Escuela de
Computación
UCV

Minería de Datos

Clase :Transformación de loa datos

Facultad de Ciencias UCV, Escuela de Computacion, Prof Leonardo Toglia

Transformaciones

Engloba cualquier proceso que modifique la forma de los datos, con el objetivo de mejorar su representación.

Básicamente:

a) Cambiar el rango de las variables

- Normalización (o escalado)
- Estandarización

b) Cambiar el tipo de las variables

- Numerización
- Discretización

a) Cambiar el rango de las variables:

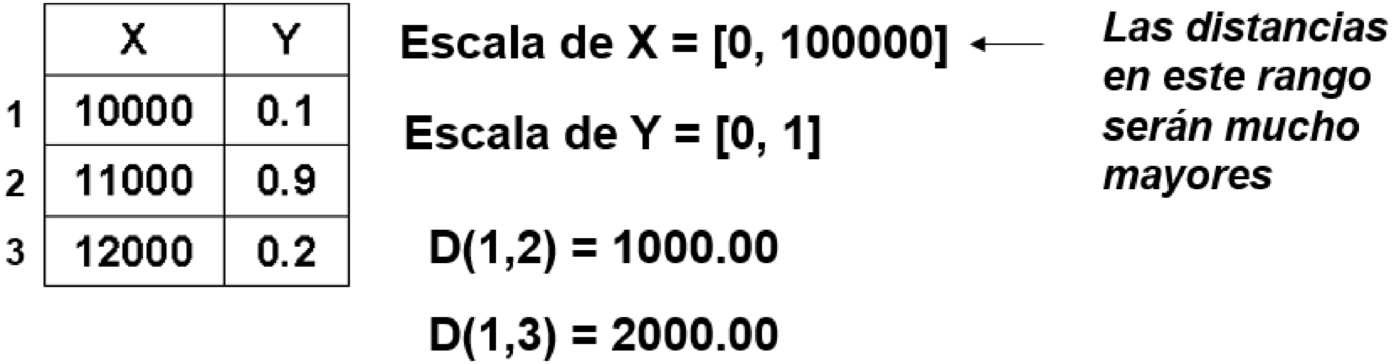
¿Por qué?

- En algunos algoritmos de aprendizaje, las diferencias en los rangos puede producir un sesgo hacia las variables de mayor rango (Ej. métodos basados en distancias).

- Al cambiar (normalizar o estandarizar) el rango de las variables, se intenta dar a los atributos un peso similar.

- Se evitan las dependencias en las unidades de medida.

Ejemplo:



	X	Y
1	10000	0.1
2	11000	0.9
3	12000	0.2

Escala de X = [0, 100000] ← ***Las distancias en este rango serán mucho mayores***

Escala de Y = [0, 1]

D(1,2) = 1000.00

D(1,3) = 2000.00

Técnicas:

- Normalización → *Balance lineal simple*
- Estandarización → *Por la media y la desviación estándar*

Balance lineal simple:

Requiere los valores máximos y mínimos de la variable sobre los datos, y los valores máximos y mínimos del rango de normalización

Sea la variable X = (x_1, x_2, …., x_n). Se quiere transformar en una variable X’ = (x1’, x2’, …., xn’) tal que

$$x_i' = R_{min} + \frac{(R_{max} - R_{min})(x_i - X_{min})}{(X_{max} - X_{min})}$$

Donde

- X_{max} = máximo de la variable
- X_{min} = mínimo de la variable
- R_{max} = valor máximo del rango de normalización
- R_{min} = valor mínimo del rango de normalización

Si la escala o rango a mapear es [0, 1] la expresión se simplifica a

$$x_i' = R_{min} + \frac{(x_i - X_{min})}{(X_{max} - X_{min})}$$

Ejemplo:

	X
1	10
2	2
3	4
4	11
5	6

$$\xrightarrow{\text{Rango} = [1,\ 5]}$$

X’
4.5
1
1.8
5
2.7

$X_{min} = 2$

$X_{max} = 11$

$$\text{Para } x_1 = 10 \longrightarrow x_1' = 1 + \frac{(5-1)\,(10-2)}{(11-2)} = 4.5$$

Estandarización:

Requiere la media y la desviación estándar (observada) de la variable

Sea la variable X = (x_1, x_2, …., x_n). Se quiere transformar en una variable X’ = (x1’, x2’, …., xn’) tal que

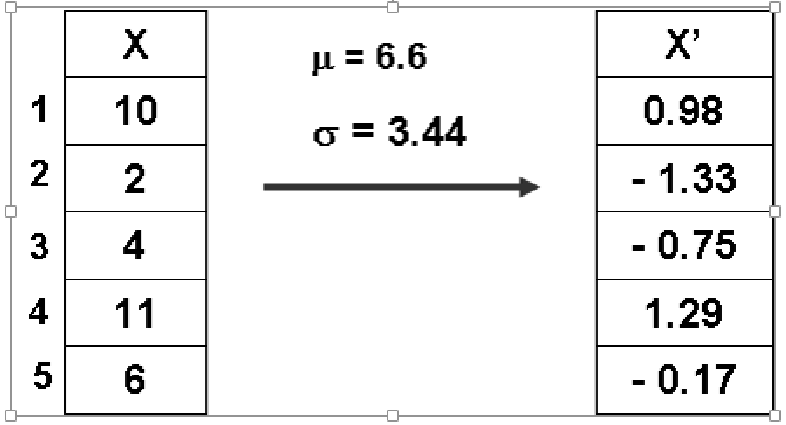
$$x_i' = \frac{x_i \ - \ \mu}{\sigma}$$

Donde

μ = *media de la variable*

σ = *desviación estándar de la variable*

Ejemplo:



$\mu = 6.6$

$\sigma = 3.44$

	X	X’
1	10	0.98
2	2	- 1.33
3	4	- 0.75
4	11	1.29
5	6	- 0.17

Con esté método se obtiene una nueva variable con:

media = 0

desviación estándar = 1

$$\text{Para } x_1 = 10 \longrightarrow x_1' = \frac{(10 \ - \ 6.6)}{3.44} = 0.98$$

- Cambiar el tipo de las variables:

  ¿Por qué?

  - El algoritmo de minería que se va a utilizar no admite valores categóricos (Ej: redes neuronales, máquinas de soporte vectorial, entre otros).
  - El algoritmo de aprendizaje no acepta valores numéricos (Ej: algunas técnicas de árboles de decisión, entre otros).

Técnicas:

Numerización → *Variable nominal a numérica*

Discretización → *Variable numérica a nominal*

Numerización: transformación de una variable nominal en una numérica

- ***Numerización 1-de-n*** ➡ Una variable nominal que puede asumir n valores $\{a_1, a_2, \ldots, a_n\}$, se transforma en n variables numéricas que pueden asumir los valores 0, 1 (variables binarias)

Ejemplo:

Tipo_inmueble

- Apartamento
- Casa
- Terreno
- Oficina

Se transforma en 4 variables

- Inmueble_apartamento
- Inmueble_casa
- Inmueble_terreno
- Inmueble_Oficina

Si en el conjunto de datos se tiene

Tipo_inmueble
Casa
Apartamento
Casa
Terreno
Oficina

Se transforma en

Inmueble_ apartamento	Inmueble_ casa	Inmueble_ terreno	Inmueble_ oficina
0	1	0	0
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

- ***Numerización 1-de-1*** ➡ **Una variable nominal que puede asumir n valores $\{a_1, a_2, \ldots, a_n\}$, se transforma en una variable numérica que pueden asumir n valores.**

Ejemplo:

Nivel_estudios
Primaria
bachillerato
Universitaria
Postgrado

→ ***Se transforma en 1 variable numérica***

Nivel_estudios
1
2
3
4

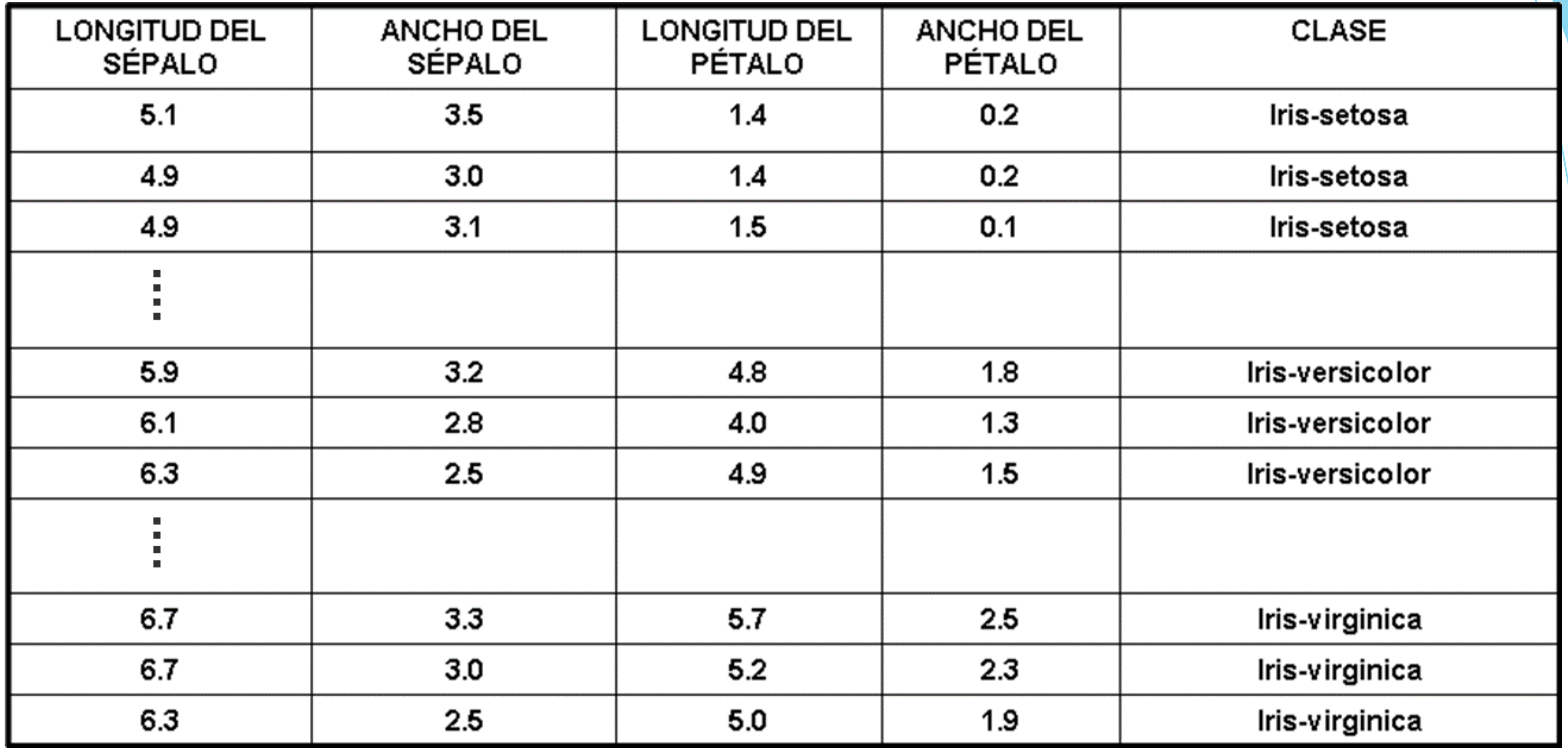
Si en el conjunto de datos se tiene

Nivel_estudios
Primaria
Universitaria
Bachillerato
Postgrado
Universidad

→ ***Se transforma en***

Nivel_estudios
1
3
2
4
3

Ejemplo: conjunto de datos IRIS



LONGITUD DEL SÉPALO	ANCHO DEL SÉPALO	LONGITUD DEL PÉTALO	ANCHO DEL PÉTALO	CLASE
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.9	3.1	1.5	0.1	Iris-setosa
⋮				
5.9	3.2	4.8	1.8	Iris-versicolor
6.1	2.8	4.0	1.3	Iris-versicolor
6.3	2.5	4.9	1.5	Iris-versicolor
⋮				
6.7	3.3	5.7	2.5	Iris-virginica
6.7	3.0	5.2	2.3	Iris-virginica
6.3	2.5	5.0	1.9	Iris-virginica

Problema de clasificación →

Derivar un modelo que permita predecir la clase de planta IRIS a partir de la información suministrada por las 4 variables que la caracterizan.

Longitud del pétalo	Ancho del pétalo	Longitud del sépalo	Ancho del sépalo	TIPO
5.1	3.5	1.4	2.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
⋮				
7.0	3.2	4.7	1.4	Iris-versicolor
⋮				
5.8	2.7	5.1	1.9	Iris-virginica

1-de-n

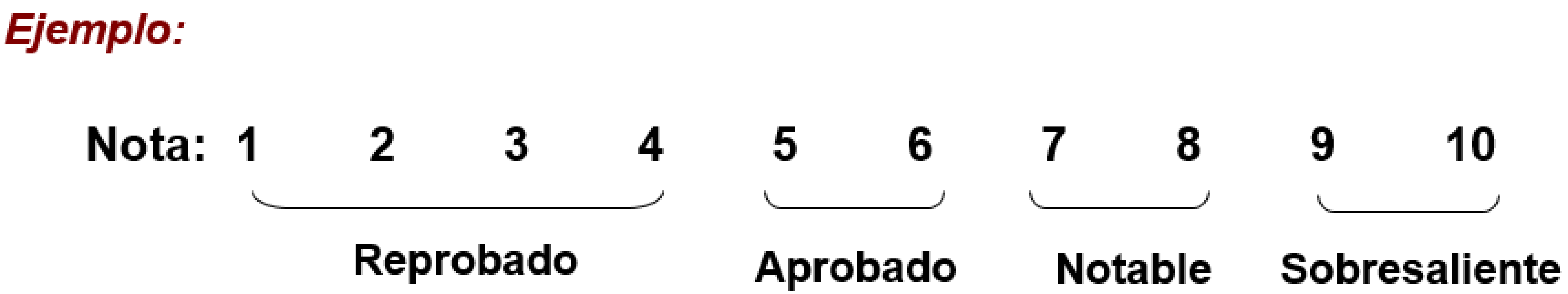
Iris_setosa	Iris_versicolor	Iris_Virgnica
1	0	0
1	0	0
⋮		
0	1	0
⋮		
0	0	1

1-de-1

Tipo
1
1
⋮
2
⋮
3

Discretización: Transformación de una variable numérica en una variable nominal ordenada (variable ordinal).

La discretización también permite integrar variables con escalas diferentes
Puede ser utilizada para tratar outliers o datos fuera de rango
Simplifica los datos originales
Los patrones resultantes de la minería son más fáciles de entender



Ejemplo:

Nota: 1 2 3 4 5 6 7 8 9 10

Reprobado (1–4) **Aprobado** (5–6) **Notable** (7–8) **Sobresaliente** (9–10)

¿Cómo discretizar?

- Ordenar los valores del atributo.
- Decidir cuántas categorías o intervalos (*cuántos puntos de división y dónde colocarlos).*
- Determinar cómo mapear los valores numéricos de la variable en las categorías (*todos los valores en un intervalo se mapean al mismo valor categórico)*

Entonces, si X es una variable numérica, el resultado de la discretización será:

$$\{(x_0, x_1], (x_2, x_3], (x_4, x_5], \ldots\ldots, (x_{n-1}, x_n)\} \Rightarrow$$

Suele sustituirse por -∞

Suele sustituirse por ∞

Conjunto de intervalos a los cuales se les puede asociar una etiqueta

¿Cómo encontrar los intervalos de discretización?

Técnicas
- **No supervisadas**
  - **Toman sólo la información suministrada por la variable a discretizar**
- **Supervisadas**
  - **Además de la información suministrada por la variable a discretizar, utilizan la etiqueta de clase**

- *Técnicas no supervisadas*

Se puede realizar la discretización por
- **Igual ancho**
  - **Dividen el rango del atributo en un número de intervalos (definidos previamente), cada uno del mismo ancho**
- **Igual frecuencia**
  - **Tratan de colocar el mismo número de objetos en cada intervalo**

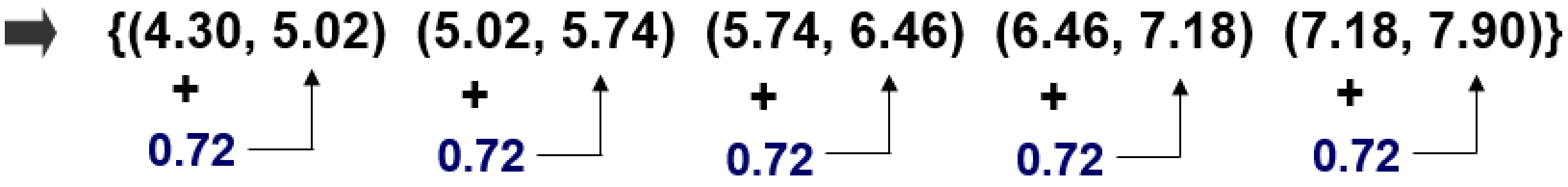
Ejemplo: *discretización de la variable Longitud de Pétalo (conjunto de datos IRIS) por igual ancho*

- ☞ **Se selecciona el número de intervalos** ➡ **No. Intervalos = 5**

- ☞ **Se construyen los intervalos (*Ej. por igual ancho*):**

$$\text{Ancho del intervalo} = \frac{X_{max} - X_{min}}{\text{No. intervalos}}$$

$$\text{Si } X_{min} = 4.3 \quad X_{max} = 7.9 \;\Rightarrow\; \text{Ancho del intervalo} = \frac{7.9 - 4.3}{5} = 0.72$$



➡ **{(4.30, 5.02) (5.02, 5.74) (5.74, 6.46) (6.46, 7.18) (7.18, 7.90)}**

☞ **Luego, se puede asignar una etiqueta a cada intervalo (que dependerá del dominio del problema)**

$$\{\underbrace{(4.30, 5.02]}_{\text{Muy corto}}\ \ \underbrace{(5.02, 5.74]}_{\text{Corto}}\ \ \underbrace{(5.74, 6.46]}_{\text{Medio}}\ \ \underbrace{(6.46, 7.18]}_{\text{Largo}}\ \ \underbrace{(7.18, 7.90]}_{\text{Muy largo}}\}$$

¿Cómo se asignan los valores numéricos a los intervalos?

Todos los valores numéricos de un intervalo se mapean al mismo valor categórico:

Longitud del pétalo
5.44
6.02
7.24
4.90
7.43
5.80

→

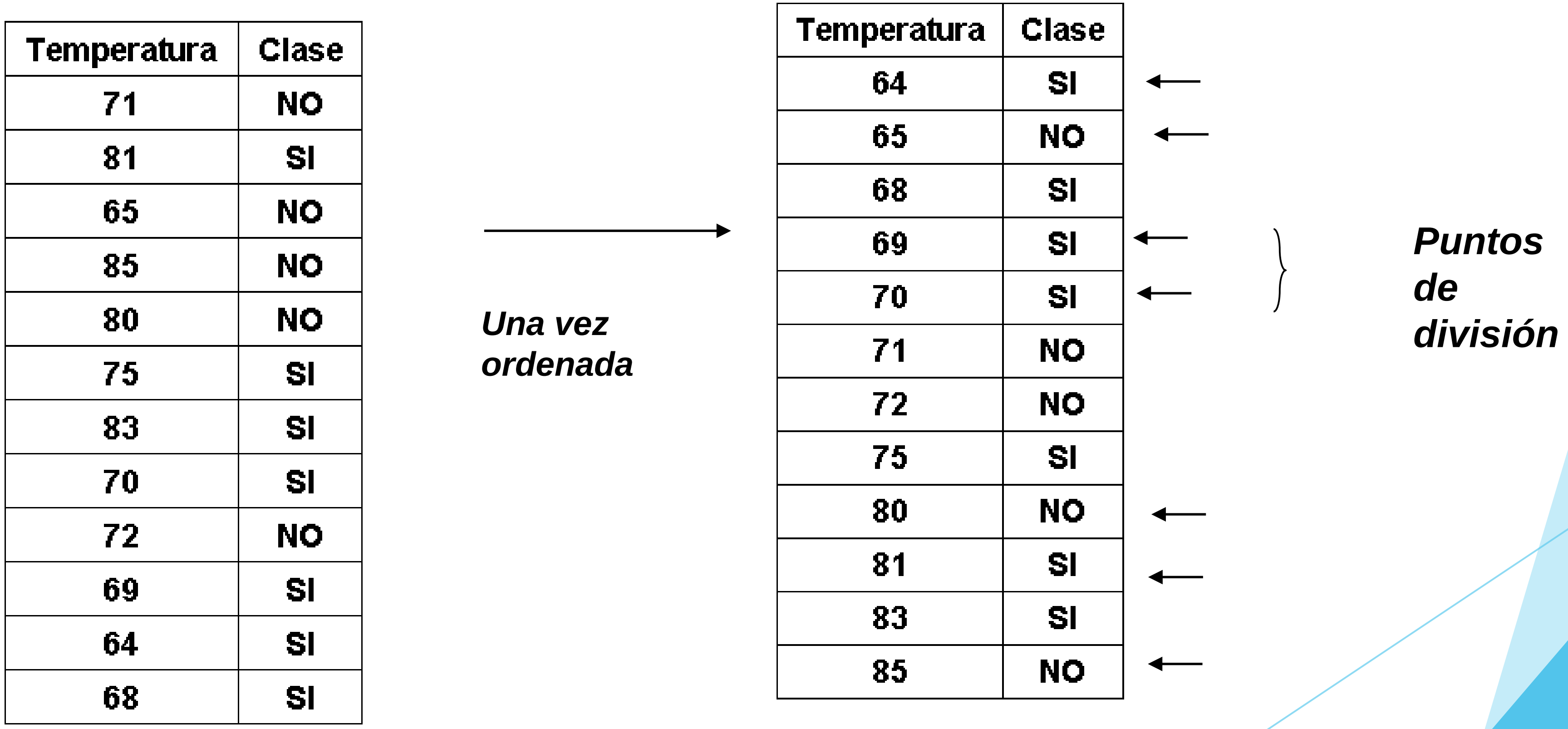
Se transforma en

Longitud del pétalo
Corto
Medio
Muy largo
Muy corto
Muy largo
Medio

- ***Técnicas supervisadas***

Transformaciones

Utilizan información de la clase para determinar dónde colocar los puntos de división para definir los intervalos.
Esquema más utilizado: colocar los puntos de división de manera tal que se maximice la pureza de los intervalos



Temperatura	Clase
71	NO
81	SI
65	NO
85	NO
80	NO
75	SI
83	SI
70	SI
72	NO
69	SI
64	SI
68	SI

Una vez ordenada

Temperatura	Clase
64	SI
65	NO
68	SI
69	SI
70	SI
71	NO
72	NO
75	SI
80	NO
81	SI
83	SI
85	NO

Puntos de división

¿Cómo hacerlo?

Transformaciones

Una forma: utilizar enfoques basados en la entropía.

Entropía:

Sean

K = No. de clases
m = No. de valores
m_i = No. de valores en el intervalo i-ésimo de una partición
m_{ij} = No. de valores de la clase j en el intervalo i-ésimo

Se define la entropía del intervalo i-ésimo

$$e_i = - \sum_{i\ k=1} p_{ij} \log_2 p_{ij}$$

Donde p_{ij} es la probabilidad de la clase j en el intervalo i, y se calcula como

$$p_{ij} = \frac{m_{ij}}{m_i}$$

Ejemplo:

Temperatura	64	65	68	69	70	71	72
Clase	Si	No	Si	Si	Si	No	No

Intervalo 1

Intervalo 2

Punto de división

☞ **Entropía del intervalo 1:**

$$e_1 = - (1/5)\log_2 (1/5) \; - (4/5)\log2 (4/5) \; = \; 0.7219$$

☞ **Entropía del intervalo 2:**

$$e_2 = - (2/2)\log2 (2/2) \; - (0/2)\log2 (2/2) \; = \; 0$$

- La entropía de un intervalo es una medida de su *pureza*

➡ - Intervalos que sólo contienen valores pertenecientes a una sola clase (totalmente puros) tienen una entropía igual a cero.

- Si las clases de los valores de un intervalo se presentan con igual frecuencia entonces la entropía es máxima (el intervalo es lo más impuro posible)

- Como pueden haber muchas maneras de colocar los puntos de división

➡ Se debe calcular la entropía total de una partición

- La entropía total E de una partición será el promedio ponderado de las entropías individuales

$$E = \sum_{i=1}^{n} w_i e_i$$

No. de intervalos

No. de valores en el intervalo i-ésimo de una partición

$$w_j = \frac{m_i}{m}$$

No. de valores

La mejor partición será la que tenga la entropía total más baja

Ejemplo:

Temperatura	64	65	68	69	70	71	72
Clase	Si	No	Si	Si	Si	No	No

p1

p2

$$e_i = - \sum p_{ij} \log_2 p_{ij}$$

$$p_{ij} = \frac{m_{ij}}{m_i}$$

$$e_1 = - (1/5)\log_2 (1/5) \ - (4/5)\text{log2} (4/5) \ = \ 0.7219$$

$$e_2 = - (2/2)\text{log2} (2/2) \ - (0/2)\text{log2} (2/2) \ = \ 0$$

Entropía de la partición 1 = (5/7)*0.7219 + (2/7)*0 = 0.5156

Ejemplo:

Temperatura	64	65	68	69	70	71	72
Clase	Si	No	Si	Si	Si	No	No

p1 ↑ p2

$$e_i = - \sum p_{ij} \log_2 p_{ij}$$

$$p_{ij} = \frac{m_{ij}}{m_i}$$

$$e_1 = - (1/3)\log_2 (1/3) \ - (2/3)\mathrm{log2}\ (2/3) \ = \ 0.9183$$

$$e_2 = - (2/4)\mathrm{log2}\ (2/4) \ - (2/4)\mathrm{log2}\ (2/4) \ = 1$$

Entropía de la partición 2 = (3/7)*0.9183 + (4/7)*1 = 0.9650

¿Cuál escoger? → **Partición 1**

c) Otras transformaciones

- Es posible aplicar otras transformaciones para cambiar un atributo o conjunto de atributos en otros
- Pueden revelar características más interesantes e informativas para la tarea de minería.

Ejemplos:

- Transformaciones funcionales: *se aplica una función matemática simple a cada valor de la variable (Log (X), X^k, e^X, entre otras).*
- Transformada de Fourier: *Permite pasar de una representación en el domino del tiempo a otra representación en el dominio de la frecuencia (Ej. En procesamiento de señales)*

Importante: en general, al mejorar la representación de los datos, el proceso de minería puede ser más eficiente y/o los patrones más fáciles de entender

Selección de variables

En el conjunto de datos se pueden encontrar:

- Características redundantes

  Duplican información en dos o más atributos (Ejemplo: atributos correlacionados)

- Características irrelevantes

  No aportan información útil (Ejemplo: Números de identificación)

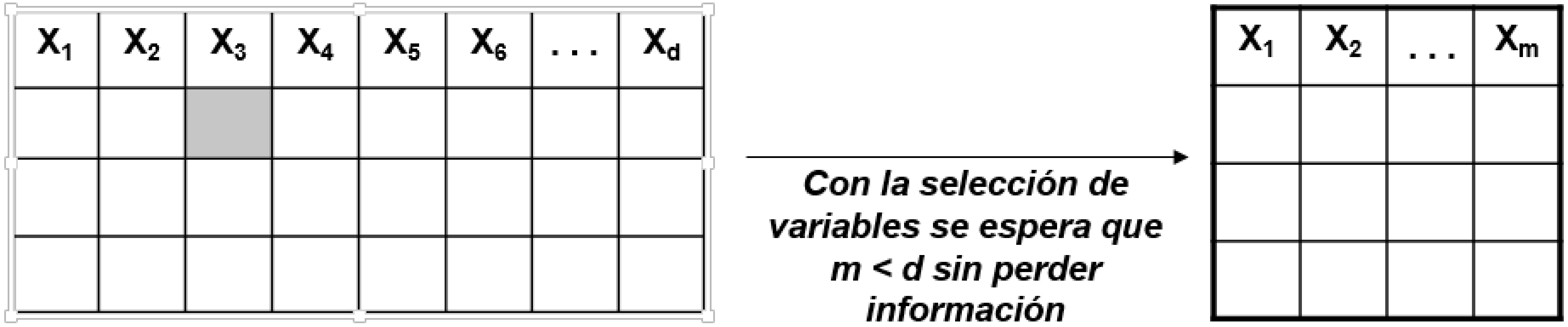
¿Por qué seleccionar?

- Reduce la dimensionalidad de los datos
- Mejora el rendimiento de predicción
- Modelos más comprensible
- Mejora la visualización de lo datos
- Reduce el tiempo de estimación de modelos

Problema de la selección de variables

Seleccionar, a partir de d variables originales, un subconjunto (pequeño) de m características ($m < d$) que, idealmente, es el necesario para explicar el problema bajo estudio (variables más informativas)

Se obtiene de esta forma un espacio de entrada de baja dimensionalidad y con la máxima información



X_1	X_2	X_3	X_4	X_5	X_6	. . .	X_d
		▨					

Con la selección de variables se espera que $m < d$ sin perder información

X_1	X_2	. . .	X_m

¿Cómo seleccionar?

a) Utilizando conocimiento del dominio

b) Mediante técnicas de exploración de datos (gráficos, análisis de correlación, entre otros)

c) Selección automática

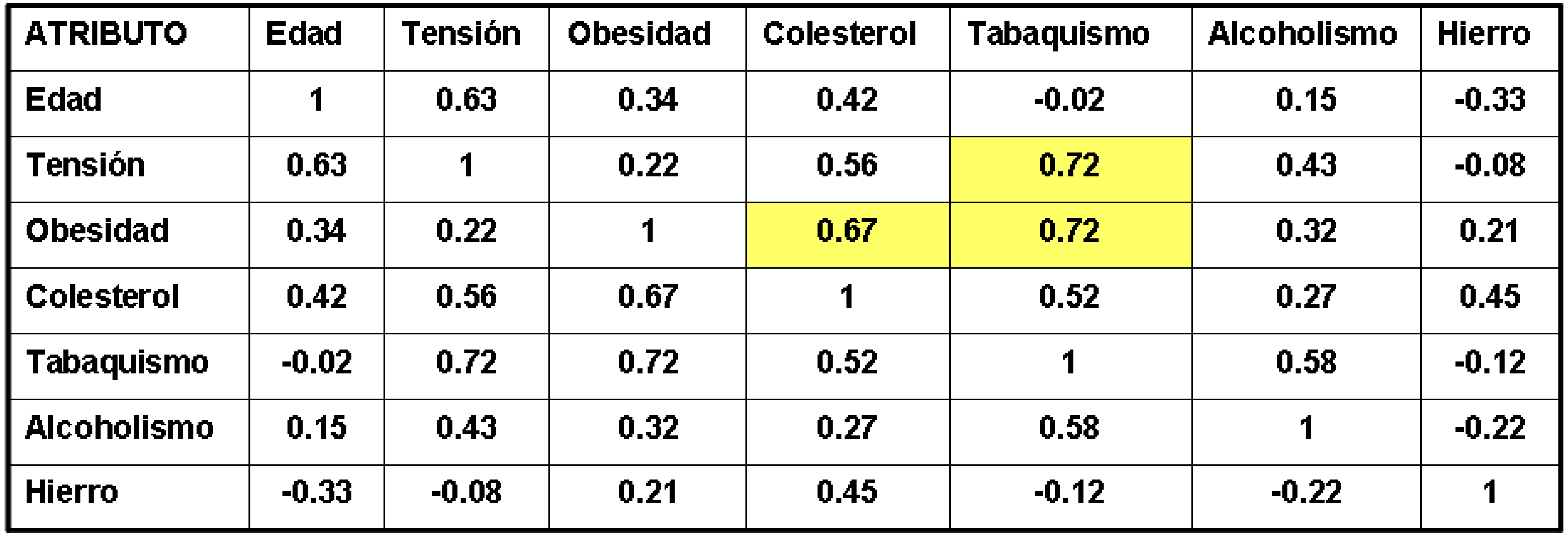
a) Utilizar conocimiento del dominio

- Los expertos pueden aportar conocimiento para identificar los atributos relevantes.
- Eliminar características irrelevantes como números de identificación, nombres, entre otros

b) Análisis exploratorio. Análisis de correlación

Permite visualizar cuáles atributos están estrechamente relacionados (linealmente).

Ejemplo: Conjunto de datos de donantes de sangre de un hospital. Se quiere determinar cuáles son las variables más correlacionadas



ATRIBUTO	Edad	Tensión	Obesidad	Colesterol	Tabaquismo	Alcoholismo	Hierro
Edad	1	0.63	0.34	0.42	-0.02	0.15	-0.33
Tensión	0.63	1	0.22	0.56	0.72	0.43	-0.08
Obesidad	0.34	0.22	1	0.67	0.72	0.32	0.21
Colesterol	0.42	0.56	0.67	1	0.52	0.27	0.45
Tabaquismo	-0.02	0.72	0.72	0.52	1	0.58	-0.12
Alcoholismo	0.15	0.43	0.32	0.27	0.58	1	-0.22
Hierro	-0.33	-0.08	0.21	0.45	-0.12	-0.22	1

Atributos más correlacionados:

- Obesidad - tabaquismo
- Obesidad - colesterol
- Tensión - tabaquismo

¿Cómo utilizar esta información?

Se pueden utilizar sólo aquellas variables que son más fiables
Ej: Tensión y Obesidad
Eliminar tabaquismo