Gestión de datos: Entrega I

Jon Zorrilla Gamboa y Rubén Martínez Gijón

1. Enunciado, datos de partida y condiciones

En esta práctica, se realizará un análisis de un conjunto de datos real anonimizado. Para ello, trataremos de obtener información sobre el profesor Álvaro Ortigosa a través de un conjunto de datos público de la UAM sobre PDI, proveniente del siguiente enlace *https://www.universidata.es/datasets/uam-personal-pdi*. Lo que haremos será, para cada campo del fichero (con algunas excepciones):

- Inferir el valor más probable de cada campo para el caso del profesor Ortigosa.
- Calcular el nivel de certeza (%) de que cada valor inferido corresponda realmente al profesor Ortigosa.
- Una breve descripción del proceso de inferencia utilizado en cada caso.

Antes de nada, tenemos ciertos datos de entrada y condiciones:

- El profesor Ortigosa estaba en 2020 en la UAM, se encuentra en el dataset.
- El profesor Ortigosa era de género masculino en 2020.
- El profesor Ortigosa pertenecía al Departamento de Ingeniería Informática en 2020.

Además, se puede usar cualquier otra información que pueda obtenerse por otra vía adicional, especificando en la respuesta tanto la información usada como su origen.

Cabe mencionar que para realizar esta práctica, se requiere del permiso explícito del profesor Ortigosa para inferir en sus datos, permiso que él nos ha concedido.

2. Análisis del conjunto de datos

Según la información que podemos obtener sobre este conjunto de datos, las variables pivote son el género y la unidad responsable. En el conjunto de datos, hay 4 atributos asociados a estas dos variables: *cod_genero* y *des_genero*, asociados al código de género y la descripción, y *cod_unidad_responsable*

y *des_unidad_responsable*, asociados al código de unidad responsable y su descripción. Además, según el enunciado, no trabajaremos con los atributos cuyos nombres empiezan por *cod_, lat_* y *lon_*. En nuestro caso, la primera variable ya la conocemos, pues el porfesor Ortigosa es de género masculino, es decir, *cod_genero = "H"*. Para conocer la unidad responsable, sabiendo que el profesor Ortigosa pertenecía al Departamento de Ingeniería Informática, obtenemos lo siguiente: *cod_unidad_responsable = 55001545*.

Por otro lado, si estudiamos los atributos que nos provee el conjunto de datos, podemos ver que aparecen los siguientes campos: *num_quinquenios*, *num_sexenios* , *anio_expedicion_titulo_doctor*, *des_categoria_cuerpo_escala*, *des_titulo_doctorado*, *anio_lectura_tesis* y *des_area_conocimiento*. En internet, podemos obtener la siguiente información sobre cada uno de estos campos, en el enlace adjunto [1].

- En 2018, Ortigosa contaba con 3 quinquenios y 4 sexenios.
- Recibió el título de doctor en el año 2000.
- En 2018, era profesor contratado Doctor.
- Pertenecía al área de conocimiento de "Lenguajes y Sistemas Informáticos".
- Era Profesor Contratado Doctor.

Una vez tenemos las variables pivote, podemos reducir mucho el conjunto de datos. De hecho, ahora existen únicamente 54 candidatos hombres que pertenezcan al Departamento de Ingeniería informática en 2020.

3. Análisis

En esta sección, estudiaremos campo a campo la probabilidad de que el profesor Ortigosa pertenezca a cada uno de ellos. Para ello, una vez simplificado el conjunto de datos mediante las variables pivote, calcularemos los valores posibles para cada atributo (siempre y cuando exista más de un valor posible), y veremos la probabilidad de pertenecer a cada uno dividiendo la cantidad de apariciones de dicho dato entre los datos totales.

A continuación, una vez nos hemos quedado con los datos que nos interesa estudiar, calcularemos la probabilidad para cada dato que tenemos en el dataset. Esta probabilidad se calculará como la probabilidad a priori; es decir, dividiendo el dato más probable para cada atributo entre la cantidad total de datos que tenemos.

[1] https://portalcientifico.uam.es/es/ipublic/researcher/261195

Tabla *1*: Tabla con campo, dato más probable y probabilidad para el profesor Ortigosa para cada uno de los casos del conjunto de datos.

Campo	Dato más probable	Probabilidad
des_universidad	Universidad Autónoma de Madrid	100 %
anio	2020	100 %
des_pais_nacionalidad	España	96.2963 %
des_continente_nacionalidad	Europa	100 %
des_agregacion_paises_nacionalidad	Europa meridional	100 %
des_comunidad_residencia	Madrid	100 %
des_provincia_residencia	Madrid	100 %
des_municipio_residencia	MADRID	50.0 %
des_genero	**H**	**100** %
anio_nacimiento	1967	7.4074 %
des_tipo_personal	Personal laboral	62.9630 %
des_categoria_cuerpo_escala	**Profesor Contratado Doctor**	**100** %
des_tipo_contrato	Contrato de Duración Determinada	38.8889 %
des_dedicacion	Dedicación a Tiempo Completo	81.4815 %
num_horas_semanales_tiempo_parcial	NaN	81.4815 %
des_situacion_administrativa	Servicio Activo	100 %
ind_cargo_remunerado	N	83.3333 %
des_titulo_doctorado	**Uno**	**100** %
des_pais_doctorado	NaN	53.7037 %
des_continente_doctorado	NaN	53.7037 %
des_agregacion_paises_doctorado	NaN	53.7037 %
des_universidad_doctorado	**Universidad Autónoma de Madrid**	**100** %
anio_lectura_tesis	**2000**	**100** %
anio_expedicion_titulo_doctor	**2000**	**100** %
des_mencion_europea	No	83.33 %
des_tipo_unidad_responsable	Departamento	100 %
des_unidad_responsable	Departamento de Ingeniería Informática	100 %
des_area_conocimiento	**Lenguajes y Sistemas Informáticos**	**100** %
anio_incorporacion_ap	NaN	37.0370 %

Continúa en la siguiente página

Tabla *1* – continuación de la página previa

First column	Second column	Third column
anio_incorpora_cuerpo_docente	Nan	37.0370 %
num_trienios	5	100 %
num_quinquenios	**3**	**100** %
num_sexenios	**4**	**100** %
num_tesis	NaN	96.2963 %
ind_investigador_principal	N	72.2222 %

Además de poner el porcentaje haciendo los cálculos de probabilidad a priori, se han indicado también en negrita las probabilidades obtenidas mediante búsqueda en la red.

Como conclusión, se puede añadir que independientemente de la información extra encontrada en la web, el sistema de anonimización de datos a veces es muy eficiente, y otras veces no. Por ejemplo, para el campo *anio_nacimiento*, obtenemos una probabilidad muy baja de acertar, mientras que para el campo *des_pais_nacionalidad* se obtiene un porcentaje de acierto muy alto. En general, cuantas más opciones diferentes existan para cada campo en el conjunto de datos, mayor será la anonimización. Mientras que si la mayoría de entradas tienen ciertas cosas en común, podremos hallar el valor correcto con mayor probabilidad. De todas formas, al existir tantos campos, es muy difícil conocer toda la información que queremos con exactitud.