

Minería de datos

Tema 1: Proceso de minería de datos



Agenda:

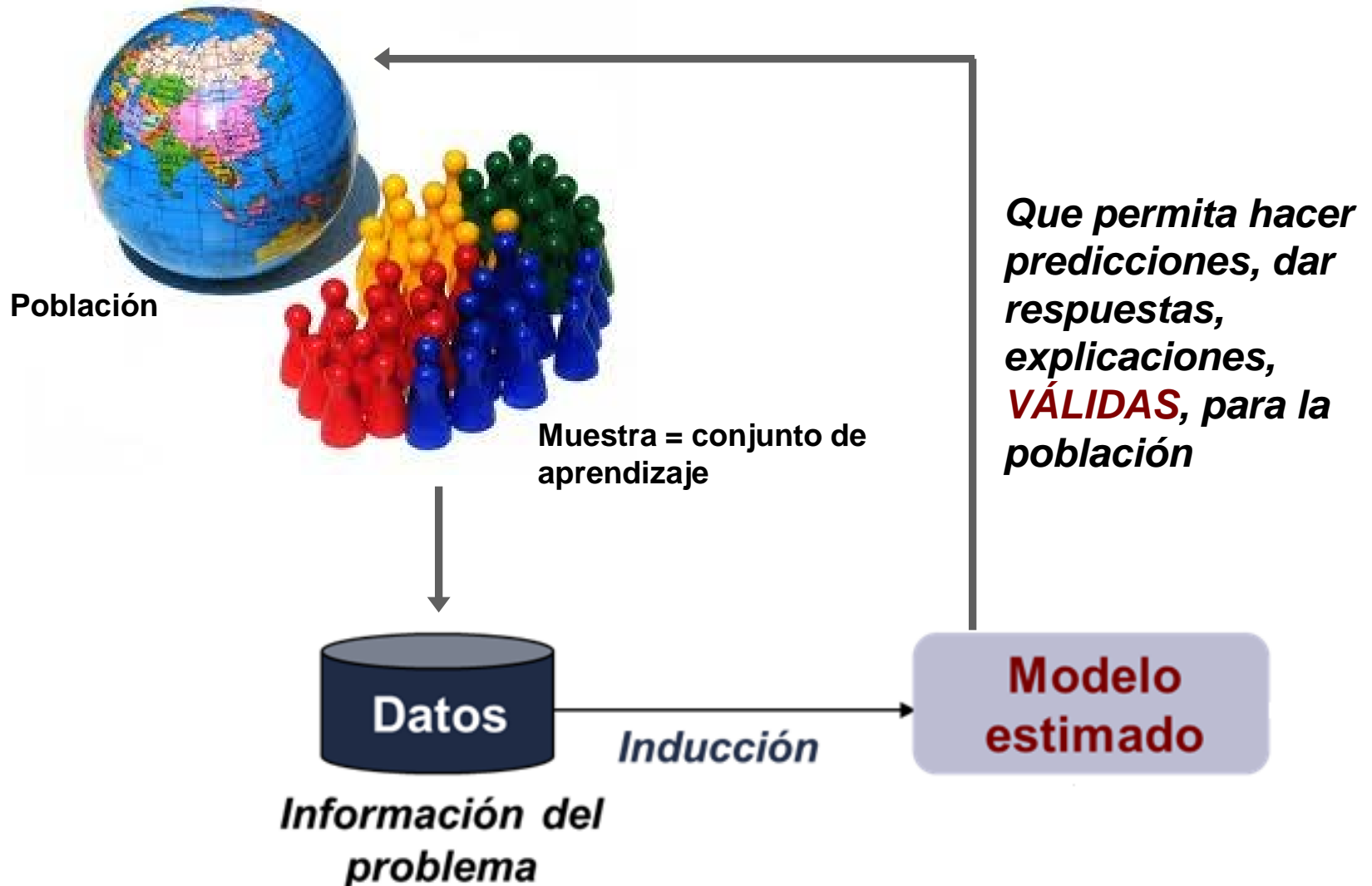
- **Proceso de minería de datos**
- **Ejemplos**
- **Metodología CRISP-DM**
- **Herramientas de MD: Weka**

Para los siguientes problemas, indique qué tarea de minería de datos puede ser adecuada para su resolución:

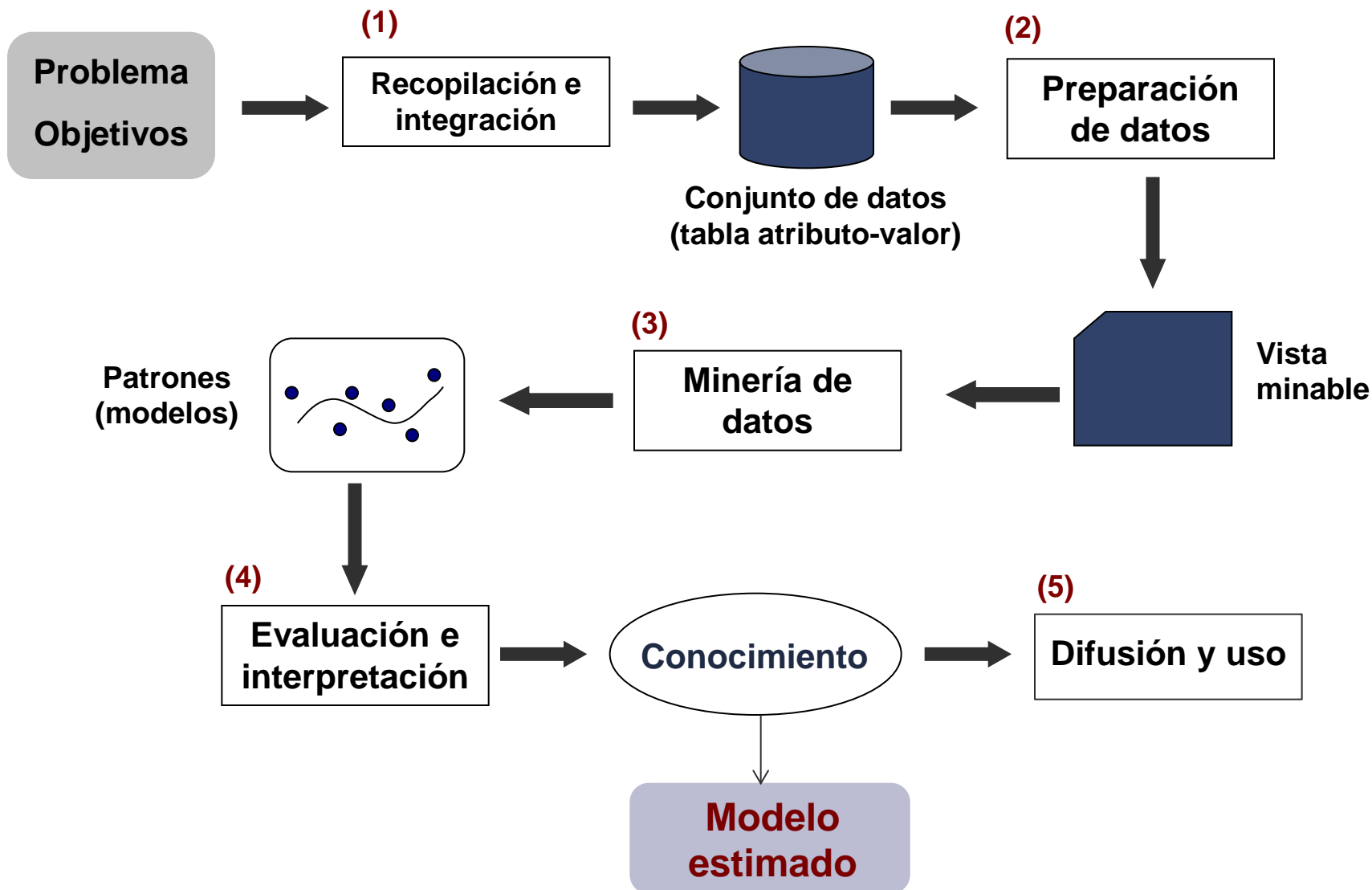
- **Determinar la aptitud física de la tierra para un determinado cultivo.**
- **Indicar la cantidad de lluvia a corto plazo a partir de datos climatológicos**
- **Determinar la categoría de un Servicio Web.**
- **identificar las fallas de un servicio de telefonía móvil a partir de los registros de los usuarios**
- **Identificar brotes de gripe a partir de los Tweets de los usuarios**
- **Determinar el costo de un nuevo contrato en una compañía a partir de los costos correspondientes a contratos anteriores.**
- **Determinar cuáles son los itinerarios más seguidos por los visitantes de un sitio Web**
- **Estimar el tamaño del software a partir de mediciones realizadas en el ámbito de la especificación de requisitos.**
- **Indicar la presencia de intrusos o eventos maliciosos en un sistema móvil.**
- **Determinar las preferencias de compra de clientes en un sistema de comercio electrónico.**
- **Determinar la calidad del agua de un repositorio basado en indicadores ecológicos.**
- **Determinar el rendimiento físico de deportistas para una competencia.**
- **Desarrollar un buscador Web de noticias adaptado a las preferencias de los usuarios**

Proceso de minería de datos

Importante:

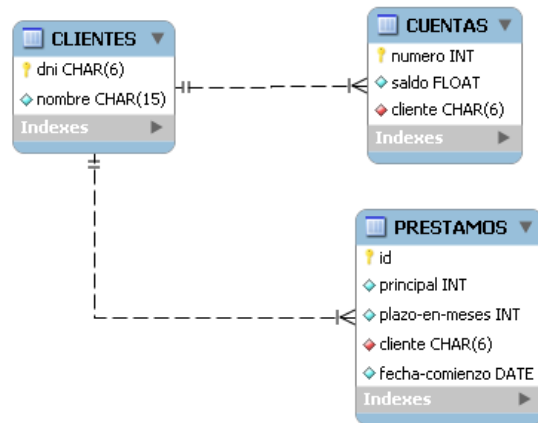


Proceso de minería de datos

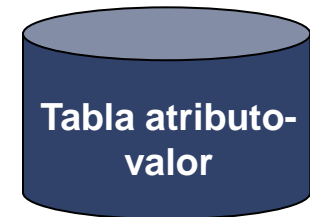


Recolección/integración de los datos

Recopilación e integración:



Conjunto de datos

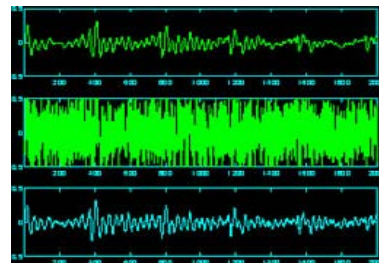
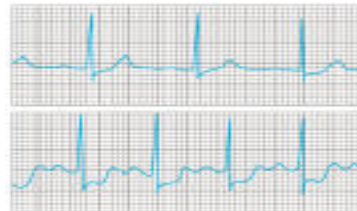
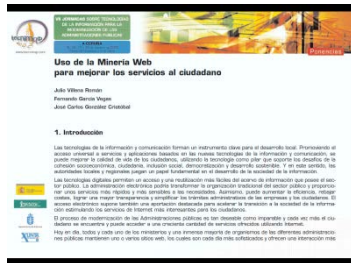
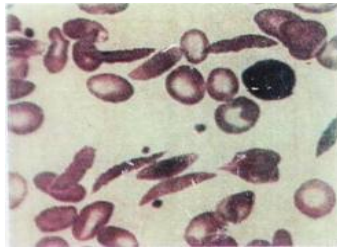


No	Contenido canasta
1	Brócoli, pimienta, maíz
2	Espárragos, calabaza, maíz
3	Maíz, Tomates, frijoles, calabazas
4	Pimienta, maíz, tomates, frijoles
5	Frijoles, espárragos, frijoles, tomates
6	Calabaza, espárragos, frijoles, tomates
7	Tomates, maíz
8	Brócoli, tomates, pimienta
9	Calabaza, espárragos, frijoles
10	Frijoles, maíz
11	Pimienta, brócoli, frijoles, calabaza
12	Espárragos, frijoles, calabaza
13	Calabaza, maíz, espárragos, frijoles
14	Maíz, pimienta, tomates, frijoles, brócoli

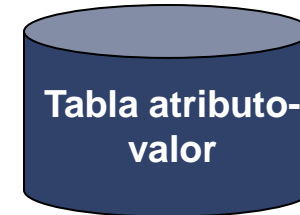
	A	B	C	D	E	F	G	H
1	DULCES DE LA CLASE							Grupo:
2	COLOR	Naranja	Café	Amarillo	Verde	Rojo	Azul	TOTAL
3	Total de la Clase	40	104	83	67	29	40	363
4	Juan	2	8	5	4	2	3	24
5	María	1	7	8	3	2	3	24
6	Antonio	3	6	4	5	4	3	25
7	Andrea	4	7	4	6	1	2	24
8	Carlos	3	7	3	5	3	3	24
9	Jorge	3	6	7	5	1	1	23
10	Luisa	3	6	6	6	1	3	25
11	Mary	4	7	6	4	2	2	25
12	Guillermo	2	6	7	3	2	3	23
13	Felipe	1	7	8	4	1	4	25
14	Carmen	2	8	4	5	1	3	23
15	Cristina	4	6	6	3	3	3	25
16	Gustavo	3	8	4	5	2	2	24
17	Ana	1	8	5	4	2	4	24
18	Pedro	4	7	6	5	2	1	25

Recolección/integración de los datos

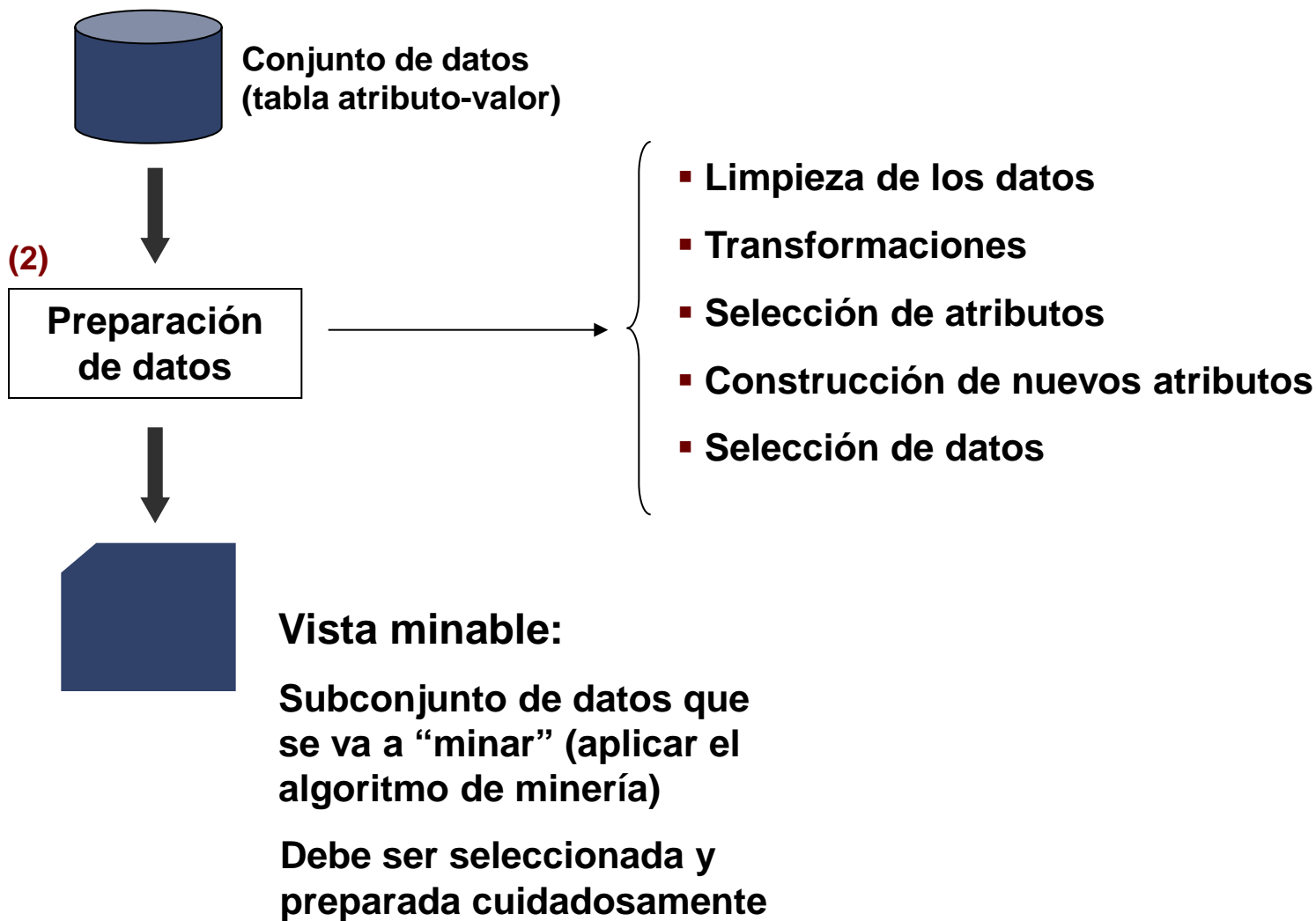
Recopilación y pre-procesamiento:



Conjunto de datos



Preparación de los datos



a) Limpieza de los datos

- De:
- Campos obsoletos o redundantes
 - Valores inconsistentes
 - Datos erróneos o con ruido
 - Valores ausentes
 - Valores fuera de rango o anómalos (*outliers*)

Objetivo:
*mejorar la
calidad de los
datos*

b) Transformaciones

Engloba cualquier proceso que modifique la forma de los datos

- Cambiar el rango de las variables
- Cambiar el tipo de las variables
- Cambiar un atributo o conjunto de atributos en otros

Objetivo:
*mejorar la
representación
de los datos*

c) Selección de atributos

Para determinar los atributos más relevantes

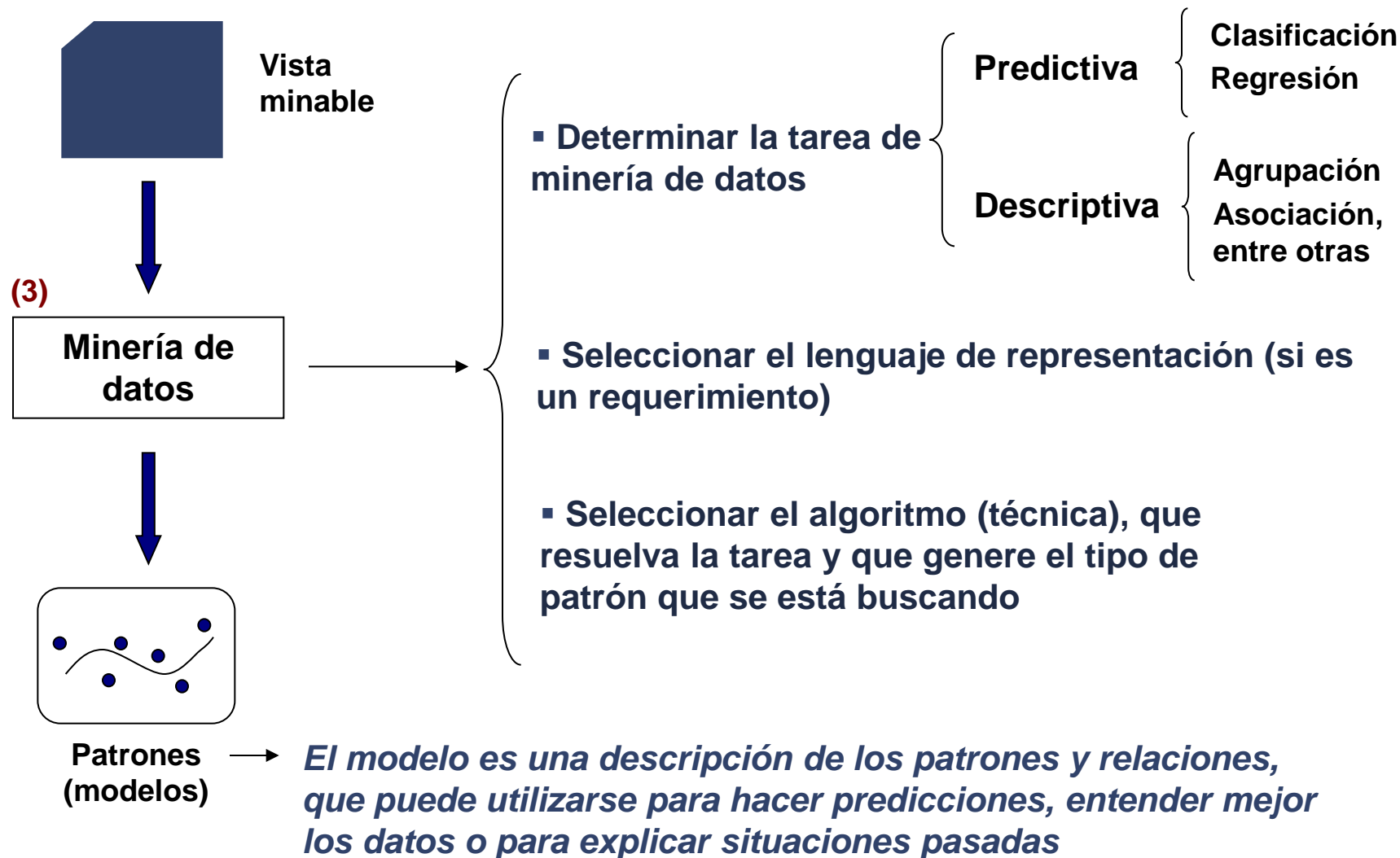
d) Construcción de nuevos atributos

Que puedan resultar más informativos

e) Selección de datos

Para obtener una muestra representativa

Objetivo: Reducir el tamaño de los datos y obtener los más informativos



Algunas técnicas:

Técnicas	Tarea			
	Clasificación	Regresión	Agrupación	Asociación
Basadas en árboles de decisión (ID3, C4.5, CART, CHAID, REPTree, entre otros)	✓	✓		
Basadas en reglas (PRISM, RIPPER, CN2, entre otros)	✓			
K-vecinos más cercanos	✓	✓		
Redes bayesianas	✓			
Redes neuronales	✓	✓	✓	
Máquinas de soporte vectorial	✓	✓		
K-medias, K-medoide			✓	
Apriori				✓
Algoritmos genéticos	✓	✓	✓	✓

1) Decisión sobre si jugar un deporte en un día determinado basado en datos climáticos

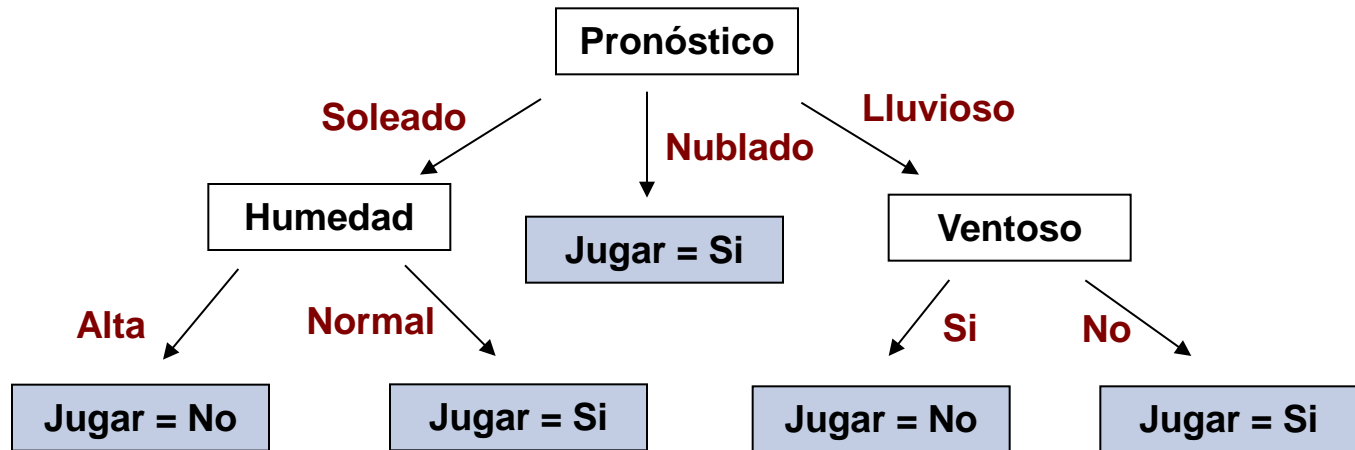
- Tarea: Clasificación
- Algoritmo: C4.5 (*árboles de decisión como lenguaje de representación*)

Vista minable:

ID	PREVISIÓN	TEMPERATURA	HUMEDAD	VENTOSO	JUGAR
1	Soleado	Caliente	Alta	No	No
2	Soleado	Caliente	Alta	Si	No
3	Nublado	Caliente	Alta	No	Si
4	Lluvioso	Suave	Alta	No	Si
5	Lluvioso	Fría	Normal	No	Si
6	Lluvioso	Fría	Normal	Si	No
7	Nublado	Fría	Normal	Si	Si
8	Soleado	Suave	Alta	No	No
9	Soleado	Fría	Normal	No	Si
10	Lluvioso	Suave	Normal	No	Si
11	Soleado	Suave	Normal	Si	Si
12	Nublado	Suave	Alta	Si	Si
13	Nublado	Caliente	Normal	No	Si
14	Lluvioso	Suave	Alta	Si	No

Minería de datos - ejemplos

➔ *Modelo obtenido:*



Nodo raíz y nodos internos	➔	Particiones sobre atributos
Arcos	➔	Posibles valores del atributo.
Nodos hojas	➔	Predicciones/Clasificaciones

- Otros algoritmos: ID3, C5.0, CART, CHAID, . . .

2) Identificar caracteres manuscritos a partir de imágenes

- Tarea: Clasificación
- Algoritmo: Red neuronal perceptron (*funciones lineales como lenguaje de representación*)

$$\text{Discriminante lineal: } \begin{cases} ax_1 + bx_2 + c \geq 0 \rightarrow \text{Clase 1} \\ ax_1 + bx_2 + c < 0 \rightarrow \text{Clase 2} \end{cases}$$

Conjunto de datos:

Imagen 1

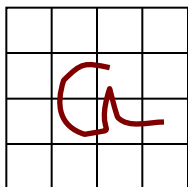


Imagen 2

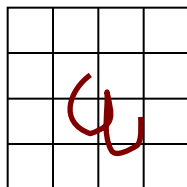


Imagen 3

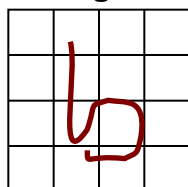


Imagen 4

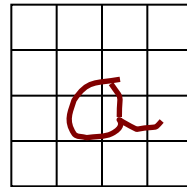
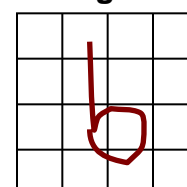


Imagen 8



Clase \rightarrow "a"

"a"

"b"

"a"

"b"

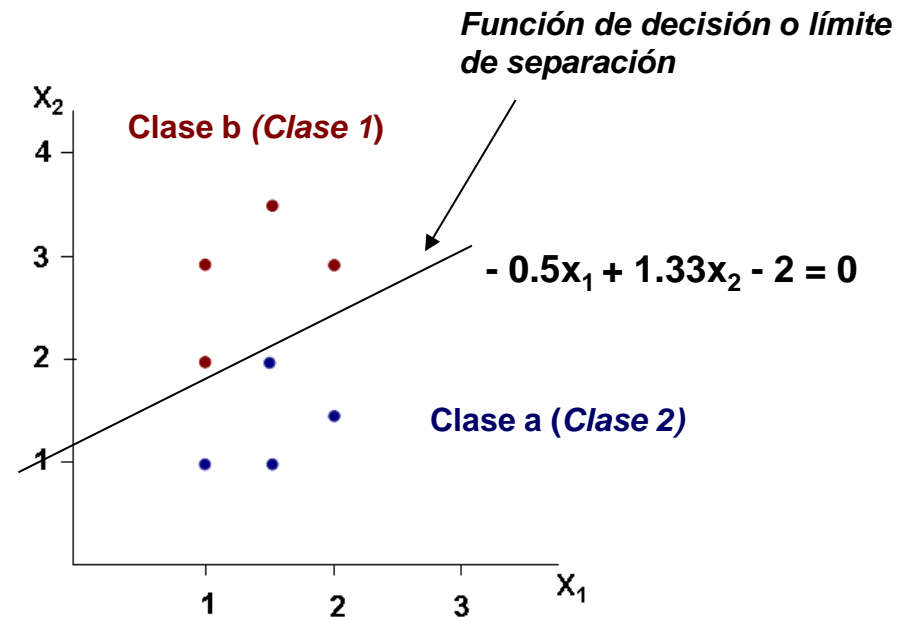


Vista minable:

X_1	X_2	CLASE
1.0	1.0	a
1.5	1.0	a
1.5	3.5	b
2.0	1.5	a
1.0	2.0	b
1.0	3.0	b
1.5	2.0	a
2.0	3.0	b

➡ **Modelo obtenido:**

$$\begin{cases} -0.5x_1 + 1.33x_2 - 2 \geq 0 \rightarrow \text{Clase b} \\ -0.5x_1 + 1.33x_2 - 2 < 0 \rightarrow \text{Clase a} \end{cases}$$



- Otras técnicas: máquinas de soporte vectorial, discriminante de Fisher...

3) Analizar las peticiones de servicios médicos que se realizan en un centro hospitalario.

- **Tarea: Análisis de asociación**
- **Algoritmo: Apriori (*reglas de asociación como lenguaje de representación*)**

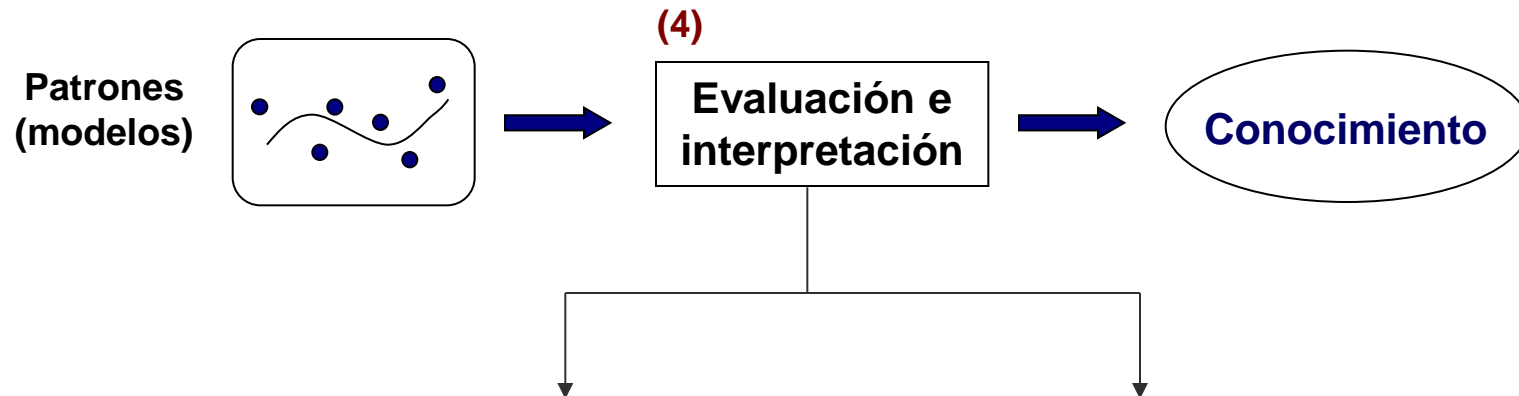
Vista minable:

Paciente	Perfil 20	Orina	Heces	Colesterol	VIH	Glicemia
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

➔ *Modelo obtenido:*

- Si Colesterol (3) entonces Heces (3) conf: (1)
 - Si Orina (4) entonces Perfil 20 (3) conf: (0.75)
 - Si Heces (4) entonces Orina (3) conf: (0.75)
 - Si Heces (4) entonces Colesterol (3) conf: (0.75)
-
- Otros algoritmos: TERTIUS, AprioriDP, ECLAT, FP-growth, . . .

Evaluación e interpretación



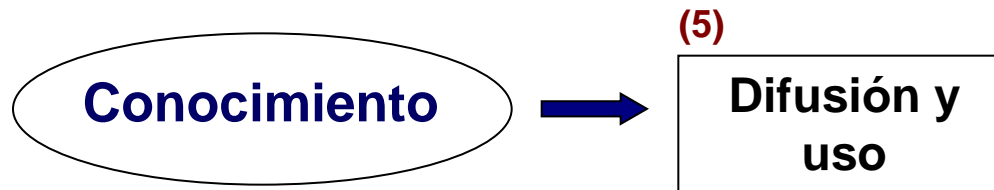
Las medidas de evaluación dependen del tipo de tarea a realizar

Existen varios criterios, ya que en general los patrones minados deben poseer tres cualidades:

- **Precisos**
- **Comprensibles**
- **Interesantes (útiles y novedosos)**

Es importante contrastar el conocimiento adquirido con cualquier conocimiento previo que esté disponible

Verificación con los expertos

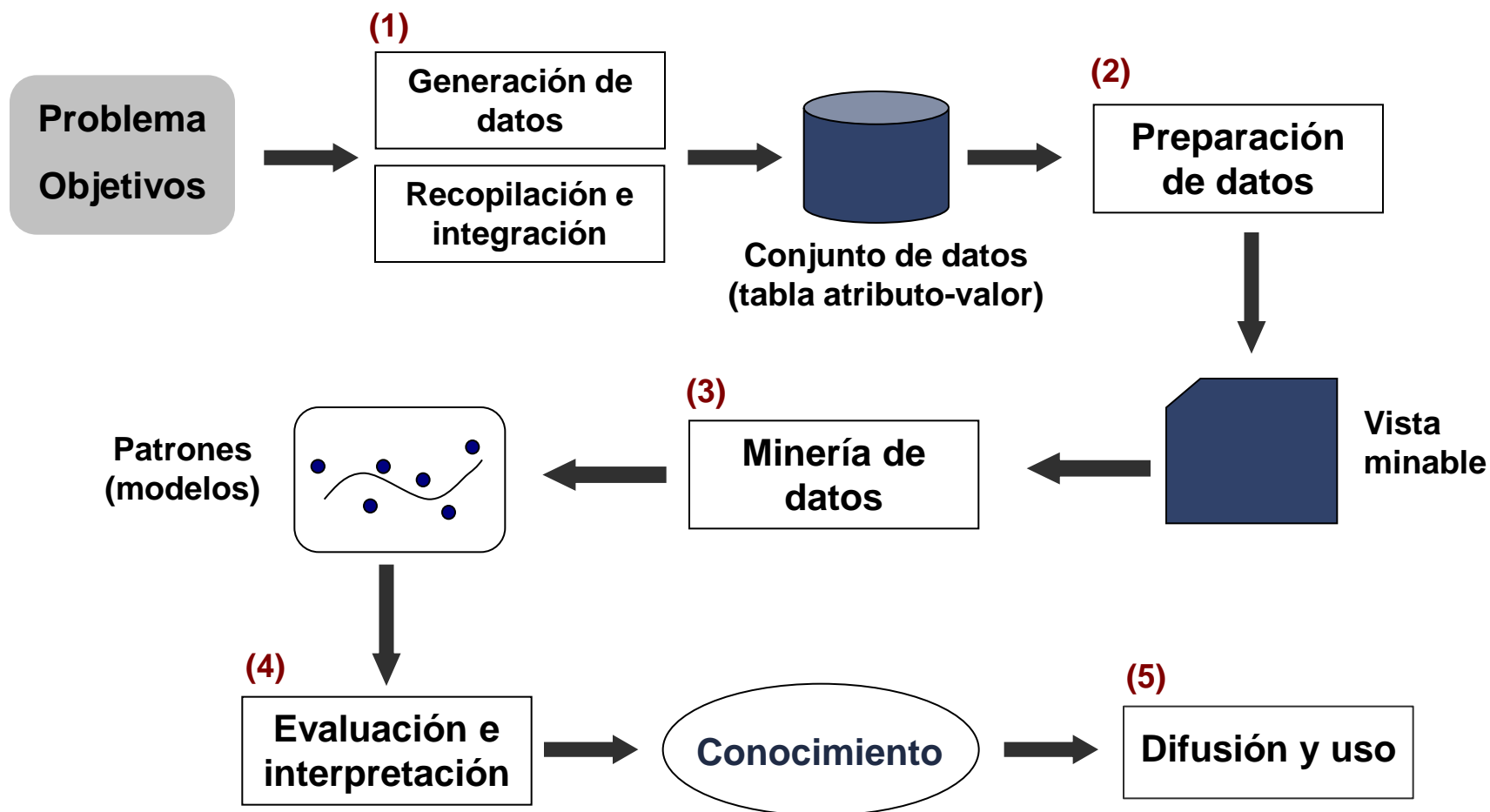


El modelo una vez validado puede:

- Utilizarse para el desarrollo de aplicaciones de apoyo a la toma de decisiones
- Aplicarse a otros conjuntos de datos
- Ser utilizado por otras aplicaciones

Importante —→ Su difusión

- Intranet de la organización
- Eventos
- Publicaciones, entre otras



Problema 1: Renovación de pólizas de seguro HCM

- Uno de los principales tipos de póliza es el HCM (hospitalización, cirugía y maternidad), el cual permite proteger a los asegurados en estos tres ramos. Las aseguradoras deben garantizar su funcionamiento y solvencia económica, su única fuente de ingresos es percibida mediante el pago de las primas de las pólizas por parte de sus asegurados.
- Importante para las empresas es mantener a sus clientes satisfechos con los servicios prestados para de esta forma maximizar las ganancias que puedan presentarse a raíz del pago mensual de primas. Si la empresa de seguros no es capaz de mantener sus clientes, con el pasar del tiempo incurrirá en pérdidas. Sin embargo, debido a la gran cantidad de asegurados que posee una empresa aseguradora, no es posible contactar con todos y cada uno de los titulares para indicarle los beneficios de renovar su póliza
- El uso de una herramienta que permita conocer si un cliente renovará o no su póliza permitiría a la empresa detectar que factores afectan la toma de esta decisión.

Objetivo: Determinar los clientes que podrían renovar una póliza HCM y los factores que influyen en esta decisión

Tarea de minería de datos: Clasificación

(1) Recolección de datos:

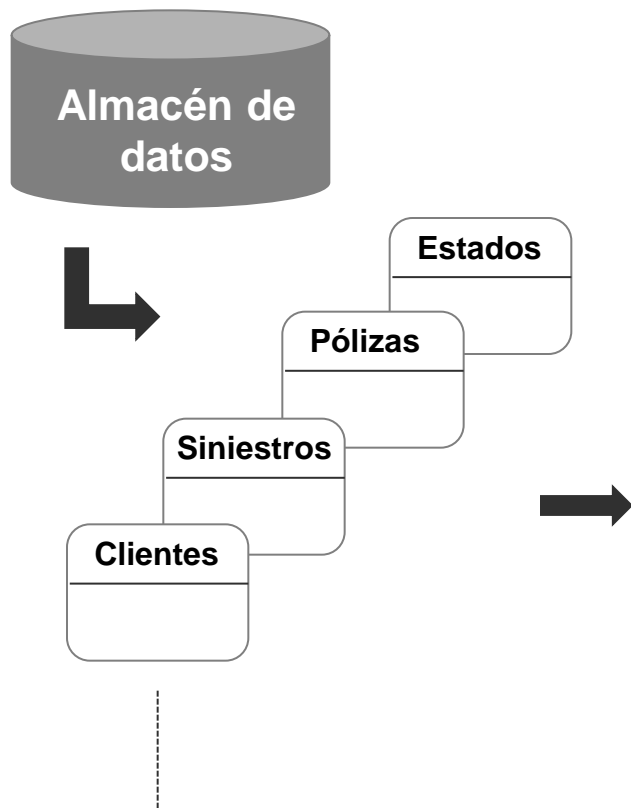


Tabla atributo – valor (extracto):

R2	ASEGURADOS	R2	HIJOS	R2	PADRES	R2	HERMANOS	R2	RENOVO
	1	N		N		N		S	
	1	N		N		N		S	
	1	N		N		N		S	
	2	S		N		N		S	
	1	N		N		N		S	
	1	N		N		N		S	
	2	S		N		N		S	
	2	S		N		N		S	
	2	N		S		N		S	
	1	N		N		N		S	
	1	N		N		N		S	
	1	N		N		N		S	
	1	N		N		N		N	
	1	N		N		N		N	
	1	N		N		N		S	

(2) Limpieza y preparación del conjunto de datos:

- Eliminación de atributos no informativos (identificadores).
- Eliminación de variables con ausencias y valores inconsistentes
- Selección de variables con apoyo de los expertos y con la herramienta WEKA
- Sobre-muestreo de los datos para aumentar la clase minoritaria

Resultado de la fase de preparación de datos:

➔ **Vista minable =** *Tabla atributo-valor con 107046 registros (filas) y 10 variables, incluyendo la clase (Renovó= Si o No)*

Vista minable (extracto):

	STSPOL Nominal	EDAD Numeric	REGION Nominal	CANT_COBERTURAS Numeric	ASEGURADOS Numeric	HIJOS Nominal	PADRES Nominal	HERMANOS Nominal	RENOVO Nominal
	N	41.0	ANZOA...	12.0	1.0	N	N	N	S
	N	28.0	CARAB...	12.0	1.0	N	N	N	N
	N	52.0	ANZOA...	12.0	1.0	N	N	N	S
	N	26.0	TACHIRA	12.0	1.0	N	N	N	S
	R	29.0	DISTRI...	12.0	1.0	N	N	N	S
	N	57.0	LARA	13.0	3.0	S	N	N	N
	R	31.0	CARAB...	10.0	2.0	N	S	N	S
	N	31.0	TACHIRA	13.0	1.0	N	N	N	N
	R	39.0	ARAGUA	13.0	2.0	S	N	N	S
	N	29.0	ARAGUA	12.0	2.0	N	S	N	S
	R	57.0	DISTRI...	9.0	2.0	N	S	N	S
	R	51.0	DISTRI...	12.0	1.0	N	N	N	S
	N	45.0	DISTRI...	12.0	3.0	S	N	N	S
	R	51.0	CARAB...	11.0	1.0	N	N	N	S
	R	36.0	ANZOA...	13.0	1.0	N	N	N	S
	N	65.0	ARAGUA	11.0	1.0	N	N	N	S
	R	29.0	DISTRI...	10.0	1.0	N	N	N	S
	N	40.0	MONA...	12.0	3.0	S	N	N	S
	R	55.0	LARA	12.0	2.0	N	N	N	S
	N	27.0	DISTRI...	13.0	1.0	N	N	N	S
	R	66.0	DISTRI...	12.0	2.0	N	N	N	S
	N	46.0	DISTRI...	12.0	3.0	S	N	N	S
	N	34.0	TACHIRA	14.0	1.0	N	N	N	N
	R	33.0	BOLIVAR	12.0	1.0	N	N	N	S

(3) Minería de datos:

- Tarea: Clasificación
- Algoritmo: C4.5



Modelo obtenido:

```

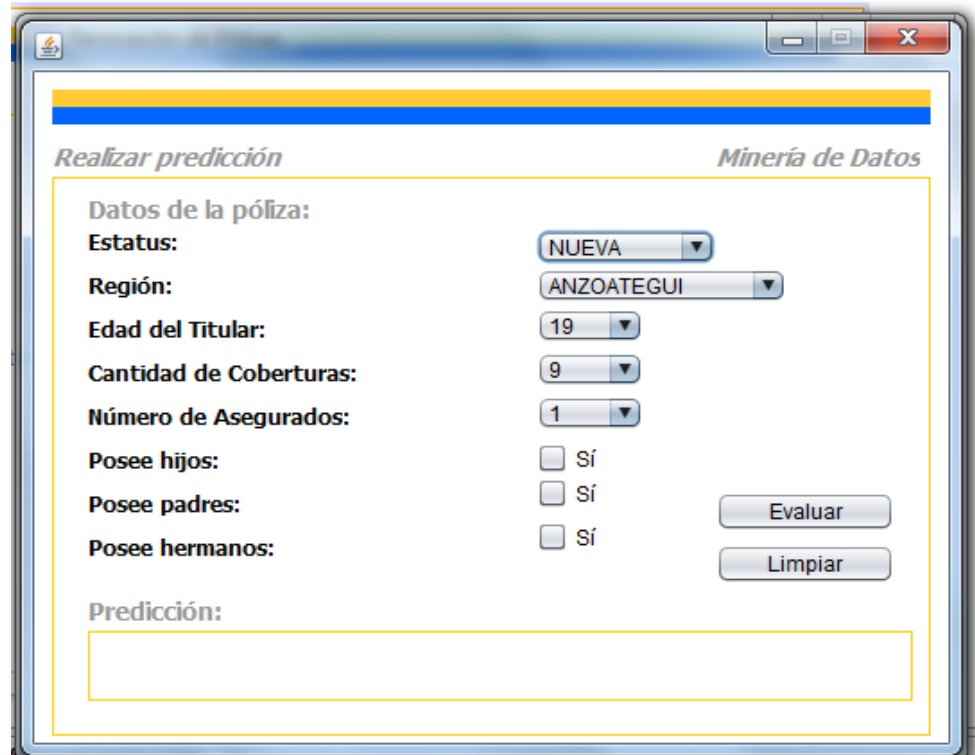
STSPOL = N
| CANT_COBERTURAS <= 10
| | EDAD <= 28
| | | ASEGURADOS <= 1
| | | | REGION = ANZOATEGUI: N (0.0)
| | | | REGION = CARABOBO
| | | | | CANT_COBERTURAS <= 9: S (17.0/4.0)
| | | | | CANT_COBERTURAS > 9: N (164.0/35.0)
| | | | REGION = TACHIRA: S (17.0)
| | | | REGION = DISTRITO CAPITAL: S (86.0/23.0)
| | | | REGION = LARA: S (5.0/1.0)
| | | | REGION = ARAGUA
| | | | | EDAD <= 27: N (17.0/5.0)
| | | | | EDAD > 27: S (4.0)
| | | | REGION = MONAGAS: S (2.0)
| | | | REGION = BOLIVAR: N (4.0)
| | | | REGION = ZULIA: S (12.0/5.0)
| | | | REGION = MERIDA: S (4.0/1.0)
| | | | REGION = PORTUGUESA: N (0.0)
| | | ASEGURADOS > 1: S (28.0/2.0)
| | EDAD > 28: S (667.0/155.0)
| CANT_COBERTURAS > 10: S (28306.0/4445.0)
STSPOL = R: S (44246.0)
  
```

(4) Evaluación e interpretación:

- Medida de rendimiento: exactitud predictiva
- Técnica de evaluación: validación cruzada de 10 particiones
- Instancias correctamente clasificadas: 84,96%

(5) Difusión y uso:

Se desarrolló una aplicación en lenguaje JAVA, que le indica al usuario si la póliza será o no renovada, a partir de los valores de la variables seleccionadas.



Realizar predicción *Minería de Datos*

Datos de la póliza:

Estatus: NUEVA ▼

Región: ANZOATEGUI ▼

Edad del Titular: 19 ▼

Cantidad de Coberturas: 9 ▼

Número de Asegurados: 1 ▼

Posee hijos: ☐ Sí

Posee padres: ☐ Sí

Posee hermanos: ☐ Sí

Evaluar

Limpiar

Predicción:

Problema 2: Evaluación de la morfología de espermatozoide humano

- La evaluación de la morfología espermática suministra información valiosa para la toma de decisiones en los procedimientos de reproducción asistida.
- Sin embargo, en la práctica resulta ser una tarea compleja, difícil de enseñar, con gran variabilidad y dificultad para replicar los resultados.
- Con el fin de apoyar a los profesionales de la salud en esta actividad, sería de gran utilidad contar con sistemas que, de manera automática, realicen la evaluación morfológica a partir de imágenes digitales de muestras de semen.

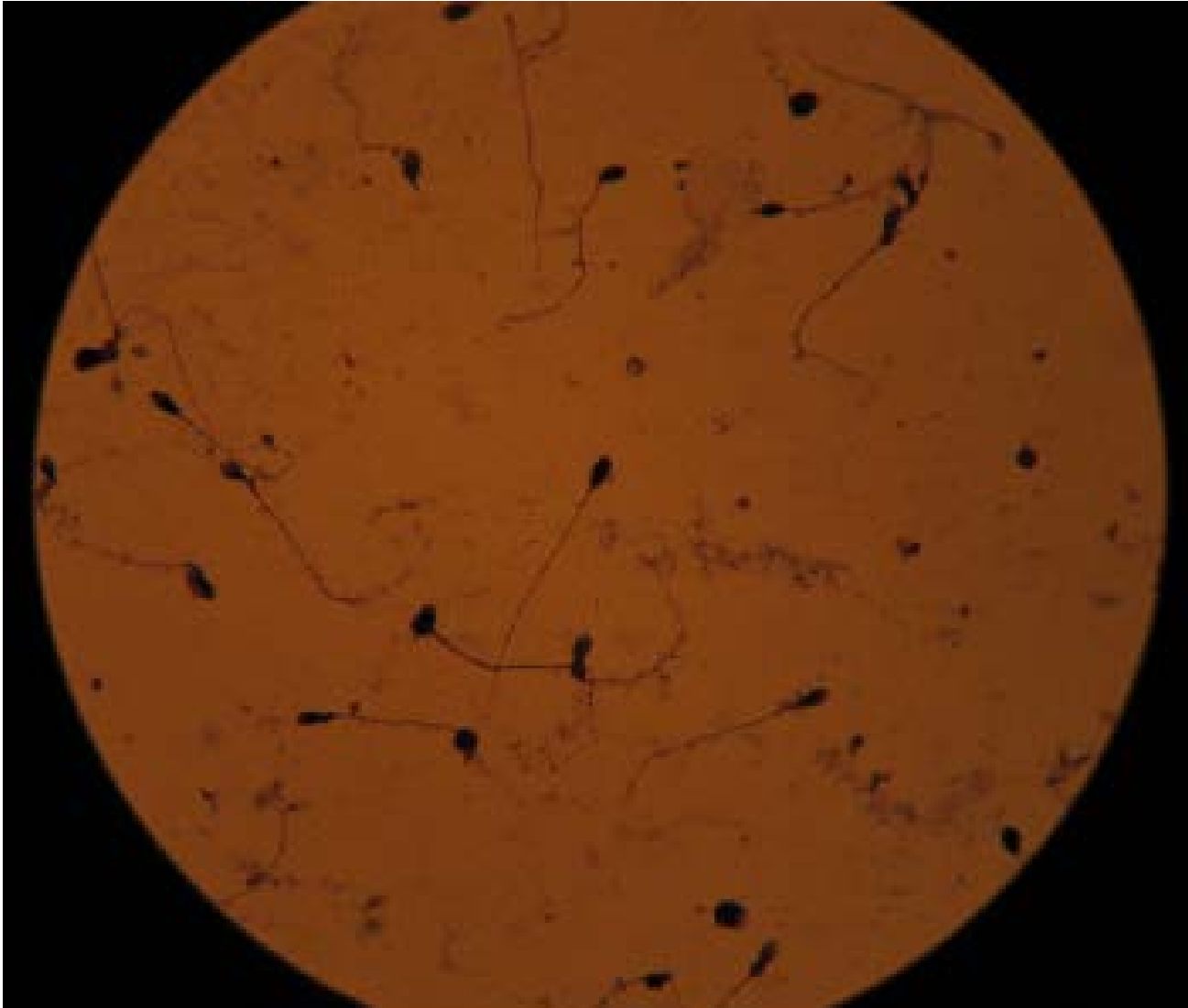
Objetivo: Realizar, de manera automática, la evaluación de la morfología de cabeza de espermatozoide humano a partir de imágenes digitales

Tarea de minería de datos: Clasificación

(1) Generación de los datos (tabla atributo-valor):

- ➡ Se utilizaron 12 láminas de muestras de semen preparadas siguiendo los procedimientos convencionales de tinción.
- ➡ Conjunto de datos: 224 imágenes digitales capturadas en formato JPEG con una resolución de 1600x1200 píxeles.

Una imagen:



➡ Se desarrollaron algoritmos para la segmentación de los espermatozoides en la imagen y para la extracción de características.

Variables	Descripción
Área cabeza	Número de píxeles en la cabeza
Área acrosoma	Número de píxeles en región acrosómica
Área post-acrosoma	Número de píxeles en región post-acrosómica
Eje horizontal cabeza	Longitud en píxeles del eje horizontal de la cabeza
Eje vertical cabeza	Longitud en píxeles del eje vertical de la cabeza
Elipticidad	División entre eje horizontal y eje vertical de la cabeza
Diferencia Ei-Cr	Diferencia en píxeles entre contorno real y contorno de elipse ideal calculada a partir de los ejes de la cabeza
Área cabeza	Número de píxeles en la cabeza
Clasificación	Normal, Anormal

➡ Sobre la base de estos algoritmos se construyó una herramienta de adquisición de conocimiento para generar el conjunto de datos (tabla atributo-valor).

➡ Con esta herramienta, y con la participación de expertos en la realización de este tipo de examen, se generó un conjunto de 868 datos.

CME - Clasificador de Morfología Espermática

Archivo Ver Ir a Procesamiento Ayuda

Clasificación Estadísticas

Ir a Capture_00014.JPG Vista Imagen Original RGB

Imagen Procesada

Clasificación

- ☐ No Clasificado
- ☐ Normal
- ☒ Anormal
- ☐ No es un espermatozoide

Cabeza: Amorfa

Cuello: Normal

Cola: Normal

Observaciones:

Guardar

Características

Atributo	Valor
ID	726_13_0
Área de Cabeza	598
Área de Acro...	312
Área de Post...	286
Eje horizontal	37
Eje vertical	21
Elipticidad	1.761904761...
Diferencia ent...	92
Perímetro ide...	189.0184327...

Espermatozoide

Segmentación

Herramienta →

Inicio CME - Clasificador de ... 07:09 p.m.

Tabla atributo-valor (extracto):

Relation: Morfologia.txt

No.	AreaCabeza Numeric	AreaAcrosoma Numeric	AreaPostAcrosoma Numeric	CocienteAcrosomaAtotal Numeric	EjeHorizontal Numeric	EjeVertical Numeric	Elipticidad Numeric	DiferenciaEiCr Numeric	Clasificación Nominal
1	427.0	175.0	252.0	0.41	34.0	18.0	1.89	122.0	Anormal
2	397.0	165.0	232.0	0.42	35.0	14.0	2.5	60.0	Anormal
3	511.0	325.0	186.0	0.64	29.0	22.0	1.32	135.0	Normal
4	566.0	327.0	239.0	0.58	36.0	19.0	1.89	62.0	Anormal
5	369.0	210.0	159.0	0.57	27.0	16.0	1.69	74.0	Normal
6	369.0	152.0	217.0	0.41	28.0	17.0	1.65	47.0	Anormal
7	462.0	262.0	200.0	0.57	32.0	19.0	1.68	69.0	Anormal
8	386.0	203.0	183.0	0.53	30.0	15.0	2.0	46.0	Anormal
9	435.0	167.0	268.0	0.38	31.0	19.0	1.63	72.0	Anormal
10	439.0	194.0	245.0	0.44	31.0	18.0	1.72	82.0	Anormal
11	342.0	124.0	218.0	0.36	29.0	14.0	2.07	51.0	Anormal
12	365.0	160.0	205.0	0.44	34.0	14.0	2.43	94.0	Anormal
13	515.0	279.0	236.0	0.54	32.0	20.0	1.6	85.0	Normal
14	425.0	138.0	287.0	0.32	27.0	20.0	1.35	89.0	Anormal
15	437.0	248.0	189.0	0.57	28.0	19.0	1.47	57.0	Anormal
16	424.0	138.0	286.0	0.33	30.0	19.0	1.58	54.0	Anormal
17	379.0	154.0	225.0	0.41	25.0	18.0	1.39	70.0	Normal
18	429.0	265.0	164.0	0.62	31.0	18.0	1.72	95.0	Anormal
19	447.0	200.0	247.0	0.45	29.0	19.0	1.53	72.0	Anormal
20	550.0	312.0	238.0	0.57	42.0	18.0	2.33	127.0	Anormal
21	510.0	245.0	265.0	0.48	31.0	22.0	1.41	88.0	Anormal
22	456.0	222.0	234.0	0.49	31.0	19.0	1.63	129.0	Anormal
23	472.0	162.0	310.0	0.34	31.0	20.0	1.55	81.0	Anormal
24	551.0	313.0	238.0	0.57	36.0	19.0	1.89	88.0	Anormal
25	499.0	211.0	288.0	0.42	32.0	19.0	1.68	67.0	Anormal

(2) Limpieza y preparación del conjunto de datos:

- Detección y eliminación de valores extremos (*outliers*)
- Generación de una nueva variable que representa el cociente entre los atributos Área Acrosoma y Área Cabeza, ya que según la OMS el acrosoma de un espermatozoide normal debería ocupar entre 40% y 70% de área de la cabeza.
- Selección de atributos, dando como resultado que la característica “Ei – Cr” podía no ser considerada.

Resultado de la fase de preparación de datos:

➔ Vista minable = *Tabla atributo-valor con 832 registros (filas) y 9 variables, incluyendo la clase (normal o anormal).*

(3) Minería de datos:

- Tarea: Clasificación
- Algoritmo: RIPPER



Modelo obtenido:

**(Acrosoma/Atotal \geq 0.51) and (Elipticidad \leq 1.55) and
(Area_cabeza \leq 556) and (Area_cabeza \geq 493) \Rightarrow**

Clasificacion = Normal (55.0/16.0)

**(Area_post-acrosoma \leq 216) and (Elipticidad \leq 1.75) and
(Area_acrosoma \geq 205) and (Area_acrosoma \leq 274) \Rightarrow**

Clasificacion = Normal (100.0/42.0)

\Rightarrow Clasificacion = Anormal (677.0/57.0)

(4) Evaluación e interpretación:

- Medida de rendimiento: exactitud predictiva
- Técnica de evaluación: validación cruzada de 10 particiones.
- Resultados obtenidos: porcentaje de aciertos o exactitud predictiva del 82 %
- Sin embargo, la matriz de confusión muestra que el mayor porcentaje de errores se presenta al clasificar instancias normales (falsos negativos).

(5) Difusión y uso:

- Los resultados obtenidos hasta los momentos han sido difundidos a través de eventos en el área de inteligencia artificial e informática.

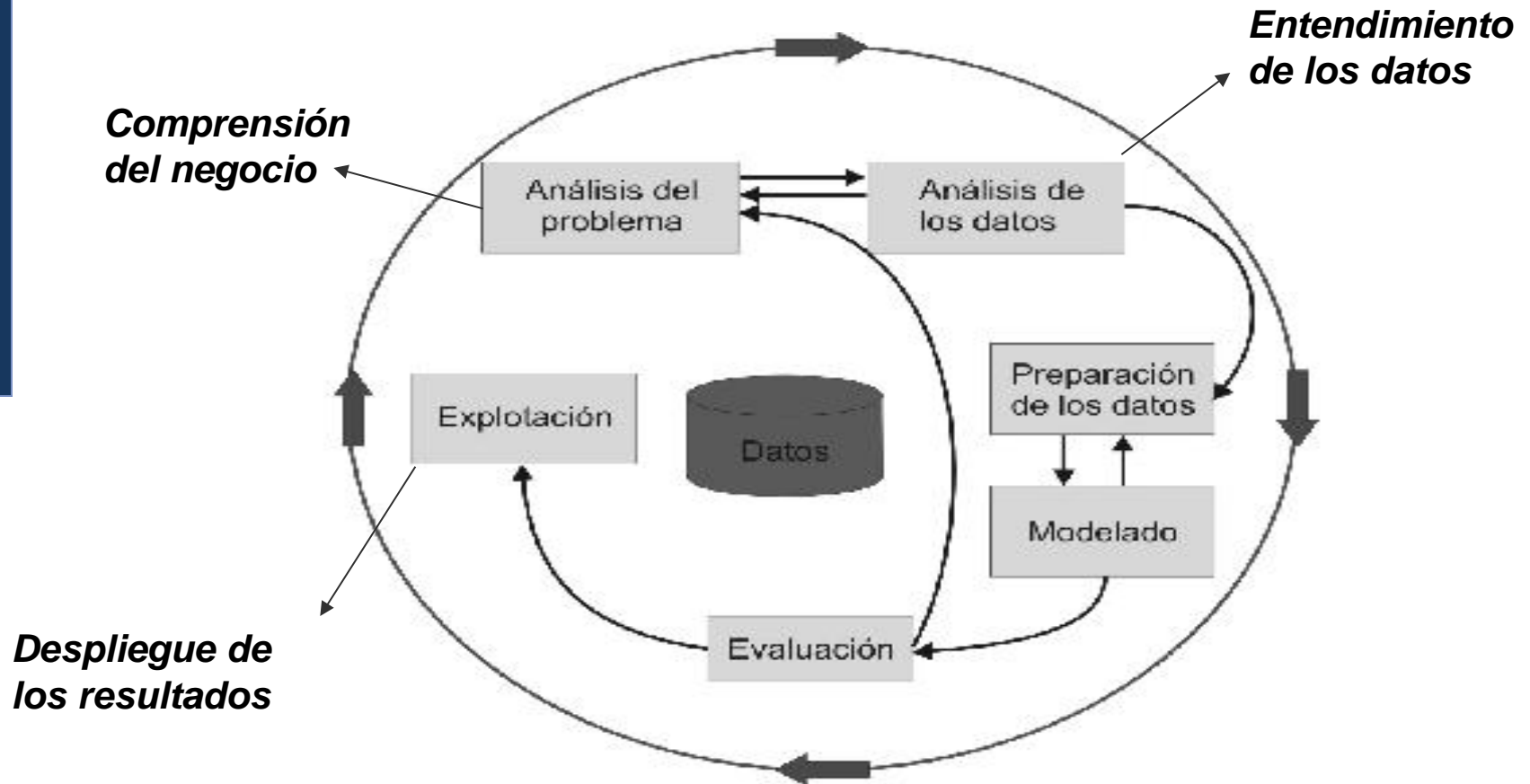
- **Surgen ante la necesidad de una aproximación sistemática para la realización de proyectos de Minería de Datos en las organizaciones.**
- **Se basan en los pasos que deben llevarse a cabo para el descubrimiento de conocimiento a partir de datos.**
- **Facilitan la planificación y dirección de proyectos**
- **Actualmente las más utilizadas son:**

☞ **CRISP- DM (*CRoss Industry Standard Process for Data Mining*):** propuesta por un consorcio de empresas europeas.

☞ **SEMMA (*Sample, Explore, Modify, Model, Assess*):** propuesta por SAS Institute

Metodología CRISP-DM:

- El proceso está organizado en seis fases, que se comunican de manera iterativa.



- Cada fase se estructura en varias tareas.

(1)

Comprensión del negocio



*Se establecen los
objetivos y
requerimientos
desde una
perspectiva no
técnica*

- *Determinar los objetivos del negocio (Contexto inicial, objetivos, criterios de éxito)*
- *Evaluar la situación (Inventario de recursos, requerimientos, supuestos, terminologías propias del negocio,...)*
- *Establecer los objetivos de la minería de datos (objetivos y criterios de éxito)*
- *Generar el plan del proyecto (plan, herramientas, equipo y técnicas)*

(2)

Comprensión de los datos



Familiarización con los datos tomando en cuenta los objetivos del negocio

- *Recopilación inicial de los datos*
- *Descripción de los datos*
- *Exploración de los datos*
- *Verificación de la calidad de los datos*

(3)

Preparación de los datos



Se obtiene la vista minable

- *Selección de los datos*
- *Limpieza de datos*
- *Construcción de datos*
- *Integración de datos*
- *Formateo de datos*

(4)

Modelado



Se aplican las técnicas de minería de datos a la vista minable

- *Selección de la técnica de modelado*
- *Diseño de la evaluación*
- *Construcción del modelo*
- *Evaluación del modelo*

(5)

Evaluación



De los modelos obtenidos en la fase de modelado, para determinar si son útiles a las necesidades del negocio

- *Evaluación de resultados*
- *Revisar el proceso*
- *Establecer los siguientes pasos o acciones*

(6)

Explotación



Explotar la utilidad de los modelos obtenidos, mediante su integración en las tareas de toma de decisiones de la organización

- *Planificación del despliegue*
- *Planificación de la monitorización y del mantenimiento*
- *Generación de informe final*
- *Revisión del proyecto*

- Proyecto de la Universidad de Waikato
- WEKA es una colección de algoritmos de aprendizaje automático para las tareas de minería de datos.
- Los algoritmos pueden ser aplicados directamente a un conjunto de datos o ser llamados desde código Java.
- Contiene herramientas para el pre-procesamiento de los datos y visualización, clasificación, regresión, agrupación y reglas de asociación.
- También permite la incorporación de nuevos algoritmos de aprendizaje
- Sitio Web: <http://www.cs.waikato.ac.nz/ml/weka/>

Formato de archivos para Weka:

Ejemplo: conjunto de datos
weather.arff →

@relation weather

@attribute outlook {sunny, overcast, rainy}

@attribute temperature real

@attribute humidity real

@attribute windy {TRUE, FALSE}

@attribute play {yes, no}

@data

sunny,85,85,FALSE,no

sunny,80,90,TRUE,no

overcast,83,86,FALSE,yes

rainy,70,96,FALSE,yes

rainy,68,80,FALSE,yes

rainy,65,70,TRUE,no

overcast,64,65,TRUE,yes

sunny,72,95,FALSE,no

sunny,69,70,FALSE,yes

rainy,75,80,FALSE,yes

sunny,75,70,TRUE,yes

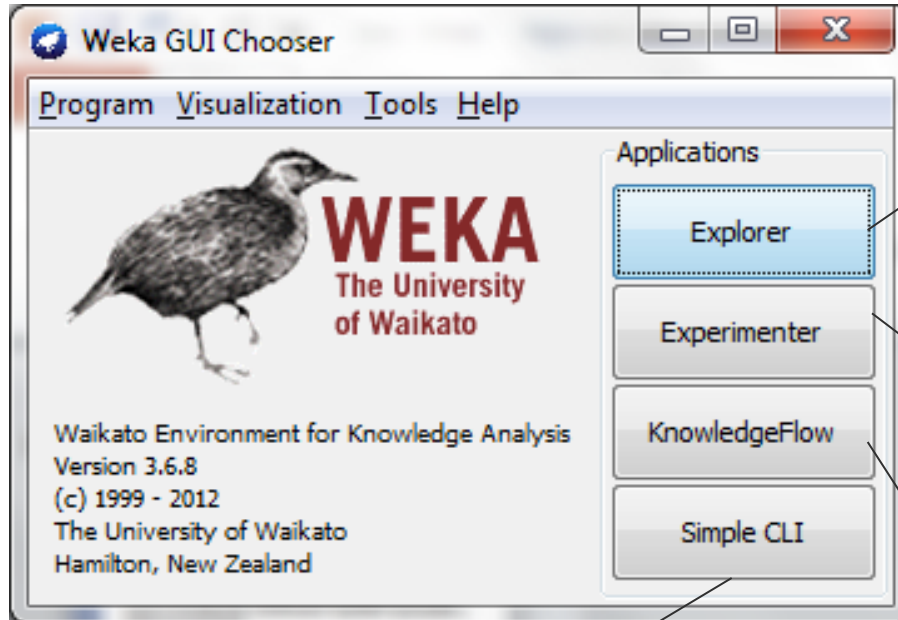
overcast,72,90,TRUE,yes

overcast,81,75,FALSE,yes

rainy,71,91,TRUE,no

Herramientas de MD: Weka

Interfaces:



A través de esta interfaz se pueden realizar diversas operaciones (preparación y visualización de datos, selección de atributos, clasificación, regresión, agrupación y análisis de asociación), sobre un solo archivo de datos

Interfaz de usuario para comparar el rendimiento de algoritmos de aprendizaje.

Interfaz basada en componentes que tiene una funcionalidad similar a Explorer.

Esta interfaz permite introducir comandos para realizar operaciones de forma directa, a través de consola