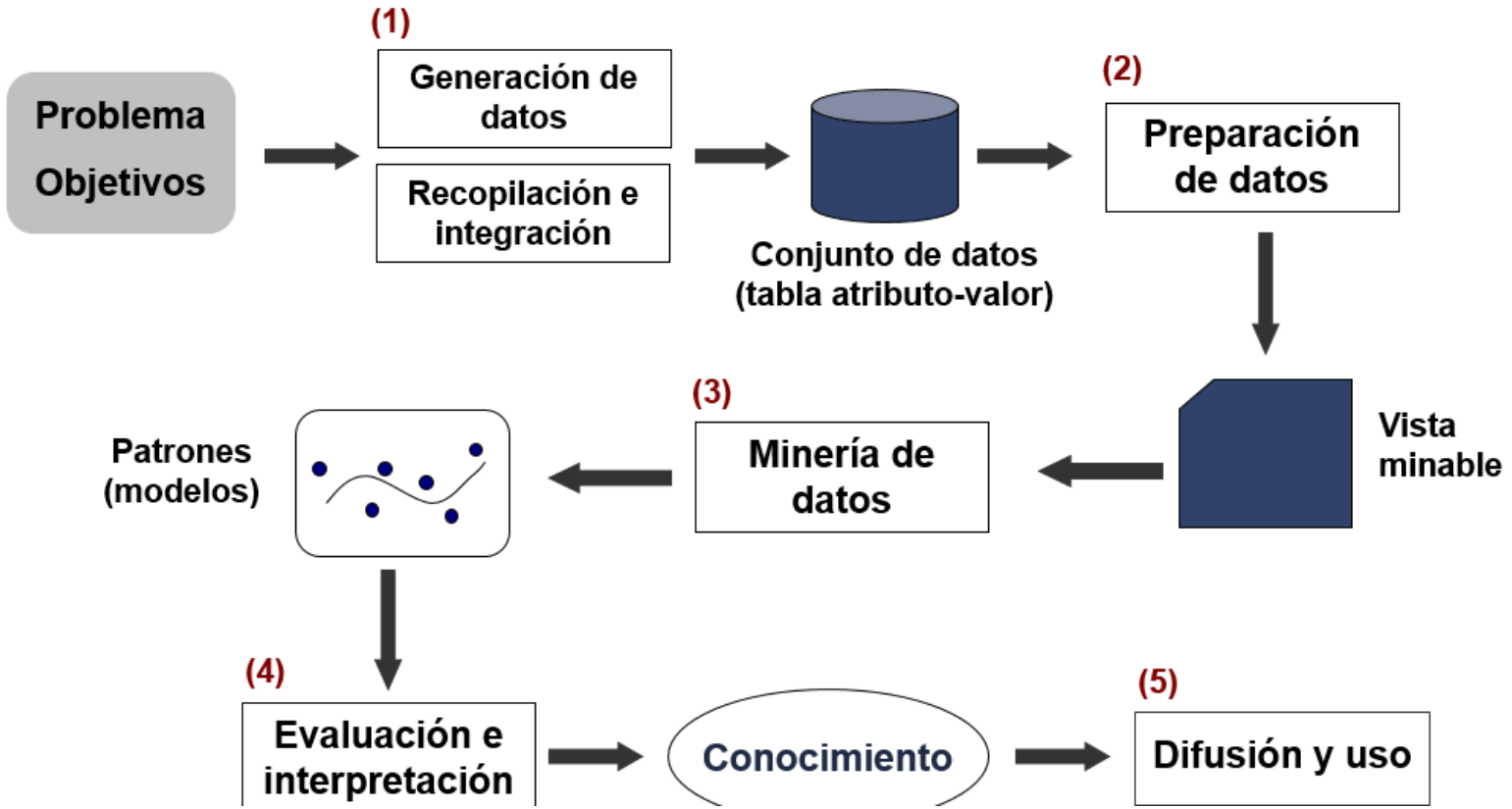
Escuela de Computación UCV

Minería de Datos

Clase : Metodologías , Preparación de los datos

Proceso de minería de datos:



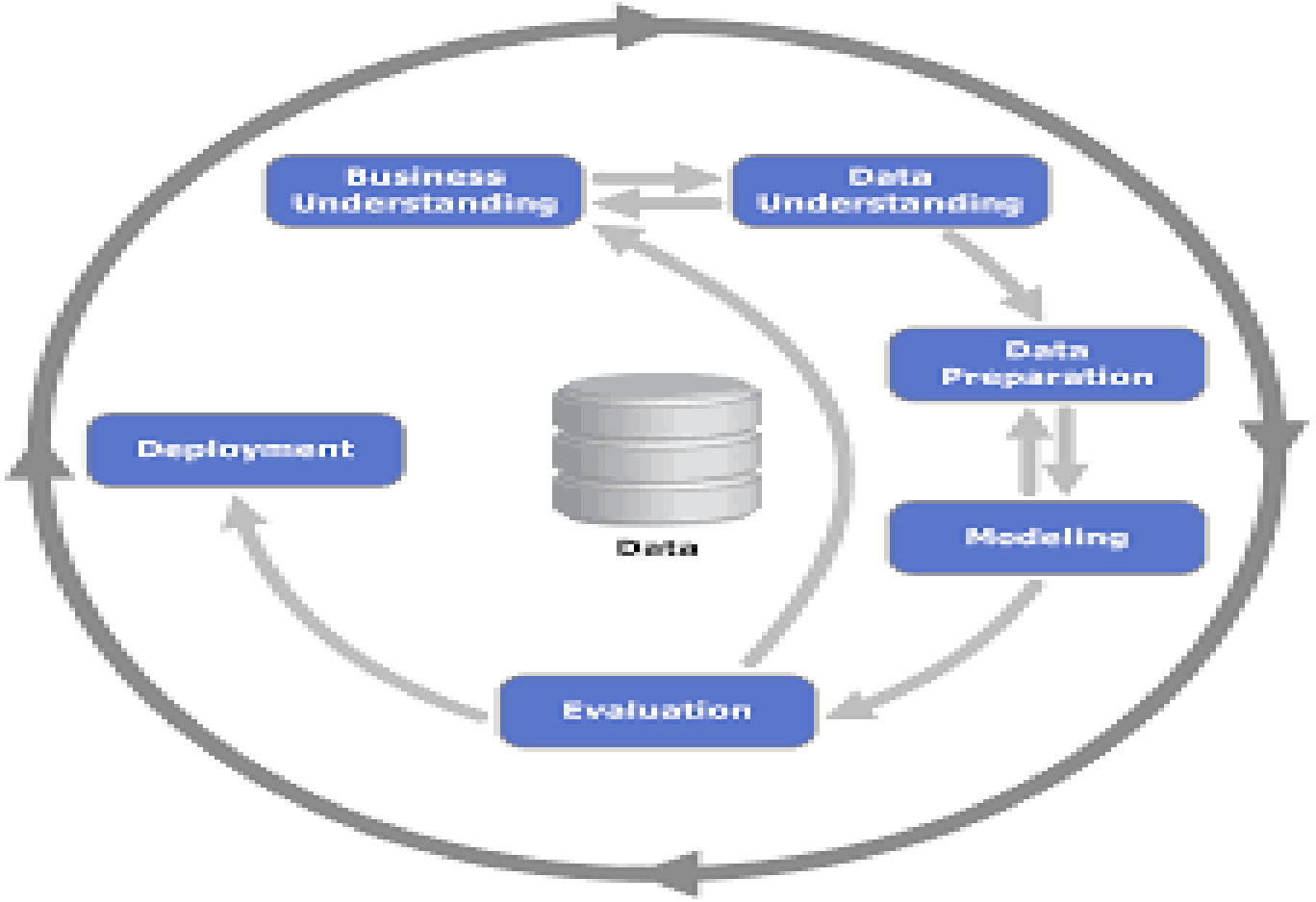
Metodologías para minería de datos:

- 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.
- Facilita 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.



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)*

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*

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*

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*

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*

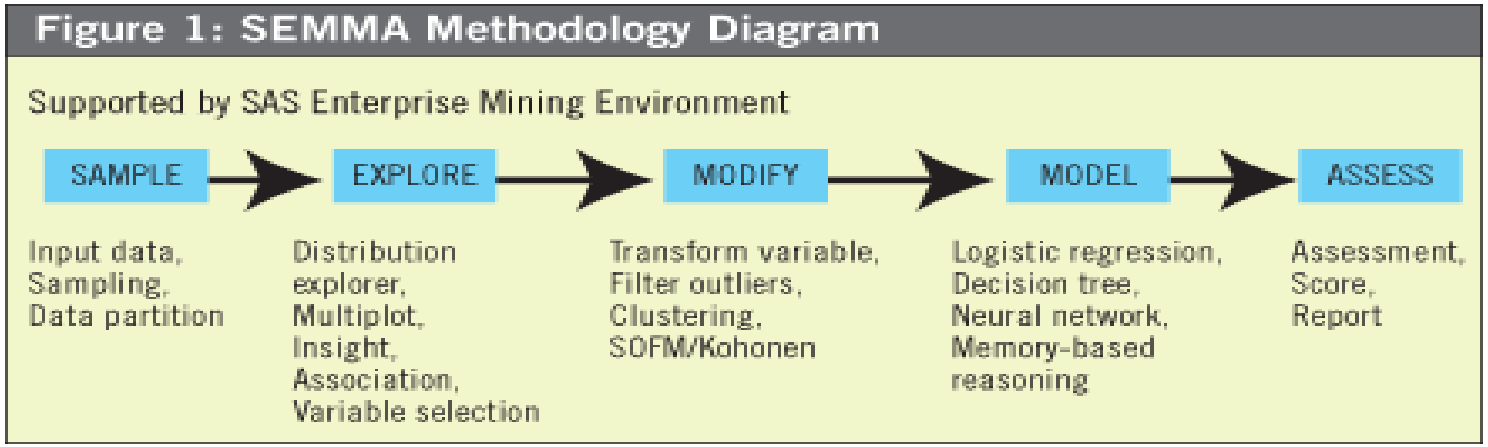
Explotación

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*

- *En general, los proyectos de minería de datos no culminan con la implantación del modelo. Se deben documentar y presentar los resultados de manera comprensible con el fin de lograr un incremento del conocimiento en la organización.*
- *La fase de explotación debe asegurar el mantenimiento de la aplicación y la difusión de los resultados*

Metodología SEMMA



Sample → Tomar un subconjunto de datos que sea lo suficientemente grande para ser una muestra representativa pero no demasiado grande para que el conjunto de datos se pueda procesar fácilmente

Explore → Esta etapa consiste en la exploración de los datos mediante la búsqueda de anomalías y posibles patrones.

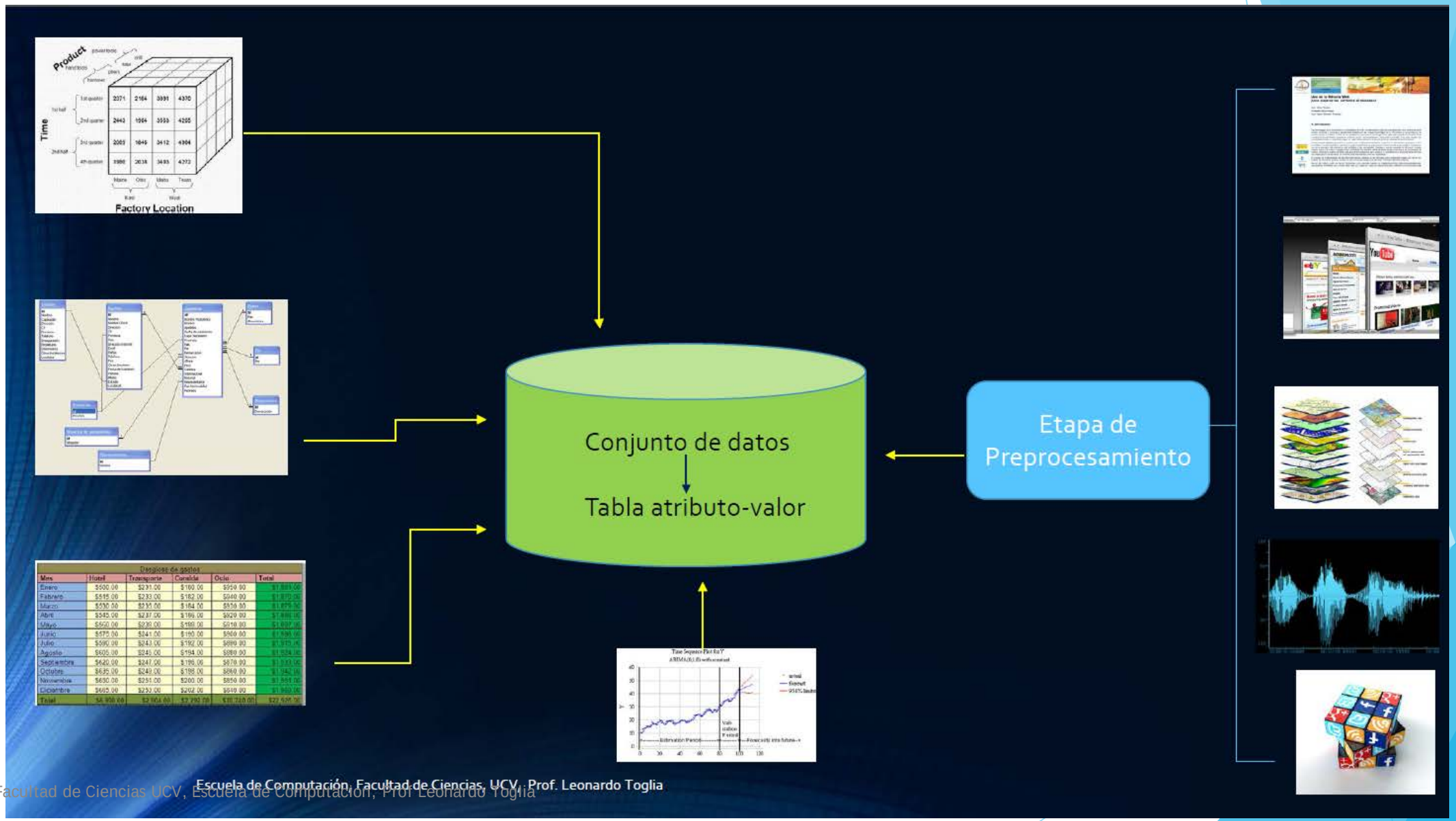
Modify → Crear y transformar variables, o eliminar las innecesarias

Model → Seleccionar y aplique un modelo que se adapte mejor a su situación ya sus datos

Assess

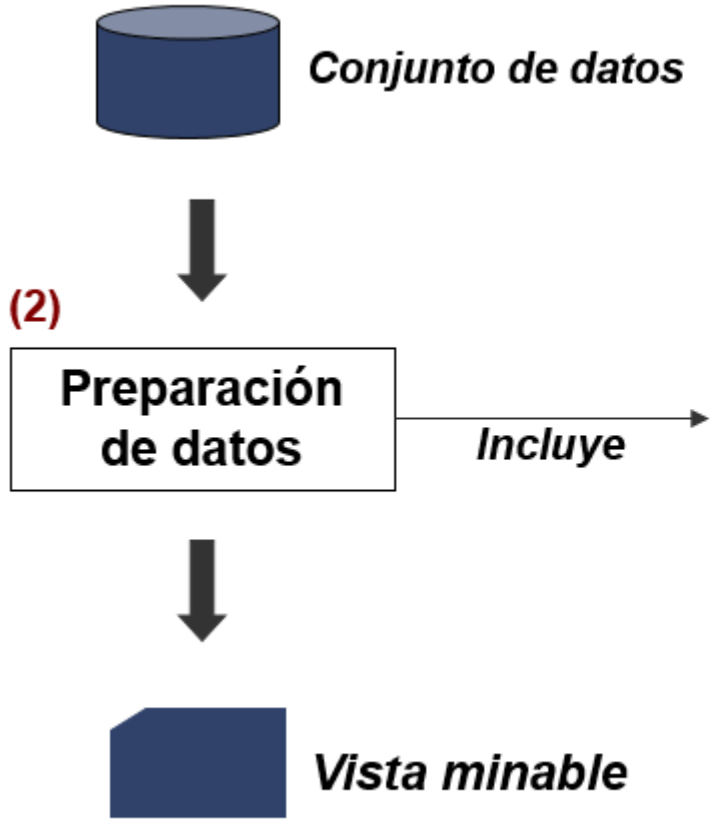
Determinar si los resultados son útiles y confiables.
Probar los resultados con datos conocidos u otra muestra

Preparación de los datos



Escuela de Computación, Facultad de Ciencias, UCV , Prof. Leonardo Toglia

¿Como preparar los datos para mejorar los resultados de la minería de datos?



a) Limpieza de los datos

b) Transformaciones

c) Selección de atributos

d) Construcción de nuevos atributos

e) Selección de datos

Las técnicas de preparación al aplicarse antes del paso de MD:

- Mejora la calidad de los patrones minados (en exactitud y/o comprensibilidad)
- Mejora la eficiencia del paso de minería (tiempo requerido para obtener los patrones)

Nota :

Calidad en los datos → Calidad en las decisiones

≈ 60 – 70 % del esfuezo en MD se dedica a la preparación de los datos

¿Cuáles criterios se utilizan para determinar la calidad de los datos?

- Exactitud (*accuracy*). Los datos no contienen errores, los valores son los esperados, datos exactos.
- Completitud (*completeness*). Todos los datos relevantes y de interés están registrados.
- Consistencia (*consistency*). No hay discrepancia en los datos, son consistentes a través de diferentes fuentes

Existen otros criterios como :

- *Puntualidad*: los datos se encuentran actualizados
- *Interpretabilidad*: facilidad en la comprensión de los datos
- *Integridad*: los datos son confiables.

Sin embargo, debido a:

- Fallas en los instrumentos/procedimientos de recolección
- Errores al registrar los datos (ejemplo, de forma manual)
- Integración de datos desde diferentes fuentes
- Diferentes definiciones para un atributo
- Suministro de información incorrecta por parte de los usuarios
- Registros duplicados
- Atributos de interés no están totalmente disponibles o no fueron considerados oportunamente
- Mal funcionamiento de equipos, entre otros…..

Los conjuntos de datos pueden contener:

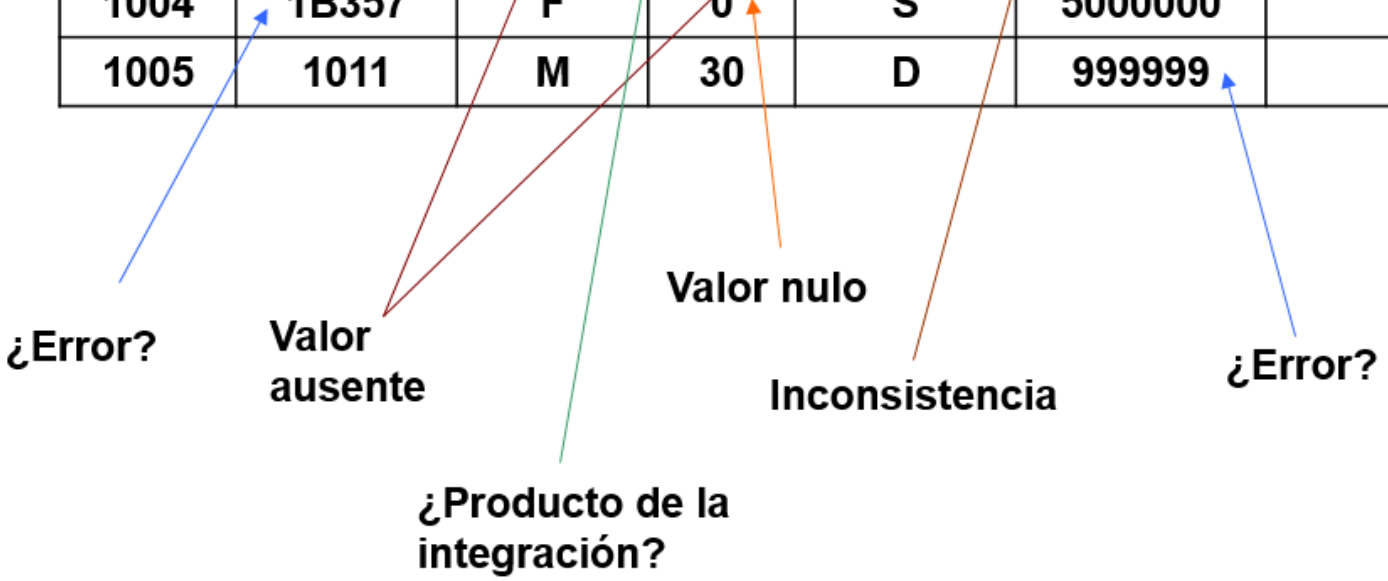
- Valores ausentes
- Valores anómalos (*outliers*)
- Valores no consistentes, datos mal clasificados
- Errores, ruido

El objetivo de la limpieza de los datos:

Detección y corrección de estos errores

Ejemplo

ID Cliente	Código postal	Género	Edad	Estado civil	Ingreso	Cantidad de transacciones
1001	1020-A	M	C	C	750000	5000
1002	1041	F	40		- 400000	4000
1003	1038		45	S	10000000	7000
1004	1B357	F	0	S	5000000	1000
1005	1011	M	30	D	999999	3000



Para empezar, elaborar un tabla resumen de las características de cada variable o atributo:

ATRIBUTO	TIPO	# TOTAL	# NULS	# DIST	MEDIA	DESV.	MODA	MIN	MAX
Código postal	Nominal	10320	150	1672	-	-	"46003"	-	-
Sexo	Nominal	10320	23	6	-	-	"V"	-	-
Estado civil	Nominal	10320	317	8	-	-	Casado	-	-
Edad	Numérico	10320	4	66	42,3	12,5	37	18	87
Total póliza p/a	Numérico	17523	1325	142	737,24	327	680	375	6200
Asegurados	Numérico	17523	0	7	1,31	0,25	1	0	10
Matrícula	Nominal	16324	0	16324	-	-	-	-	-
Modelo	Nominal	16324	1321	2429	-	-	"O. Astra"	-	-

Valores ausentes:

- El valor de la variable no se conoce para algunas instancias

Ejemplo:

Edad	Fecha de nacimiento	Sueldo	Zona
49	10/07/1999		Santa Mónica
	04/02/2005	2500,00	Boleita

- Algunos algoritmos de MD pueden manejar estos datos incompletos
- En general, es mejor realizar previamente el tratamiento de estos datos

Los valores ausentes pueden deberse a errores de medición, mal funcionamiento de equipos, cambios en los procedimientos de recolección, la información no fue suministrada, omisiones involuntarias, entre otras

Pueden representar características relevantes o pueden no existir en la realidad

¿Cómo detectarlos?

En general, el valor ausente se representa con "?". El campo también puede estar vacío.

Las herramientas de minería de datos pueden generar información de cada variable (tabla resumen), donde se indica el porcentaje de ausencias.

¿Cómo tratarlos?

- Ignorar el dato ausente: si la técnica de minería es robusta a las ausencias.

- Crear un nuevo atributo lógico que indique si el valor de la variable original era nulo o no.

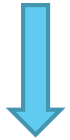
- Por eliminación:

  - *Eliminación de registros*: omitir del análisis los registros con valores ausentes.

  - *Eliminación de atributos*: sobre todo cuando la proporción de nulos es muy alta.

- Por sustitución: Reemplazar el valor ausente por otro valor de acuerdo con ciertos criterios.

Importante: si se realiza una imputación (sustitución), este valor debe estar lo más cercano posible al valor real



La distribución resultante debe ser lo más parecida a la distribución original

Algunas técnicas de imputación:

- Reemplazar el valor ausente por una constante indicada por los expertos o utilizando conocimiento del dominio.

- Sustitución por un valor de tendencia central.

- Reemplazar el valor ausente por un valor generado a partir de la distribución observada de la variable.

- Asignación del más parecido. Imputa el valor ausente de una variable, con el valor que ese atributo toma en casos similares (utilizando técnicas como los vecinos más cercanos).

- Estimación del valor. A través de modelos predictivos, mediante interpolación, entre otros.

Ejemplo:

ID	x1	x2	x3
1	3	2	4
2	1	1	5
3	3	2	1
4	--	2	2
5	--	1	3
6	4	3	4
7	1	1	1
8	3	3	5

a) Por eliminación de registros

Se eliminan los registros 4 y 5

b) Por sustitución por un valor de tendencia central

Para distribuciones simétricas (normal) se recomienda utilizar la media, para distribuciones sesgadas algunos autores recomiendan utilizar la mediana

Media de la variable x1 = 2.5

↓

Se sustituye por este valor →

ID	x1	x2	x3
1	3	2	4
2	1	1	5
3	3	2	1
4	2.5	2	2
5	2.5	1	3
6	4	3	4
7	1	1	1
8	3	3	5

c) Asignación del más parecido

Se calculan las distancias del registro con ausencias, a los demás registros.

Por ejemplo con:

$$\text{Distancia Manhattan} = \sum_{i=1}^{d} \left| (x_i - y_i) \right|$$

Para el registro 4: mas parecido = registro 3

$$d_{Manhattan}(I4,I3) = |(x_{42} - x_{32})| + |(x_{43} - x_{33})| = |(2 - 0)| + |(2 - 1)| = 1$$

Para el registro 5: mas parecidos = registros 1, 2, 7

$$d_{Manhattan}(I5,I1) = |(x_{52} - x_{12})| + |(x_{53} - x_{13})| = |(1 - 2)| + |(3 - 4)| = 2$$

$$d_{Manhattan}(I5,I2) = |(x_{52} - x_{22})| + |(x_{53} - x_{23})| = |(1 - 1)| + |(3 - 5)| = 2$$

$$d_{Manhattan}(I5,I2) = |(x_{52} - x_{72})| + |(x_{53} - x_{73})| = |(1 - 1)| + |(3 - 1)| = 2$$

Valor de x1 para el registro 1 = 3

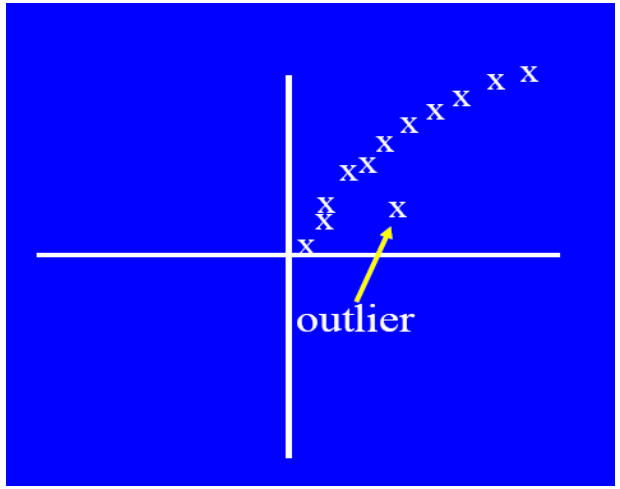
Valor de x1 para el registro 2 = 1

Valor de x1 para el registro 7 = 1

Media = 1.67

ID	x1	x2	x3
1	3	2	4
2	1	1	5
3	3	2	1
4	1	2	2
5	1.67	1	3
6	4	3	4
7	1	1	1
8	3	3	5

Valores anómalos (outliers):



definiciones:

- Datos que poseen características que son diferentes al resto de los datos.
- Valores de atributos que son inusuales con respecto a sus valores típicos.
- Valores extremos que yacen cercanos a los límites del rango de los datos, o que están fuera de la tendencia de los datos.

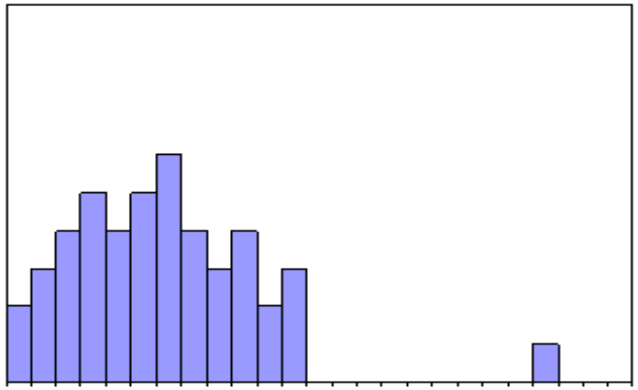
Los outliers pueden ser valores de interés, pueden representar errores pero no siempre.

- *Afectan la normalización de los datos.*

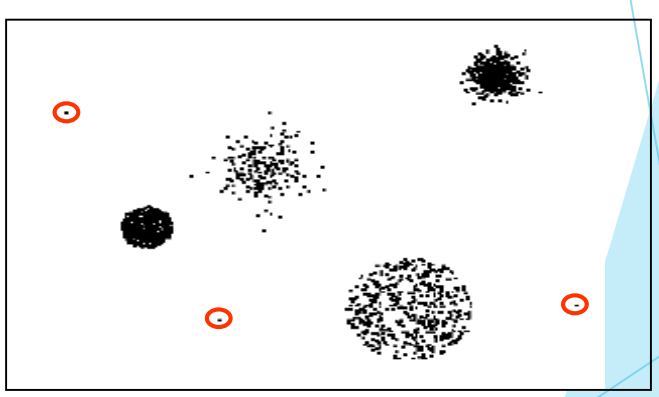
Es importante identificarlos ya que pueden generar problemas de representación de datos, a pesar que el valor sea válido y no represente ningún error.

¿Cómo detectarlos?

Técnicas clásicas de análisis exploratorio de datos (gráficos)

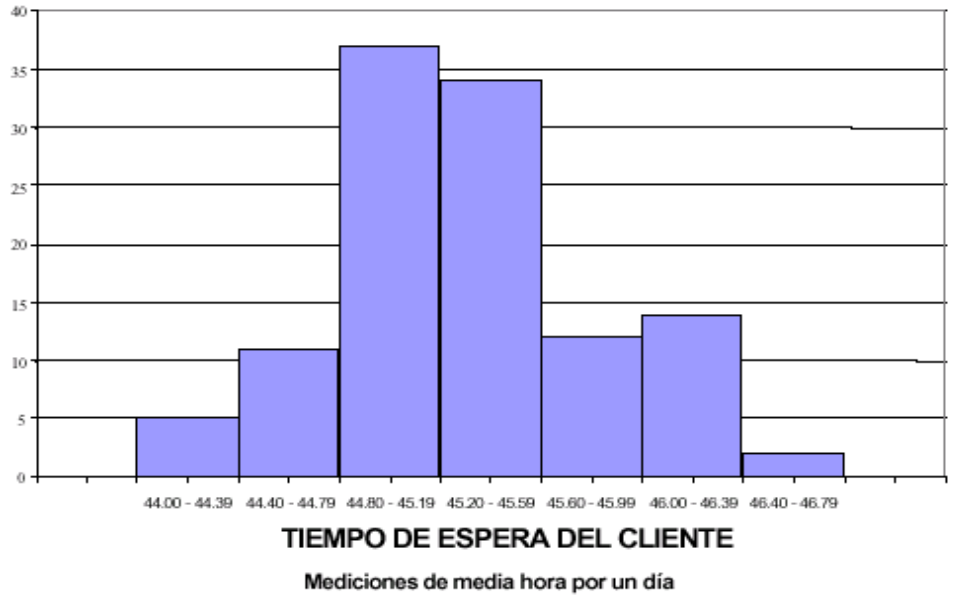


Técnicas de agrupación



Ejemplo: Histogramas

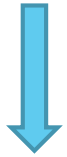
- Representación gráfica de una variable en forma de barras.
- La superficie de cada barra es proporcional a la frecuencia de los valores representados.



En el eje vertical se representan las frecuencias, y en el eje horizontal los valores de las variables.

¿Cómo tratarlos?

Una vez detectados, para su tratamiento se pueden utilizar algunas de las técnicas utilizadas para tratar valores ausentes.



- Ignorar el valor extremo: si la técnica de minería es robusta a estos datos
- Por eliminación: de registros (filas) o de columnas (atributos)
- Por discretización: Si se transforma un atributo continuo en uno discreto, los outliers caerán en las categorías extremas.

Valores no consistentes, errores:

Edad	Fecha de nacimiento	Sueldo	Zona
49	10/07/1999	3000,00	Santa Mónica
5	04/02/2005	2500,00	Boleita

Edad	Fecha de nacimiento	Peso
49	10/07/1999	65
25	04/02/1985	-57

- Discrepancias en los datos
- Registros duplicados
- Dos o más registros con los mismos valores en los atributos, pero diferente valor en el atributo clase

Nota: Importante detectarlos y corregir el problema

conocimiento acerca de los datos

¿Cuáles son los tipos de datos y dominio de los atributos?

¿Cuál es el rango de valores permitido para los atributos?

¿Los valores observados caen dentro del rango esperado?

¿Los datos son simétricos o sesgados?

¿Los datos son completos? (cubren todos los casos requeridos)

¿Hay dependencias entre los atributos?

¿Existen valores ausentes?, ¿cómo se representan?, ¿Con qué frecuencia se presentan?

Cuando los datos provienen de diferentes fuentes, ¿Los significados de los datos son iguales? ¿Tienen la misma unidad de medida?

¿Existen datos redundantes?

¿Los datos son consistentes?

- En general, las herramientas de MD facilitan el proceso de "conocer" los datos y verificar su calidad, y suministran filtros para la limpieza de los datos y corrección de errores.

- Otras están dirigidas a la limpieza y preparación de datos, por ejemplo Google Refine (gratuita)

Una vez que se realiza la limpieza de los datos

- Existen tipos de Tratamiento a los datos más elaborados:

  - Transformaciones

    Engloba cualquier proceso que modifique la forma de los datos.

  - Selección de atributos

    Para determinar los atributos más relevantes.

  - Construcción de nuevos atributos

    Que puedan resultar más informativos.

  - Selección de datos

    Para obtener una muestra representativa.