Práctica anonimización de datos

Rocío Rico Sanche-Mateos,

Iker Villegas Labairu

23 de diciembre de 2022

Introducción y objetivos

Se nos plantea realizar, a partir de un conjunto de datos recogidos en el fichero de datos abiertos del PDI de la Universidad Autónoma de Mardrid correspondiente al año 2020, los siguientes procedimientos referentes al profesor Ortigosa:

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

Datos de partida y condiciones

Para realizar dicho estudio se nos dan una serie de suposiciones que tendremos que considerar. De esta forma, aseguramos lo siguiente:

- El profesor Ortigosa estaba en 2020 en la UAM (es decir, aparece en el fichero).

- El profesor Ortigosa era de género masculino en 2020.
- El profesor Ortigosa pertenecía al Departamento de Ingeniería Informática en 2020.

Por otra parte, además de dichas suposiciones, se han recogido otras informaciones adicionales que procedemos a enumerar a continuacion, junto con su procedencia:

1. El profesor Ortigosa pertenece al Departamento Universitario de Ingeniería Informática, dentro del área de conocimiento de lenguaje y sistemas informáticos. Además ostentó durante el año 2020 el cargo de director del Instituto de Ciencias Forenses y de la Seguridad en la UAM. Por otro lado, a lo largo de su carrera ha dirigido 3 tesis doctorales. Dicha información se ha obrenido a través del portal científico de la UAM, al cuál se accede mediante el siguiente enlace.

2. Del mismo link que el anterior, se puede obtener la información de que en el 2020 el profesor Ortigosa paso de *Profesor Contratado Doctor* a *Profesor Titular Universidad* en el periodo de tiempo del 9-Nov al 21-Dic.

3. A través de la búsqueda de la tesis del profesor Ortigosa en el portal web de la biblioteca de la UAM, asumimos que dispone de una tesis doctoral expedida en la Universidad Autónoma de Madrid en el año 2000. (enlace).

4. El profesor Ortigosa en el año 2022 ostentaba el puesto de profesor asociado. Además, se empezó su labor como profesor en la administración pública en 1995 y comenzó sus actividades como docente en

la UAM en el año 2001. Esta información se ha obtenido a través del perfil de la red social Linkedin del profesor (perfil de Linkedin).

Además, para llevar a cabo este estudio, contamos con el permiso explícito del profesor Ortigosa, tal y como se muestra en el siguiente mensaje comunicado a través del Moodle del Posgrado de la UAM:

> Dejo constancia que autorizo a los estudiantes de esta asignatura a intentar reidentificar mis datos personales a partir del fichero de datos sobre PDI publicado por la UAM.
>
> Alvaro Ortigosa

Primeros pasos

En primer lugar, para poder llevar a cabo el estudio, se han recogido los datos a través del fichero de datos abiertos del PDI de la UAM correspondientes al año 2020 localizado en el siguiente enlace. El fichero descargado será un archivo excel de extensión `.xlsx`.

A continuación, para llevar a cabo el tratamiento de los datos, utilizaremos el lenguaje de programación *R* a través de *Rstudio*.

De este fichero de datos, nos encargaremos de eliminar aquellas columnas que comienzan con *lon_*, *lat_* y *cod_*, tal y como se puede ver en el archivo de código adjunto.

En segundo lugar, en el siguiente enlace se indica que los campos pivote son: *des_unidad_responsable* y *des_genero*. Estas variables conservan la relación respecto a todos los bloques de coherencia, que son respecto a los cuales vamos a calcular el nivel de certeza. Conocemos, gracias a los datos de partida, el valor exacto correspondiente a cada campo pivote:

Departamento de Ingeniería Infor- mática y *Hombre* respectivamente, de modo que, podemos eliminar todas las filas del dataset que no coincidan con estas dos columnas reduciendo, así, la incertidumbre.

Análisis de las variables

Para llevar a cabo el estudio de qué valores se corresponderán con los del profesor Ortigosa y su porcentaje de certeza, analizaremos cada una de las variables dentro de su correspondiente bloque de coherencia.

Bloque de coherencia 1

Este bloque agrupa las siguientes variables:

- *des_universidad*: Nombre de la Universidad a la que pertenece el empleado. En este caso solo existe una posibilidad y es la Universidad Autónoma de Madrid, por lo que con un 100 % de certeza el profesor Ortigosa pertenece a esta.
- *anio*: Año natural al que se refieren los datos. En este caso solo existe una posibilidad y es 2020, por lo que con un 100 % de certeza el profesor Ortigosa ha trabajado en la UAM en el año 2020.

De este modo, tenemos que ambos valores pertenecerán a dicho profesor con un 100 % de probabilidad.

Bloque de coherencia 2

Este bloque agrupa las siguientes variables:

- *des_pais_nacionalidad*: Nombre del país correspondiente a la nacionalidad del empleado. Podemos asumir nacionalidad española con un

96, 296 % de probabilidad frente a un 3, 704 % para italiano. De este modo tomaremos que el profesor ortigosa tendrá nacionalidad española con un 96, 296 %.

- *des_ continente_ nacionalidad*: Nombre del continente correspondiente a la nacionalidad del empleado. En este caso solo existe una posibilidad y es Europa. En este caso, tendremos una certeza del 100 %
- *des_ agregacion_ paises_ nacionalidad*: Nombre de la agregación de países correspondiente a la nacionalidad del empleado. En este caso solo existe una posibilidad y es Europa meridional, ya que observando los datos, es la única opción posible. Así, para esta variable, también tendremos un 100 % de certeza.

Bloque de coherencia 3

Este bloque agrupa las siguientes columnas del dataset:

- *des_ comunidad_ residencia*: Nombre la comunidad autónoma del domicilio del empleado. En este caso solo existe una posibilidad y es Madrid (100 % de certeza).
- *des_ provincia_ residencia*: Nombre de la provincia del domicilio del empleado. En este caso solo existe una posibilidad y es Madrid (100 % de certeza).
- *des_ municipio_ residencia*: Nombre del municipio del domicilio del empleado. No podemos asumir nada pero con lo que menos nos arriesgamos es afirmando que sea de Madrid con un nivel de certeza del 50 %.

MUNICIPIO	PROBABILIDAD
Alcala de Henares	0.0185
Alcobendas	0.0740
Alcorcón	0.0185
Algete	0.0185
Boalo	0.0185
Colmenar Viejo	0.1111
Colmenarejo	0.0185
Fuenlabrada	0.0185
Galapagar	0.0185
Hoyo de Manzanares	0.0185
Madrid	0.5000
Moralzarzal	0.0185
Pozuelo de Alarcón	0.0185
San Sebastián de los Reyes	0.0740
Tres Cantos	0.0555

Bloque de coherencia 4

Este bloque agrupa la variable *año_ nacimiento* (año de nacimiento del empleado). Para los siguientes porcentajes se han efectuado los siguientes razonamientos: Si suponemos que el primer trabajo del profesor Ortigosa fue como investigador y profesor adjunto ordinario en UNICEN en 1995, basándonos en 1, nos podemos imaginar que como mínimo ha tenido que cursar una carrera universitaria y un máster para acceder a ese trabajo por lo que suponemos una edad mínima de 23 años, es decir, un año de nacimiento menor o igual a 1972. Además, tiene que estar fuera de los años de jubilación de España que se encuentra en torno a los 65, es decir, su año de nacimiento tiene que ser mayor que 1972. Con estos argumentos hemos creado la tabla 0.1.

Aludiendo a la juventud de nuestro profesor apostamos por 1967 como valor más probable (nivel de certeza del 14,81 %).

Año de nacimiento	Probabilidad
1958	0.0370
1960	0.0370
1961	0.0370
1962	0.0740
1963	0.0740
1964	0.1111
1965	0.1481
1966	0.0740
1967	0.1481
1968	0.0740
1969	0.0740
1970	0.0370
1971	0.0370
1972	0.0370

Tabla 0.1: Tabla de nacimentos

Bloque de coherencia 5

Este bloque agrupa las siguientes variables:

- *des_ categoria_ cuerpo_ escala*: Código de la categoría/cuerpo/escala del empleado. Para extraer bien los datos de esta variable nos tenemos que fijar en la sección 1 (Datos generales sobre el conjunto de datos) que nos da el portal de datos donde nos concreta que los datos se actualizan anualmente y en diciembre. Esta información es importante porque en el 2020 el profesor Ortigosa paso de *Profesor Contratado Doctor* a *Profesor Titular Universidad* en el periodo de tiempo del 9-Nov al 21-Dic. Con esto entendemos que la nueva información se vería reflejada ya en el dataset del año 2021 porque si se actualizan anualmente y en diciembre, quiere decir que esta tabla de datos fue creada antes o en diciembre, es decir, antes del cambio de categoría. Con este conocimiento, reducimos la incertidubre del resto de varia-

bles. Así, establecemos con un 100 % de certeza de que será Personal Contratado Doctor.

- *des_ tipo_ personal*:Descripcion del tipo de personal del empleado. En este caso solo existe una posibilidad y es Personal Laboral.
- *des_ tipo_ contrato*:Descripcion del tipo de contrato del empleado. Podemos asumir contrato indefinido o fijo con un $81,256\,\%$ de probabilidad frente a un $18,75\,\%$ para contrato de duración determinada.
- *des_ dedicacion*:Descripcion del regimen de dedicacion del empleado. En este caso solo existe una posibilidad y es Dedicación a tiempo completo.
- *num_ horas_ semanales_ tiempo_ parcial*:Horas semanales en contrato del profesorado a tiempo parcial. En este caso no existe información, solo hay NaN.
- *des_ situacion_ administrativa*:Descripcion de la situacion administrativa del empleado. En este caso solo existe una posibilidad y es Servicio activo.

Bloque de coherencia 6

Este bloque agrupa las siguientes variables:

- *ind_ cargo_ remunerado*: indicador de si la persona al que hace referencia el registro ocupa algún cargo unipersonal remunerado económicamente en la Universidad. Sus posibles valores serán "S"para "Síz "N"para "No". Con los datos de los que disponesmos el profesor Ortigosa, por 1, desde el 2017 hasta la actualidad (y, por tanto, durante el año 2020 también), ostenta el cargo de director del Instituto de

Ciencias Forenses y de la Seguridad. Por tanto la variable tomará el valor de "Sçon un 100 % de seguridad.

Realizando un conteo de los registros que poseen como valor "S.[en] la variable *ind_ cargo_ remunerado*, vemos que aparecen 45 con "Nz 9 con "S". De este modo sí hay registros que poseen dicho valor en la columna, y, por tanto, tendremos una certeza del 100 % de que sea el correspondiente al profesor Ortigosa.

Bloque de coherencia 7

Disponemos de las siguientes columnas:

- *des_ titulo_ doctorado*: variable que indica si el empleado tiene túlulo de doctor y, en caso afirmativo, la cantidad. Como hemos supuesto en 2, el profesor Ortigosa dispone de un título de doctorado.
- *des_ pais_ doctorado*: nombre del pais donde se ha obtenido el título de doctorado. A excepción de que dicha variable figure con un valor desconocido, podemos asumir que el país sería España, al localizarse la UAM en la misma. Hemos supuesto por 2 que en dicha universidad ha obtenido su título de doctor.
- *des_ continente_ doctorado*: nombre del continente correspondiente al pais en el cuál se ha obtenido el título de Doctorado. Por la variable anterior, al situarse España dentro de Europa, podemos suponer, en un primer momento, que el continente donde se ha sacado el doctorado sea Europa.
- *des_ agregacion_ paises_ doctorado*: nombre de la agregación de paises correspondiente al país en el que se ha obtenido el título de doctorado.

En principio, no se han encontrado datos que nos den alguna idea de su valor, por lo cuál, lo consideraremos desconocido a falta de realizar una inferencia en nuestros datos.

- *des_universidad_doctorado*: nombre de la universidad en la que se ha obtenido el título de doctorado, o su clasificación si dicha universidad no es española. Por 2, hemos asumido que dicha universidad es la Universidad Autónoma de Madrid.

- *anio_lectura_tesis*: año en el cuál ha tenido lugar la lectura de la tesis, si así procede. Justo como hemos tomado en 2, el año de lectura de la tesis tuvo lugar en el 2000.

- *anio_expedicion_titulo_doctor*: año en el que se ha producido la expedición de título de doctor, si procede el caso.

- *des_mencion_europea*: variable donde se indica si alguno de los títulos de doctorado tiene mención europea/internacional. No disponemos de datos que arrojen luz sobre esta variable, esperaremos a realizar una inferencia sobre los datos con las variables conocidas de este bloque.

Basandonos en nuestras suposiciones para las variables anteriores, realizamos una selección de los registros que cumplen las siguientes condiciones:

- $des\_titulo\_doctorado = Uno$
- $des\_universidad\_doctorado = Universidad\ Autónoma\ de\ Madrid$
- $anio\_lectura\_tesis = 2000$

De esta forma, se han obtenido dos registros identicos que adoptan los siguientes valores para las variables a las que no le habíamos asignado valor:

- $des\_pais\_doctorado = España$
- $des\_continente\_doctorado = Europa$
- $des\_agregacion\_paises\_doctorado = Europa\ meridional$
- $anio\_expedicion\_titulo\_doctor = 2000$
- $des\_mencion\_europea = No$

Por lo tanto, podemos concluir que los estos valores para todas las variables de este bloque tendrán 100 % de certeza de pertenecer al profesor Ortigosa.

Bloque de coherencia 8

- *des_tipo_unidad_responsable*: tipo de unidad en la que está adscrito el PDI. Tal y como se ha consultado en 1, dicho tipo de unidad será *Departamento*.
- *des_area_conocimiento*: área de conocimiento en la cuál se encuentra. A través de 1, vemos que su área de conocimiento es *Lenguajes y sistemas informáticos*.

Filtrando las entradas de nuestra tabla de datos, tomando ambos valores para las variables, vemos que hay varias entradas que las cumplen. Y, por ello, vemos que tanto *Departamento*, como *Lenguajes y sistemas informáticos*, serán con una certeza del 100 % los valores del profesor Ortigosa para las columnas correspondientes.

Bloque de coherencia 9

- *anio_incorporacion_ap*: año en el que el PDI funcionario de carrera se incorporó por primera vez a la administración pública. Tal y como se ha supuesto en 3, asumiremos que el profesor Ortigosa se incorporó por primera vez a la administración pública en 1995.

- *anio_incorporacion_cuerpo_docente*: año de incorporación del funcionario al cuerpo docente universitario. Por 3, hemos asumido que se incorporó como docente en la UAM en el año 2001

- *num_trienios*: número de trienios con fecha 31 de diciembre. No se han encontrado datos al respecto.

- *num_quinquenios*: número de quinquenios a fecha 31 de diciembre. No se han encontrado datos al respecto.

- *num_sexenios*: número de sexenios con fecha 31 de diciembre. No se han encontrado datos al respecto.

De este modo, filtrando las columnas anteriores de nuestros datos para los casos donde $anio\_incorporacion\_ap = 1995$ y $anio\_incorporacion\_cuerpo\_docente = 2001$, obtenemos un registro donde:

- $num\_trienios = 8$

- $num\_quinquenios = 5$

- $num\_sexenios = 2$

De esta manera, podremos asumir que todos los anteriores valores pertenecen al profesor Ortigosa con una certeza del 100 %.

Bloque de coherencia 10

- *num_ tesis*: número de tesis doctorales dirigidas o codirigidas por dicho profesor que hayan sido leidas a lo largo del año. Tal y como asumimos en 1, el profesor Ortigosa ha dirigido 3 tesis a lo largo de su trayectoria, luego sabemos que dicho valor en nuestros datos tiene que ser menor o igual que 3.

Tras realizar un filtro a nuestros registros, vemos que únicamente hay 2 casos con valor 1 y 52 con valor `Nan`. De este modo, lo más probable será suponer que el número de tesis doctorales dirigidas por el profesor Ortigosa sean desconocidas para el año 2020. Esta suposición nos aporta un $96,296\,\%$ de certeza.

Bloque de coherencia 11

- *ind_ investigador_ principal*: si ha sido investigador principal (IP) de algún proyecto y/o contrato de investigación durante el año. Para esta variable, no se han encontrado ningún proyecto o contrato donde figure el profesor Ortigosa como investigador principal durante el 2020.

Realizando un conteo de los datos para esta variable obtenemos que aparecen 39 con valor "N", y 15 con "S". Luego podremos suponer que el profesor ortigosa no ha sido investigador principal con una certeza del $72,222\,\%$.

Resumen de los datos

Como resumen, hemos creado la siguiente tabla que muestran los valores para cada variables que hemos inferido como los más probables para el profesor Ortigosa con su nivel de certeza correspondiente:

Variable	**Valor inferido**	**Certeza** (%)
des_universidad	Universidad Autónoma de Madrid	100
anio	2020	100
des_pais_nacionalidad	España	96,296
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
des_genero	Hombre	100
anio_nacimiento	1967	14,81
des_tipo_personal	Personal Laboral	100
des_categoria_cuerpo_escala	Personal Contratado Doctor	100
des_tipo_contrato	Contrato Indefinido o Fijo	81,256
des_dedicacion	Dedicación a Tiempo Completo	100
num_horas_semanales_tiempo_parcial	NaN	100
des_situacion_administrativa	Servicio Activo	100
ind_cargo_remunerado	S	100
des_titulo_doctorado	Uno	100
des_pais_doctorado	España	100
des_continente_doctorado	Europa	100
des_agregacion_paises_doctorado	Europa meridional	100
des_universidad_doctorado	Universidad Autónoma de Madrid	100
anio_lectura_tesis	2000	100
anio_expedicion_titulo$_{d}$*octor*	2000	100
des_mencion_europea	No	100
des_tipo_unidad_responsable	Departamento	100
des_unidad_responsable	Departamento de Ingeniería Informática	100
des_area_conocimiento	Lenguaje y Sistemas Informáticos	100
anio_incorporacion_ap	1995	100
anio_incorpora_cuerpo_docente	2001	100
num_trienios	8	100
num_quinquenios	5	100
num_sexenios	2	100
num_tesis	NaN	96,296
ind_investigador_principal	N	72,222

Tabla 0.2: Resumen variables