

PROGRAMACIÓN SOBRE GRANDES VOLUMENES DE DATOS

Machine Learning

Magister - Efraín Alberto Oviedo
eaoc46@gmail.com

UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA

ESPECIALIZACIÓN EN ANALÍTICA Y CIENCIA DE DATOS



**UNIVERSIDAD
DE ANTIOQUIA**

1 8 0 3

AGENDA

1. Machine Learning

- 2. Preparación de datos
- 3. Análisis Predictivo
- 4. Análisis Descriptivo

Machine Learning

- El Machine Learning o aprendizaje automático es una disciplina orientada a crear **sistemas** que puedan **aprender por sí solos**, con el fin de **extraer información no trivial** de **grandes volúmenes de datos** por medio de la identificación de patrones complejos.
- Spark implementa el aprendizaje automático a través del módulo MLlib que cuenta con un gran número de algoritmos que permiten crear modelos para el aprendizaje automático.
- Pueden identificarse dos grandes ramas en el aprendizaje automático, a saber, el **aprendizaje supervisado** y el **aprendizaje NO supervisado**.

Tipos de Análisis

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

Análisis Predictivo

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

Análisis Descriptivo

Análisis Descriptivo



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

● Agrupamiento / Clustering

● Asociación

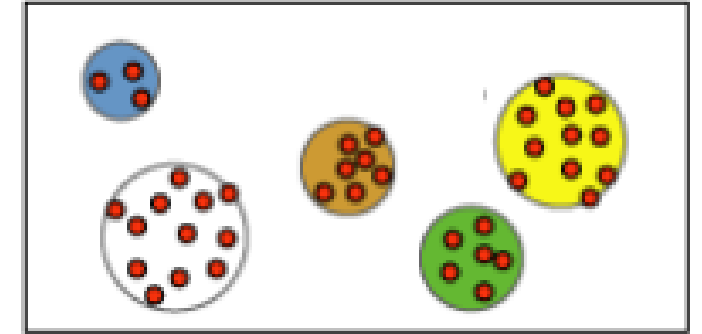
● Selección de Factores

Análisis Descriptivo

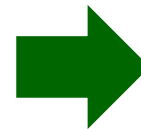
● Agrupamiento / Clustering

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

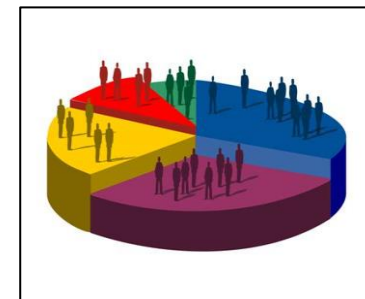
Ejemplos: Diseñar estrategias de mercadeo según el tipo de cliente. Detección de anomalías identificando datos que se alejen de los centroides de agrupación.



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

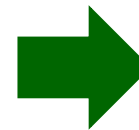


Asociación

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



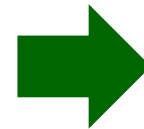
● Selección de Factores

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



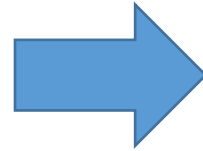
Factores seleccionados

Atributo 1
Atributo 4
Atributo 6

Técnicas

Análisis Predictivo

- Clasificación
- Regresión

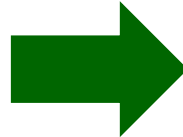


Técnicas Supervisadas

- 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

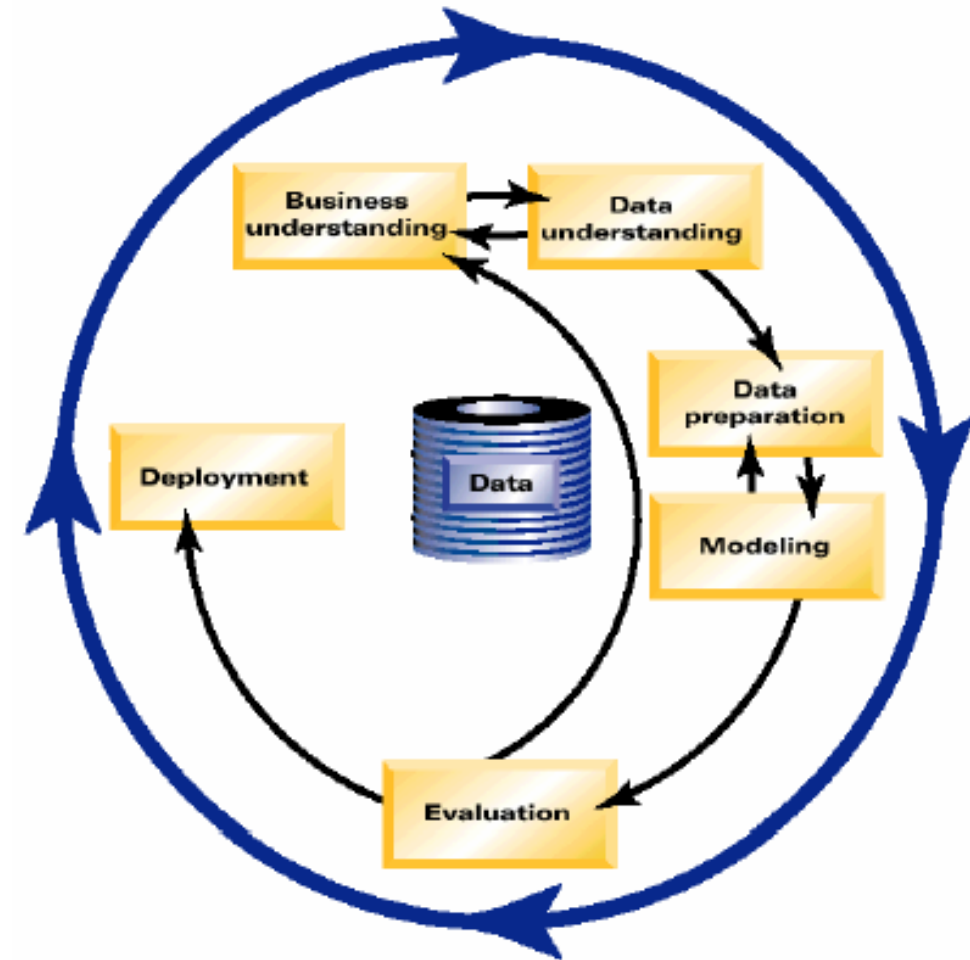
- Clustering
- Asociación
- Selección de Factores



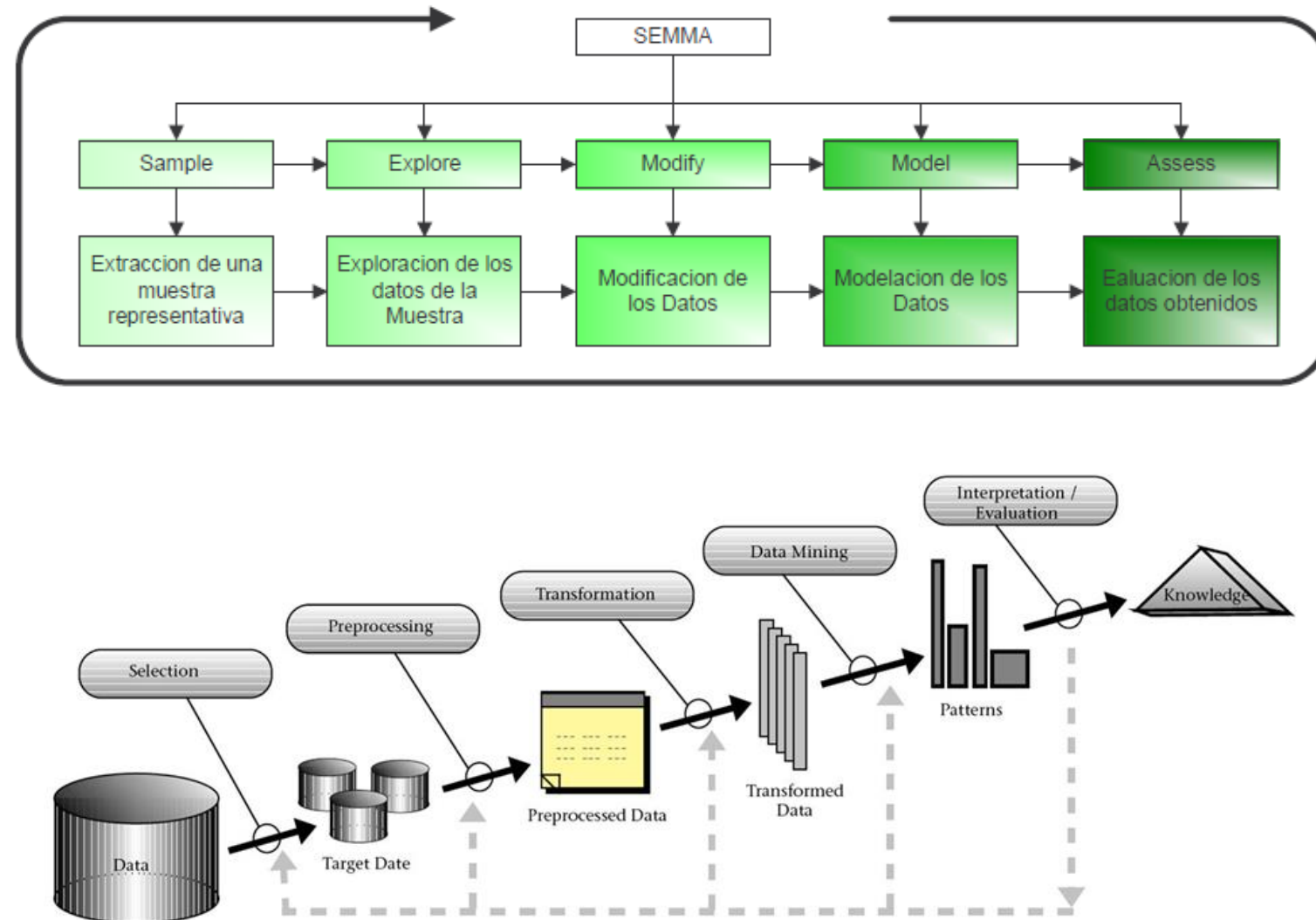
Técnicas NO Supervisadas

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

Metodologías



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



Metodologías

- CRISP-DM

- SEMMA

- KDD



AGENDA

1. Machine Learning
- 2. Preparación de datos**
3. Análisis Predictivo
4. Análisis Descriptivo

Tipos de Variables

● Variables Numéricas (cuantitativas)

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

● Variables Categóricas (cualitativas)

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

● Cadenas de Caracteres (string)

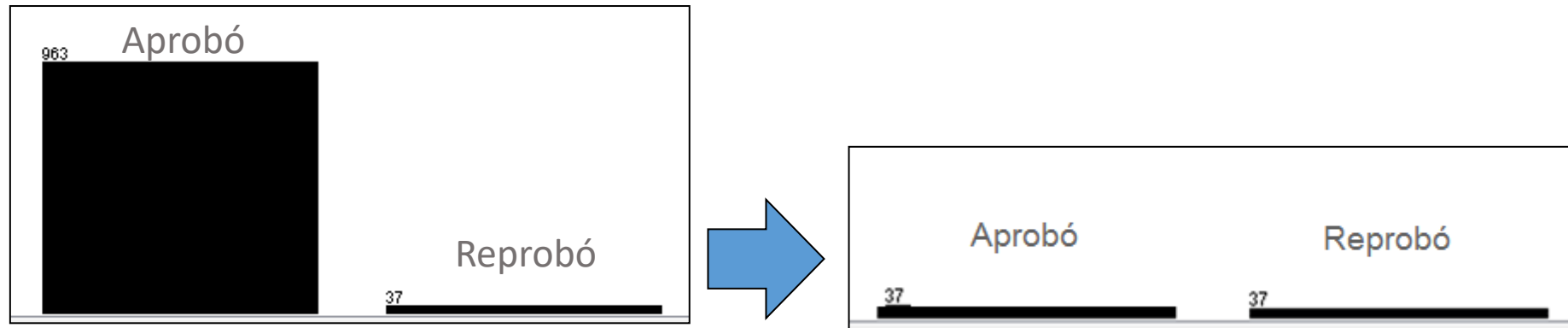
● Fechas (date)

Requisitos Mínimos

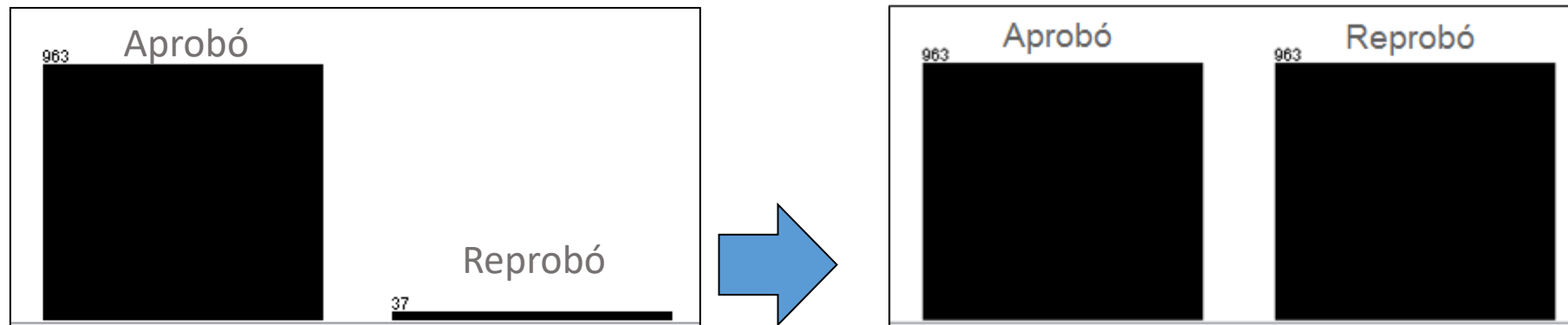
1. Identificación de variables
 - Propiedades del conjunto de datos
 - Tipos de datos (verificar carga correcta de datos)
2. Tratamiento de duplicados
 - Eliminar variables duplicadas (columnas)
 - Eliminar registros duplicados (filas)
 - Eliminar variables irrelevantes (ID, cedula, nombre, teléfono)
3. Análisis univariable
 - Variables numéricas: estadística descriptiva, histogramas, box plot
 - Variables categóricas: tabla de frecuencias y diagrama de barras
4. Análisis bivariable
 - Correlaciones entre las variables predictoras deben ser menores a 0.7
 - Correlaciones con la variable objetivo debe ser mayor a 0.3
5. Tratamiento de outliers (eliminar registros, eliminar variables, imputar, predecir)
6. Tratamiento de datos nulos (eliminar registros, eliminar variables, imputar o predecir)
7. Transformación de variables desde reglas del negocio
 - Discretización o Binning: convertir de número a categoría
 - Crear variables Dummy: convertir de categoría a número
8. Creación de variables (fecha y otros)
9. Reducción de variables (en caso de ser necesario)
10. Balanceo de la variable objetivo (sólo en clasificación)
11. Transformación de datos para el método

Balanceo de Datos (Clasificación)

- Selección aleatoria de datos



- Adicionar registros cercanos a la media de los datos



AGENDA

1. Machine Learning
2. Preparación de datos
- 3. Análisis Predictivo**
4. Análisis Descriptivo

Análisis Predictivo

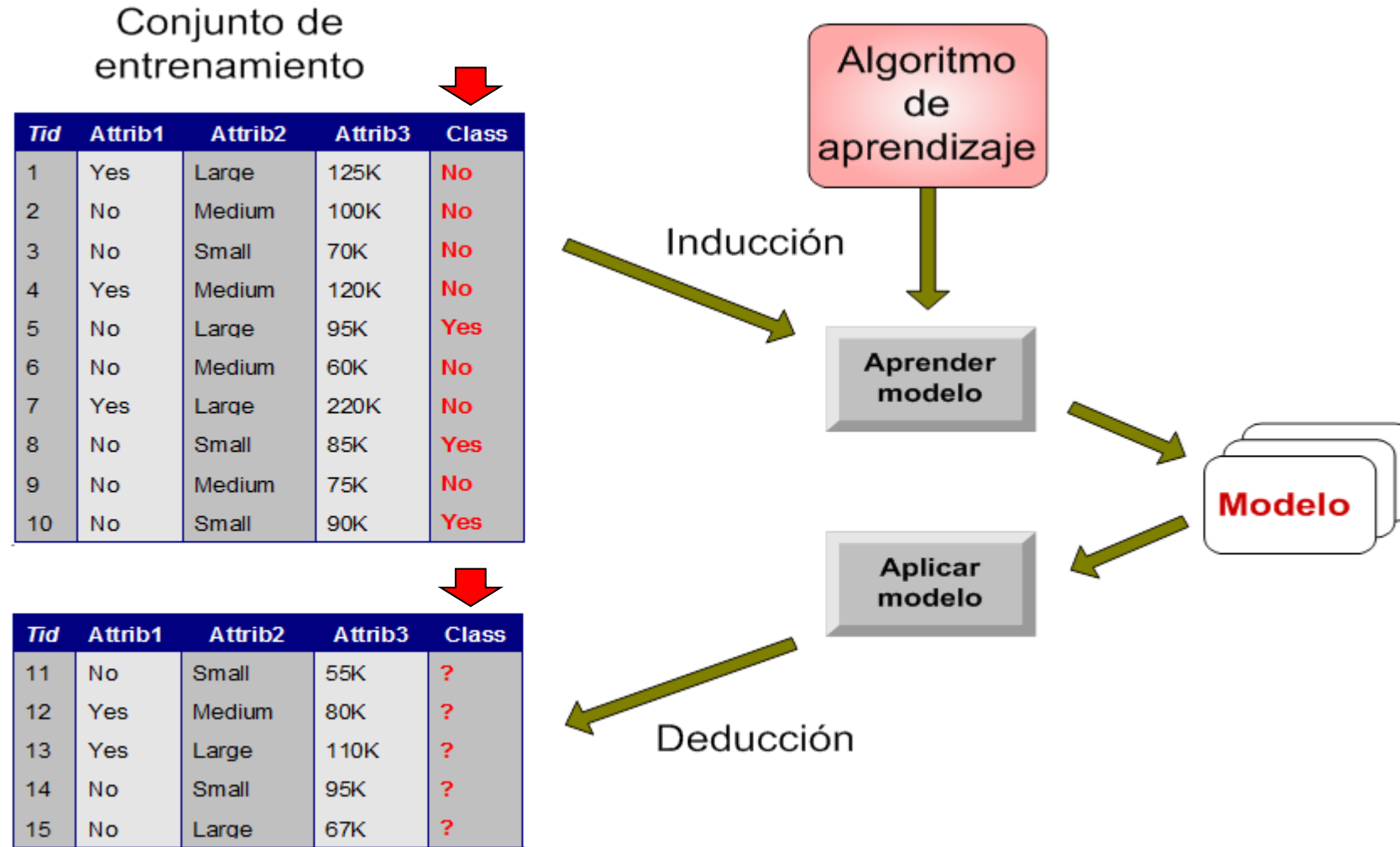
Análisis Predictivo



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

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

Clasificación




Conjunto de predicción (futuro)

Regresión

Conjunto de
entrenamiento

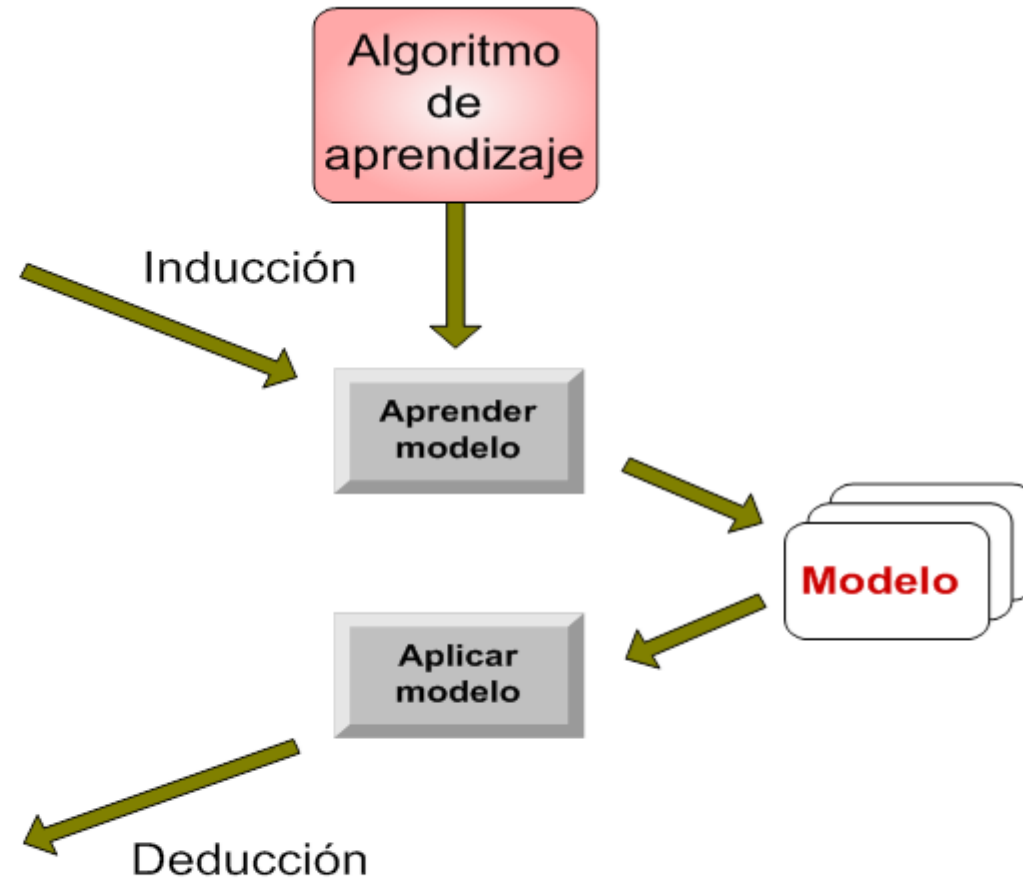


Tid	Attrib1	Attrib2	Attrib3	Pred.
1	Yes	Large	125K	78
2	No	Medium	100K	80
3	No	Small	70K	56
4	Yes	Medium	120K	40
5	No	Large	95K	39
6	No	Medium	60K	65
7	Yes	Large	220K	67
8	No	Small	85K	98
9	No	Medium	75K	76
10	No	Small	90K	45

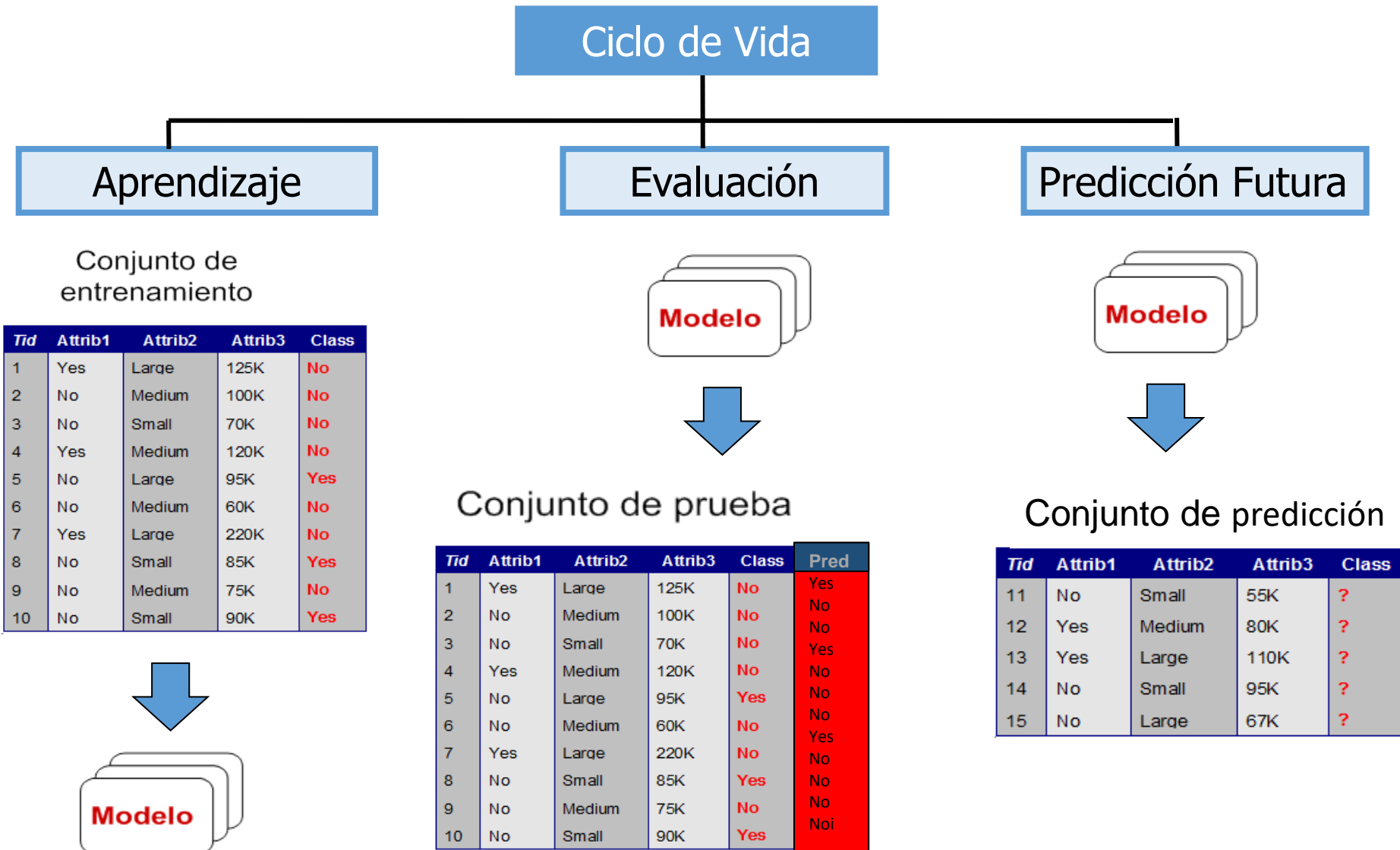


Tid	Attrib1	Attrib2	Attrib3	Pred.
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Conjunto de predicción (futuro)



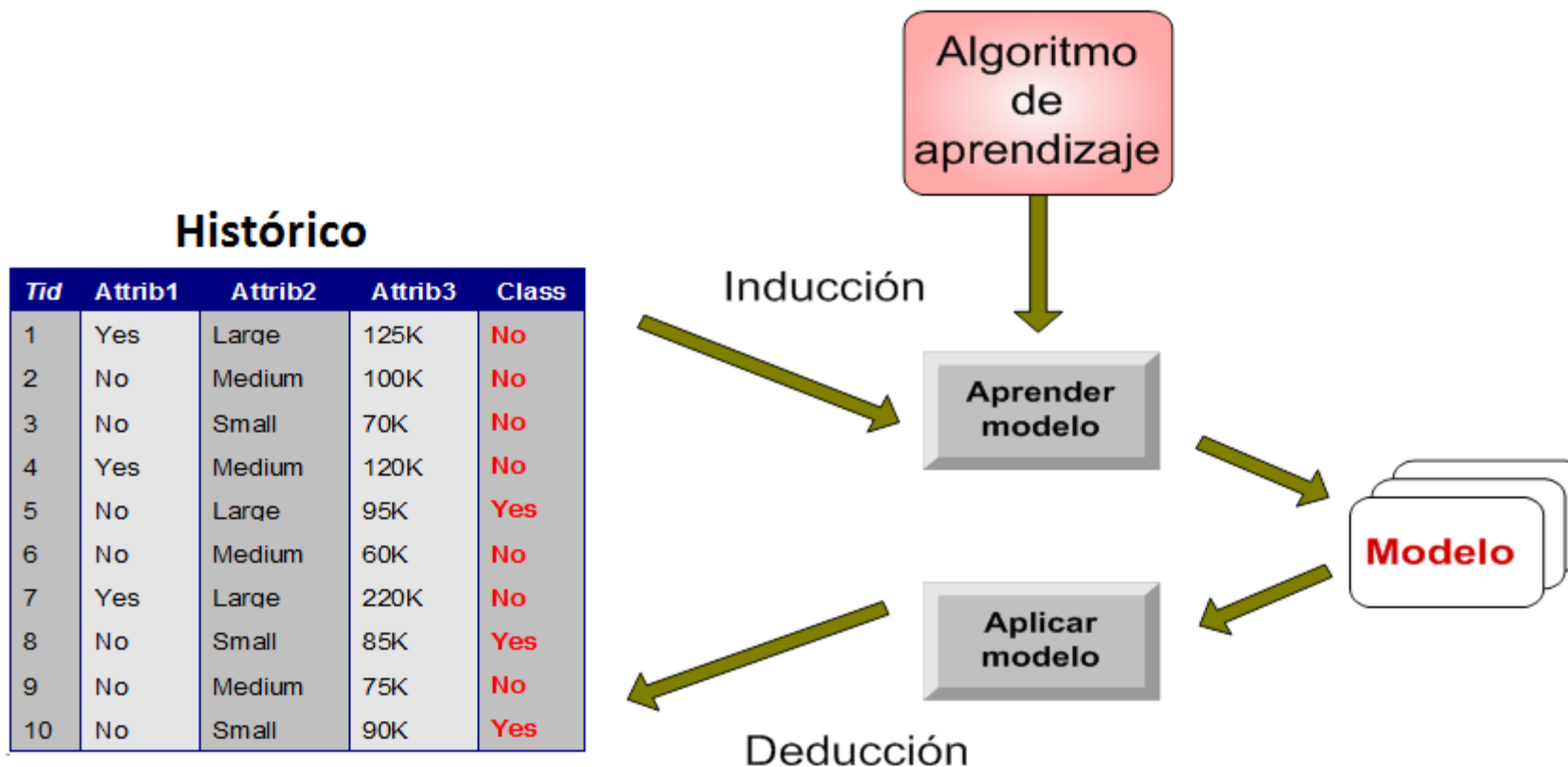
Ciclo de vida: Clasificación y Regresión



Cómo obtener los conjuntos de Entrenamiento y de Prueba

- Evaluar el conjunto de entrenamiento
- División de Datos 70-30 (Split)
- División de Datos 70-15-15
- Validación Cruzada (K-fold Cross Validation)

Evaluar el conjunto de Entrenamiento



Conjunto de
entrenamiento

Conjunto de prueba

División de Datos 70-30

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

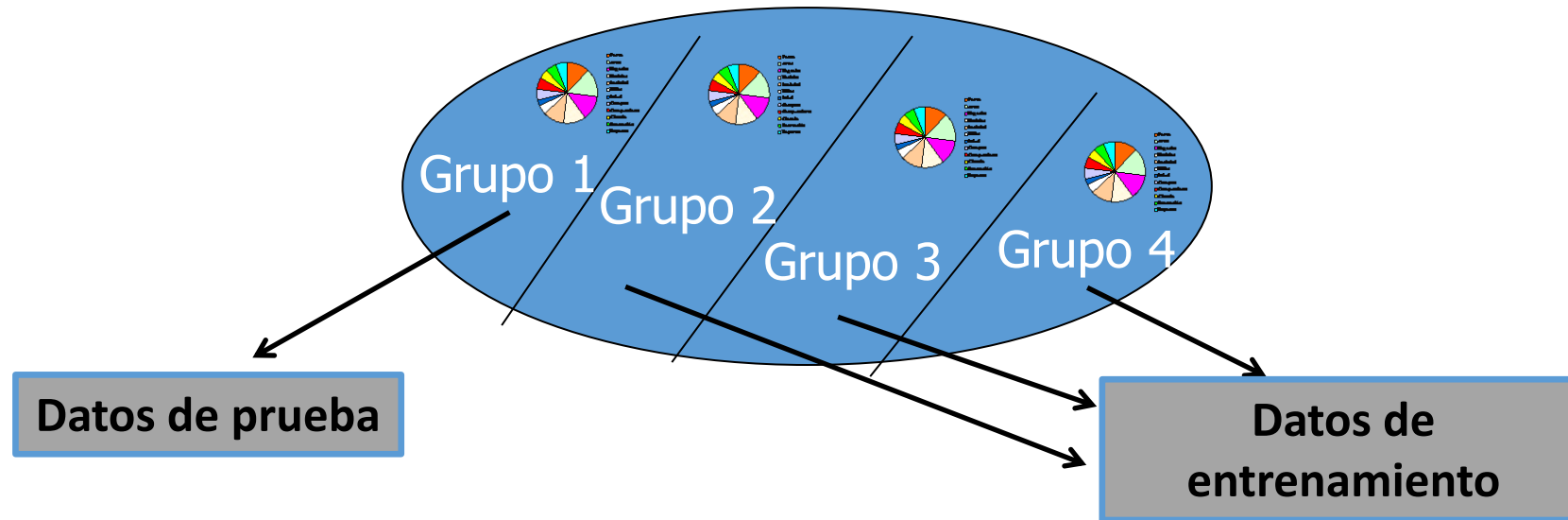
Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

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

Validación Cruzada (K-fold Cross Validation)



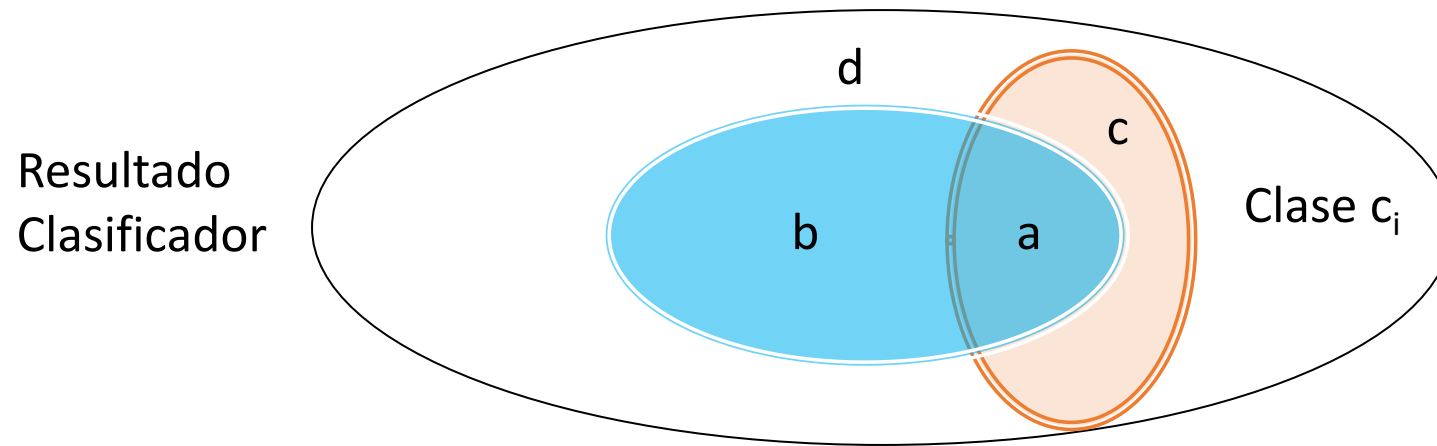
1. Aleatoriamente se divide el conjunto de datos en k subgrupos.
2. Se usan k-1 subgrupos en entrenamiento y el otro subgrupo en prueba.
3. Se repite el experimento k veces.

Medidas de Evaluación - Regresión

Mediciones de Error:

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

Medidas de Evaluación - Clasificador



Clase c_i : Clase REAL

Medidas de Evaluación - Clasificador

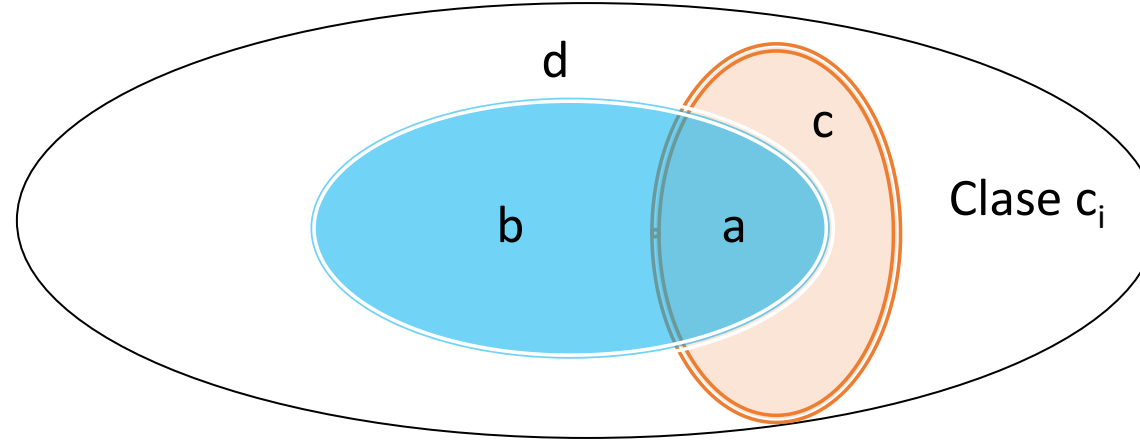


$$\text{Precision } p = \frac{a}{a+b}$$



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

Resultado
Clasificador



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

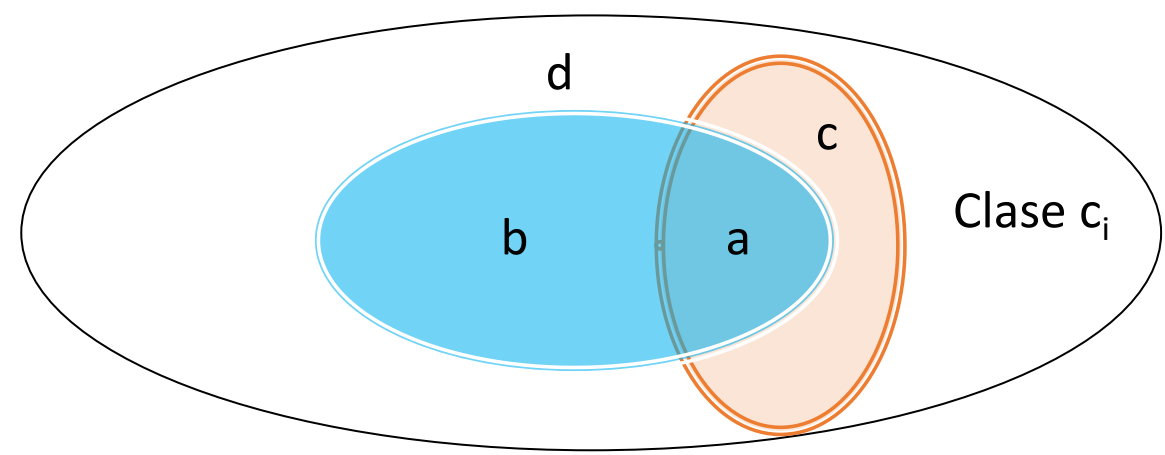


$$\text{Media Armónica } f1 = \frac{2pr}{p+r}$$

Medidas de Evaluación – Clasificador: 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

Resultado Clasificador



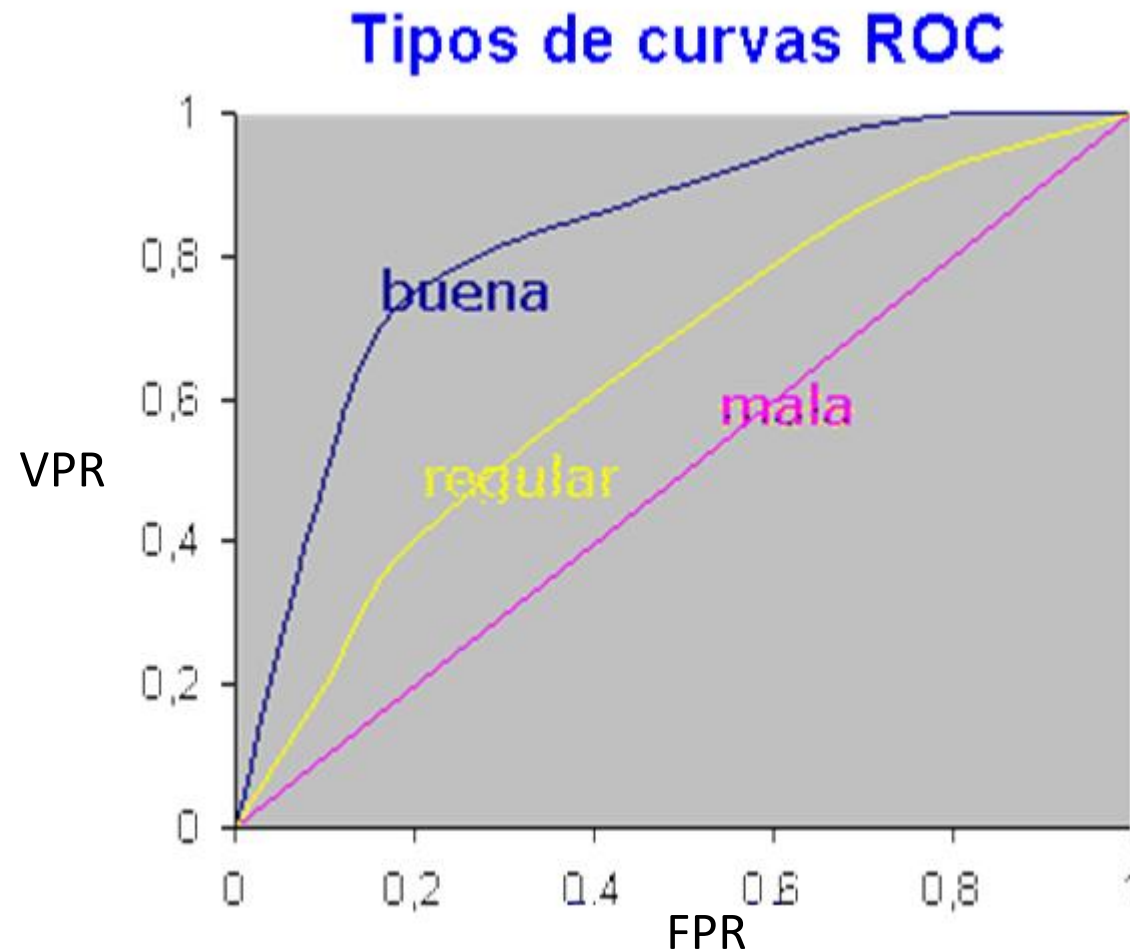
Medidas de Evaluación – Clasificador: Curva ROC (Receiver Operating Characteristic)

- Razón de verdaderos positivos

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

- Razón de falsos positivos

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



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]

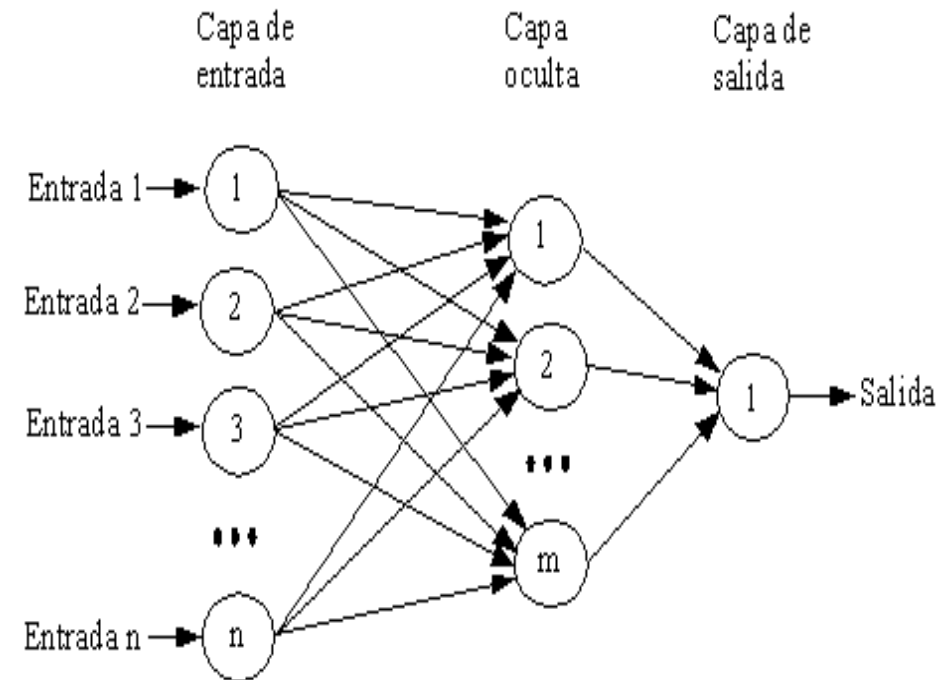
Árboles de Decisión
[Apte, 1997]

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

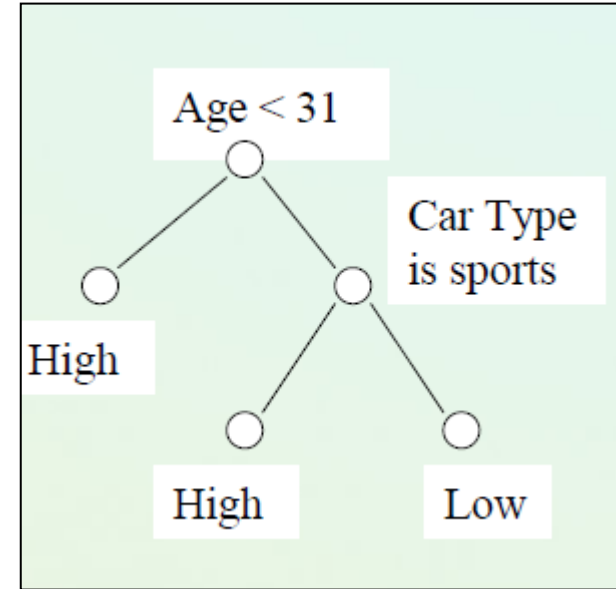
[Joachims, 1998]

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

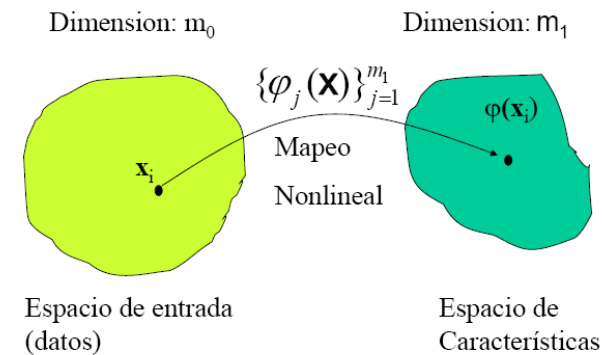
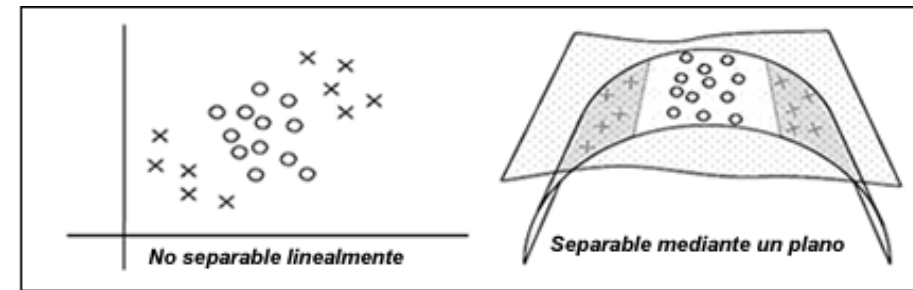
[Joachims, 1998]

Métodos de Regresión

[Yang, 1999]

Métodos basados en Ejemplos

[Yang, 1999]



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

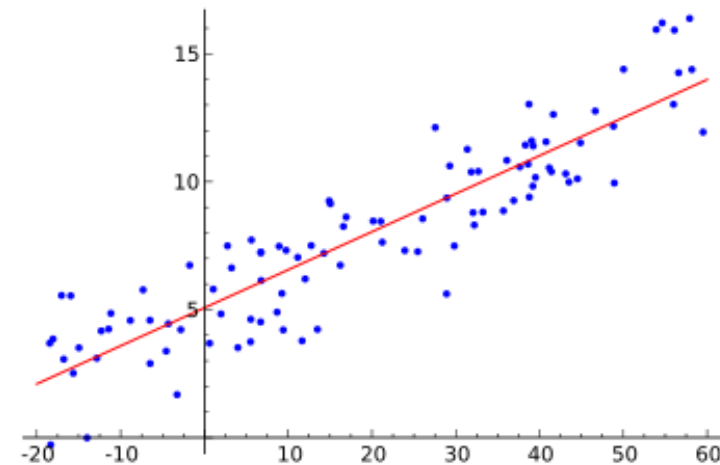
[Joachims, 1998]

● 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

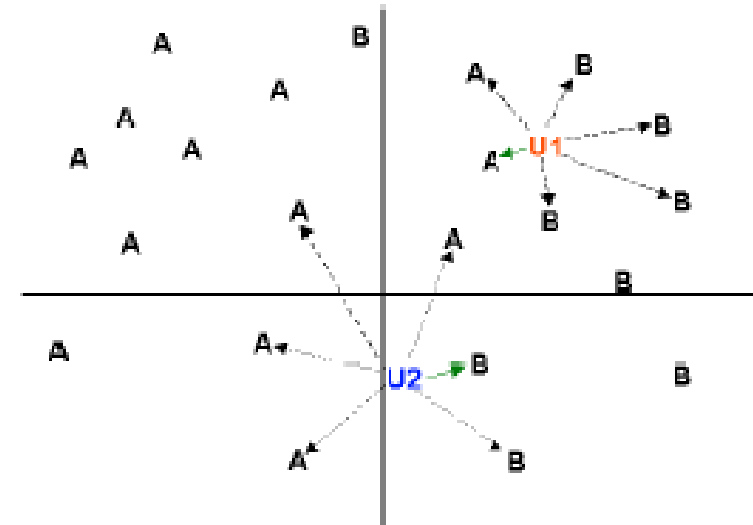
[Joachims, 1998]

Métodos de Regresión

[Yang, 1999]

Métodos basados en Ejemplos

[Yang, 1999]



AGENDA

1. Machine Learning
2. Preparación de datos
3. Análisis Predictivo
- 4. Análisis Descriptivo**

Análisis Descriptivo

Análisis Descriptivo



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

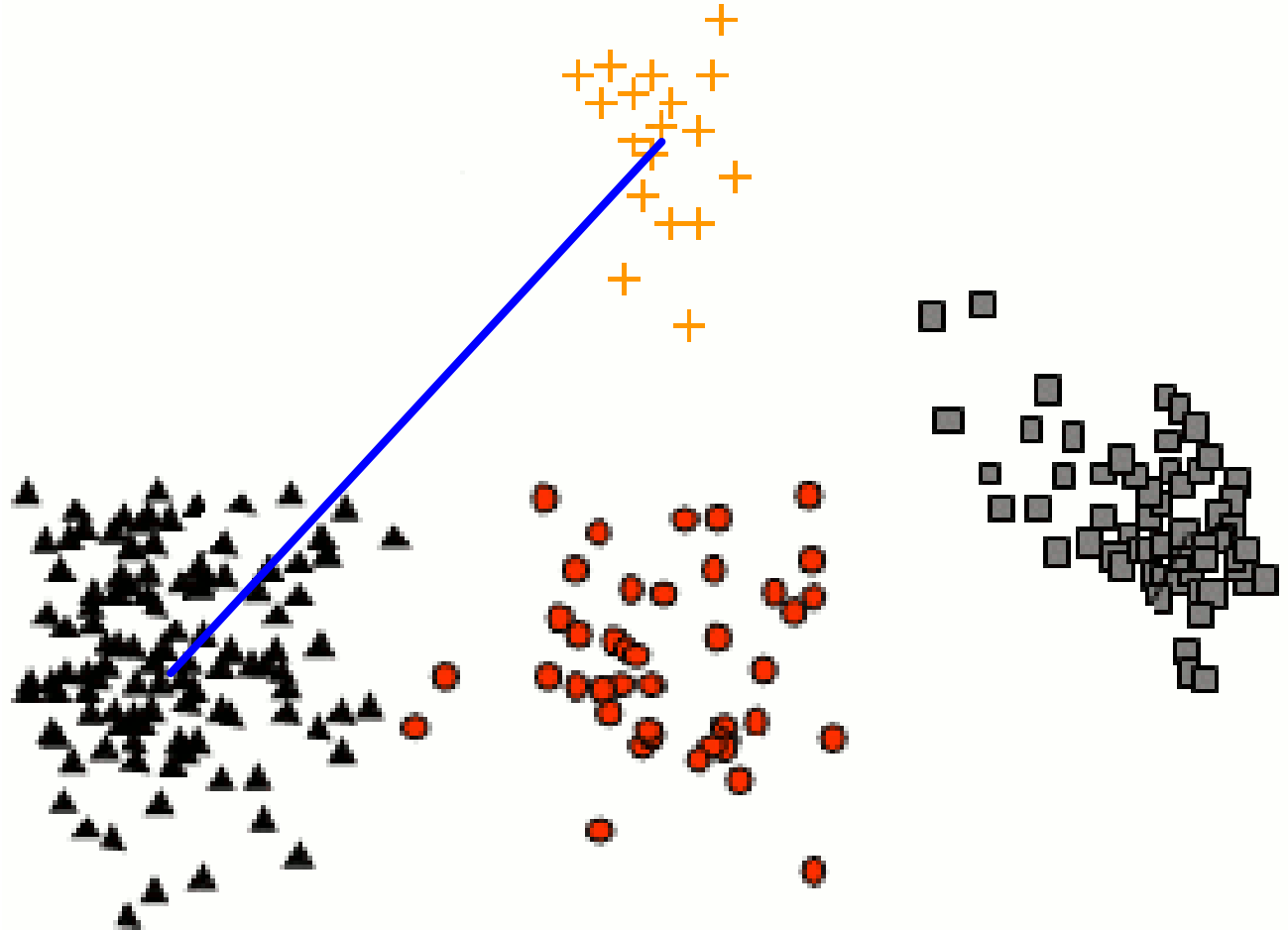
● Agrupamiento / Clustering

● Asociación

● Selección de Factores

Clustering

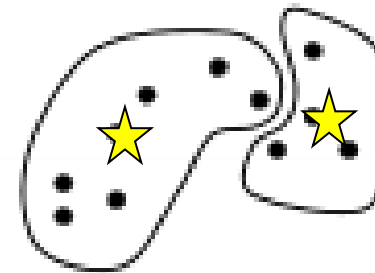
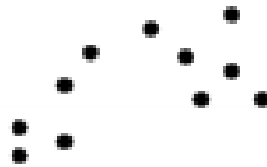
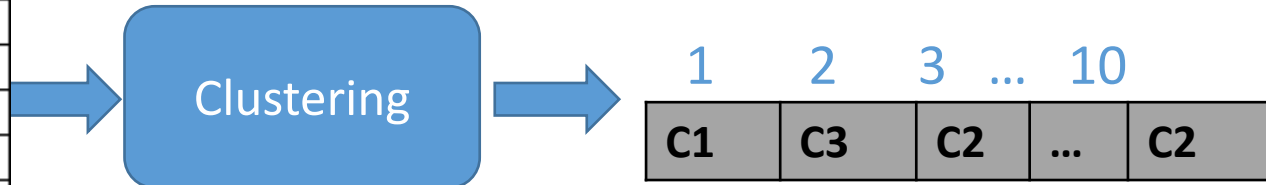
Una medida de distancia determina la similitud entre los datos.



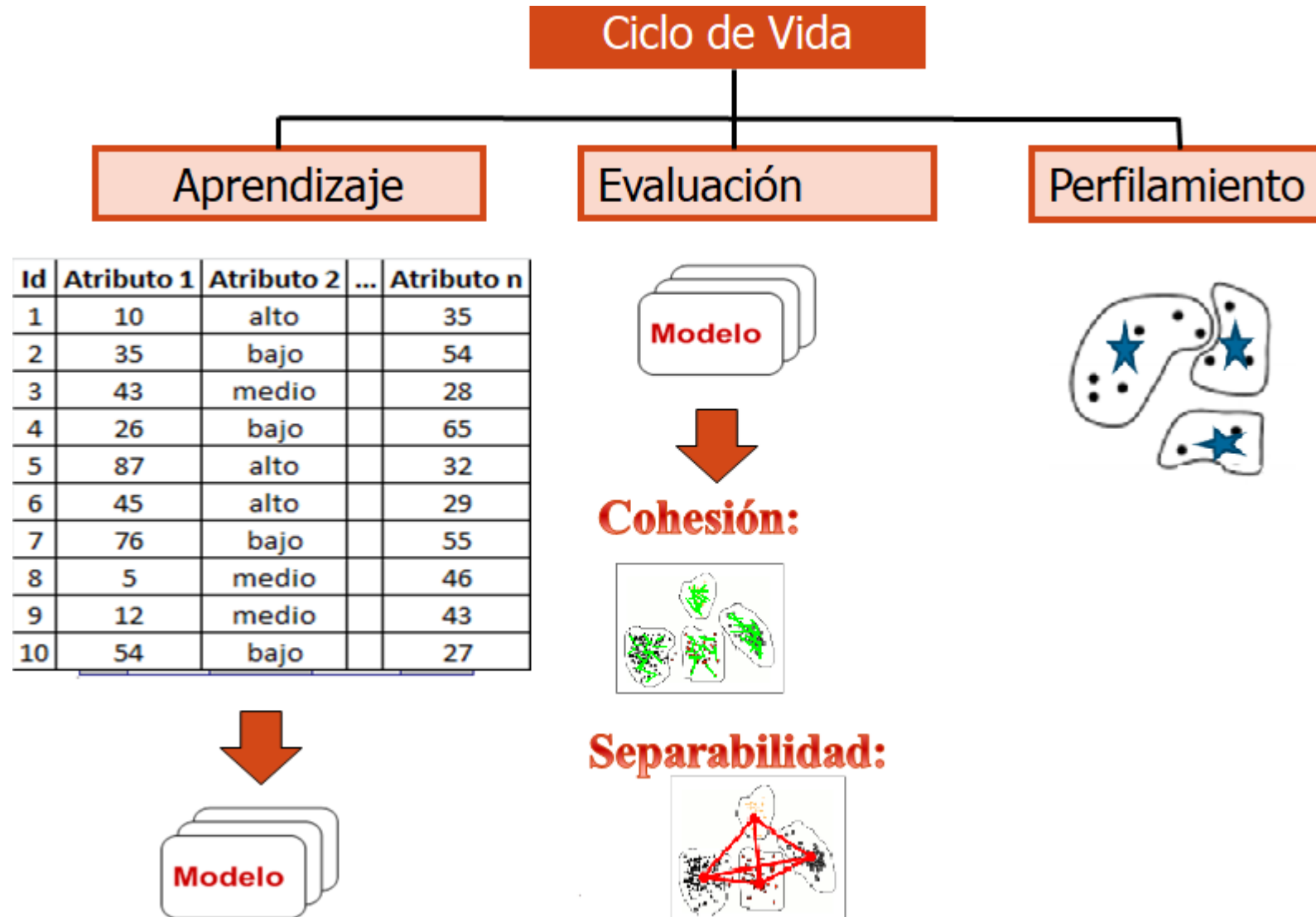
Los objetos en un grupo deben ser similares o relacionados entre ellos.

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

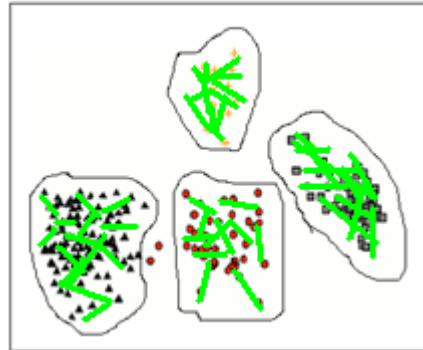


Evaluación

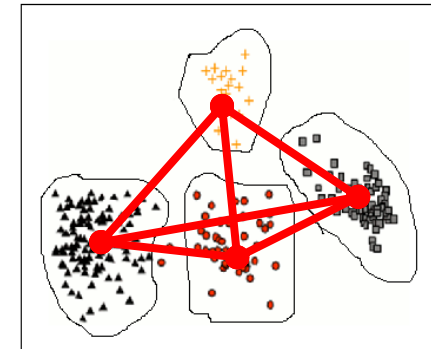
- Validación Interna: Evalúa la calidad de los clusters basado en medidas de distancia, algunos índices son:

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

Compactness:



Separability:



Índices para Validación Interna

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

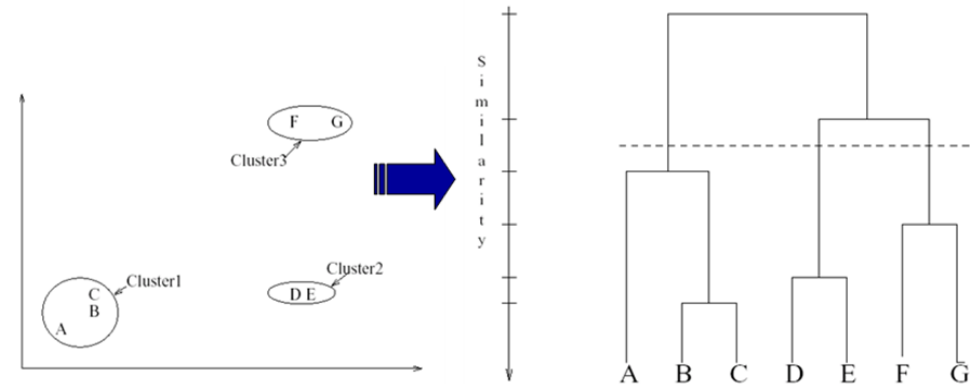


Métodos Jerárquicos

Métodos Particionales

Redes Neuronales

Métodos Probabilísticos



Métodos de Clustering

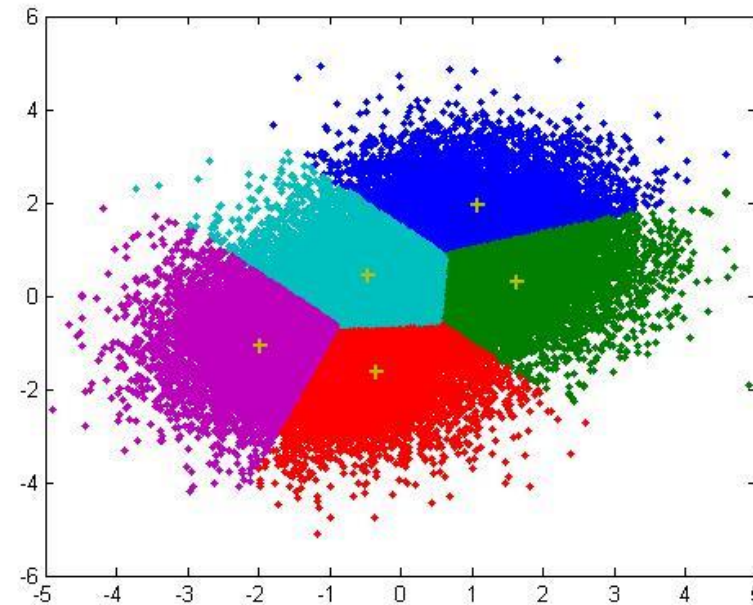
Métodos Jerárquicos



Métodos Particionales

Redes Neuronales

Métodos Probabilísticos



[Steinley, 2006]

Métodos de Clustering

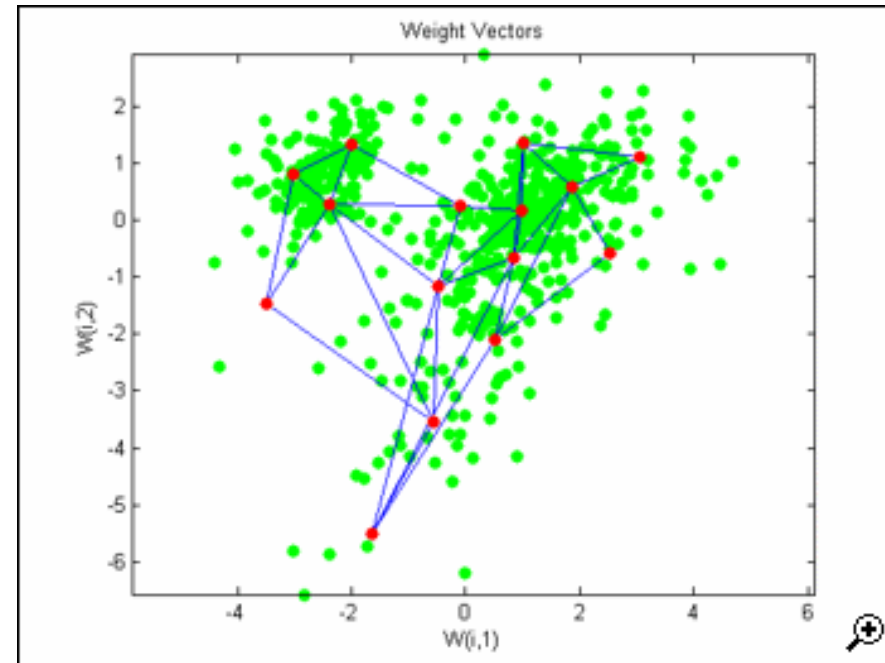
Métodos Jerárquicos

Métodos Particionales



Redes Neuronales

Métodos Probabilísticos



[Vesanto and Alhoniemi, 2000]

Métodos de Clustering

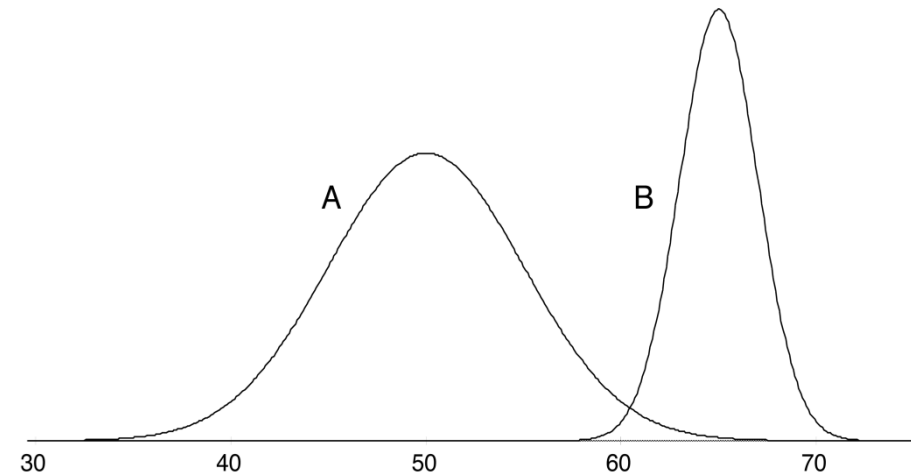
Métodos Jerárquicos

Métodos Particionales

Redes Neuronales



Métodos Probabilísticos

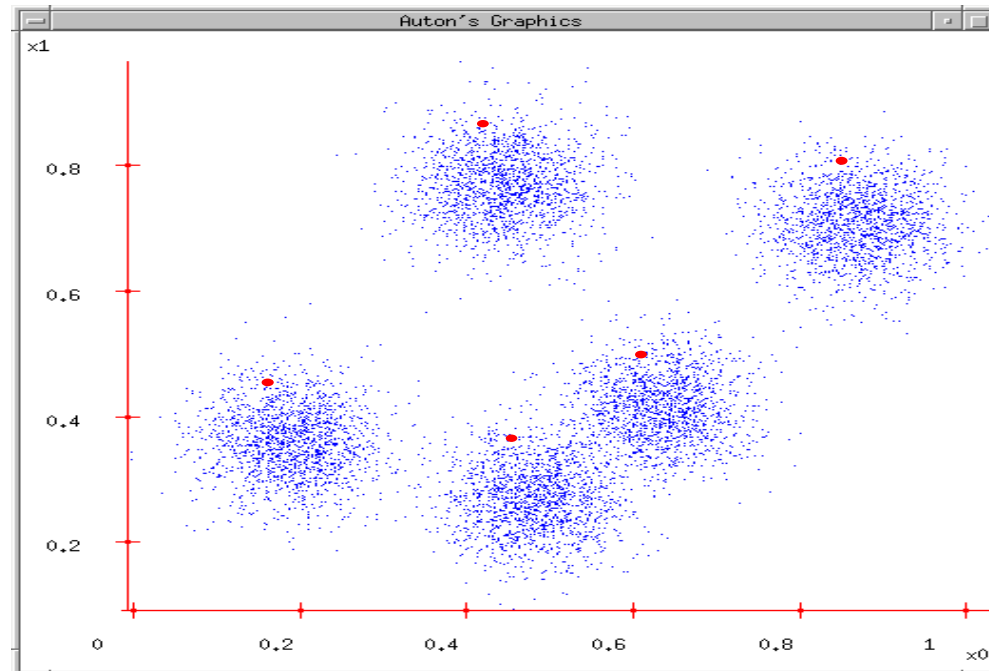


[Francois et al., 2006]

Métodos Particionales



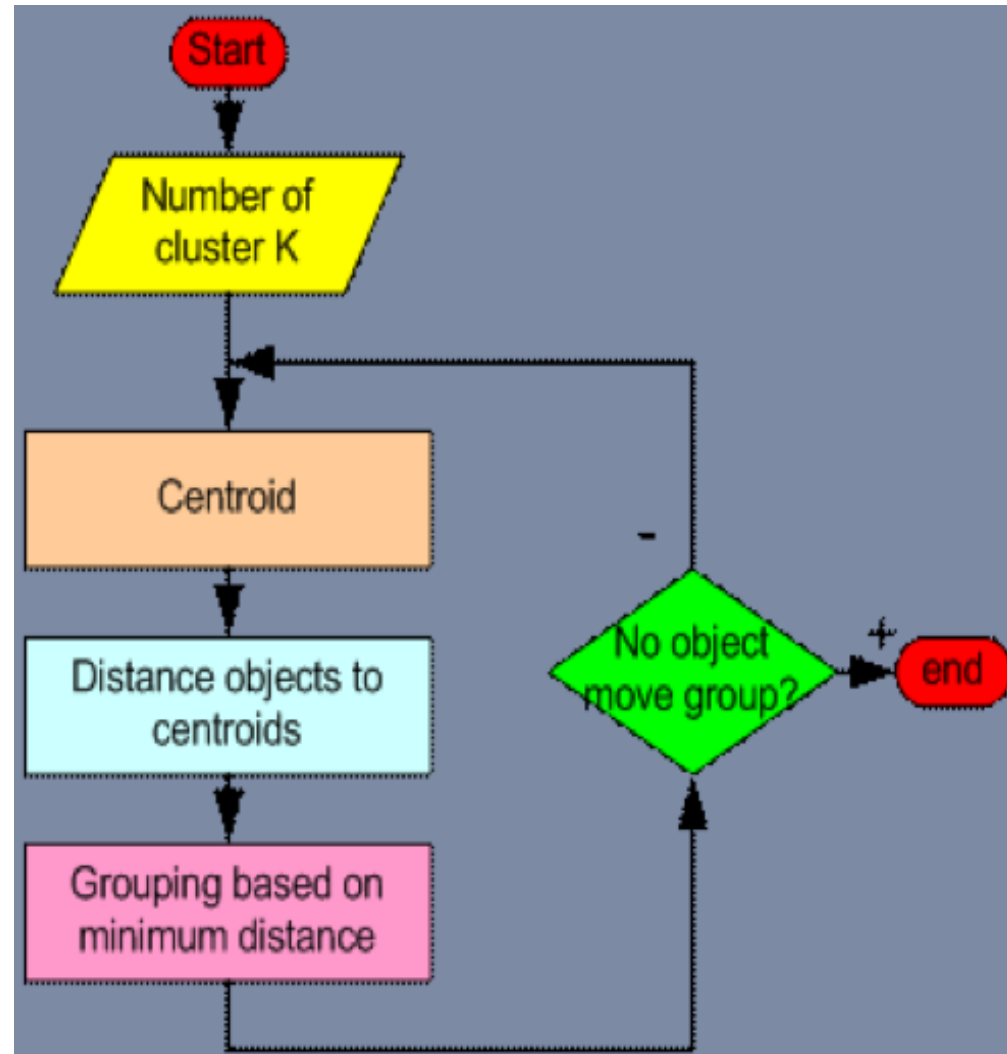
Divide el conjunto de datos en un número predefinido de grupos. K-Means es el método más comúnmente utilizado, la idea del método es definir k centroides, uno por clúster, y los datos son asociados al centroide más cercano.



Medidas de Similitud basadas en Distancia

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

K-Means



K-Means: Limitaciones

- Sensible a los centroides iniciales, ya que converge a óptimos locales.
- Requiere especificar el número de clusters.
- Se afecta por datos “ruidosos”.
- No es aplicable a datos categóricos.