

Escalamiento y Encoding de variables

1. Proceso de descubrimiento de conocimiento.
2. Escalamiento de variables numéricas.
3. Encodeo (Encoding) de variables categoricas.

Proceso de
descubrimiento
de conocimiento

Proceso de Descubrimiento de Conocimiento o KDD

En esta definición se introducen las propiedades deseables del conocimiento extraído:

Válido: los patrones deben ser precisos para nuevos datos y no solo para aquellos que se han utilizado en su concepción, con un cierto grado de incertidumbre.

Novedoso: debe aportar algo que previamente se desconocía.

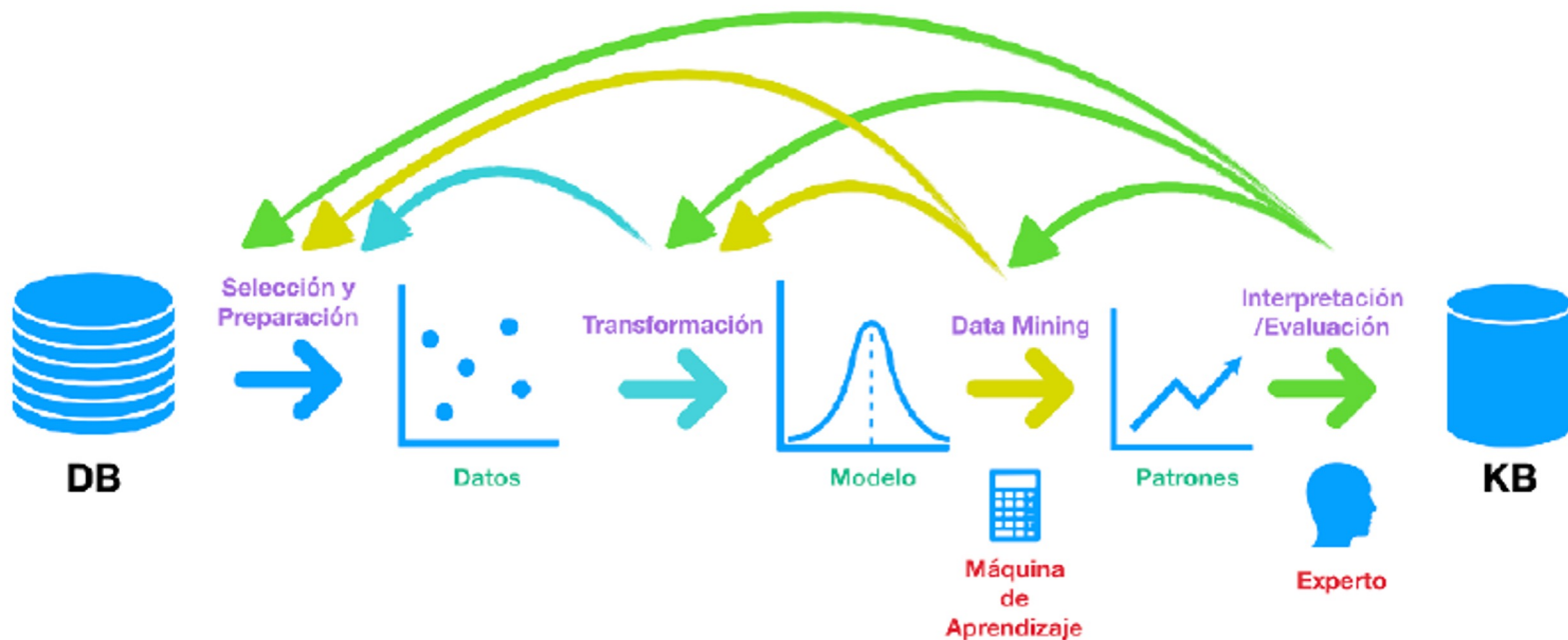
Potencialmente útil: que debe devolver algún tipo de beneficio.

Comprensible: la información incomprensible no aporta conocimiento en cuanto a su utilidad.



Proceso de Descubrimiento de Conocimiento o KDD

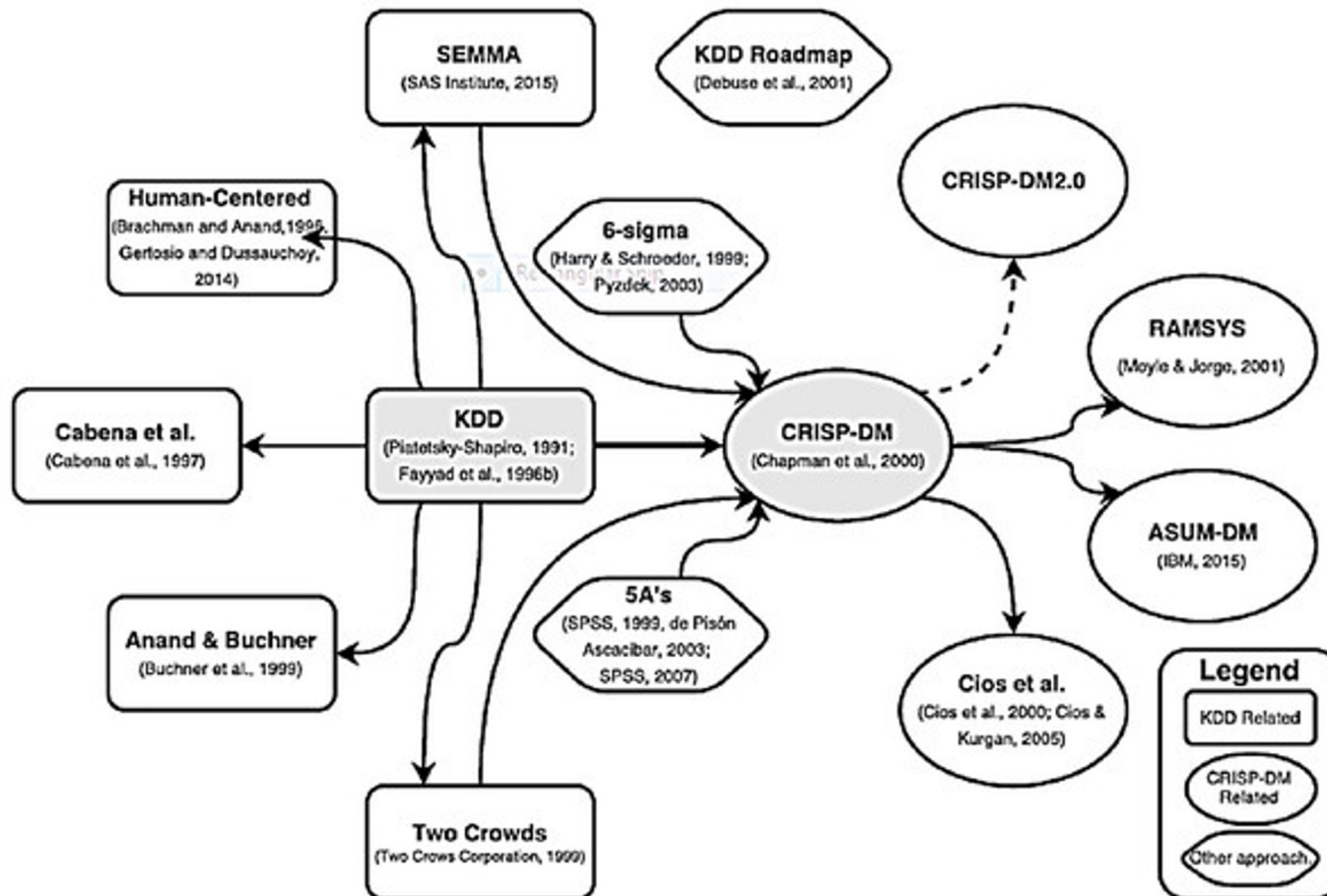
KDD se compone de diferentes fases y tareas



Metodología CRISP-DM

- CRISP-DM (Cross-Industry Standard Process For Data Mining: Procedimiento Industrial Estándar para realizar Minería de Datos), es creada en el 2000 por el grupo de empresas SPSS, NCR y Daimler Chrysler.
- Es de distribución libre lo que le permite estar en constante desarrollo por la comunidad internacional. Además resulta independiente de la herramienta que se utilice para llevar a cabo el proceso de MD.
- Es ampliamente usado por la industria. El modelo consiste en seis fases definidas de manera cíclica¹ : análisis del problema, comprensión de datos, preparación de datos, modelado, evaluación y despliegue.

Evolución de la metodología:



FUENTE : <https://doi.org/10.7717/peeri-cs.267/fig-2/>

Evolution of data mining process and methodologies, as presented in Martnez-Plumed et al. (2017)

Escalamiento de variables numéricas

Sea X una matriz de datos $n \times D$ consistente de n vectores x_i ($i \in 1, 2, \dots, n$). D es la dimensión de cada vector e indica el total de variables con las que estamos trabajando.

Sea entonces, x_{ij} el elemento de la fila i -ésima (observación) y de la columna j -ésima (variable).

La normalización y la estandarización de los datos busca llevar a una escala común las distintas variables con las que estamos trabajando.

Esto es necesario en muchas ocasiones ya que al estudiar fenómenos con variables heterogéneas, las variables con valores más grandes tendrán mayor ponderaciones por sobre las variables con valores pequeños.

Normalización:

La normalización busca dejar los datos entre un intervalo [a, b], el cual comúnmente es el [0, 1]. Para esto se usa la siguiente ecuación:

$$x_{new} = \frac{x_{old} - x_{min}}{x_{max} - x_{min}}$$

Este procedimiento se hace a nivel de cada variable. Entonces, el dato antiguo se le resta el mínimo y se divide por el rango (máximo - mínimo).

Es sensible a los outliers. ¿Por que?

Estandarización:

La estandarización busca dejar los datos centrados con media 0 (cero) y desviación estándar 1. Para esto se usa la siguiente ecuación:

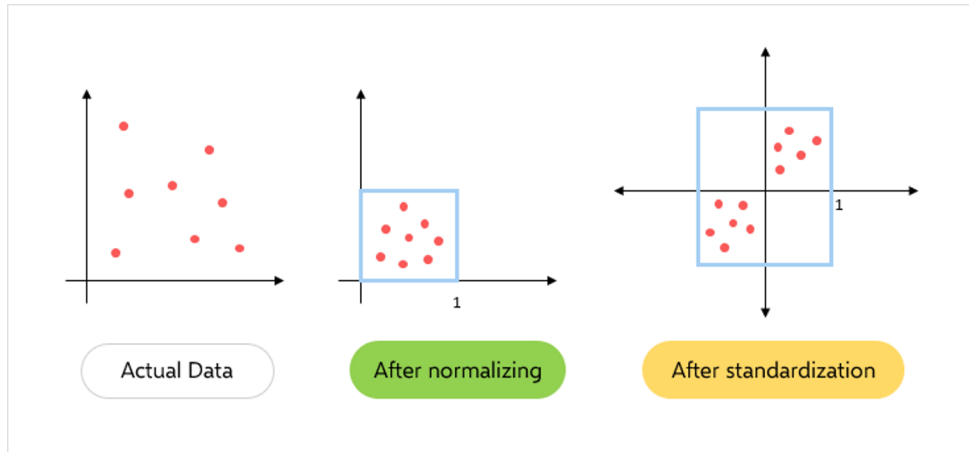
$$x_{new} = \frac{x_{old} - \mu_{old}}{\sigma_{old}}$$

Este procedimiento se hace a nivel de cada variable. Entonces, el dato antiguo se le resta la media y se divide por la desviación estándar.

De esta manera podemos ver en cuántas desviaciones estándar se distribuyen los datos.

Es más resistente a outliers, pero no es muy interpretable para datos que se alejan mucho de una distribución Normal.

Un pequeño ejemplo de cómo quedarían los datos después de una normalización o estandarización.



Un término más general para este proceso es: **Features Scaling**

En python usaremos las funciones:

- » **MinMaxScaling()**
- » **StandardScaling()**

Encodeo (Encoding)
de variables
categóricas.

Encoding

- » Existen variables categóricas ordinales y nominales.
- » Si quisiéramos hacer análisis exploratorio de datos no tendríamos muchos problemas en trabajarlas tal como están.
- » Pero si quisiéramos hacer cosas más complejas como:
 - Aplicar estadística inferencial o más aún,
 - Usar técnicas de Machine Learning para predecir comportamientos, tendríamos que transformarlas.
- » Para esto, veremos 2 técnicas de encoding que nos permitirán seguir trabajando con estas variables categóricas.

- » **Label encoding** es una técnica muy simple, pero que requiere que la variable categórica sea ordinal.
- » Se deben ordenar los distintos valores y a estos se les asigna un valor numérico que respete ese orden. Por ejemplo:
- » Suponiendo que las distintas respuestas son:

Muy desacuerdo: 0

En desacuerdo: 1

De Acuerdo: 2



Muy de acuerdo: 3

Encuesta	Label Encoding
Muy de acuerdo	3
De acuerdo	2
En desacuerdo	1

- » **One hot encoding** es una técnica un poco más compleja, pero permite trabajar con variables categóricas nominales.
- » Al transformar una variable categorica nominal, se deben crear n variables adicionales. Una por cada valor.

- » Estas variables nuevas son variables binarias.

- » Indican con un 1 la presencia de la categoría y con 0 la ausencia.

	Cat	Dog	Zebra
	1	0	0
	0	1	0
	0	0	1

- » En el ejemplo, la variable podria ser: tipo de animal.

Encoding

El encoding de las variables pretender transformar las variables categóricas a numéricas, tratando de mantener el sentido que originalmente tienen.



En python usaremos las funciones:

- » `LabelEncoder()`
- » `OneHotEncoder()`