO



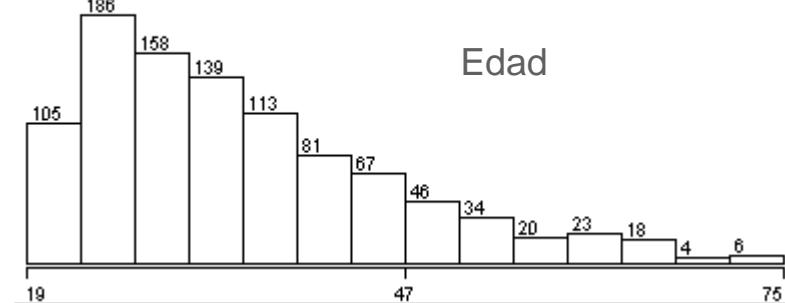
Id	Nombre	Estrato	Sexo	Enfermedad	Colegio_U	Activo_Web	Asistencia	Entregas_Completas	Trabaja	Examen_Final
1	Ana Perez	3	F	SI	SI	BAJA	0,1	0,45	NO	APROBADO
2	Mauricio Rios	2	M	NO	NO	MEDIA	0,45	0,6	NO	APROBADO
3	Samuel Ochoa	4	M	NO	NO	ALTA	0,5	0,75	SI	DESAPROBADO
4	Emilia Oviedo	3	F	NO	NO	ALTA	0	0,5	SI	DESAPROBADO
5	Carmen Reyes	4	F	SI	NO	MEDIA	0,65	0,85	NO	APROBADO
6	Alberto Arenas	4	M	NO	NO	BAJA	0,1	0	NO	DESAPROBADO
7	Maria Betancur	3	M	NO	NO	MEDIA	0,2	0,9	NO	APROBADO
8	Manuel Regino	2	M	NO	NO	MEDIA	0,3	0,8	SI	DESAPROBADO
9	Sara Merino	2	F	NO	SI	BAJA	0,35	0,7	NO	APROBADO
10	Pepe Corrales	2	M	SI	NO	BAJA	0,75	0,5	SI	DESAPROBADO
11	Raul Poveda	4	M	NO	SI	ALTA	0,7	0,6	NO	APROBADO
12	Cristina Carrillo	3	F	NO	NO	MEDIA	0,9	0,8	NO	APROBADO
13	Olga Arias	2	F	NO	NO	ALTA	0,2	0,25	NO	DESAPROBADO
14	Yaned Cardona	4	F	SI	NO	ALTA	0,2	0,2	NO	DESAPROBADO
15	Paula Quintero	3	F	NO	NO	BAJA	0,9	0,8	NO	APROBADO
16	Jeronimo Martinez	2	M	NO	NO	MEDIA	1	1	NO	APROBADO

2. Descripción de Datos

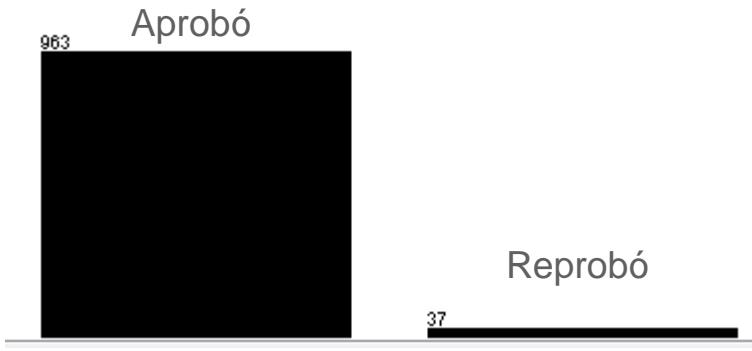
- **Variables numéricas:** estadística descriptiva y visualización con histogramas de frecuencia, cajas de bigote, dispersión.

Edad

Statistic	Value
Minimum	19
Maximum	75
Mean	35.546
StdDev	11.375



- **Variables categóricas:** histogramas de frecuencia.



3. Limpieza de Datos

- Datos faltantes (nulos)
- Datos atípicos (outliers)
- Registros duplicados

3. Limpieza de Datos

Datos faltantes (nulos)

Id	Nombre	Estrato	Sexo	Enfermedad	Colegio_U	Activo_Web	Asistencia	Entregas_Completas	Trabaja	Examen_Final
1	Ana Perez	3	F	SI	SI	BAJA	0,1	0,45	NO	APROBADO
2	Mauricio Rios	2	M	NO	NO	MEDIA	0,45	0,6	NO	APROBADO
3	Samuel Ochoa	4	M	NO	NO	ALTA	0,5	0,75	SI	DESAPROBADO
4	Emilia Oviedo	3	F	NO	NO	ALTA	0	0,5	SI	DESAPROBADO
5	Carmen Reyes	4	F	SI	NO	MEDIA	0,65	0,85	NO	APROBADO
6	Alberto Arenas	4	M	NO	NO	BAJA	0,1	0	NO	DESAPROBADO
7	Maria Betancur	3	M	NO	NO	MEDIA	0,2	0,9	NO	APROBADO
8	Manuel Regino	2	M	NO	NO	MEDIA	0,3	0,8	SI	DESAPROBADO
9	Sara Merino		F	NO	SI	BAJA	0,35	0,7		APROBADO
10	Pepe Corrales	2	M	SI	NO	BAJA	0,75	0,5	SI	DESAPROBADO
11	Raul Poveda	4	M	NO	SI	ALTA	0,7	0,6	NO	APROBADO
12	Cristina Carrillo	3	F	NO	NO	MEDIA	0,9	0,8	NO	APROBADO
13	Olga Arias	2	F	NO	NO	ALTA	0,2	0,25	NO	DESAPROBADO
14	Yaned Cardona	4	F	SI	NO	ALTA	0,2	0,2	NO	DESAPROBADO
15	Paula Quintero	3	F	NO	NO	BAJA	0,9	0,8	NO	APROBADO
16	Jeronimo Martinez	2	M	NO	NO	MEDIA	1	1	NO	APROBADO

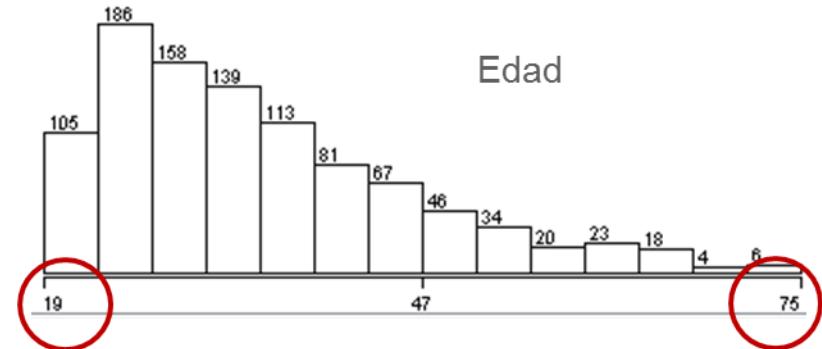
- Eliminar los registros
- Eliminar la variable
- Llenar los campos manualmente
- Usar una constante que identifique el valor nulo
- Llenar los campos con la media o la moda
- Predecir el valor más probable con minería de datos

3. Limpieza de Datos

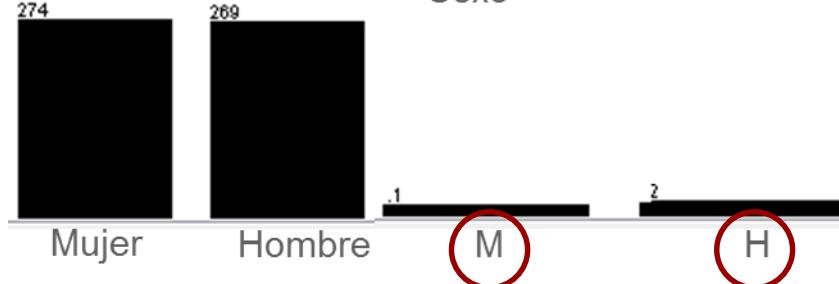
- **Datos atípicos (outliers):** valores por fuera del rango esperado.

Edad

Statistic	Value
Minimum	19
Maximum	75
Mean	35.546
StdDev	11.375



Sexo



SOLUCIONAR

3. Limpieza de Datos

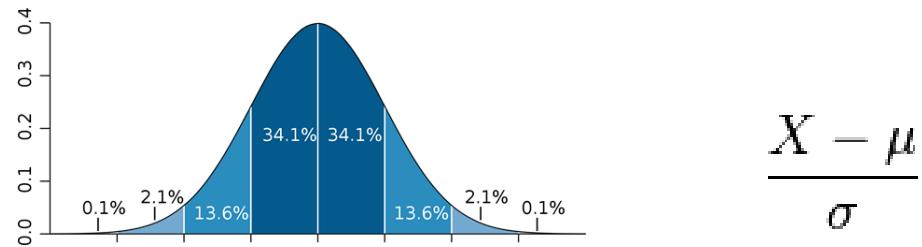
● Registros duplicados

Id	Nombre	Estrato	Sexo	Enfermedad	Colegio_U	Activo_Web	Examen_Final
1	Ana Perez	3	F	SI	SI	BAJA	APROBADO
2	Mauricio Rios	2	M	NO	NO	MEDIA	APROBADO
3	Rios Mauricio	2	M	NO	NO	MEDIA	APROBADO
4	Emilia Oviedo	3	F	NO	NO	ALTA	DESAPROBADO
5	Carmen Reyes	4	F	SI	NO	MEDIA	APROBADO
6	Alberto Arenas	4	M	NO	NO	BAJA	DESAPROBADO

SOLUCIONAR

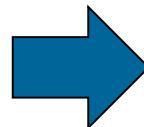
4. Transformaciones de Datos

● Normalización de variables numéricas

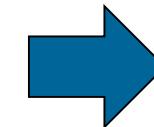


● Discretización de variables numéricas

Edad Numérica
12
17
24
15
7



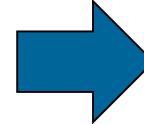
0 – 12 → Niño
13 – 17 → Adolescente
18 en adelante → Adulto



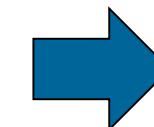
Edad Categórica
Niño
Adolescente
Adulto
Adolescente
Niño

● Conversión de categórica a numérica

Edad Categórica
Niño
Adolescente
Adulto
Adolescente
Niño



Niño → 1
Adolescente → 2
Adulto → 3



Edad Numérica
1
2
3
2
1

5. Selección de Variables

- Eliminar variables irrelevantes para la minería
- Eliminar variables redundantes

● Eliminar variables irrelevantes para la minería:

Información Irrelevante:

- **Nombre**
- **Teléfono**
- **Dirección**
- **Documento de Identificación**

5. Selección de Variables

● Eliminar variables redundantes:

Variables numéricas:

- Edad
- Año de nacimiento

Variables categóricas:

- Edad={niño, adolescente, adulto}
- Mayor de edad={si, no}

Variables numéricas/categóricas:

- Edad
- Mayor de edad={si, no}

ELIMINAR REDUNDANCIAS

6. Análisis de Correlaciones

- Se eliminan variables que estén correlacionadas entre sí (correlación>0,7).
- En análisis predictivo, cada variable debe tener correlación con la predicción (correlación>0,3)

Estrato	Asistencia	Entregas_Completas	Examen_Final
3	0,1	0,45	4,5
2	0,45	0,6	3
4	0,5	0,75	4,8
3	0	0,5	3,7
4	0,65	0,85	5
4	0,1	0	5
3	0,2	0,9	5
2	0,3	0,8	4,6
2	0,35	0,7	4,3
2	0,75	0,5	5
4	0,7	0,6	4,8
3	0,9	0,8	5
2	0,2	0,25	2
4	0,2	0,2	2,5
3	0,9	0,8	5
2	1	1	4

	Estrato	Asistencia	Entregas_Completas	Examen_Final
Estrato	1			
Asistencia	-0,10710653	1		
Entregas_Completas	-0,23797091	0,60077523	1	
Examen_Final	0,27766372	0,36597988	0,432598726	1

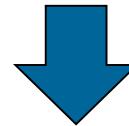
Análisis de correlaciones para datos numéricos

7. Reducción de Variables

- Análisis de Componentes Principales: reducción de dimensión (menos variables):

Matriz de covarianzas
ó
Matriz de correlaciones

Correlation matrix														
1	-0.52	-0.52	0.26	0.15	0.04	-0.24	0.2	0.03	0.13	0.1	-0.15	0.1	Estrato=2	
-0.52	1	-0.45	-0.4	0.08	-0.02	-0.16	0.03	0.13	-0.08	0.21	0.08	-0.32	Estrato=3	
-0.52	-0.45	1	0.13	-0.23	-0.02	0.42	-0.24	-0.16	-0.06	-0.32	0.08	0.22	Estrato=4	
0.26	-0.4	0.13	1	0.29	0.16	-0.13	0.26	-0.13	0.14	0.14	-0.29	0.13	Sexo	
0.15	0.08	-0.23	0.29	1	0.09	0.08	0.15	-0.23	0.06	0.23	0	-0.07	Enfermedad	
0.04	-0.02	-0.02	0.16	0.09	1	-0.02	0.37	-0.37	0.11	0.04	-0.28	0.42	Colegio_U	
-0.24	-0.16	0.42	-0.13	0.08	-0.02	1	-0.52	-0.45	-0.29	-0.37	-0.23	0.49	Activo_web=ALTA	
0.2	0.03	-0.24	0.26	0.15	0.37	-0.52	1	-0.52	0.31	0.63	0.15	-0.42	Activo_web=MEDIA	
0.03	0.13	-0.16	-0.13	-0.23	-0.37	-0.45	-0.52	1	-0.03	-0.29	0.08	-0.05	Activo_web=BAJA	
0.13	-0.08	-0.06	0.14	0.06	0.11	-0.29	0.31	-0.03	1	0.6	0.13	-0.46	Asistencia	
0.1	0.21	-0.32	0.14	0.23	0.04	-0.37	0.63	-0.29	0.6	1	-0.07	-0.58	Entregas_completas	
-0.15	0.08	0.08	-0.29	0	-0.28	-0.23	0.15	0.08	0.13	-0.07	1	-0.65	Trabaja	
0.1	-0.32	0.22	0.13	-0.07	0.42	0.49	-0.42	-0.05	-0.46	-0.58	-0.65	1	Examen_Final	



Ranked attributes:

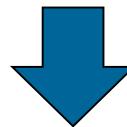
0.7553	1	-0.491	Activo_web=MEDIA	-0.464	Entregas_completas...
0.5778	2	-0.45	Estrato=3	-0.434	Activo_web=BAJA...
0.4376	3	-0.575	Estrato=2	+0.439	Estrato=3...
0.3219	4	-0.509	Estrato=4	-0.474	Trabaja...
0.2339	5	-0.657	Enfermedad	+0.535	Colegio_U...
0.1547	6	0.48	Trabaja	-0.38	Asistencia...
0.0841	7	0.506	Sexo	-0.373	Activo_web=ALTA...
0.03	8	0.582	Colegio_U	+0.508	Asistencia...

7. Reducción de Variables

- Análisis de Componentes Principales: reducción de dimensión (menos variables):

Matriz de covarianzas
ó
Matriz de correlaciones

Correlation matrix															
1	-0.52	-0.52	0.26	0.15	0.04	-0.24	0.2	0.03	0.13	0.1	-0.15	0.1		Estrato=2	
-0.52	1	-0.45	-0.4	0.08	-0.02	-0.16	0.03	0.13	-0.08	0.21	0.08	-0.32		Estrato=3	
-0.52	-0.45	1	0.13	-0.23	-0.02	0.42	-0.24	-0.16	-0.06	-0.32	0.08	0.22		Estrato=4	
0.26	-0.4	0.13	1	0.29	0.16	-0.13	0.26	-0.13	0.14	0.14	-0.29	0.13		Sexo	
0.15	0.08	-0.23	0.29	1	0.09	0.08	0.15	-0.23	0.06	0.23	0	-0.07		Enfermedad	
0.04	-0.02	-0.02	0.16	0.09	1	-0.02	0.37	-0.37	0.11	0.04	-0.28	0.42		Colegio_U	
-0.24	-0.16	0.42	-0.13	0.08	-0.02	1	-0.52	-0.45	-0.29	-0.37	-0.23	0.49		Activo_web=ALTA	
0.2	0.03	-0.24	0.26	0.15	0.37	-0.52	1	-0.52	0.31	0.63	0.15	-0.42		Activo_web=MEDIA	
0.03	0.13	-0.16	-0.13	-0.23	-0.37	-0.45	-0.52	1	-0.03	-0.29	0.08	-0.05		Activo_web=BAJA	
0.13	-0.08	-0.06	0.14	0.06	0.11	-0.29	0.31	-0.03	1	0.6	0.13	-0.46		Asistencia	
0.1	0.21	-0.32	0.14	0.23	0.04	-0.37	0.63	-0.29	0.6	1	-0.07	-0.58		Entregas_completas	
-0.15	0.08	0.08	-0.29	0	-0.28	-0.23	0.15	0.08	0.13	-0.07	1	-0.65		Trabaja	
0.1	-0.32	0.22	0.13	-0.07	0.42	0.49	-0.42	-0.05	-0.46	-0.58	-0.65	1		Examen_Final	

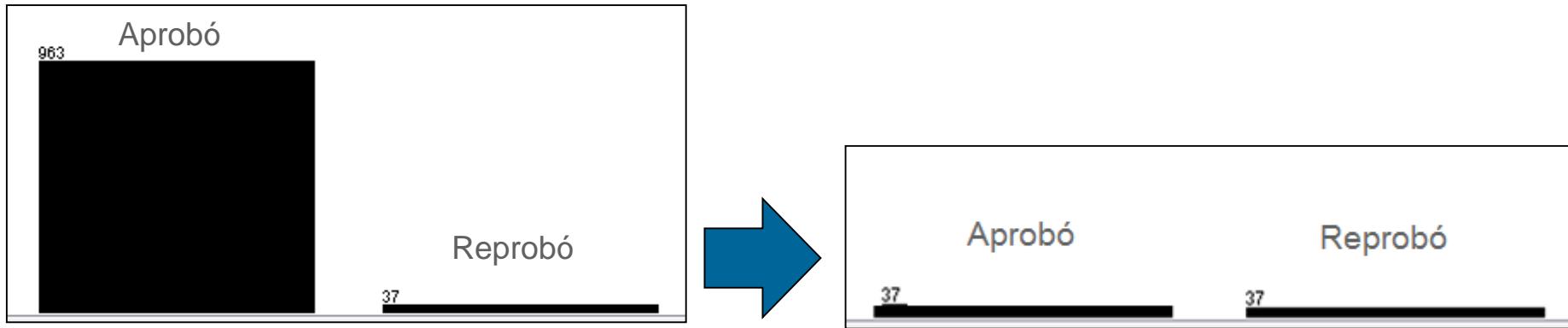


Ranked attributes:

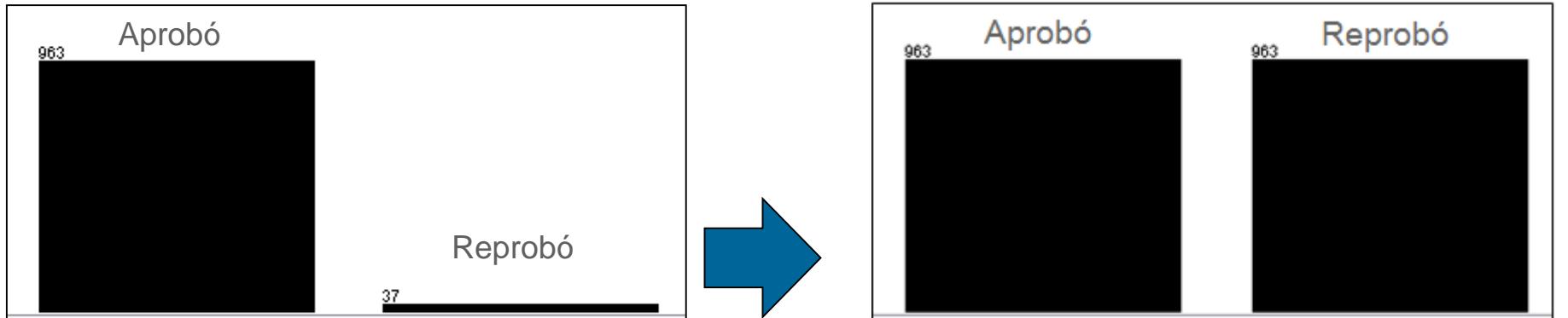
```
1 12 Trabaja
1 11 Entregas_completas
1 4 Sexo
1 3 Estrato=4
1 2 Estrato=3
1 5 Enfermedad
1 6 Colegio_U
1 7 Activo_web=ALTA
1 10 Asistencia
1 9 Activo_web=BAJA
1 8 Activo_web=MEDIA
1 1 Estrato=2
```

8. Balanceo de Datos

- Selección aleatoria de datos



- Adicionar registros cercanos a la media de los datos



AGENDA



1. Introducción
2. Metodologías
3. Herramientas
4. Análisis de Casos
5. Preparación de Datos
- 6. Técnicas y Algoritmos**
7. Minería de Texto
8. Bibliografía

Técnicas y Algoritmos de Minería de Datos

- Técnicas Supervisadas
- Técnicas No Supervisadas

Técnicas y Algoritmos de Minería de Datos

● Técnicas Supervisadas

- Clasificación
- Regresión

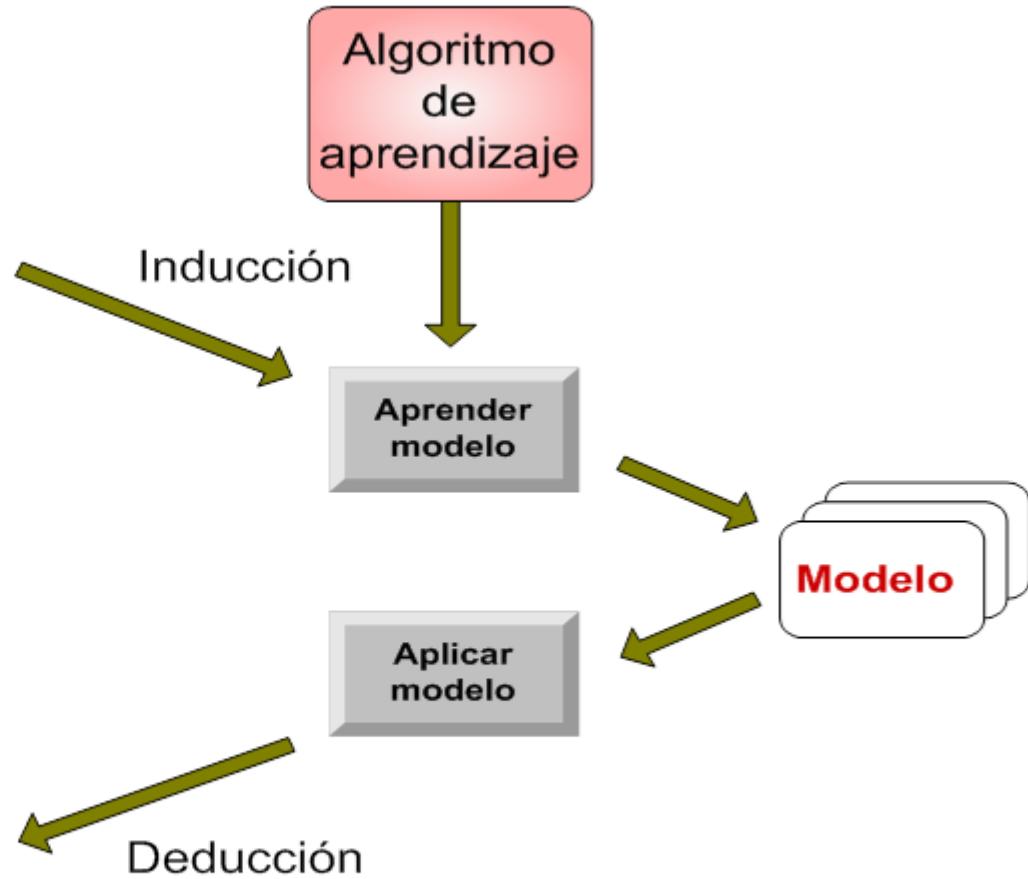
● Técnicas No Supervisadas

Clasificación

Conjunto de entrenamiento

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?



Conjunto de validación

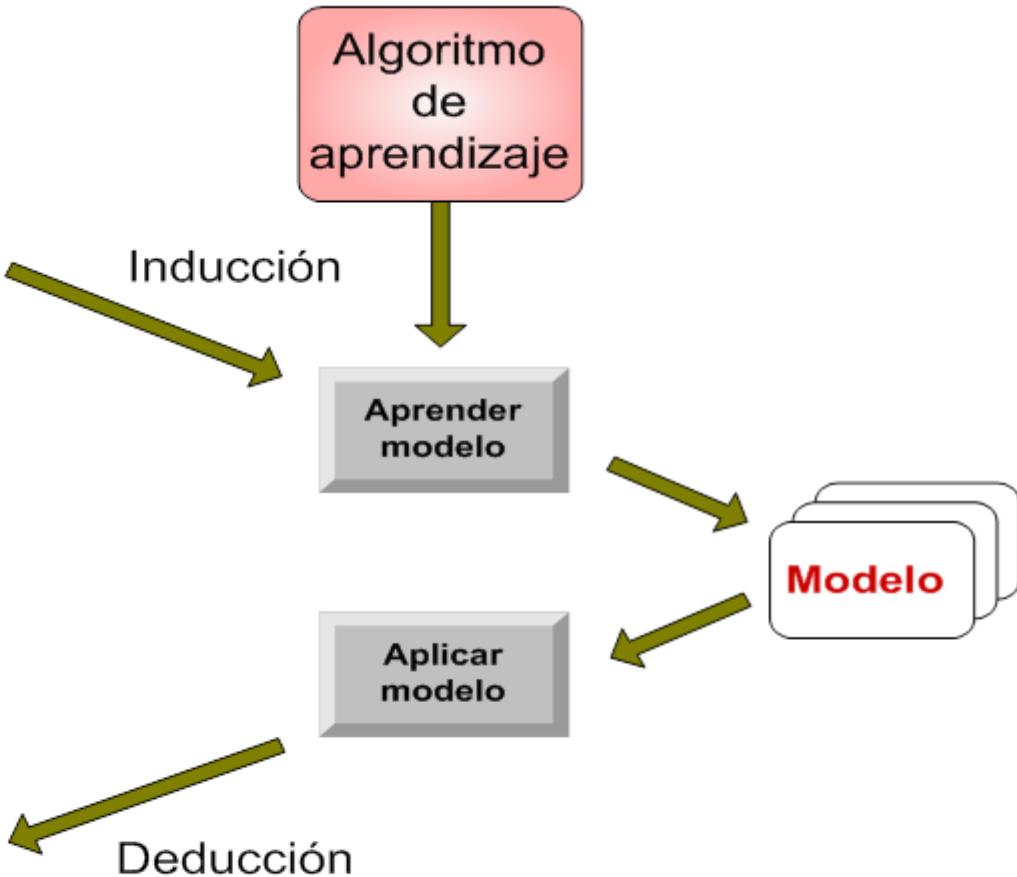
Regresión

Conjunto de entrenamiento

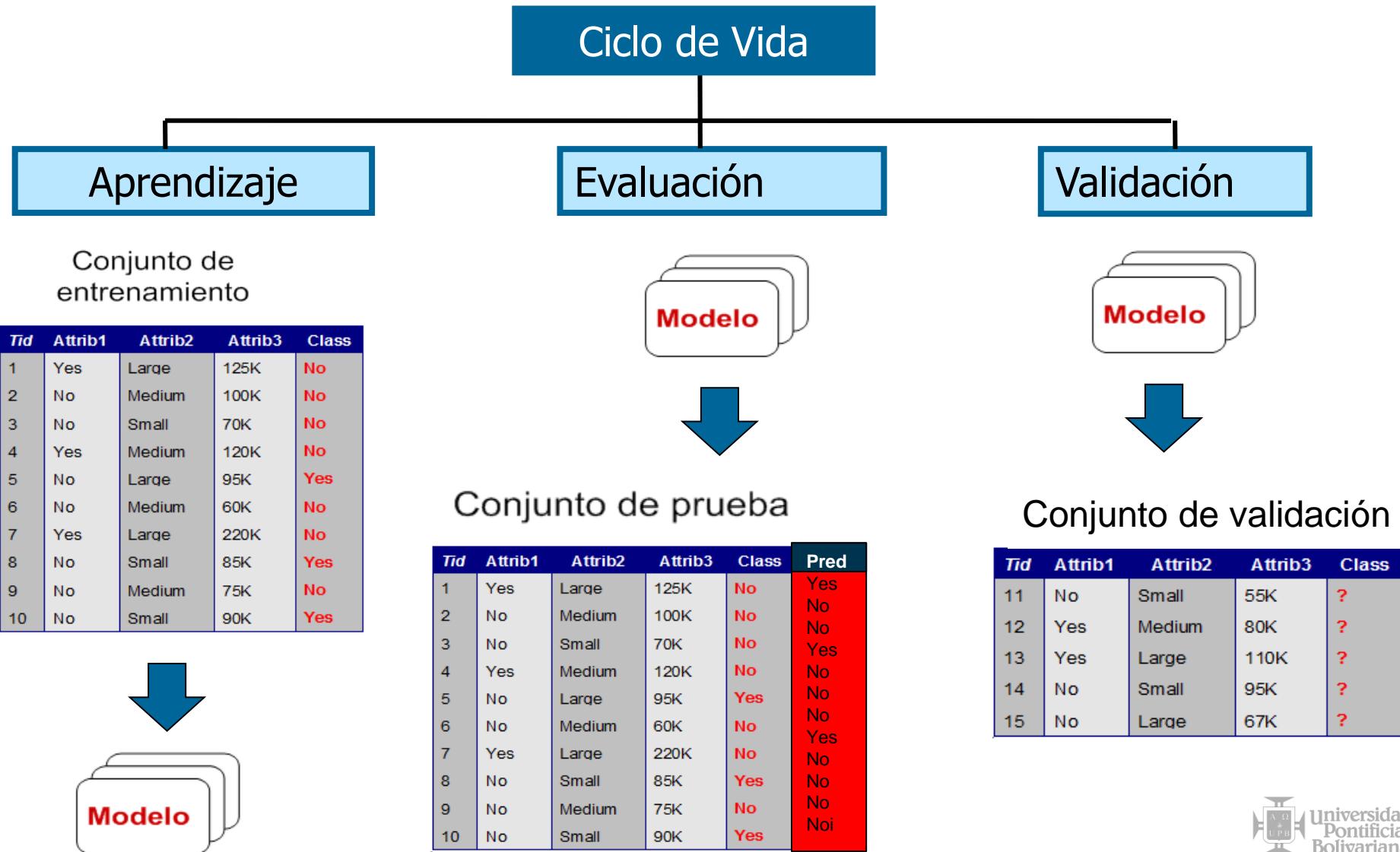
Tid	Attrib1	Attrib2	Attrib3	Pred.
1	Yes	Large	125K	78
2	No	Medium	100K	80
3	No	Small	70K	56
4	Yes	Medium	120K	40
5	No	Large	95K	39
6	No	Medium	60K	65
7	Yes	Large	220K	67
8	No	Small	85K	98
9	No	Medium	75K	76
10	No	Small	90K	45

Tid	Attrib1	Attrib2	Attrib3	Pred.
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Conjunto de validación



Ciclo de Vida : Clasificación y Regresión



Modelos de Evaluación

- Evaluar el conjunto de entrenamiento
- División de Datos (Split)
- Validación Cruzada (K-fold Cross Validation)

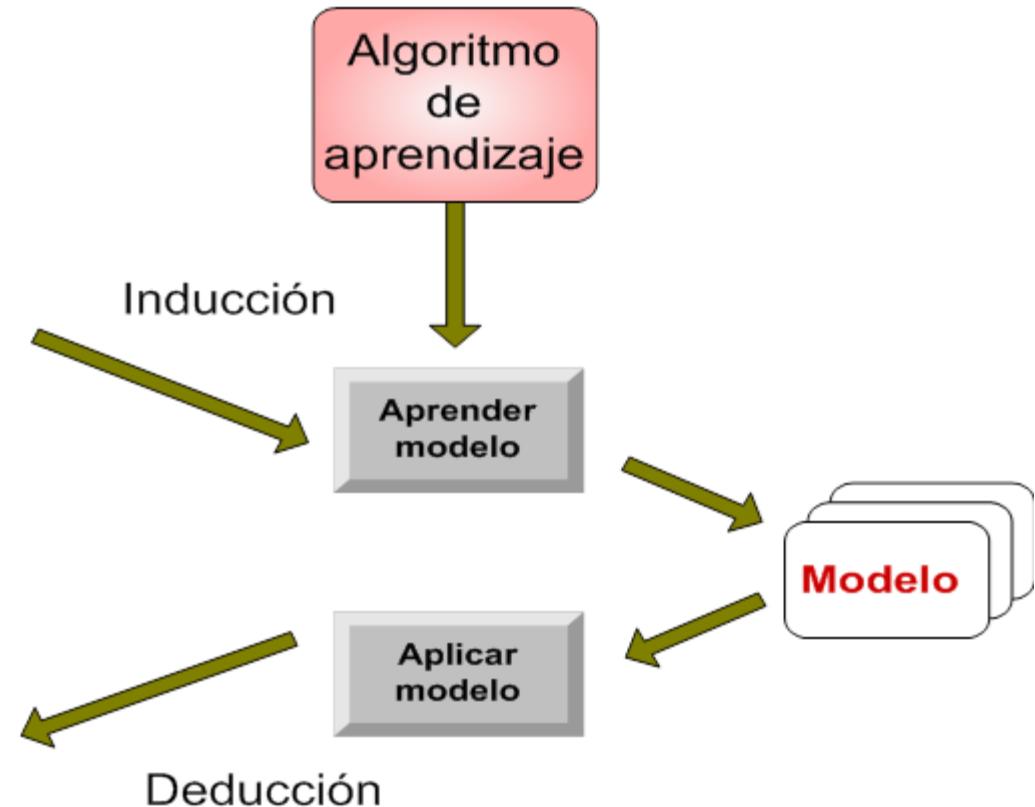
Modelo de Evaluación: Evaluar el conjunto de entrenamiento

Histórico

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Conjunto de entrenamiento

Conjunto de prueba



Modelo de Evaluación: División de Datos

Histórico

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes
11	Yes	Large	125K	No
12	No	Medium	100K	No
13	No	Small	70K	No
14	Yes	Medium	120K	No
15	No	Large	95K	Yes
16	No	Medium	60K	No
17	Yes	Large	220K	No
18	No	Small	85K	Yes
19	No	Medium	75K	No
20	No	Small	90K	Yes

70%

Conjunto de
entrenamiento

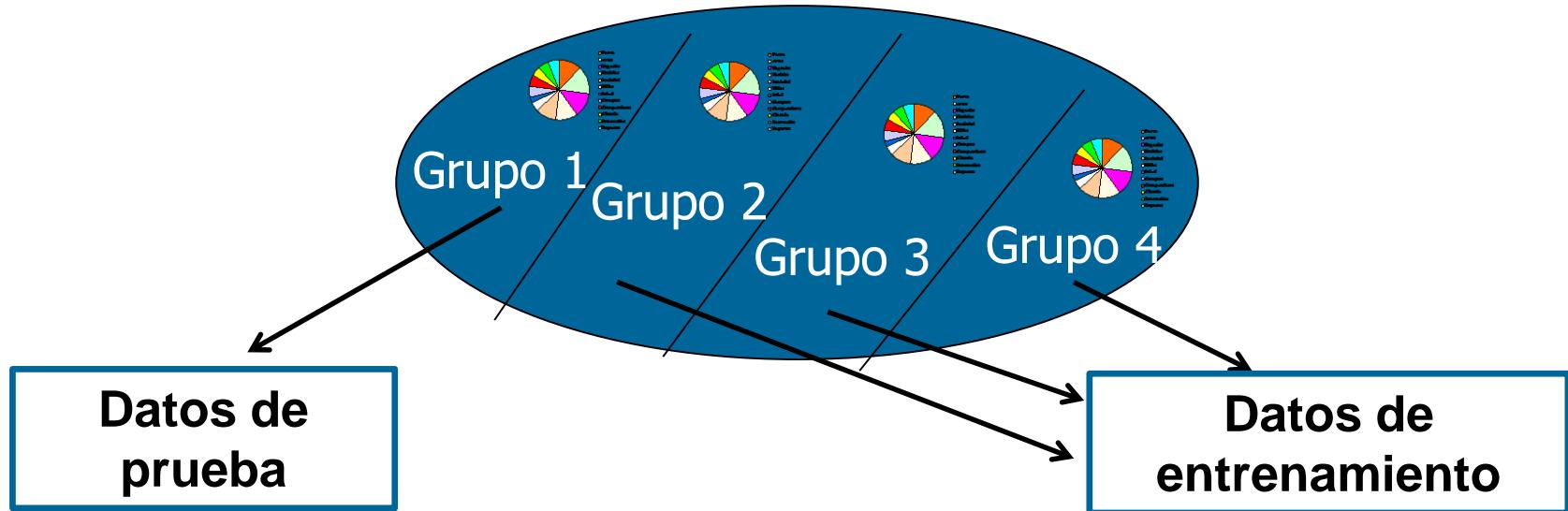
30%

Conjunto de prueba

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Modelo de Evaluación: Validación Cruzada (K-fold Cross Validation)



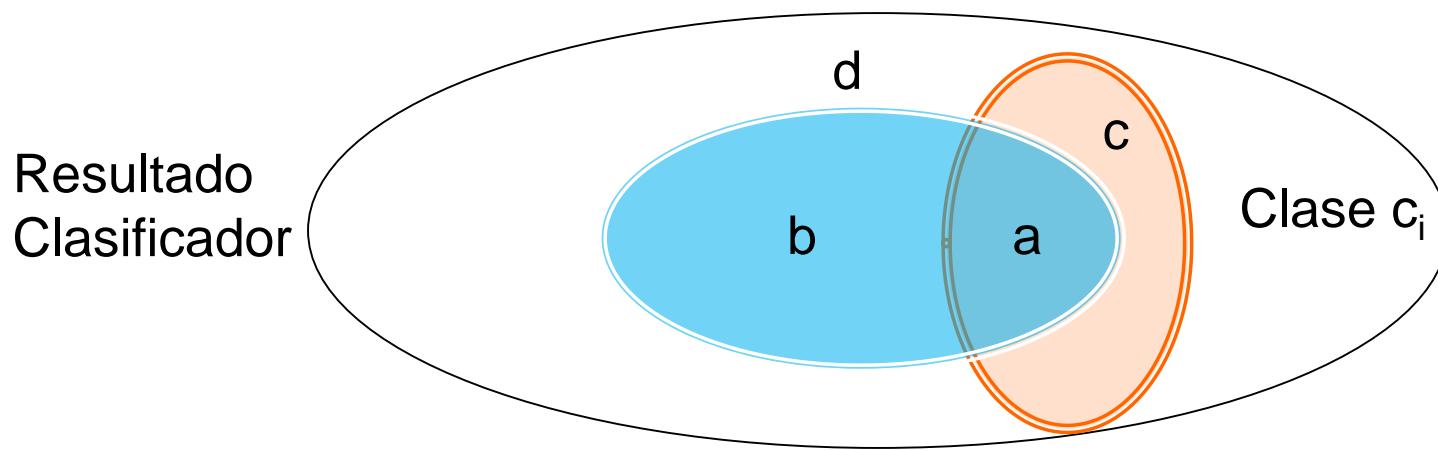
1. Aleatoriamente se divide el conjunto de datos en k subgrupos.
2. Se usan $k-1$ subgrupos en entrenamiento y el otro subgrupo en prueba.
3. Se repite el experimento k veces.

Medidas de Evaluación - Regresión

Mediciones de Error:

$$error(p) = \frac{1}{n} \sum_x (f(x) - p(x))^2$$

Medidas de Evaluación - Clasificador



Medidas de Evaluación - Clasificador

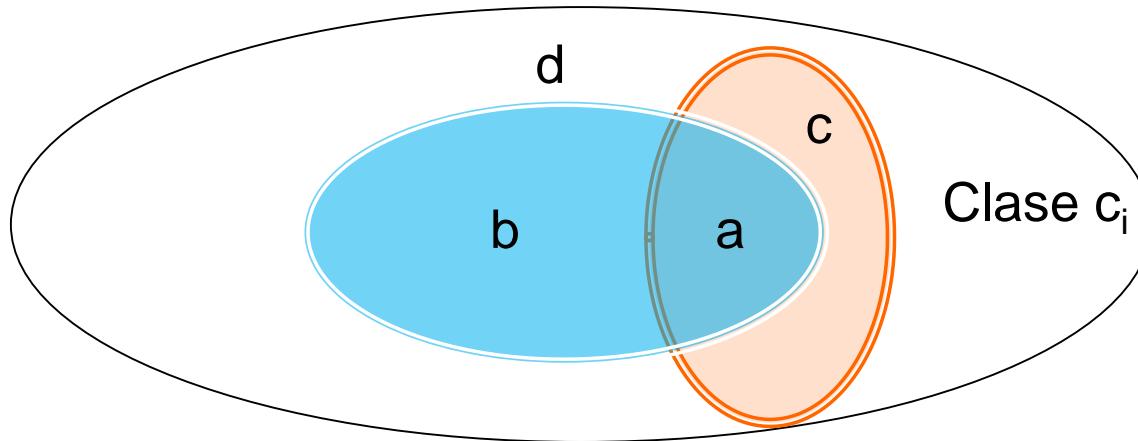


$$\text{Precision } p = \frac{a}{a+b}$$



$$\text{Cobertura } r = \frac{a}{a+c}$$

Resultado
Clasificador



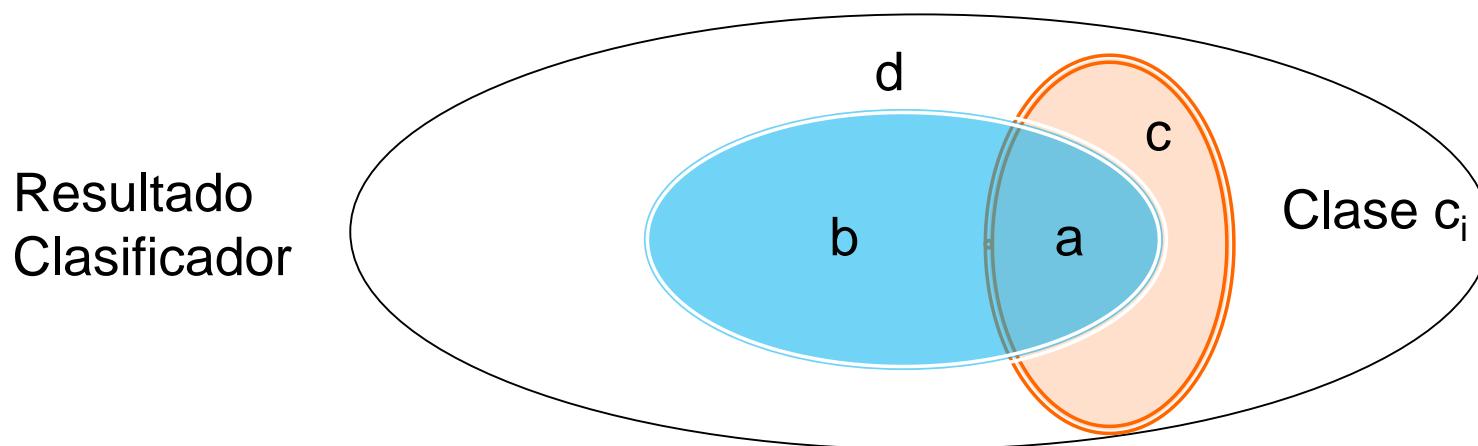
$$\text{Exactitud } e = \frac{a+d}{a+b+c+d}$$



$$\text{Media Armónica } f1 = \frac{2pr}{p+r}$$

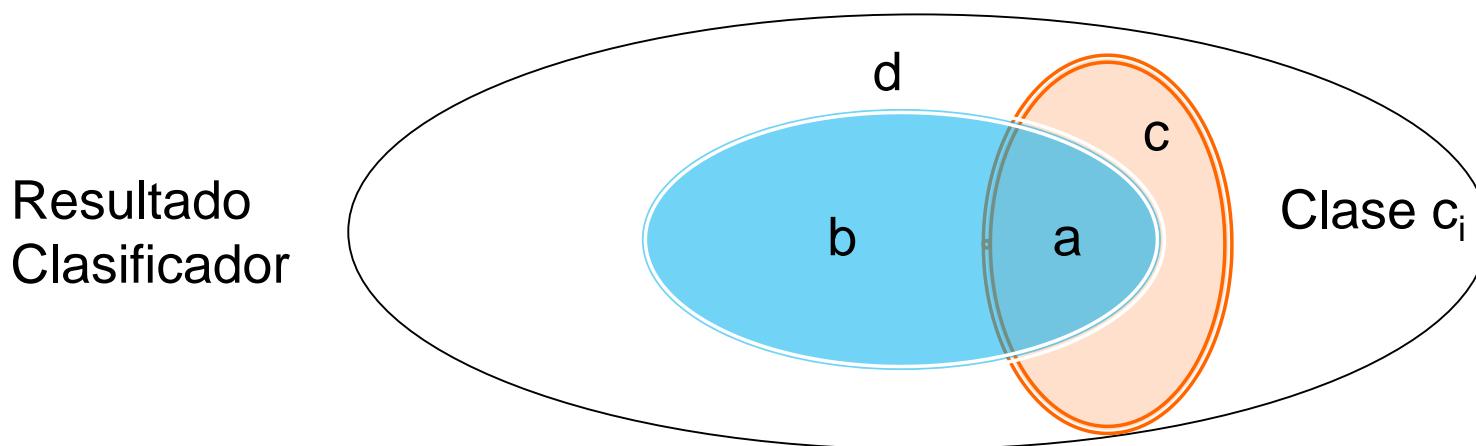
Medidas de Evaluación de la Clasificación: Matriz de Confusión

		Clase Real	
		Clase C_i	NO Clase C_i
Predicción del Clasificador	Positivos para la clase C_i	Verdaderos Positivos (VP)	Falsos Positivos (FP)
	Negativos para la clase C_i	Falsos Negativos (FN)	Verdaderos Negativos (VN)



Medidas de Evaluación de la Clasificación: Matriz de Confusión

		Clase Real	
		Clase C_i	NO Clase C_i
Predicción del Clasificador	Positivos para la clase C_i	Verdaderos Positivos (VP) a	Falsos Positivos (FP) b
	Negativos para la clase C_i	Falsos Negativos (FN) c	Verdaderos Negativos (VN) d



Medidas de Evaluación de la Clasificación: Matriz de Confusión

		Clase Real	
		Clase C _i	NO Clase C _i
Predicción del Clasificador	Positivos para la clase C _i	Verdaderos Positivos (VP) a	Falsos Positivos (FP) b
	Negativos para la clase C _i	Falsos Negativos (FN) c	Verdaderos Negativos (VN) d

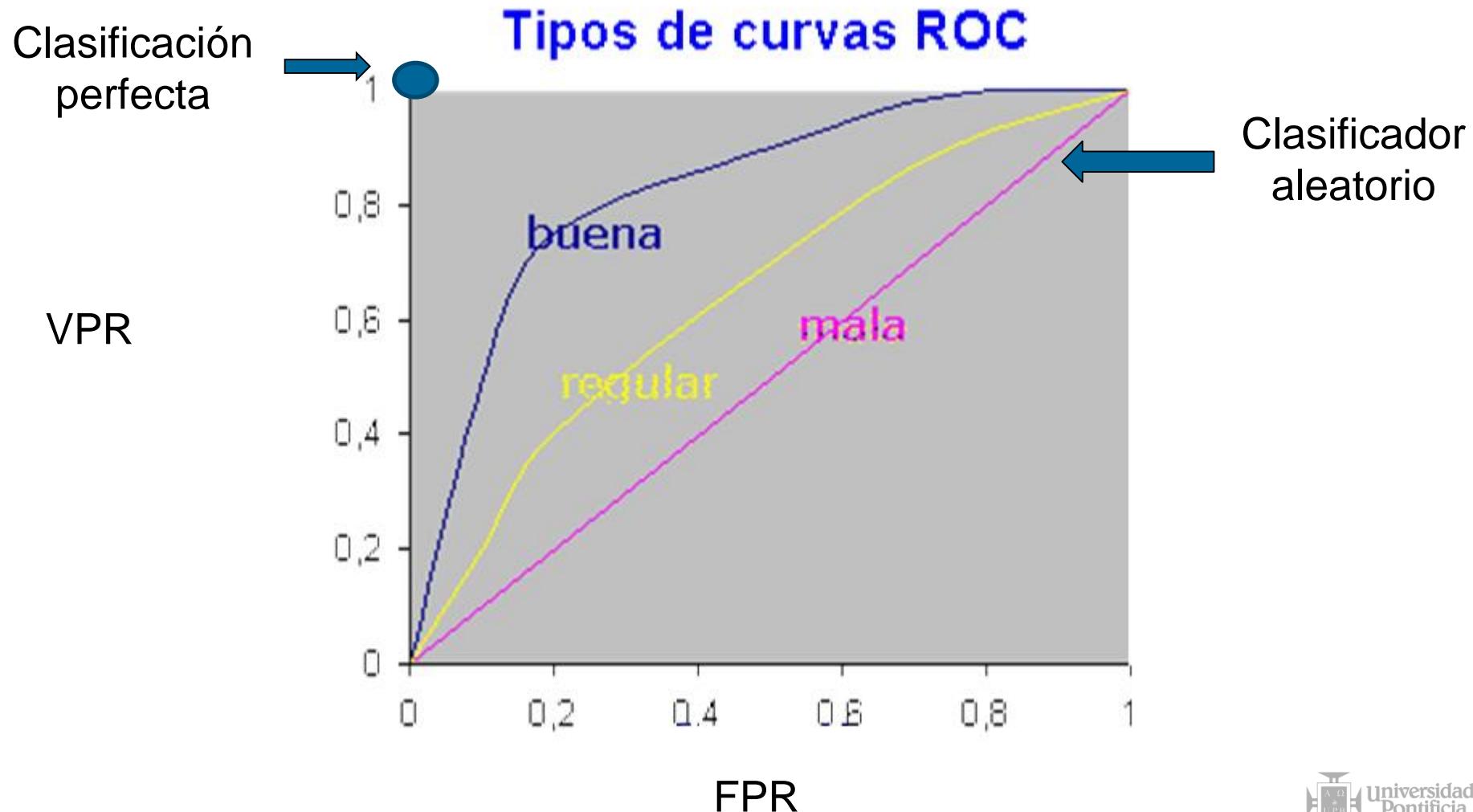
● Razón de verdaderos positivos

$$VPR = \frac{VP}{VP + FN} = \frac{a}{a + c}$$

● Razón de falsos positivos

$$FPR = \frac{FP}{FP + VN} = \frac{b}{b + d}$$

Medidas de Evaluación de la Clasificación: Curva ROC (Receiver Operating Characteristic)



Métodos Supervisados

- Redes Neuronales
[Wiener et al., 1995]
- Árboles de Decisión
[Apte, 1997]
- Métodos Probabilísticos
[Lewis, 1998] [Wettig et al., 2002]
- Máq. de Soporte Vectorial
[Joachims, 1998]
- Métodos de Regresión
[Yang, 1999]
- Métodos basados en Ejemplos
[Yang, 1999]

Métodos Supervisados

Redes Neuronales

[Wiener et al., 1995]

Árboles de Decisión

[Apte, 1997]

Métodos Probabilísticos

[Lewis, 1998] [Wettig et al., 2002]

Máq. de Soporte Vectorial

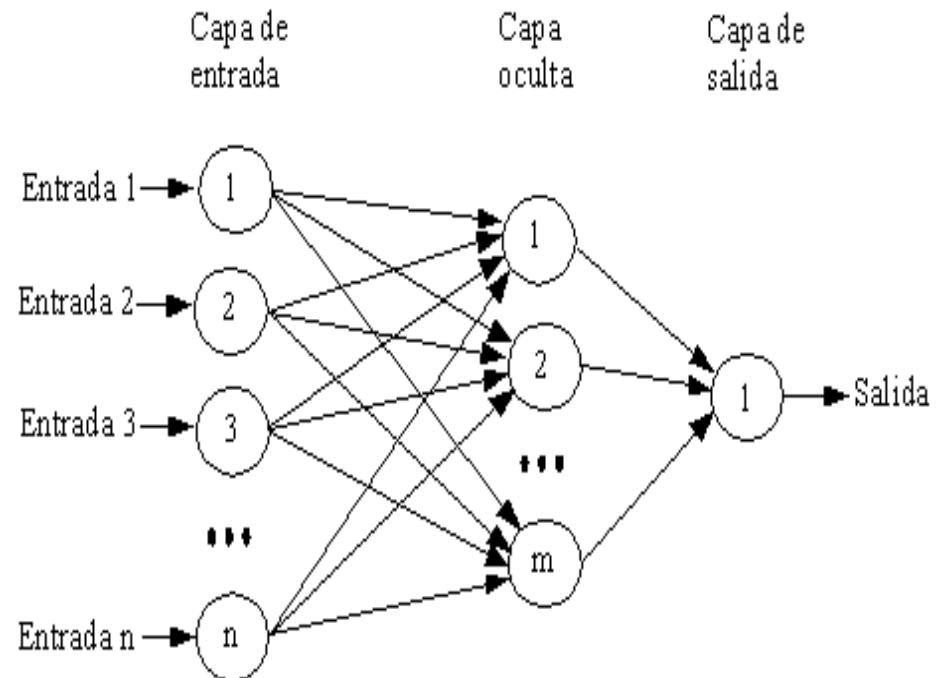
[Joachims, 1998]

Métodos de Regresión

[Yang, 1999]

Métodos basados en Ejemplos

[Yang, 1999]



Métodos Supervisados

Redes Neuronales

[Wiener et al., 1995]

Árboles de Decisión

[Apte, 1997]

Métodos Probabilísticos

[Lewis, 1998] [Wettig et al., 2002]

Máq. de Soporte Vectorial

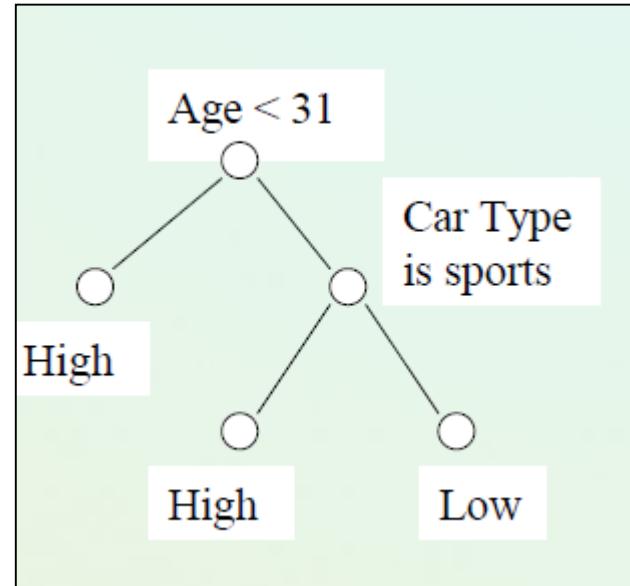
[Joachims, 1998]

Métodos de Regresión

[Yang, 1999]

Métodos basados en Ejemplos

[Yang, 1999]



Métodos Supervisados

Redes Neuronales

[Wiener et al., 1995]

Árboles de Decisión

[Apte, 1997]

Métodos Probabilísticos

[Lewis, 1998] [Wettig et al., 2002]

Máq. de Soporte Vectorial

[Joachims, 1998]

Métodos de Regresión

[Yang, 1999]

Métodos basados en Ejemplos

[Yang, 1999]

$$P(c_j | d) = \frac{P(c_j)P(d | c_j)}{P(d)}$$

Métodos Supervisados

Redes Neuronales

[Wiener et al., 1995]

Árboles de Decisión

[Apte, 1997]

Métodos Probabilísticos

[Lewis, 1998] [Wettig et al., 2002]

Máq. de Soporte Vectorial

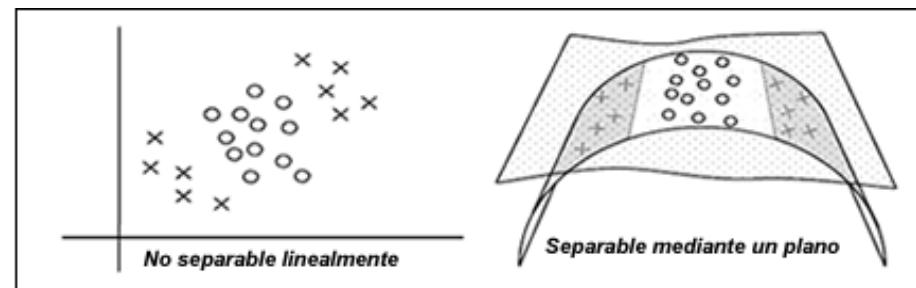
[Joachims, 1998]

Métodos de Regresión

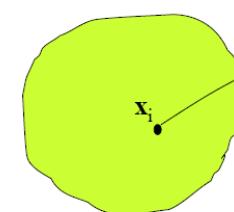
[Yang, 1999]

Métodos basados en Ejemplos

[Yang, 1999]

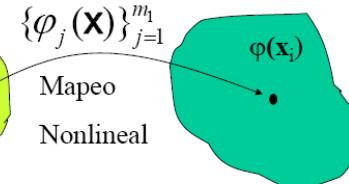


Dimension: m_0



Espacio de entrada
(datos)

Dimension: m_1



Espacio de
Características

Métodos Supervisados

Redes Neuronales

[Wiener et al., 1995]

Árboles de Decisión

[Apte, 1997]

Métodos Probabilísticos

[Lewis, 1998] [Wettig et al., 2002]

Máq. de Soporte Vectorial

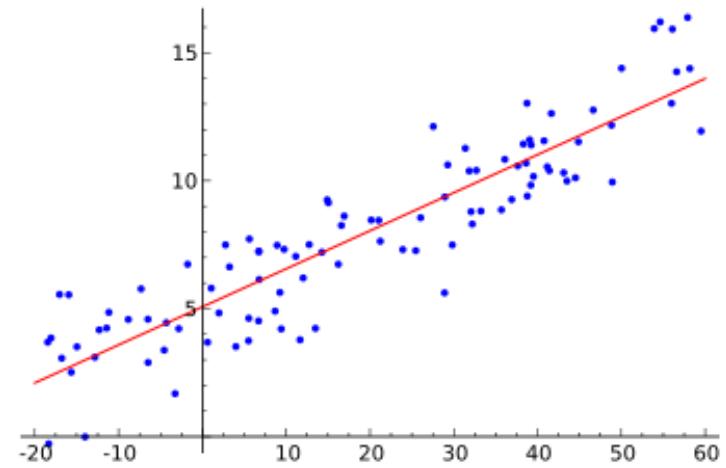
[Joachims, 1998]

Métodos de Regresión

[Yang, 1999]

Métodos basados en Ejemplos

[Yang, 1999]



Métodos Supervisados

Redes Neuronales

[Wiener et al., 1995]

Árboles de Decisión

[Apte, 1997]

Métodos Probabilísticos

[Lewis, 1998] [Wettig et al., 2002]

Máq. de Soporte Vectorial

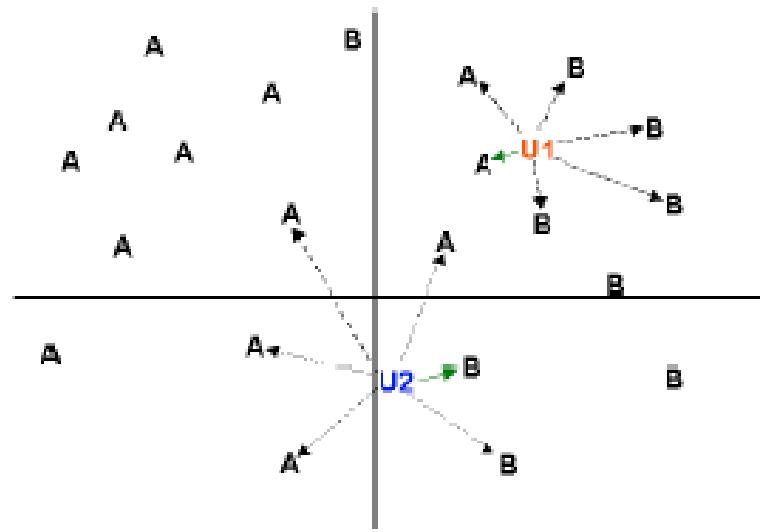
[Joachims, 1998]

Métodos de Regresión

[Yang, 1999]

Métodos basados en Ejemplos

[Yang, 1999]



Algoritmos Supervisados

Método	Algoritmos
Redes Neuronales	Perceptrón, backpropagation, madaline, redes de base radial, hopfield.
Reglas y Arboles de Decisión	C4.4, CLS, ID3, See5/C5.0
Métodos Probabilísticos	EM, clasificador de naive bayes, regresion lineal, regresión gausiana
Máquinas de Soporte Vectorial	SVM, SVMlight, transductiveSVM, multiclassSVM
Métodos de Regresión	SimpleLinear Regresion, Logistic Regresion, CARS
Métodos basados en Ejemplos	KNN

Otros Métodos Supervisados

Método	Algoritmos
Reglas Semánticas	tableaux, resolution, datalog
Modelos Ocultos de Markov	HMM, viterbi, HMM jerarquico,
Método Rocchio	rocchio
Métodos Evolutivos	AG (algoritmos genéticos)

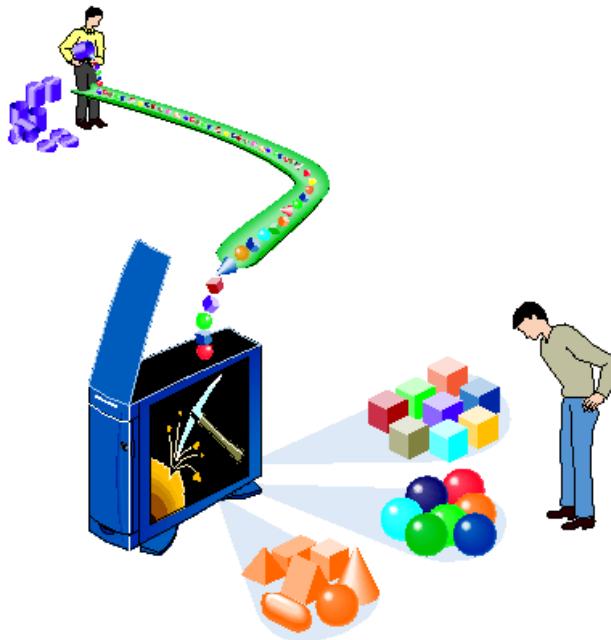
Técnicas y Algoritmos de Minería de Datos

● Técnicas Supervisadas

● Técnicas No Supervisadas

- Clustering
- Reglas de Asociación
- Selección de Factores

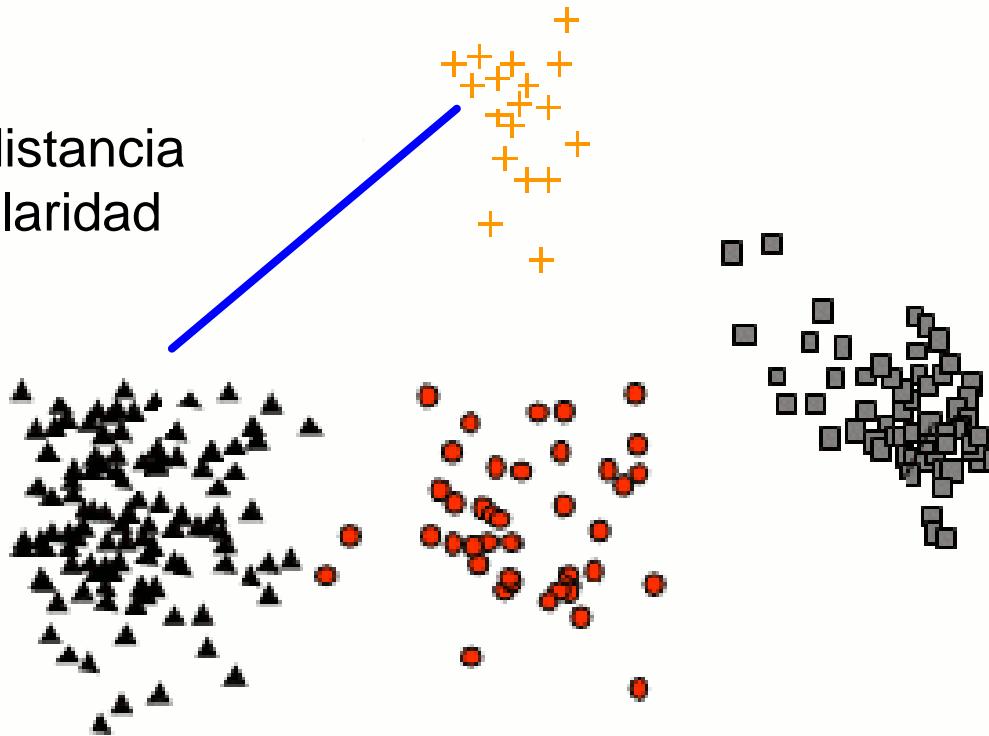
Técnicas NO Supervisadas



- **Descripción de datos**, a partir de los datos de entrada “X” se busca nuevo conocimiento.

Clustering

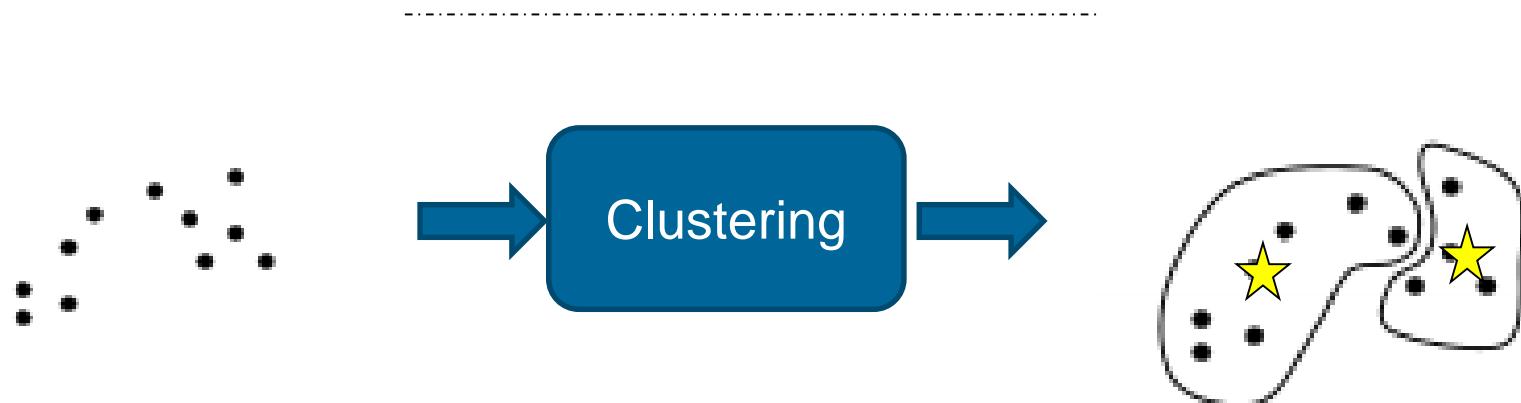
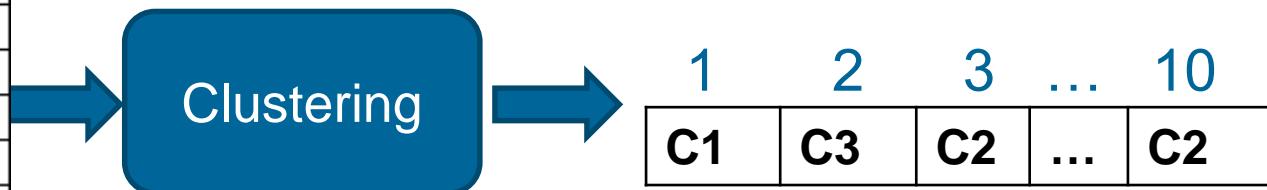
Una medida de distancia determina la similaridad entre los datos.



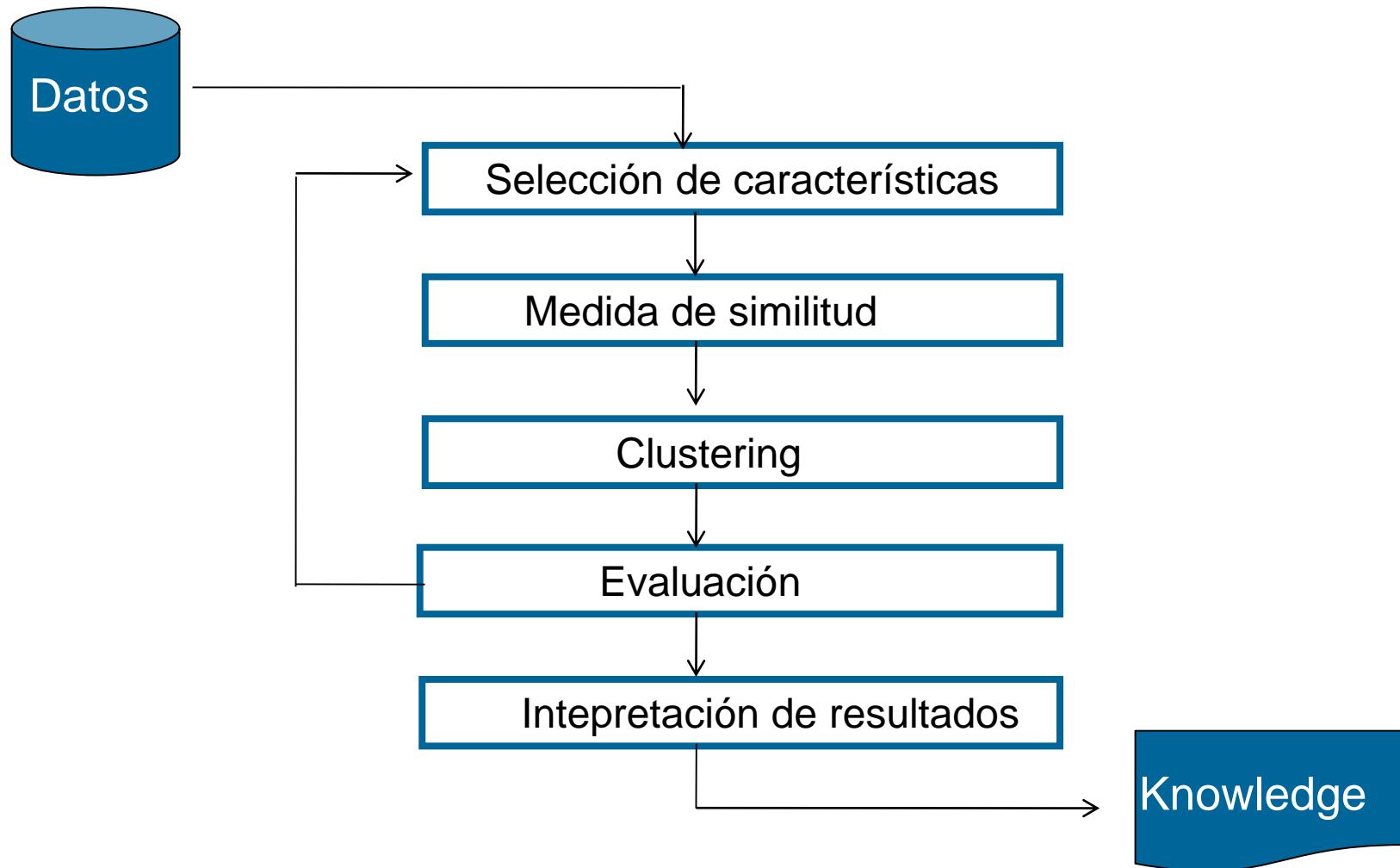
Los objetos en un grupo deben ser similares o relacionados entre ellos.

Clustering

Id	Atributo 1	Atributo 2	...	Atributo n
1	10	alto		35
2	35	bajo		54
3	43	medio		28
4	26	bajo		65
5	87	alto		32
6	45	alto		29
7	76	bajo		55
8	5	medio		46
9	12	medio		43
10	54	bajo		27



Ciclo de Vida de Clustering



Medidas de Similitud basadas en Distancia

Distance Measure	Description
Euclidean	This is the geometric distance in the multidimensional space [Jain et al., 1999]. $d(x_i, x_j) = \sqrt{\sum_{l=1}^d \ x_{il} - x_{jl}\ ^2} \quad (4.1)$
Cosine	This is the cosine of the angle between the feature vectors [Friedman et al., 2007]. $d(x_i, x_j) = \frac{x_i \cdot x_j}{\ x_i\ \times \ x_j\ } \quad (4.2)$
Manhattan	This is the sum of the differences of their corresponding components [Friedman et al., 2007]. $d(x_i, x_j) = \sum_{l=1}^d \ x_{il} - x_{jl}\ \quad (4.3)$
Chebyshev	This finds the absolute magnitude of the differences between the vectors [de Souza and de Carvalho, 2004]. $d(x_i, x_j) = \max_l(\ x_{il} - x_{jl}\) \quad (4.4)$
Mahalanobis	This is the same as the euclidean distance with the covariance matrix [Jain et al., 1999]. $d(x_i, x_j) = \sqrt{(x_i - x_j)^T S^{-1} (x_i - x_j)} \quad (4.5)$
Minkowski	This is a general case, when $p = 2$ is the euclidean distance, while $p = 1$ is the manhattan distance [Groenen and Jajuga, 2001]. $d(x_i, x_j) = \left(\sum_{l=1}^d \ x_{il} - x_{jl}\ ^p \right)^{\frac{1}{p}} \quad (4.6)$
Hamming	This is the number of features in which the vectors differ [Leszek et al., 2004]. $d(x_i, x_j) = \text{amount}_l(x_{il} \neq x_{jl}) \quad (4.7)$

Evaluación de Clustering

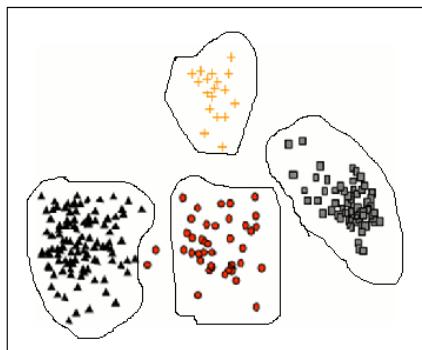
- Validación Externa: Se requiere conocer las clases reales para hacer una comparación con la agrupación obtenida.
- Validación Interna: Evalúa la calidad de los clusters basado en medidas de distancia.

Validación Externa de Clustering

- Validación Externa: Evalúa la calidad de los clusters basado en una estructura pre-clasificada, algunos índices son:

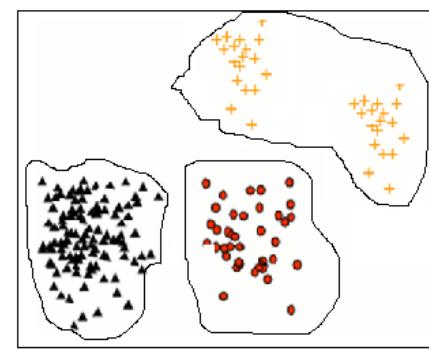
- ✓ Rand Index [Rand, 1971]
- ✓ Fowlkes and Mallows Index [Fowlkes and Mallows, 1983]
- ✓ Hubert and Arabie Index [Hubert and Arabie, 1985]
- ✓ Jaccard Index [Harper, 1999]

Comparing:



Partición real conocida

VS.



Partición obtenida por el método de clustering

Índices para Validación Externa de Clustering

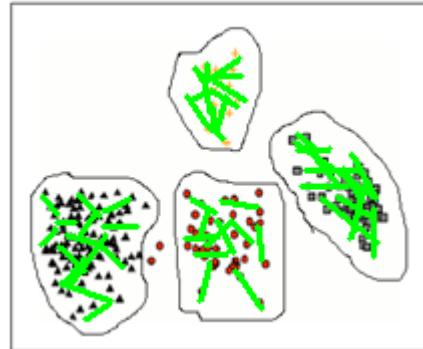
External Indexes
Rand Index [Rand, 1971]:
$R(C, C^*) = 1 + \frac{2 \sum_{pq} \binom{n_{pq}}{2} - \sum_p \binom{n_p}{2} - \sum_q \binom{n_q^*}{2}}{\binom{n}{2}}, \quad (4.8)$ A value near to 1 indicates a strong similarity between C^* and C .
Fowlkes and Mallows Index [Fowlkes and Mallows, 1983]:
$FM(C, C^*) = \frac{\frac{1}{2}(z - n)}{\left(\sum_p \binom{n_p}{2} \sum_q \binom{n_q^*}{2}\right)^{\frac{1}{2}}}, \quad (4.9)$ where $z = \sum_{pq} n_{pq}^2$. A value near to 1 indicates a strong similarity between C^* and C .
Hubert and Arabie Index [Hubert and Arabie, 1985]:
$HA(C, C^*) = (1/M) \sum_{i=1}^{n-1} \sum_{j=i+1}^n Z(i, j)Y(i, j), \quad (4.10)$ where: $M = n(n - 1)/2$ $Z(i, j) = \{1, \text{ if } x_i \text{ and } x_j \text{ belong to same cluster in } C, \text{ and } 0 \text{ otherwise}\}, \forall i, j = 1 \dots n$ $Y(i, j) = \{1, \text{ if } x_i \text{ and } x_j \text{ belong to same cluster in } C^*, \text{ and } 0 \text{ otherwise}\}, \forall i, j = 1 \dots n$ High values of this index indicate a strong similarity between C^* and C .
Jaccard Index [Harper, 1999]:
$J(C, C^*) = \frac{(z - n)}{\sum_p n_p^2 + \sum_q n_q^* - z - n}, \quad (4.11)$ where $z = \sum_{pq} n_{pq}^2$. A value near to 1 indicates a strong similarity between C^* and C .

Validación Interna de Clustering

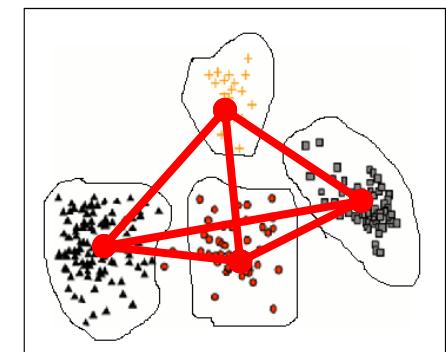
- Validación Interna: Evalúa la calidad de los clusters basado en medidas de distancia, algunos índices son:

- ✓ Dunn Index [Dunn, 1974]
- ✓ Davies-Bouldin Index [Davies and Bouldin, 1979]
- ✓ Silhouette Index [Kaufman and Rousseeuw, 1990]

Compactness:



Separability:



Índices para Validación Interna de Clustering

Dunn Index [Dunn, 1974]:

$$D(C) = \min_p \left\{ \min_{p \neq q} \left\{ \frac{d_{inter}(c_p, c_q)}{\max_{1 \leq r \leq k} \{d_{intra}(c_r)\}} \right\} \right\}, \quad (4.14)$$

High values of this index indicate a good clustering structure.

Davies and Bouldin Index [Davies and Bouldin, 1979]:

$$DB(C) = \frac{1}{k} \sum_{p=1}^k \max_{q \neq p} \left\{ \frac{d_{intra}(c_p) + d_{intra}(c_q)}{d_{inter}(c_p, c_q)} \right\} \quad (4.15)$$

Small values of this index indicate a good clustering structure.

Silhouette Index [Kaufman and Rousseeuw, 1990]:

$$S(C) = \frac{\sum_i sil_i}{n} \quad \text{and} \quad sil_i = \frac{(b_i - a_i)}{\max(a_i, b_i)}, \quad (4.16)$$

where a_i is the average distance of object x_i to all other objects in the same cluster, and b_i is the minimum of average distance of object x_i to all objects in other clusters. High values of this index indicate a good clustering structure.

Métodos de Clustering

- Métodos Jerárquicos [Xu and Wunsch, 2005] [Filippone et al., 2008]
- Métodos Particionales [Steinley, 2006]
- Redes Neuronales [Vesanto and Alhoniemi, 2000]
- Métodos Probabilísticos [Francois et al., 2006]

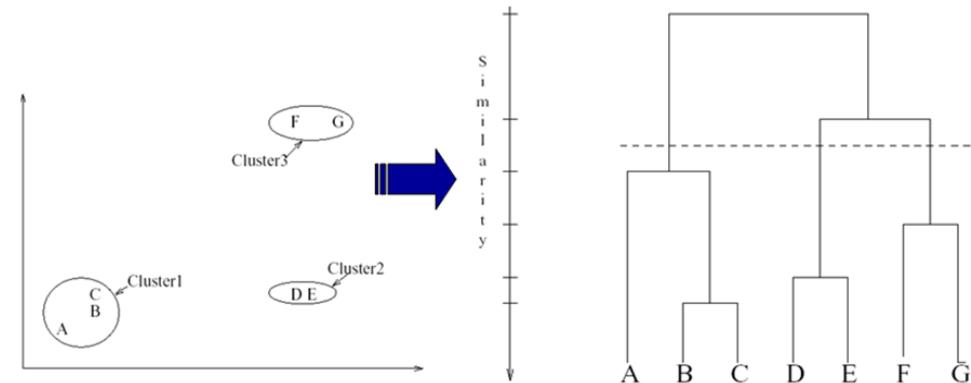
Métodos de Clustering

Métodos Jerárquicos

Métodos Particionales

Redes Neuronales

Métodos Probabilísticos



[Xu and Wunsch, 2005] [Filippone et al., 2008]

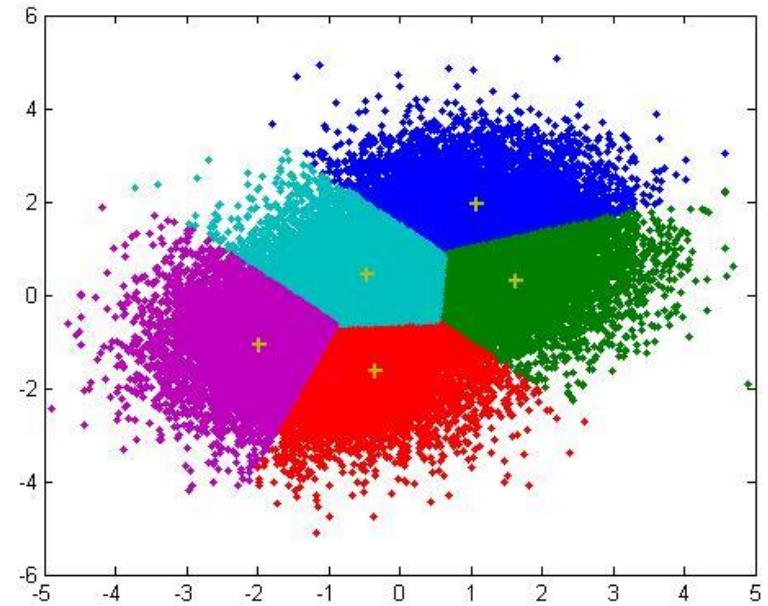
Métodos de Clustering

Métodos Jerárquicos

Métodos Particionales

Redes Neuronales

Métodos Probabilísticos



[Steinley, 2006]

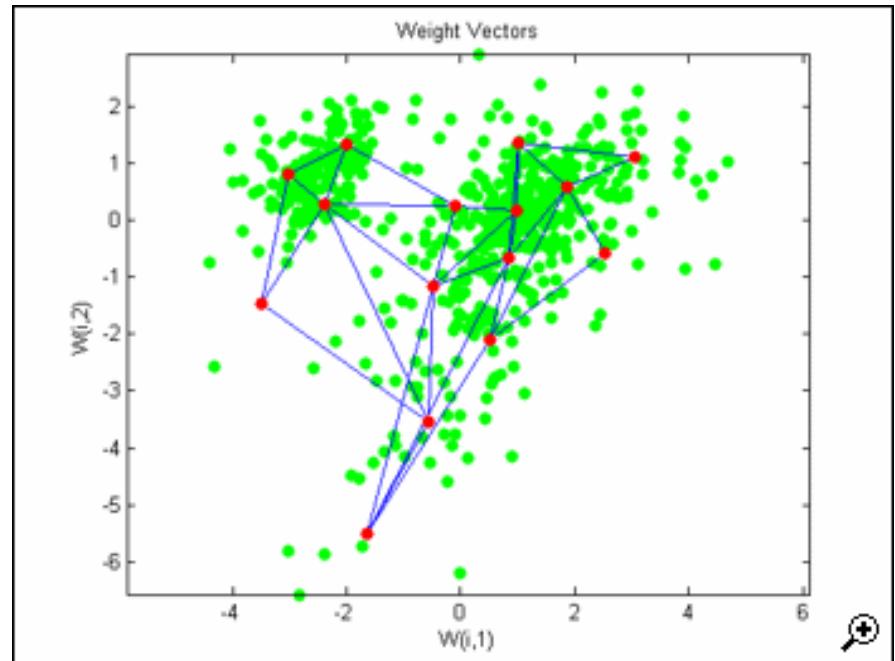
Métodos de Clustering

Métodos Jerárquicos

Métodos Particionales

Redes Neuronales

Métodos Probabilísticos



[Vesanto and Alhoniemi, 2000]

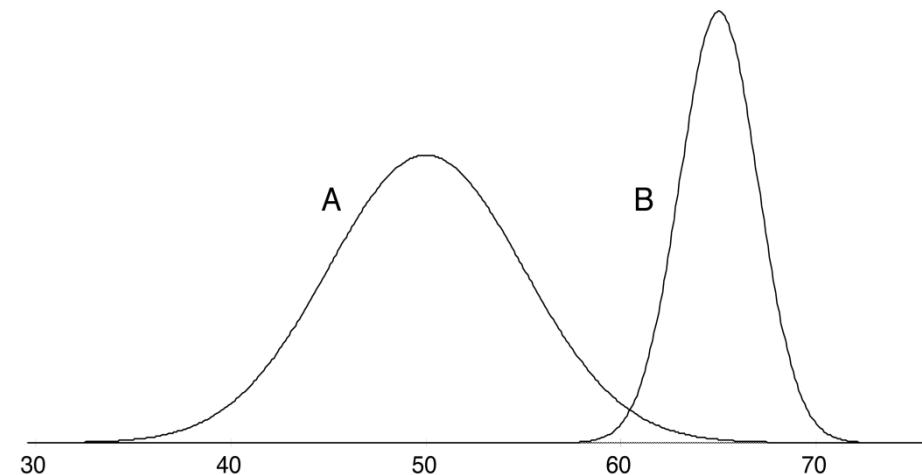
Métodos de Clustering

Métodos Jerárquicos

Métodos Particionales

Redes Neuronales

Métodos Probabilísticos



[Francois et al., 2006]

Otros Métodos de Clustering

- Métodos Difusos
- Métodos basados en Grafos
- Métodos Evolutivos
- Métodos basados en Kernel
- Métodos Espectrales

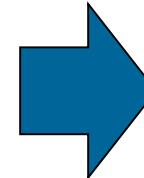
Algoritmos de Clustering

Approach	Algorithms
Hierarchical Method	Agglomerative: Single linkage, complete linkage, group average linkage, median linkage, centroid linkage, Ward's method, BIRCH, CURE, ROCK
	Divisive: Divisive analysis (DIANA), monothetic analysis (MONA)
Partitional Method	K-means, iterative self-organizing data analysis technique (ISODATA), genetic -means algorithm (GKA), partitioning around medoids (PAM)
Fuzzy Method	Fuzzy -means (FCM), mountain method (MM), possibilistic -means clustering algorithm (PCM), fuzzy c-shells (FCS), Kernel Fuzzy c-means, LAMDA
Neural Networks Method	Learning vector quantization (LVQ), self-organizing feature map (SOFM), ART, simplified ART (SART), hyperellipsoidal clustering network (HEC), self-splitting competitive learning network (SPLL)
Probabilistic Method	Gaussian mixture density decomposition (GMDD), AutoClass
Graph - based Method	Chameleon, delaunay triangulation graph (DTG), highly connected subgraphs (HCS), clustering identification via connectivity kernels (CLICK), cluster affinity search technique (CAST)
Evolutionary Method	Genetically guided algorithm (GGA), TS clustering, SA clustering
Kernel Method	Kernel -means, support vector clustering (SVC), KFCM
Spectral Method	Meila and Shi algorithm, Kannan et al Spectral algorithm

Reglas de Asociación

Los datos se reconocen como transacciones y se buscan tendencias (co-ocurrencias).

Tid	Artículos
1	Pan, leche, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, cerveza
4	Pan, leche, pañales, cerveza
5	Pan, leche, huevos, cerveza



Reglas de asociación

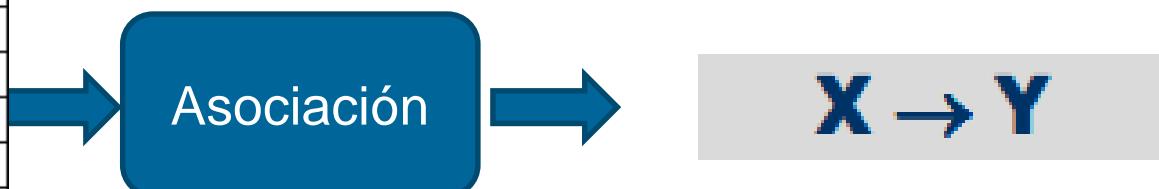
$\{\text{pañales}\} \rightarrow \{\text{cerveza}\}$

$\{\text{leche, pan}\} \rightarrow \{\text{huevos}\}$

Permite descubrir recurrencias en los datos y relaciones entre las variables. Los datos deben ser categóricos.

Reglas de Asociación

Id	Atributo 1	Atributo 2	...	Atributo n
1	10	alto		35
2	35	bajo		54
3	43	medio		28
4	26	bajo		65
5	87	alto		32
6	45	alto		29
7	76	bajo		55
8	5	medio		46
9	12	medio		43
10	54	bajo		27



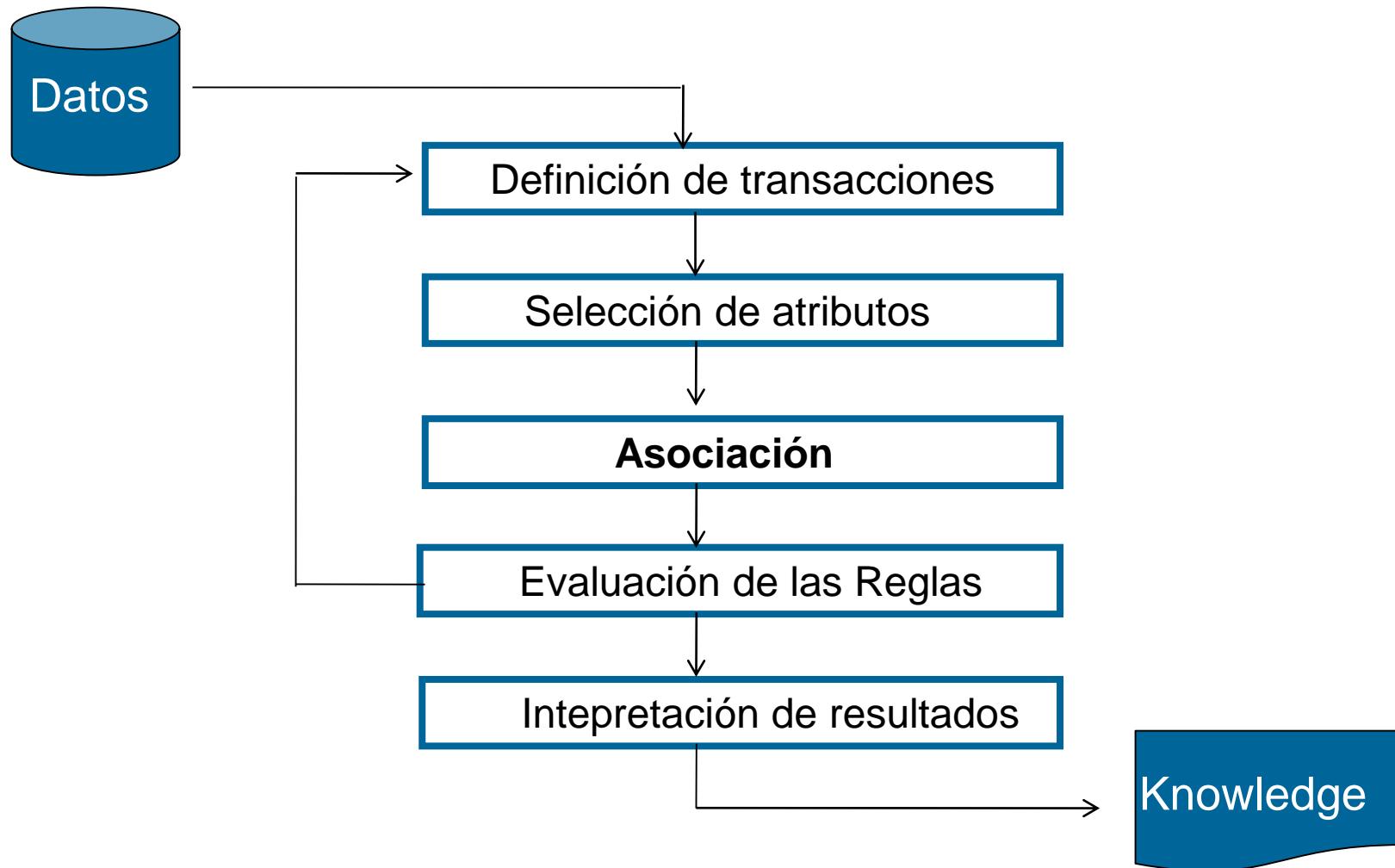
Transacciones

TID	Artículos
1	Pan, leche, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, cerveza
4	Pan, leche, pañales, cerveza
5	Pan, leche, huevos, cerveza



$\{\text{pañales}\} \rightarrow \{\text{cerveza}\}$
 $\{\text{cerveza}\} \rightarrow \{\text{pañales}\}$
 $\{\text{pan, leche}\} \rightarrow \{\text{huevos}\}$
 $\{\text{pan}\} \rightarrow \{\text{leche, huevos}\}$

Ciclo de Vida de las Reglas Asociación



Evaluación de las Reglas de Asociación

- **Soporte de la Regla:** Fracción de las transacciones que contiene los elementos o ítems de la regla:

$$\text{supp } (X \rightarrow Y) = \text{supp } (X \cup Y)$$

- **Confianza de la Regla:** Fracción de las transacciones de X en las que aparece Y:

$$\text{conf } (X \rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)}$$

Se buscan reglas con **supp>= Min_supp** y **conf>= Min_conf**

Evaluación de las Reglas de Asociación

$$\begin{aligned} \text{supp}(\{\text{cerveza}\} \rightarrow \{\text{pañales}\}) \\ = \text{supp}(\{\text{pañales}, \text{cerveza}\}) \\ = 3/5 = 0.6 = 60\% \end{aligned}$$

Tid	Artículos
1	Pan, leche, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, cerveza
4	Pan, leche, pañales, cerveza
5	Pan, leche, huevos, cerveza

$$\begin{aligned} \text{conf } (\{\text{cerveza}\} \rightarrow \{\text{pañales}\}) \\ = 3/4 = 0.75 = 75\% \end{aligned}$$

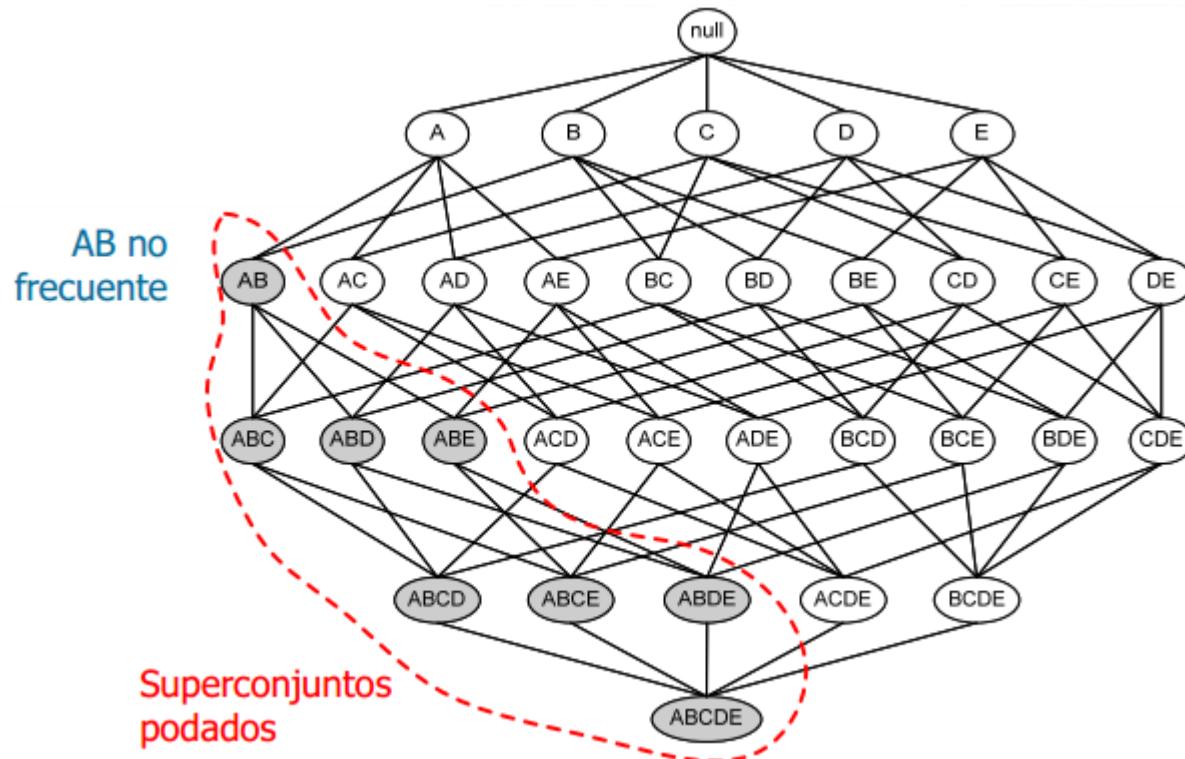
Tid	Artículos
1	Pan, leche, huevos
2	Pan, pañales, cerveza
3	Leche, pañales, cerveza
4	Pan, leche, pañales, cerveza
5	Pan, leche, huevos, cerveza

Algoritmos para Reglas de Asociación

- Apriori [Agrawal et al., 94]
- GSP (Generalized Sequential Patterns) – modificación a Apriori
- SPADE – modificación a Apriori

Algoritmos para Reglas de Asociación

- Apriori [Agrawal et al., 94]



Selección de Factores

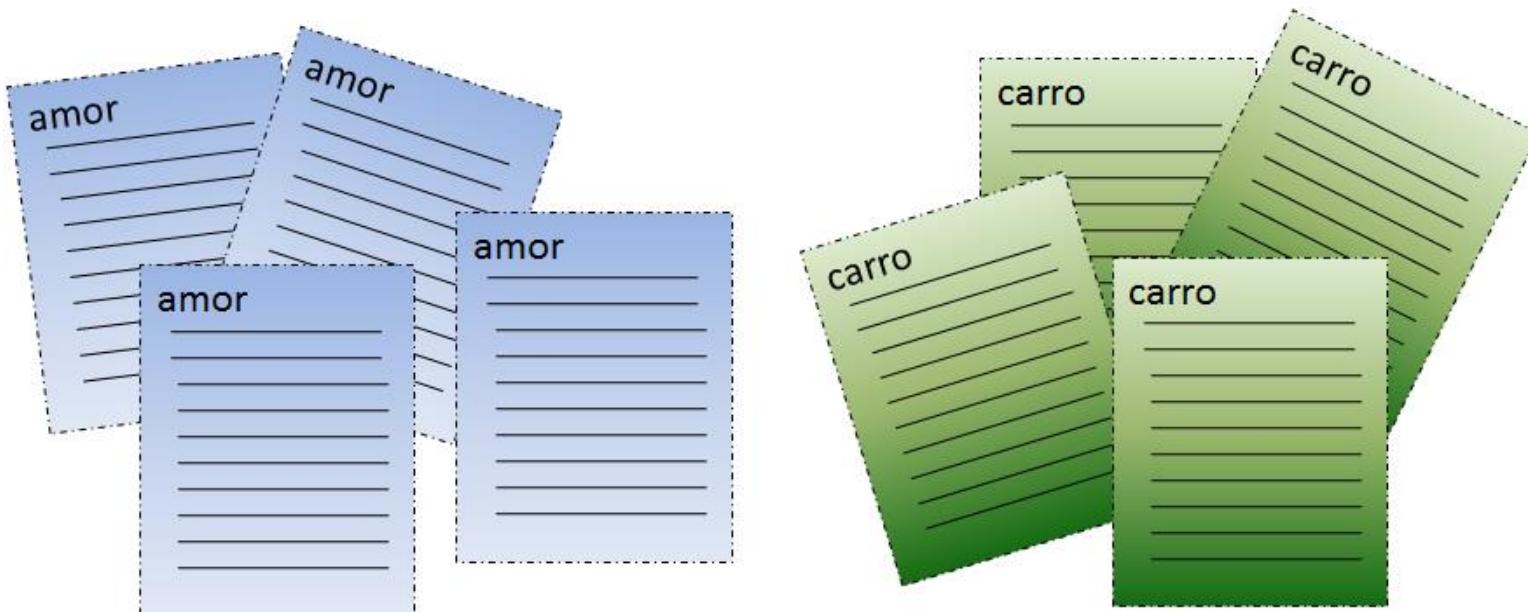
- Análisis de Componentes Principales
- Árboles de Decisión
- Análisis de Correlaciones
- Reglas de Asociación
- Regresión

AGENDA



1. Introducción
2. Metodologías
3. Herramientas
4. Análisis de Casos
5. Preparación de Datos
6. Técnicas y Algoritmos
- 7. Minería de Texto**
8. Bibliografía

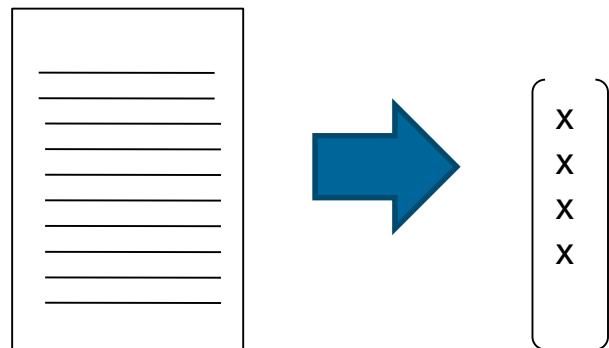
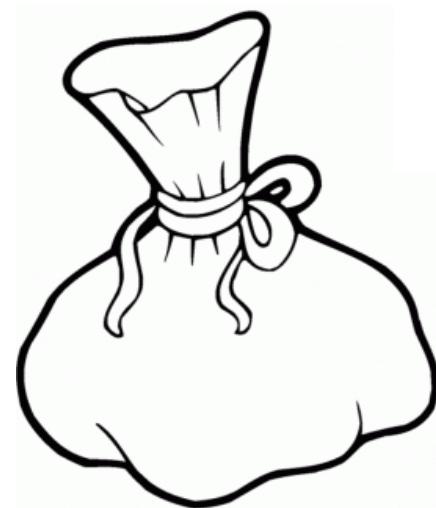
Minería de Texto



Indexamiento de Texto

Bolsa de Palabras:

- (1) Eliminar stopwords.
- (2) Reducir las palabras a las raíces.
- (3) Calcular vector de características.

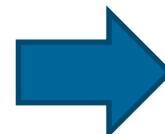


Indexamiento de Texto

Bolsa de Palabras:

(1) Eliminar stopwords.

Doc 1: La casa es bonita
Doc 2: La niña es bonita
Doc 3: La casota es bonita y grande
Doc 4: Es una bonita niñita



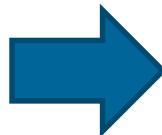
Doc 1: casa bonita
Doc 2: niña bonita
Doc 3: casota bonita grande
Doc 4: bonita niñita

Indexamiento de Texto

Bolsa de Palabras:

- (1) Eliminar stopwords.
- (2) Reducir las palabras a las raíces.

Doc 1: casa bonita
Doc 2: niña bonita
Doc 3: casota bonita grande
Doc 4: bonita niñita

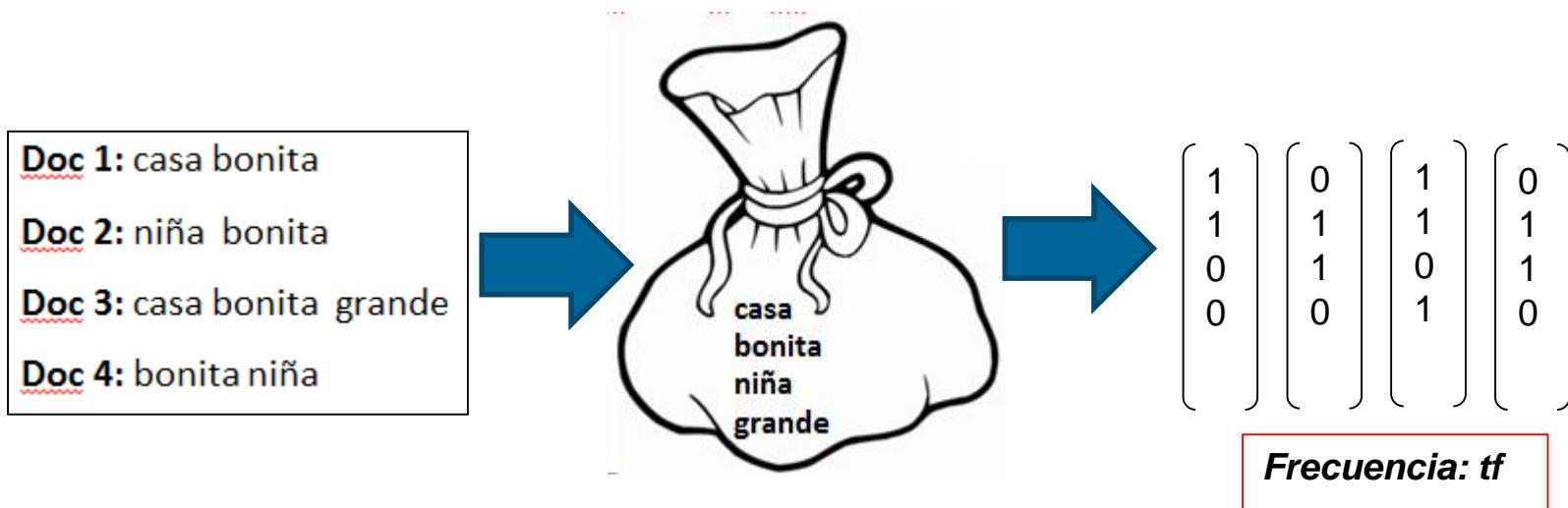


Doc 1: casa bonita
Doc 2: niña bonita
Doc 3: casa bonita grande
Doc 4: bonita niña

Indexamiento de Texto

Bolsa de Palabras:

- (1) Eliminar stopwords.
- (2) Reducir las palabras a las raíces.
- (3) Calcular vector de características



Indexamiento de Texto

Bolsa de Palabras:

- (1) Eliminar stopwords.
- (2) Reducir las palabras a las raíces.
- (3) Calcular vector de características

$$idf(t_i) = \log\left(\frac{D}{df(t_i)}\right)$$

Doc 1: casa bonita

Doc 2: niña bonita

Doc 3: casa bonita grande

Doc 4: bonita niña

$$idf(casa) = \log\left(\frac{4}{2}\right) = 0.3$$

$$idf(bonita) = \log\left(\frac{4}{4}\right) = 0$$

$$idf(niña) = \log\left(\frac{4}{2}\right) = 0.3$$

$$idf(grande) = \log\left(\frac{4}{1}\right) = 0.6$$



$$\begin{pmatrix} 0.3 \\ 0 \\ 0.3 \\ 0.6 \end{pmatrix}$$

idf

Indexamiento de Texto

Bolsa de Palabras:

- (1) Eliminar stopwords.
- (2) Reducir las palabras a las raíces.
- (3) Calcular vector de características

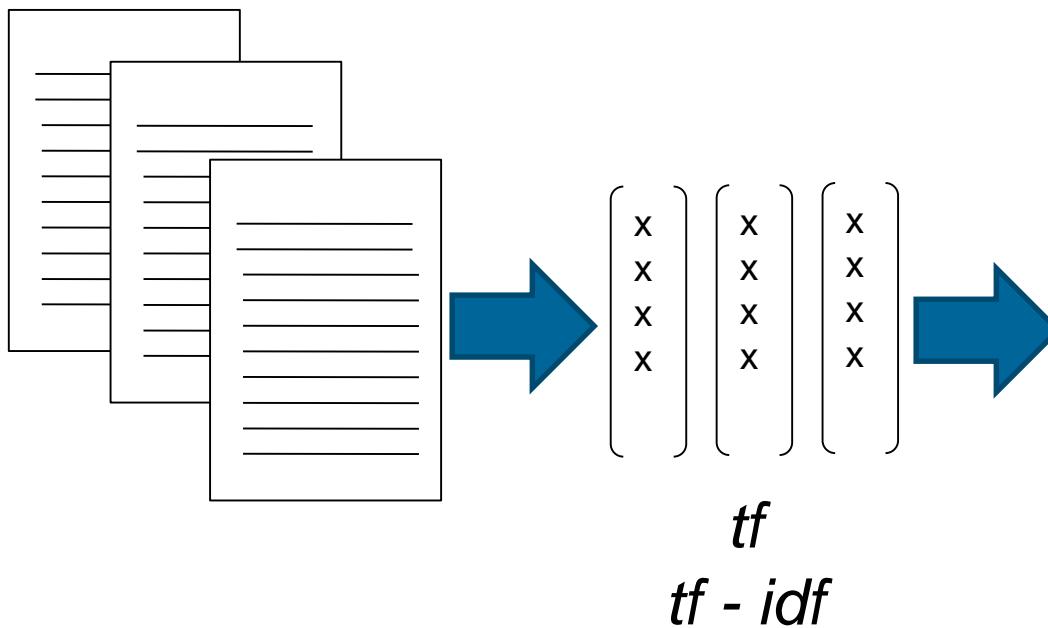
Doc 1: casa bonita
Doc 2: niña bonita
Doc 3: casa bonita grande
Doc 4: bonita niña



$$\begin{pmatrix} 0.3 \\ 0 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0.3 \\ 0 \end{pmatrix} \begin{pmatrix} 0.3 \\ 0 \\ 0 \\ 0.6 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0.3 \\ 0 \end{pmatrix}$$

$$Tf * idf$$

Minería de Texto



Técnicas Supervisadas

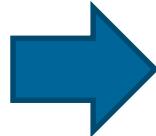
- Redes Neuronales
- Reglas de Decisión
- Árboles de Decisión
- Métodos Probabilísticos
- Máq. de Soporte Vectorial
- Métodos de Regresión
- Modelos Ocultos de Markov
- Métodos basados en Ejemplos

Técnicas NO Supervisadas

- Métodos Jerárquicos
- Métodos Particionales
- Redes Neuronales
- Métodos Probabilísticos
- Métodos Difusos
- Métodos Evolutivos
- Métodos basados en Kernel
- Métodos de reglas

Aplicaciones de la Minería de Texto

Análisis Predictivo



Comentario	Percepción
○ "Que capaciten mas a la señorita de la ventanilla (4) es totalmente despota e indiferente"	
○ "Evaluar a su personal, todos no están en la capacidad de atención al cliente"	
○ "Por lo pronto sigan el mismo servicio que emplean"	
○ "Me voy satisfecho de la atención siempre fue rápida las soluciones"	
○ "Deben implementar otro stand donde sacar ticket debido a que solo cuenta con 1 y es incomodo"	
○ "Con total conformidad"	

Análisis Descriptivo



Minería de Texto con Weka

- Enfoque: Bolsa de Palabras
- Se debe tener una lista de stopwords (palabras vacías) en el idioma del texto.
- Se debe configurar el algoritmo de reducción de raíces (stemming). Para esto se debe instalar el paquete “snowball stemmer”.

Minería de Texto con Weka: Clustering

File: **simple.arff**

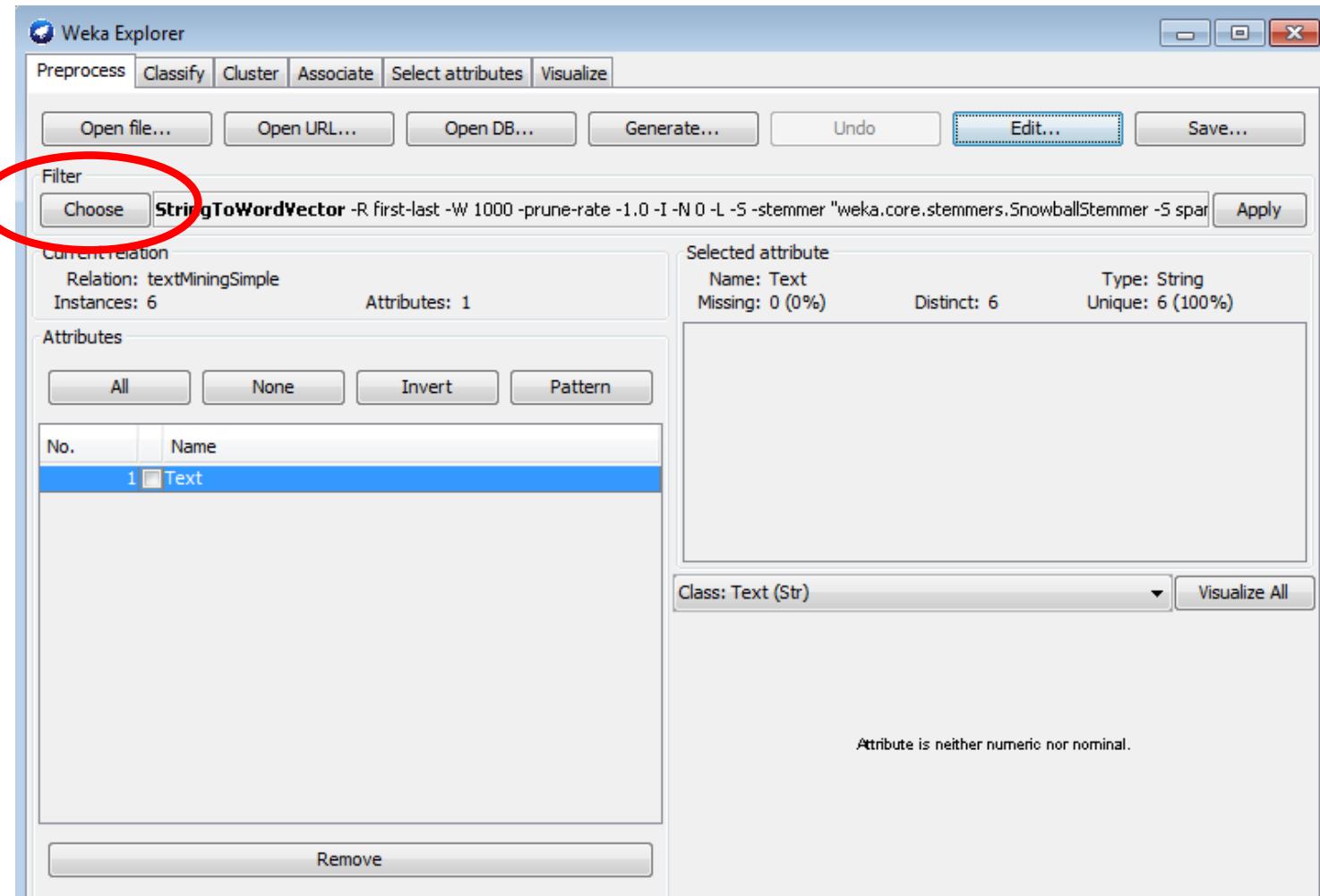
```
@relation simple
```

```
@attribute Text string
```

```
@data
```

```
'La casa es bonita'  
'La niña es bonita'  
'Es bonita la casa'  
'Es un bonito niño'  
'Una vieja en pelota'  
'una Pelota vieja'
```

Minería de Texto con Weka: Clustering



`weka.filters.unsupervised.attribute.StringToWordVector`

About

Converts String attributes into a set of attributes representing word occurrence (depending on the tokenizer) information from the text contained in the strings.

Mo

Capabilità

IDFTransform True

TFTransform False

attributeIndices 1

attributeNamePrefix

debug True

doNotCheckCapabilities False

rateOnPerClassBasis False

invertSelection False

CaseTokens True

normalizeDeclLength No normalization

outputWordCounts False

periodicPruning -1.0

saveDictionaryInBinaryForm False

stemmer Choose SnowballStemmer -S spanish

wordsToKeep 1000

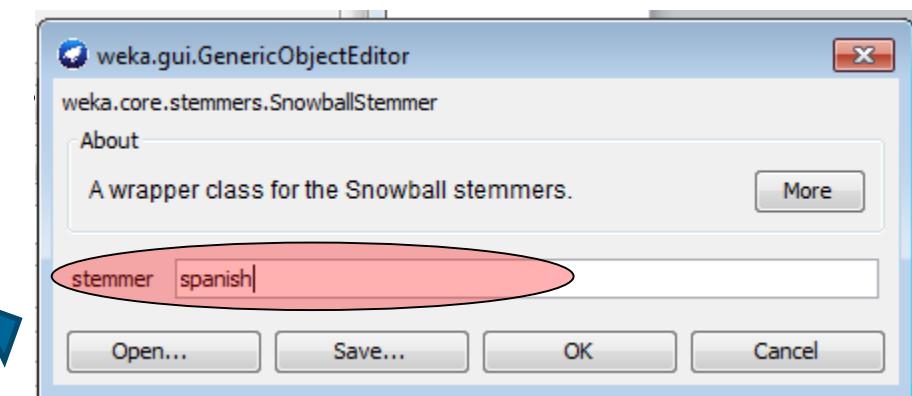
[Open](#)

[Save](#)

oh

Can

Stemmers/SnowballStemmer



Minería de Texto con Weka: Clustering

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter Choose **StringToWordVector -R first-last -W 1000 -prune-rate -1.0 -I -N 0 -L -5 -stemmer "weka.core.stemmers.SnowballStemmer -S spar** Apply

Current relation Relation: textMiningSimple Instances: 6 Attributes: 1

Selected attribute Name: Text Type: String Missing: 0 (0%) Distinct: 6 Unique: 6 (100%)

Attributes

All None Invert Pattern

No.	Name
1	Text

Class: Text (Str) Visualize All

Attribute is neither numeric nor nominal.

Remove

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter Choose **StringToWordVector -R first-last -W 1000 -prune-rate -1.0 -I -N 0 -L -5 -stemmer "weka.core.stemmers.SnowballStemmer -S spar** Apply

Current relation Relation: textMiningSimple-weka.filters.unsupervised.attribute.String... Instances: 6 Attributes: 5

Selected attribute Name: bonit Type: Numeric Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Attributes

All None Invert Pattern

No.	Name
1	bonit
2	cas
3	niñ
4	pelot
5	viej

Class: viej (Num) Visualize All

Remove

Status OK

No.	Text
1	La casa es bonita
2	La niña es bonita
3	Es bonita la casa
4	Es un bonito niño
5	Una vieja en pelota
6	una Pelota vieja

No.	bonit Numeric	cas Numeric	niñ Numeric	pelot Numeric	viej Numeric
1	0.405...	1.098...	0.0	0.0	0.0
2	0.405...	0.0	1.098...	0.0	0.0
3	0.405...	1.098...	0.0	0.0	0.0
4	0.405...	0.0	1.098...	0.0	0.0
5	0.0	0.0	0.0	1.098...	1.098...
6	0.0	0.0	0.0	1.098...	1.098...

Minería de Texto con Weka: Clustering

Simple K-means

Cluster centroids:

Attribute	Full Data	Cluster#		
		0	1	2
		(6)	(2)	(2)
<hr/>				
bonit	0.2703	0.4055	0.4055	0
cas	0.3662	0	1.0986	0
niñ	0.3662	1.0986	0	0
pelot	0.3662	0	0	1.0986
viej	0.3662	0	0	1.0986

Minería de Texto con Weka: Clasificación

File: **sentimientos.arff**

@relation sentimientos

@attribute Text string

@attribute sentimiento {positivo, negativo}

@data

'es el mejor profesor del mundo, es muy muy bueno',positivo

'muy regular el profesor',negativo

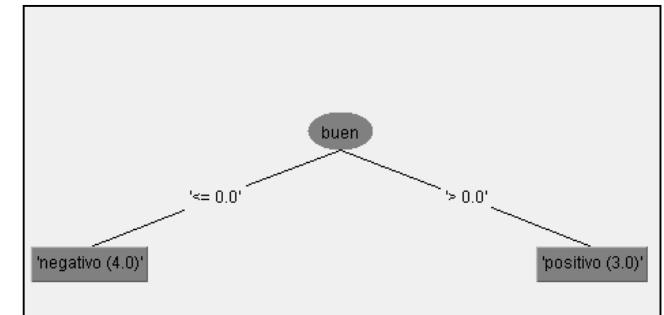
'El profesor no hace nada',negativo

'muy bueno, ha mejorado muchas cosas el profesor',positivo

'El profesor, peor es nada',negativo

'Cuando terminara ese profesor',negativo

'es bueno el profesor, me enseña mucho',positivo



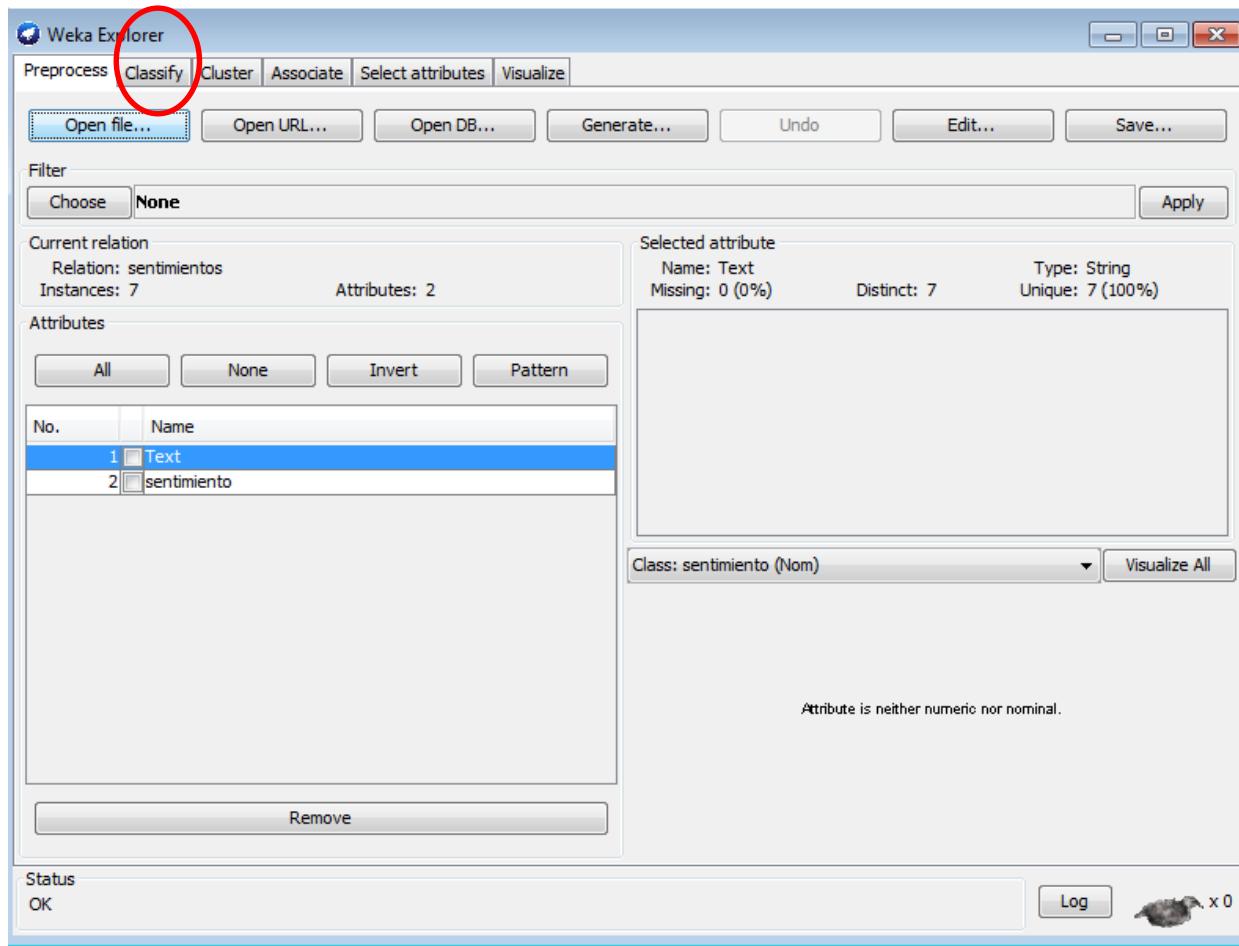
Minería de Texto con Weka: Clasificación

File: **sentimientos.arff**

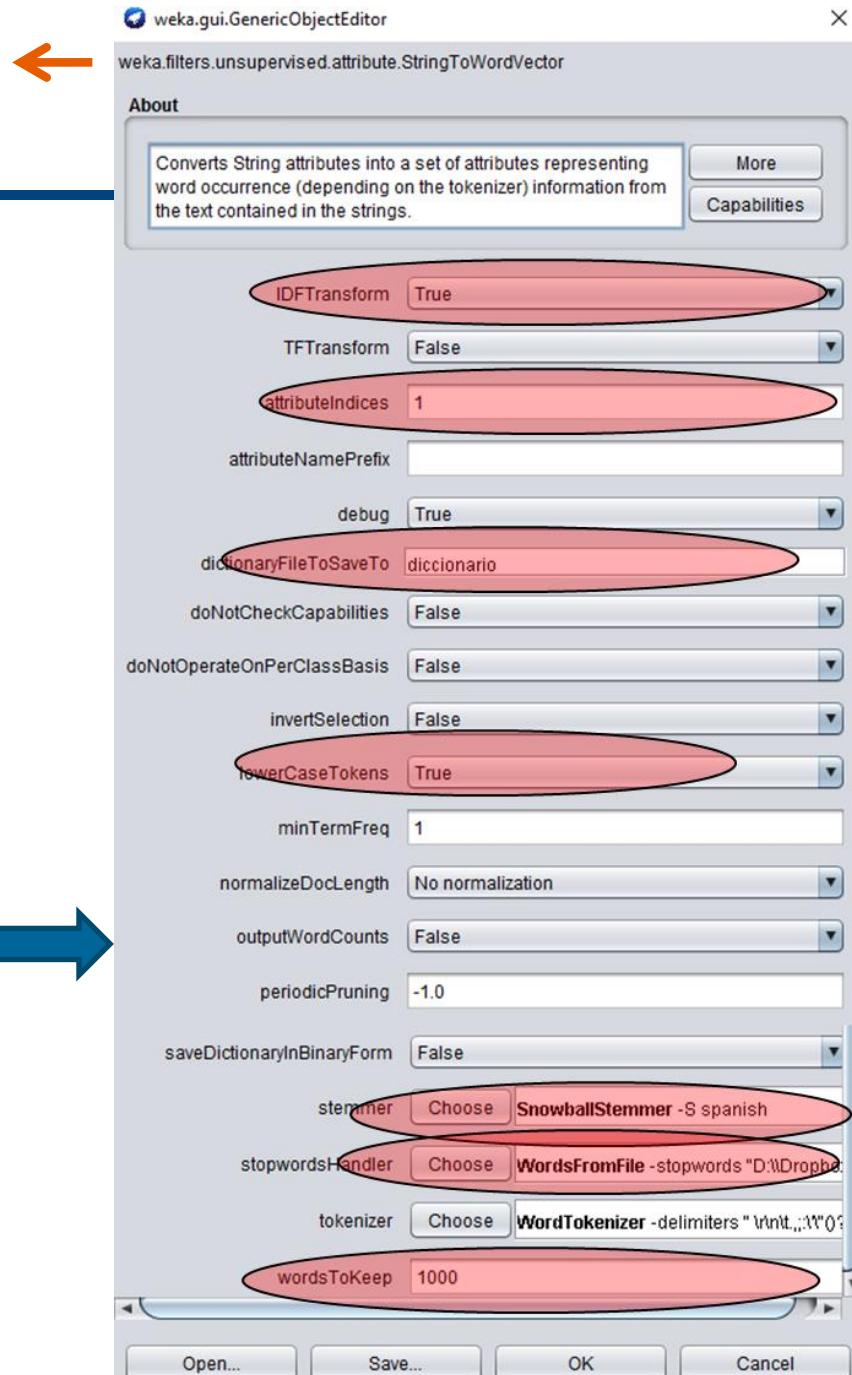
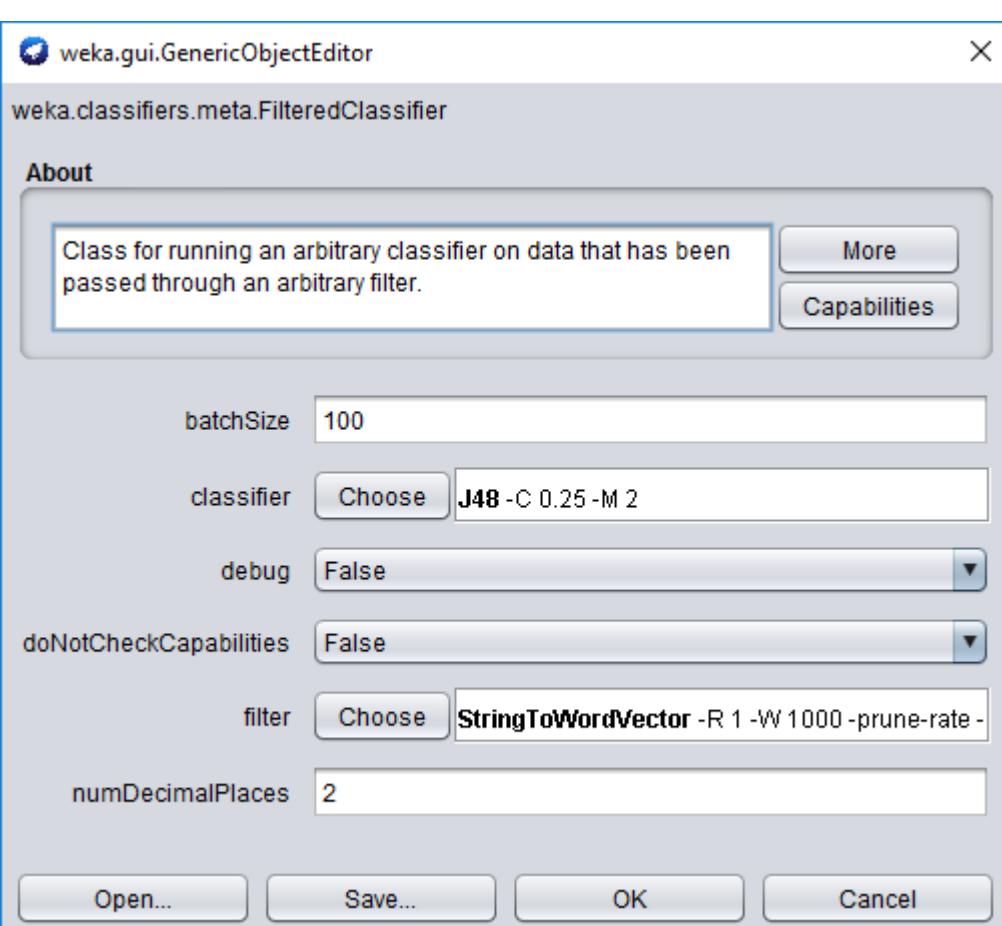
```
@relation sentimientosTest  
  
@attribute Text string  
@attribute sentimiento {positivo, negativo}  
  
@data  
'esa profesora es buena en lo que hace',?  
'es la peor profesora que he tenido',?
```

Minería de Texto con Weka: Clasificación

File: **sentimientos.arff**



Filters/unsupervised/attribute/StringToWordVector



AGENDA



1. Introducción
2. Metodologías
3. Herramientas
4. Análisis de Casos
5. Preparación de Datos
6. Técnicas y Algoritmos
7. Minería de Texto
- 8. Bibliografía**

Referencias

- [Abraham, 2003] ABRAHAM A. "i-Miner: A Web Usage Mining Framework Using Hierarchical Intelligent Systems". En: The IEEE Int. Conf. on Fuzzy Systems, FUZZ-IEEE, 2003.
- [Apte, 1997] APTE C. y WEISS S. "Data Mining with Decision Trees and Decision Rules". En: Future Generation Computer Systems, 1997.
- [Bennett et. al., 2005] BENNETT P., DUMAIS S. y HORVITZ E. "The combination of text classifiers using reliability indicators". En: Information Retrieval, 8, p. 67-100, 2005.
- [Bergman, 2001] BERGMAN M. "The Deep Web: Surfacing Hidden Value". En: Journal of Electronic Publishing from the University of Michigan, Julio 2001.
- [Breiman, 1996] BREIMAN L. "Bagging predictors". Machine Learning, p. 123-140, 1996.
- [Carr et al., 1999] CARR L., HALL W. y DE ROURE D. "The Evolution of Hypertext Link Services". En: ACM Computing Surveys, Vol. 31, No 4es, 1999.
- [Cohen, 1999] COHEN W. y SINGER Y. "Context sensitive learning methods for text categorization". En: ACM Trans. Inform. Systems, vol 17, No 2, pag 141 - 173, 1999.
- [Chakrabarti et al., 1998] CHAKRABARTI S., DOM B. y INDYK P. "Enhanced hypertext categorization using hyperlinks". En: Proceedings of SIGMOD-98, ACM International Conference on Managementof Data (Seattle, US), p. 307–318, 1998.
- [Chakrabarti, 2000] CHAKRABARTI S. "Data mining for hypertext: a tutorial survey".En: ACM SIGKDD Explorations Newsletter, enero 2000.
- [Chang et al., 2006] CHANG F., LIN C. y LU C. "Adaptive Prototype Learning Algorithms". En: Journal of Machine Learning Research, vol. 7, pp. 2125-2148, October, 2006.
- [Demsar, 2006] DEMSAR J. "Statistical Comparisons of Classifiers over Multiple Data Sets". En: Journal of Machine Learning Research, enero 2006.
- [Dietterich, 2000] DIETTERICH T. "Ensemble Methods in Machine Learning". En: First International Workshop on Multiple Classifier Systems, 2000.
- [Elsas, 2004] ELSAS J. y EFRON M. "Html tag based metrics for use in web page type classification". Submitted to ASIST Annual Meeting, Providence, USA, 2004

Referencias

- [Fang et al., 2006] FANG r., MIKROYANNIDIS A. y THEODOULIDIS B. “A voting for the Classification of Web Pages”. En: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Intelligence and Intelligent Agent Technology, 2006.
- [Feng et al., 2004] FENG H., SHI R. y CHUA T. “A bootstrapping framework for annotating and retrieving WWW images”. En: Proceedings of the 12th Annual ACM international Conference on Multimedia (New York, NY, USA, October 10 - 16), 2004.
- [Fradkin, 2005] FRADKIN D. y KANTOR P. “Methods for Learning Classifier Combinations: No Clear Winner”. En: ACM Symposium on Applied Computing, 2005.
- [Frasconi et al., 2002] FRASCONI P., SODA G. y VULLO A. “HMM for Text Categorization in Multi-Page Documents”. En: Journal of Intelligent Systems, vol 18, 2002.
- [Glover et al., 2002] GLOVER E., TSIOUTSIOULIKLIS K., LAWRENCE S., PENNOCK D. y GARY W. “Using web structure for classifying and describing web pages”. En: Proceedings of the Eleventh International World Wide Web Conference, p. 562-569, 2002.
- [Holden, 2004] HOLDEN N. y FREITAS A. “Web Page Classification with an Ant Colony Algorithm”. En: PPSN VIII, 8th International Conference on Parallel Problem Solving from Nature , septiembre 2004.
- [Hornegger et al., 1996] HORNEGGER J., NOTH E., FISCHER V., y NIEMANN H. “Semantic network meet Bayesian classifiers”. En: B. Jahne, P. Geiler, H. Hauecker, and F. Hering, editors, Mustererkennung, 1996.
- [Hunt, 1993] HUNT L. “The invention of pornography”. Zone Books. New York, 1993. ISBN 094229968X.
- [Hunter, 1999] HUNTER, C. “Filtering The Future? : Software Filters, Porn, Pics, and the Internet Content Conundrum”. Tesis en Artes en la Universidad de Pennsylvania, 1999.
- [ICBF, 2004]. Instituto Colombiano de Bienestar Familiar. “Criterios de Clasificación de Páginas en Internet con Contenidos de Pornografía Infantil”, 2004.

Referencias

- [Joachims, 1998] JOACHIMS T. “Text categorization with support vector machines: learning with many relevant features”. En: Proceedings of ECML-98, 10th European Conference on Machine Learning, 1998.
- [Joachims, 1999] JOACHIMS T. “Transductive Inference for Text Classification using Support Vector Machines”. En: International Conference on Machine Learning (ICML), 1999.
- [Kai, 2004] KAI L. y KARGER D. “Using URLs and Table Layout for Web Classification Tasks”. En: Proceedings International WWW Conference. New York, USA, 2004.
- [Kan, 2004] KAN M. “Web page categorization without the Web page”. En: Proceedings of the 13th international World Wide Web conference, p. 262-263, 2004.
- [Kester et al., 2003] KESTER H., RUSKIN D., LEE C. y ANDERSON M. “System and Method for adapting an Internet Filter”. Patente de Estados Unidos, junio, 2003.
- [Langford, 2005] LANGFORD J. “Tutorial on Practical Prediction Theory for Classification”. En: Journal of Machine Learning Research, marzo 2005.
- [Larkey, 1996] LARKEY L. y CROFT W. “Combining classifiers in text categorization”. En: Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval, p. 289-297, 1996.
- [Lewis, 1998] LEWIS D. “Naive (Bayes) at forty: The independence assumption in information retrieval”. En: Proceedings of ECML-98, 10 European Conf. on ML, 1998.
- [Li, 1998] LI Y. y JAIN A. “Classification of text documents”. En: The Computer Journal, p.41, 8, 537–546, 1998.
- [Lindemann, 2006] LINDEMANN C. y LITTIG L. “Coarse-grained Classification of Web Sites by Their Structural Properties”. En: Proceedings of the Eighth ACM international Workshop on Web information and Data Management, Virginia, USA, November, 2006.
- [Liu et al., 2006] LIU T., MOORE A. y GRAY A. “New Algorithms for Efficient High-Dimensional Nonparametric Classification”. En: Journal of M L Research, junio 2006.
- [Merz, 1999] MERZ C. “A Principal Component Approach to Combining Regression Estimates”. En: Machine Learning, 36, 9-32, 1999.

Referencias

- [Nelson, 1970] Nelson, T. "No More Teachers' Dirty Looks". Computer Decisions 9, 8, p.16-23, septiembre, 1970.
- [Nikunj, 2006] NIKUNJ C. "Ensemble Data Mining Methods". En: Encyclopedia of Data Warehousing and Mining, p. 448–453, Idea Group Reference, 2006.
- [Pant et al., 2004] PANT G., SRINIVASAN P. y MENCZER F. "Crawling the Web". En: Web Dynamics: Adapting to Change in Content, Size, Topology and Use, 2004.
- [Parmanto et al., 1996] PARMANTO B., MUNRO P. y DOYLE H. "Improving committee diagnosis with resampling techniques". En: Advances in Neural Information Processing Systems, p.882-888, 1996.
- [Peng et al., 2006] PENG T., HE F. y ZUO W. "Text Classification from Positive and Unlabeled Documents Based on GA". En: 7th International Conference on high performance computing for computational science, Rio de Janeiro, Brasil, 2006.
- [POESIA, 2004]. "A Public Open Source Environment for a Safer Internet Access". Disponible en medio electrónico: <http://www.poesia-filter.org/>, 2004.
- [Prakash, 2001] PRAKASH A. y RAVI K. "Web Page categorization Based on Document Structure". En: IEEE National Convention, Dec 2001.
- [Resnick, 1996] RESNICK, P. y MILLER J. "PICS: Internet Access Controls Without Censorship". En: Communications of the ACM, vol. 39(10), pp. 87-93, 1996.
- [Resnick et al., 2004] RESNICK P., HANSEN D. y RICHARDSON C. "Calculating Error Rates for Filtering Software". En: Comunications of the ACM, Volumen 47 No 9, Septiembre 2004.
- [Rousu et al., 2006] ROUSU J., SAUNDERS C., SZEDMAK S. y SHAWE-TAYLOR J. "Kernel-Based Learning of Hierarchical Multilabel Classification Models". En: Journal of Machine Learning Research, julio 2006.
- [Rowle et al., 2006] ROWLE H., JING Y. y BALUJA S. "Large Scale Image-Based Adult-Content Filtering". En: First International Conference on Computer Vision, 25-28; Portugal, febrero 2006.

Referencias

- [Sebastiani, 2002] SEBASTIANI F. "Machine Learning in Automated Text Categorization". En: ACM Computing Surveys, 1-47, marzo 2002.
- [Sebastiani, 2005] SEBASTIANI, F. "Text Categorization". En: Text Mining and its Applications to Intelligence, editor: Alessandro Zanasi, p 109-129, 2005.
- [Schapire, 1998] SCHAPIRE R. y SINGER Y. "Improved boosting algorithms using confidence-rated predictions". En: Proc. 11th Annu. Conf. on Comp. Learning Theory, p. 80-91, 1998.
- [Shen et al., 2000] SHEN H., OOI B. y TAN K. "Giving meaning to WWW images". En: ACM Multiemdia, LA, USA, p. 39-47, 2000.
- [Talavan, 2007] TALAVAN G. "Se supo: Todos miramos paginas prohibidas en horas de trabajo". Disponible en medio electrónico: http://www.minutouno.com/1/hoy/article/Se-supó:-todos-miramos-p%C3%A1ginas-prohibidas-en-horas-de-trabajo%5Eid_13529.htm, 2007.
- [Wanas et al., 2006] WANAS N., DARA R. y KAMEL M. "Adaptive fusion and co-operative training for classifier ensembles". En: Pattern Recogn. 39, 9, p. 1781-1794, septiembre, 2006.
- [Wettig et al., 2002] WETTIG H., GROUNWALD P., ROOS T., MYLLYMOAKI P. y TIRRI H. "On supervised learning of Bayesian network parameters". Technical Report HIIT-2002-1, Helsinki Institute for Information Technology (HIIT), 2002.
- [Wiener et al., 1995] WIENER E., PEDERSEN J. y WEIGEND A. "A neural network approach to topic spotting". En: Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval (Las Vegas, US), pp. 317–332., 1995.

Referencias

- [Xie et al., 2004] XIE Z., HSU W. y LI LEE M. “Mode committee: a novel ensemble method by clustering and local learning”. En: Tools with Artificial Intelligence, ICTAI 2004. 16th IEEE International Conference on Volume, Issue, 15-17 p. 628 – 633, Nov. 2004.
- [Yanai, 2003] YANAI K. “Generic image classificaiton using visual knowledge on the web”. En: ACM Multiemdia, Berkeley, USA. p. 167-176, 2003.
- [Yang, 1999] YANG Y. y LIU X. “A re-examination of text categorization methods”. En: Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval, 1999.
- [Yang et al., 2002] YANG Y., SLATTERY S. y GHANI R. “A study of approaches to hypertext categorization”. En: Journal of Intelligent Information Systems, vol. 18, pag.219-241, 2002.
- [Yang et al., 2003] YANG Y., ZHANG J. y KISIEL B. “A scalability analysis of classifiers in text categorization”. En: 26th ACM International Conference on Research and Development in Information Retrieval, ACM Press, New York, 2003.