Bases de Minería de datos

Juan Pablo Bermeo Moyano

KDD: Knowledge Discovery in Databases

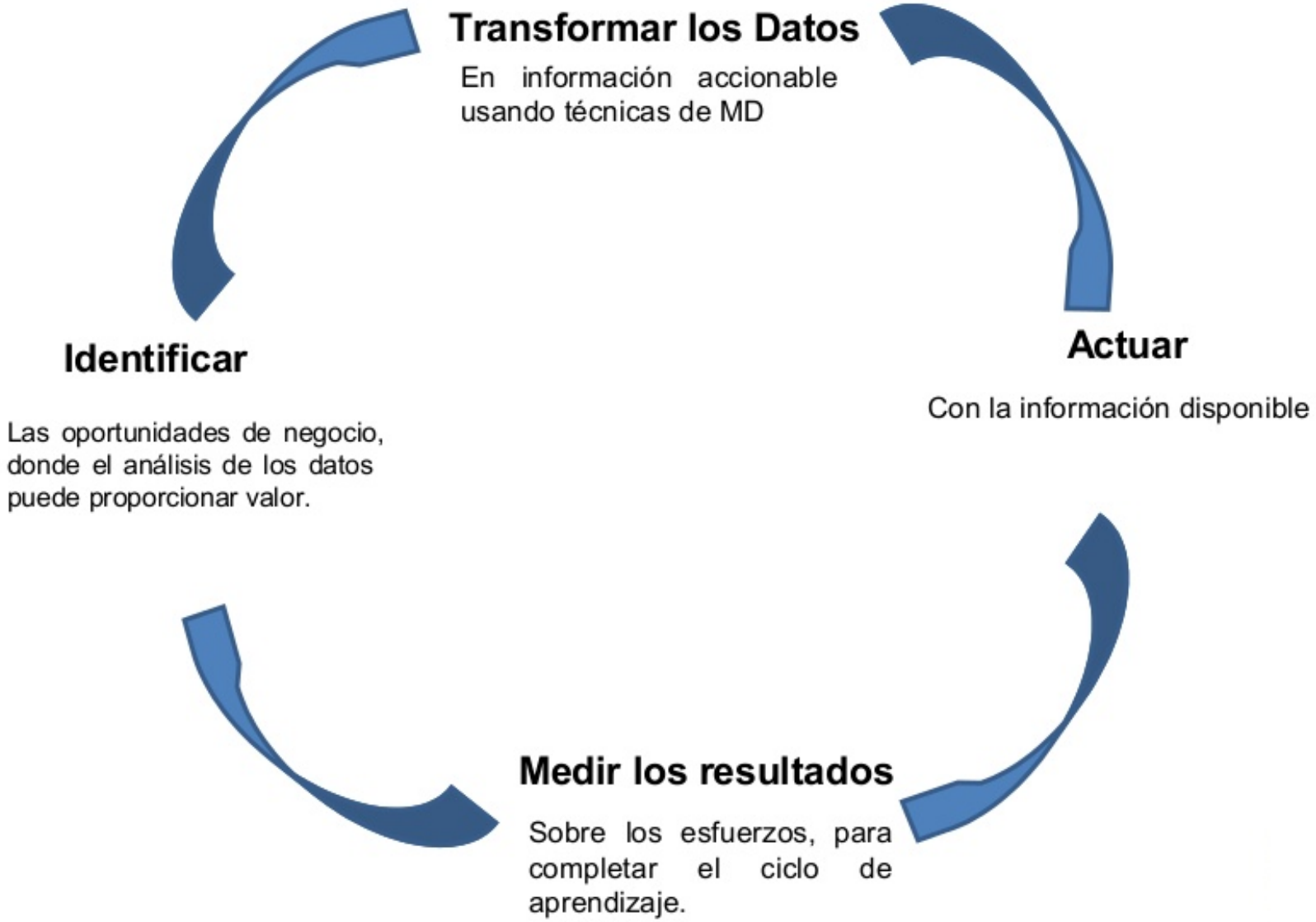
- KDD: Descubrimiento del Conocimiento en bases de datos.
- Paso 1: Comprensión del dominio del estudio y establecimiento de objetivos.
- Paso 2: Creación de un Data set objetivo.
- Paso 3: Limpieza y procesamiento de datos.
- Paso 4: Minería de datos (Data mining).
- Paso 5: Interpretación de los patrones minados.
- Paso 6: Utilización del conocimiento descubierto.
- Paso 7: Iniciar de nuevo, si no se consiguen objetivos.

¿Qué es la Minería de Datos?

- La minería de datos es la exploración y análisis de grandes cantidades de datos con el objeto de encontrar conocimiento (patrones, reglas significativas).
- Procesar datos "crudos" para realizar inferencias.
- Técnicas aplicadas para el análisis de grandes volúmenes de datos.
- Hay varios modelos híbridos basados en diferentes técnicas de minería de datos.
- "La tarea no trivial de extraer información implícita, previamente desconocida y potencialmente útil de bases de datos" (Frawley et. al. 1992).

Ciclo virtuoso de Minería de Datos

Fuente: https://es.slideshare.net/dataminingperu1/text-mining-16352175



Tipos de aplicaciones

- Los problemas de minerías de datos se pueden clasificar en las siguientes categorías:
  - Clasificación
  - Estimación
  - Pronóstico
  - Asociación
  - Agrupación o segmentación

Estadística aplicada

- Taller en Excel.
- En el archivo Ejercicio_01.xls encontrar:
- Distribución de frecuencia y representación gráfica
- Crear 10 intervalos.
- Calcular la marca de clase de cada intervalo.
- Estimar las medidas de tendencia central: media, moda y mediana, absolutas y la de los datos agrupados.
- Estimar las medidas de dispersión: varianza y desviación estándar, absolutas y la de los datos agrupados.
- Histogramas y ojivas de frecuencia.

Función de Distribución de Probabilidad PDF

- Definición:

$$f(x) = P(X = x)$$

- Propiedades:

$$f(x) \geq 0$$

- Funciones Discretas:

$$f(x) \leq 1$$

$$1 = \sum_{x=-\infty}^{+\infty} f(x)$$

- Funciones Continuas:

$$1 = \int_{-\infty}^{+\infty} f(x)dx$$

Función de Distribución Acumulada CDF

- Definición:

$$F(x) = P(X \leq x)$$

- Propiedades:

$\frac{dF(x)}{dx} \geq 0$, Siempre es creciente (inicia en 0 y termina en 1)

$$0 \leq F(x) \leq 1$$

- Funciones Discretas:

$$F(x) = \sum_{x=-\infty}^{x} f(x)$$

- Funciones Continuas:

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

La esperanza matemática

- Luego de hacer setenta mediciones de la velocidad de los vehículos en una carretera se obtiene la siguiente tabla de frecuencias.

Vo (km/h)	**20 km/h**	**40 km/h**	**60 km/h**	**80 km/h**	**100 km/h**
Frecuencia	5	10	25	20	10

- Encontrar la velocidad promedio.

Respuesta:

$$\bar{V} = \sum_{i=1}^{5} \frac{fe_i \cdot V_i}{N} = \sum_{i=1}^{5} f(V_i) \cdot V_i$$

La esperanza matemática(2)

- Variables discretas.

$$E[x] = \mu_x = \sum_{x=-\infty}^{\infty} x \cdot f(x)$$

$$E[(x-\mu_x)^2] = \sigma_x^2 = \sum_{x=-\infty}^{\infty} [x-\mu_x]^2 \cdot f(x)$$

- Variables Continuas:

$$E[x] = \mu_x = \int_{-\infty}^{\infty} f(x) \cdot x \cdot dx$$

$$E[(x-\mu_x)^2] = \sigma_x^2 = \int_{-\infty}^{\infty} [x-\mu_x]^2 \cdot f(x)\, dx$$

Simulación de Montecarlo

- Taller 2 en Excel.
- En el archivo Ejercicio_02.xls generar:
- 200 datos aleatorios que tengan una distribución de frecuencia Gaussiana con media 5 y desviación estándar.
- Utilizar las tablas para graficar los histogramas y ojivas de frecuencia.
- Comparar la media y desviación estándar calculada con la utilizada para la simulación.

Tarea: Foro

- Realizar la lectura del artículo:

Bots, Machine Learning, Servicios Cognitivos Realidad y perspectivas de la Inteligencia Artificial en España, 2018