Práctica Anonimización

Gestión de datos

Luis Antonio Ortega Andrés
Antonio Coín Castro

16 de diciembre de 2020

1. Problema a resolver y consideraciones generales

En este ejercicio pretendemos obtener toda la información estadística que podamos para reidentificar al profesor Ortigosa a partir de un *dataset* público de la UAM sobre PDI en el año 2018[1]. Este conjunto de datos se encuentra anonimizado, pero disponemos de información sobre las variables pivote y los bloques de coherencia utilizados. Nuestro propósito es explotar este conocimiento junto a otras informaciones externas sobre el profesor Ortigosa para intentar deanonimizar su información en el conjunto de datos.

La información inicial de la que partimos es la siguiente:

- El profesor se encuentra en el dataset.
- Su género es masculino.
- Pertenece al Departamento de Ingeniería Informática de la UAM.

Todas las manipulaciones, comprobaciones e inferencia realizadas para este trabajo se encuentran en el archivo adjunto `anon.py`.

1.1. Información externa

Además de la información que se proporciona, tenemos también información obtenida del perfil del profesor Ortigosa en el Portal Científico de la UAM[2], así como de menciones en torno a 2018 sobre el profesor Origosa y su relación laboral con la UAM aquí, aquí y aquí. Supondremos que estos datos son correctos y que estaban actualizados (a la fecha de redacción de este documento). En concreto, los datos que usaremos son:

- Su Área de Conocimiento a fecha de 2018 era **Lenguajes y Sistemas Informáticos**.
- Leyó su Tesis Doctoral y recibió el título de doctor en la UAM en el **año 2000**.

[1] https://www.universidata.es/datasets/uam-personal-pdi
[2] https://portalcientifico.uam.es/ipublic/agent-personal/profile/iMarinaID/04-261195

- A fecha de 2018 tenía **3 quinquenios** y **4 sexenios**.
- A fecha de 2018 era **profesor contratado doctor**.

2. Localización de variables pivote

Según la información sobre cómo ha sido anonimizado el dataset, las variables pivote son **el género** y **la unidad responsable**, representadas o bien por el código o bien por la decripción (nosotros elegiremos la primera representación para hacer los razonamientos). La primera de estas variables ya la conocemos, pues sabemos que debe ser `cod_genero="H"`.

Conocer estas variables pivote es muy importante, ya que conservan la relación respecto a todos los bloques de coherencia. Es por esto que nuestra primera tarea será encontrar la segunda de las variables pivote, `cod_unidad_responsable`. Para ello podemos utilizar alguna de las informaciones de las que disponemos. En concreto, utilizamos que conocemos cuál debe ser el área de conocimiento, suponiendo que esta información localizará unívocamente el código de la unidad (departamento) responsable. En efecto, si buscamos en el dataset aquellos elementos cuya área de conocimiento sea Lenguajes y Sistemas Informáticos, resulta que para todos ellos el código de unidad responsable es el mismo: 55001545.

```
query_pivote = \
  'des_area_conocimiento == "Lenguajes y Sistemas Informáticos"'
df.query(query_pivote)["cod_unidad_responsable"] # --> 55001545
```

La pega que podríamos tener en este razonamiento es que el campo `des_area_conocimiento` para el profesor Ortigosa fuera un valor perdido (NaN). Podemos comprobar que esto no ocurre utilizando alguna otra información adicional. Notamos que en la sintaxis de `Pandas`, para comprobar si un valor es NaN podemos comprobar si es igual a él mismo (los valores NaN son los únicos que no verifican esto). El razonamiento (que reproducimos en el archivo Python) es el siguiente, teniendo en cuenta la coherencia de las variables implicadas:

1. Existen solo dos códigos de unidad responsable cuyo área de conocimiento asociado es NaN: 55001539 y NaN.
2. No existe ninguna persona con el código 55001539, 3 quinquenios y 4 sexenios, o bien con ese código y alguno de los otros dos campos a NaN (o los dos).
3. Tampoco existe ninguna persona con el código NaN, 3 quinquenios y 4 sexenios, o bien con ese código y alguno de los otros dos campos a NaN (o los dos).

A partir de ahora, tendremos una consulta por defecto para intentar localizar información en cada bloque, que se basará en buscar por el código de unidad responsable y el género del profesor Ortigosa, es decir, las variables pivote (que ya conocemos). En los casos donde dispongamos de información adicional, modificaremos la consulta para incluirla.

```
standard_query = \
  'cod_unidad_responsable == "55001545.0" & cod_genero == "H"'
```

3. Inferencia en bloques de coherencia

Iremos ahora analizando bloque a bloque y campo a campo para encontrar los valores más probables para el profesor Ortigosa, dando información estadística sobre la predicción (número de veces que aparece el valor más probable dividido por el número total de opciones). Para ello disponemos de una función `freq` que devuelve el nivel de certeza de cada valor para un campo dado, contando únicamente dentro de aquel subconjunto que cumpla nuestra consulta estándar con las variables pivote (`standard_query`). Destacaremos en negrita la información que ya conozcamos por fuentes externas.

3.1. Bloque 1

El bloque 1 es el más sencillo, ya que todas sus variables tienen el mismo valor para todas las entradas del dataset. Por tanto, tenemos que los valores extraidos son:

Campo	Valor más probable	Nivel de certeza (%)
des_universidad	Universidad Autónoma de Madrid	100
anio	2018	100

3.2. Bloque 2

Para este bloque, como no disponemos a priori de información adicional, simplemente llamamos a la función `freq` sobre cada uno de los campos y mostramos los resultados. En lo sucesivo cuando en un bloque no comentemos nada será porque seguimos esta misma estrategia.

Campo	Valor más probable	Nivel de certeza (%)
des_pais_nacionalidad	España	98.25
des_continente_nacionalidad	Europa	100
des_agregacion_paises_nacionalidad	Europa meridional	100

3.3. Bloque 3

Campo	Valor más probable	Nivel de certeza (%)
des_comunidad_residencia	Madrid	100
des_provincia_residencia	Madrid	100
des_municipio_residencia	MADRID	49.12

3.4. Bloque 4

Campo	Valor más probable	Nivel de certeza (%)
anio_nacimiento	1967	8.77

3.5. Bloque 5

En este caso podemos usar información externa. Antes de nada, comprobamos que entre todas las entradas que tienen las variables pivote al valor correcto, ninguna tiene un valor NaN en el campo `cod_categoria_cuerpo_escala`. Una vez hecho esto, usamos que el profesor Ortigosa era Profesor Contratado Doctor, que sabemos que tiene el código "5" dentro de ese campo.

Campo	Valor más probable	Nivel de certeza (%)
des_tipo_personal	Personal laboral	100
des_categoria_cuerpo_escala	**Profesor Contratado Doctor**	100
des_tipo_contrato	Contrato Indefinido o Fijo	88.23
des_dedicacion	Dedicación a Tiempo Completo	100
num_horas_semanales_tiempo_parcial	NaN	100
des_situacion_administrativa	Servicio Activo	100

3.6. Bloque 6

Campo	Valor más probable	Nivel de certeza (%)
ind_cargo_remunerado	N	82.46

3.7. Bloque 7

Utilizamos la información de que el profesor Ortigosa leyó su tesis doctoral y recibió el título de doctor en la UAM en el año 2000. En primer lugar, encontramos todas las entradas que cumplen estos requisitos y además tienen bien las variables pivote (contando siempre que alguno de los campos pueda ser NaN). Después utilizamos la información de que tiene 1 doctorado para añadirlo a la consulta.

Campo	Valor más probable	Nivel de certeza (%)
des_titulo_doctorado	**Uno**	100
des_pais_doctorado	España	50*
des_continente_doctorado	Europa	50*
des_agregacion_paises_doctorado	Europa Meridional	50*
des_universidad_doctorado	**Universidad Autónoma de Madrid**	100
anio_lectura_tesis	**2000**	100
anio_expedicion_titulo_doctor	**2000**	100
des_mencion_europea	No	100

* El resto de valores posibles (el otro 50 %) son NaN.

4. Bloque 8

Utilizamos la información sobre su área de conocimiento.

Campo	Valor más probable	Nivel de certeza (%)
des_tipo_unidad_responsable	Departamento	100
des_area_conocimiento	**Lenguajes y Sistemas Informáticos**	100

5. Bloque 9

En este caso solo hay una persona que cumple los requisitos de los sexenios y quinquenios que sabemos que debe haber (junto con las variables pivote). Nos hemos asegurado de comprobar que estos campos no tenían valores NaN.

Campo	Valor más probable	Nivel de certeza (%)
anio_incorporacion_ap	NaN	100
anio_incorpora_cuerpo_docente	NaN	100
num_trienios	5	100
num_quinquenios	**3**	100
num_sexenios	**4**	100

6. Bloque 10

Campo	Valor más probable	Nivel de certeza (%)
num_tesis	NaN	94.74*

* El otro valor posible es 1.0

7. Bloque 11

Campo	Valor más probable	Nivel de certeza (%)
ind_investigador_principal	N	75.44