Conjuntos de Datos

Concepto, instancias y variables.

Francisco José Madrid Cuevas

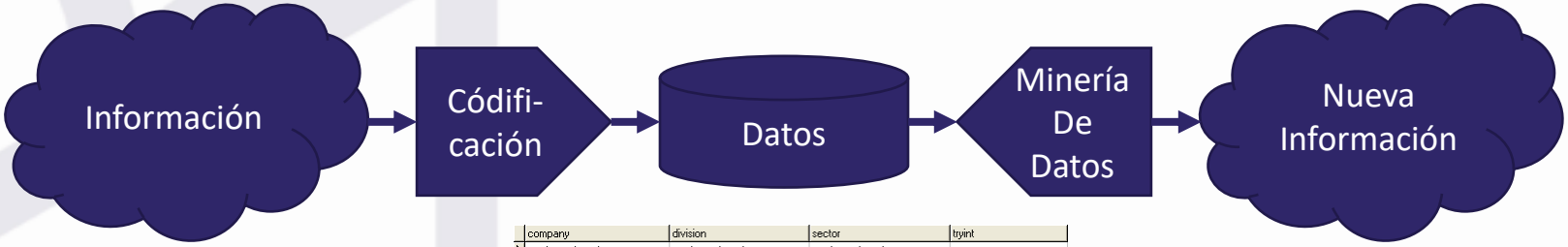
- Doctor en Informática por la Universidad Politécnica de Madrid en 2003. Desde 1996 he sido profesor a tiempo completo de la Universidad de Córdoba impartiendo docencia en Informática en varias titulaciones de Ingeniería.

Contenidos

- Conceptos:
  - Conjunto de Datos.
  - Instancia.
  - Característica.
- Tipos de Características.
- Valores perdidos.

Conjunto de datos

- El proceso de la minería de datos.



company	division	sector	tryint
00nil_Combined_Company	00nil_Combined_Division	00nil_Combined_Sector	14625
apple	00nil_Combined_Division	00nil_Combined_Sector	10125
apple	hardware	00nil_Combined_Sector	4500
apple	hardware	business	1350
apple	hardware	consumer	3150
apple	software	00nil_Combined_Sector	5625
apple	software	business	4950
apple	software	consumer	675
microsoft	00nil_Combined_Division	00nil_Combined_Sector	4500
microsoft	hardware	00nil_Combined_Sector	1890
microsoft	hardware	business	855
microsoft	hardware	consumer	1035
microsoft	software	00nil_Combined_Sector	2610
microsoft	software	business	1215
microsoft	software	consumer	1395

Conjunto de datos.

- Anatomía de un Conjunto de Datos.

ID	Nombre	F. Entrada	Est. Padres	G. Prácticas	Nota curso
23551405A	ADRIANA HERNANDEZ MONTERROZA	2021	básicos	1	6,5
45524672J	ADRIANA REY SANCHEZ	2020	superiores	3	7
51789536E	ALEJANDRO ABONDANO ACEVEDO	2021	básicos	2	8,5
38122360K	ALEXANDER CARVAJAL VARGAS	2019	medios	1	6,7
62782877F	CATALINA ACERO CARO	2020	sin estudios	4	9,2

Conjunto de datos.

- Tipos de características.
  - Nominal o Categórico: sin orden ni distancia. Si son únicos son identificadores.
  - Ordinal: inducen orden: “sin estudios”<“básicos”<“medios”<“superiores”.
  - Intervalo: inducen orden y tiene sentido medir distancia: fechas.
  - Ratio o Numérico: inducen un orden, permiten medir distancias y realizar operaciones aritméticas.

ID	Nombre	F. Entrada	Est. Padres	G. Prácticas	Nota curso
23551405A	ADRIANA HERNANDEZ MONTERROZA	2021	básicos	1	6,5
45524672J	ADRIANA REY SANCHEZ	2020	superiores	3	7
51789536E	ALEJANDRO ABONDANO ACEVEDO	2021	básicos	2	8,5
38122360K	ALEXANDER CARVAJAL VARGAS	2019	medios	1	6,7
62782877F	CATALINA ACERO CARO	2020	sin estudios	4	9,2

Conjunto de datos.

- Valores perdidos.
  - Son muy comunes.
  - Existen muchas formas de indicarlos:
    - Usando una etiqueta: ‘n.a.’, ‘n.e’, ‘nan’, …
    - Dejando el campo en blanco.
    - Utilizando un valor fuera de la escala.
    - …
  - Motivos:
    - Fallo en el aparato de medida.
    - Cambios en el diseño experimental.
    - Imposibilidad de obtener el valor en ese momento.
    - Mezcla desde distintas fuentes.
    - …
  - A veces, aportan mucha información.

¡Gracias!