

## UNIDAD 2- LAB1. ANONIMIZANDO DATOS

### Índice

#### [0. Objetivos:](#)

#### [1. Distinguir entre seudonimizacion y anonimización](#)

#### [2. La k-anonimidad](#)

#### [3. Métodos de K-anonimización](#)

#### [4. Herramientas para anonimizar los datos utilizando técnicas de K-anonimidad](#)

#### [5. Caso práctico](#)

#### [6. Conclusiones](#)

#### [7. Referencias](#)

### 0. Objetivos:

- Distinguir entre seudonimización y anonimización
- Conocer en qué consiste la técnica k-anonimización.
- Conocer métodos de anonimización
- Conocer una herramienta de ayuda a la anonimización
- Conocer el procedimiento y problemática de la anonimización
- Ver un caso práctico de anonimización.

### 1. Distinguir entre seudonimizacion y anonimización

**Anonimización:** La anonimización de datos personales, consiste en desvincular completamente los datos personales de los datos identificativos, es decir, cuando se anonimizan datos personales, se produce un nuevo conjunto de datos completamente disociado del individuo al que pertenecen, haciendo imposible que a través de esos datos anonimizados se pueda identificar o reidentificar a dicho individuo. Es un procedimiento irreversible.

**Seudonimización:** es «aquella información que, sin incluir los datos denominativos de un sujeto, permiten identificarlo mediante información adicional, siempre que esta figure por separado y esté sujeta a medidas técnicas y organizativas destinadas a

garantizar que los datos personales no se atribuyan a una persona física identificada o identificable».

## 2. La k-anonimidad

El proceso básico de anonimización consiste en disociar de los identificadores el resto de los datos más genéricos asociados a un sujeto como la fecha de nacimiento, el municipio de residencia, el género, etc. El conjunto de datos preservados serán aquellos necesarios para cumplir con el objetivo del tratamiento y, mediante su conservación y enriquecimiento, explotarlo para extraer información adicional.

Sin embargo, aunque la realización de dicho proceso de anonimización aparentemente permite mantener el anonimato, dichos datos convenientemente agrupados y cruzados con otras fuentes de información, pueden llegar a identificar a un individuo e incluso relacionarlo con categorías especiales de datos. De ahí que al conjunto de datos que no son “identificadores” pero podrían llegar a señalar de forma unívoca a un individuo se le denomina “pseudo-identificadores”, “cuasi-identificadores” o identificadores indirectos.

Existe un riesgo de que, una vez que se ha anonimizado un conjunto de datos, se pueda producir una desanonimización de éstos. Por lo tanto, es necesario tener una estimación objetiva de cuál es la probabilidad de reidentificación a partir del conjunto de cuasi-identificadores y, de esa forma, tener una medida de dicho riesgo.

La K-anonimidad es una propiedad de los datos anonimizados que permite cuantificar hasta qué punto se preserva la anonimidad de los sujetos presentes en un conjunto de datos en el que se han eliminado los identificadores. Dicho de otro modo, es una medida del riesgo de que agentes externos puedan obtener información de carácter personal a partir de datos anonimizados.

Si clasificamos los atributos de los registros según su naturaleza o tipo de información que contienen distinguimos los siguientes tipos de datos:

- Atributos clave o identificadores: son campos que identifican unívocamente a los sujetos de los datos (nombre, DNI, nº de pasaporte, teléfono, ...). Este tipo de datos deben eliminarse de los registros anonimizados.
- Cuasi-identificadores: son campos que, si bien por sí mismos y de forma aislada no identifican a un individuo, agrupados con otros atributos cuasi-identificadores pueden señalar de forma unívoca a un sujeto. Las técnicas de anonimización trabajan sobre estos datos, eliminando campos que no son necesarios para el tratamiento (en aplicación del principio de minimización), agregándolos o generalizándolos.
- Atributos sensibles: son los campos Atributos sensibles: son los campos que contienen datos que podrían tener un mayor impacto en la privacidad de un individuo concreto, entre ellos las categorías especiales de datos, y que no deben ser vinculados con el sujeto de datos al que pertenecen (enfermedades, tratamientos médicos, nivel de renta, ...). Esta información puede ser de gran interés en el objeto del tratamiento de datos, pero a menos que exista una legitimación para ello, debe mantenerse disociada de un sujeto concreto.

Se dice que un individuo es k-anónimo dentro del conjunto de datos en el que se encuentra incluido si, y sólo si, para cualquier combinación de los atributos cuasi-identificadores asociados, existen al menos otros  $K-1$  individuos que comparten con él los mismos valores para esos mismos atributos. Hay que tener en cuenta que la K-anonimidad no se centra en





los atributos sensibles de los registros, sino en los atributos cuasi-identificadores que pueden permitir la vinculación.

De este modo, la probabilidad de identificar a un individuo concreto en base a ese conjunto de cuasi-identificadores es como máximo  $1/K$ , por lo que para garantizar un bajo riesgo de reidentificación debe garantizarse un valor mínimo de  $K$  cuando se pretende llevar a cabo el diseño de un proceso de anonimización o disociación de datos

Imaginemos los siguientes conjuntos de datos en los que existen dos atributos de tipo cuasi-identificador como el “código postal” y la “edad” asociados a un atributo de tipo sensible que detalla datos de salud relativos a los sujetos de datos contenidos en el conjunto.

Código postal	Edad	Colesterol
37003	40	S
28108	44	S
24700	37	N
24700	37	N
37003	40	S
28108	44	S

*Tabla 1: 2-anonimización*

Código postal	Edad	Colesterol
37003	40	S 
28108	44	S 
24700	37	N
24700	37	N
37003	44	S 
28108	40	S 

*Tabla 2: 1-anonimización*

La tabla 1 está 2-anonimizada, puesto que cada combinación de valores de los atributos cuasi-identificadores aparece al menos en dos filas, mientras que la tabla 2 no lo está pues no existe para cada registro al menos otro que contenga valores idénticos para dichos atributos.

A la hora de diseñar un tratamiento en el que se requiera hacer uso de datos anonimizados es importante responder a las siguientes cuestiones:

- ¿Qué valor de  $K$  es adecuado?

Mayores valores de  $K$  se corresponden con requisitos de privacidad más exigentes dado que será necesaria la existencia de más sujetos dentro de un grupo que satisfagan idéntica combinación de rasgos identificativos. En la obtención de mayores valores de  $K$  se puede perder fidelidad en los datos de origen, por lo que hay que determinar si en esa pérdida de fidelidad hay o no pérdida de información que sea relevante para la finalidad del tratamiento. Si no hay pérdida de información relevante, hay que ejecutar ese proceso inicial. Si hay pérdida de información relevante, habrá que conseguir alcanzar el equilibrio entre los riesgos para los derechos y libertades de los sujetos y la potencial pérdida de fidelidad en el resultado del tratamiento.

- ¿Cómo conseguimos hacer un conjunto de datos  $K$ -anónimo? en el apartado siguiente vamos a estudiar distintas técnicas para conseguir conseguir que nuestras muestras sean conjuntos de datos  $k$ -anónimos

### 3. Métodos de K-anonimización

La Generalización y la eliminación son dos métodos no perturbativos porque logran la protección mediante la sustitución de los datos originales de los atributos por otros valores más generales sin introducir información errónea en la fuente de datos original.

- a) Generalización: La generalización consiste en hacer que el valor de los atributos cuasi-identificadores sea menos preciso, transformándolos o generalizándolos dentro de un conjunto o intervalo que comparte los mismos valores, bien mediante la creación de rangos en el caso de atributos numéricos o el establecimiento de jerarquías para los atributos nominales. De este modo, el número de registros que poseen los mismos valores para un conjunto de atributos cuasi-identificadores se puede incrementar con el objeto de satisfacer los requisitos de privacidad a la vez que sigue siendo posible cumplir con la finalidad del tratamiento.

Partiendo de la tabla 2 mostrada anteriormente, es posible transformarla en un conjunto de datos 2-anónimo realizando una generalización del atributo 'Edad' dentro de un rango numérico y del atributo 'Código postal' clasificado en una jerarquía

Código postal	Edad	Colesterol
37***	40 - 49	S
28***	40 - 49	S
24***	30 - 39	N
24***	30 - 39	N
37***	40 - 49	S
28***	40 - 49	S

- b) Eliminación: El otro método para implementar la K-anonimidad es la eliminación. En el ejemplo anterior los valores de los registros estaban bastante próximos entre sí, lo que permitía generalizar manteniendo una precisión razonable. Imaginemos que a la tabla 2 se le añaden más registros.

Para los seis primeros registros podemos hacer una generalización tal y como acabamos de ver, pero el último de los registros añadidos está fuera de rango. Intentar realizar una generalización definiendo un intervalo que lo contenga podría conllevar una

Código postal	Edad	Colesterol
37003	40	S
28108	44	S
24700	37	N
24700	37	N
37003	44	S
28108	40	S
37891	33	N
50011	13	S

pérdida de precisión tal que los datos dejarían de ser útiles para un análisis.

En estos casos, la solución pasa por suprimir o eliminar ese tipo de registros de modo que no “contaminen” el conjunto de datos y distorsionen los resultados. También los registros con valores muy poco usuales deben ser eliminados dado que aumentan significativamente la probabilidad de reidentificación.

Aplicando ambos métodos, generalización y eliminación, el nuevo conjunto de datos llegaría a ser 2-anónima:

Código postal	Edad	Colesterol
37003	40	S
28108	44	S
24700	37	N
24700	37	N
37003	44	S
28108	40	S
37891	33	N
50011	13	S

Código postal	Edad	Colesterol
37***	40 - 49	S
28***	40 - 49	S
24700	30 - 39	N
24700	30 - 39	N
37***	40 - 49	S
28***	40 - 49	S
37***	30 - 39	N

## 4. Herramientas para anonimizar los datos utilizando técnicas de K-anonimidad

Existen herramientas que permiten *K*-anonimizar el conjunto de datos que se les introduce como entrada. Algunos ejemplos de este tipo de herramientas que permiten implementar las técnicas de *K*-anonimidad son:

- **ARX Data Anonymization Tool:** ARX es una herramienta de código abierto . La herramienta soporta varias técnicas de privacidad, entre ellas la anonimidad. ARX es capaz de manejar grandes conjuntos de datos y cuenta con una interfaz gráfica multiplataforma intuitiva además de una API de integración con Java para implementar capacidades de anonimización de datos desde software desarrollado bajo este lenguaje de programación. Enlace de descarga: <https://arx.deidentifier.org/downloads/>
- **Herramienta de anonimización UTD:** Es una herramienta de código abierto desarrollada en el UT Dallas Data Security y Privacy Lab, que implementa varios métodos de anonimización para uso público por parte de investigadores. Los algoritmos se pueden usar tanto directamente contra un dataset o conjunto de datos como a través de librerías de funciones implementadas dentro de otras aplicaciones. Utiliza métodos de anonimización diferentes, entre ellos la k-anonimidad. Enlace de descarga: <http://cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php?go=download>

- **Amnesia:** Amnesia es una herramienta de anonimización de datos, que permite eliminar la información no sólo asociada a los identificadores directos como nombres o números de documentos identificativos, sino que también transforma los atributos cuasi-identificadores como la fecha de nacimiento y el código postal para mitigar los riesgos de reidentificación de los sujetos que figuran en las fuentes de datos, utilizando para ello métodos de k-anonimato. Dispone de una versión cliente y de una versión online.

Enlace de descarga: <https://amnesia.openaire.eu/installation.html>

Enlace versión online: <https://amnesia.openaire.eu/amnesia/>

En este lab utilizaremos la versión online de amnesia.

## 5. Caso práctico

Llevaremos a cabo la anonimización de datos personales propuesta sobre una muestra ficticia de 20 individuos, de ambos géneros, con edades comprendidas entre los 20 y los 60 años, a los que se les ha realizado un cribado. Hemos recogido los siguientes datos de cada individuo:

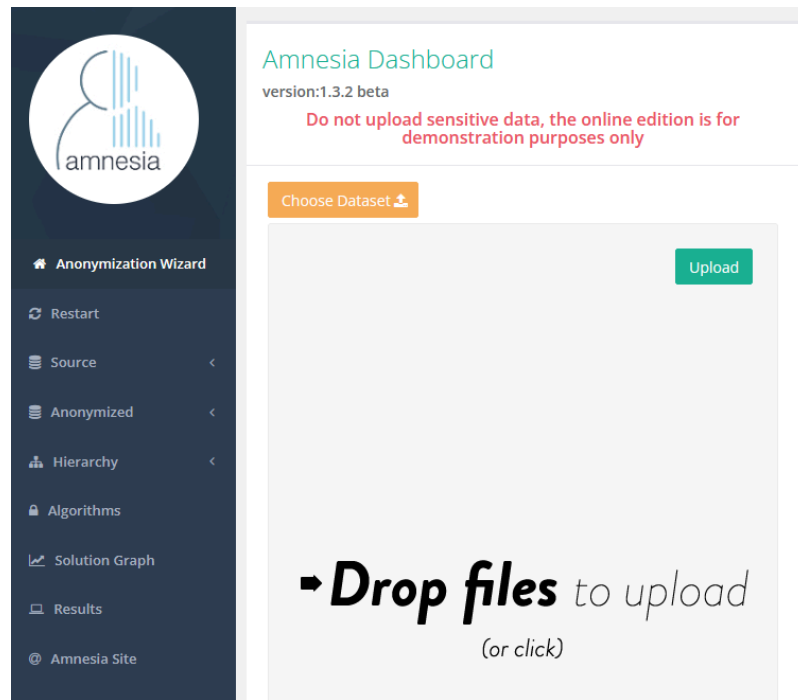
- DNI
- Nombre y apellidos
- Número de la seguridad social
- Dirección
- Código postal
- Edad
- Profesión
- Diagnóstico Positivo

DNI	Nombre	NSS	Direccion	CP	Edad	Profesion	Positivo
12348791	Andres Calamar	73648791	C. de Vicent Serra i Orvay 43	7800	22	Cientifico_de_datos	SI
45793584	Antonia Molina	45793224	Av. de les Corts Catalanes 18	8038	55	Programador_Multi	NO
15935748	Rebeca Ridruejo	14444448	C. Riu Segre 19	43006	60	Ingeniero_de_datos	SI
36987412	Felipe Romero	36985858	C. Catedratico Ferre Vidiella 4	3005	31	Ingeniero_de_datos	SI
32147985	Ramon Garcia	12127985	Av. Navarro Reverter 15	12400	40	Analista_de_datos	SI
25897461	Susana Gonzalez	99977461	C. de Josefa Valcarcel 44	28027	38	Administrador_Sist	NO
85974613	Rosario Martinez	97978441	Pl. Cervantes 1	7825	59	Cientifico_de_datos	SI
55669988	Ana Pacheco	99885566	C. Cordoba 6	8039	29	Analista_de_datos	NO
14747814	Luis Argentina	33474494	Pl. Rosalia 1	43011	35	Ingeniero_de_datos	NO
90348791	Noelia Sanchis	87917364	Av. Corts Valencianes 15	3000	42	Programador_Web	SI

Vamos a seguir los siguientes pasos:

1. Antes de empezar a trabajar con la herramienta, vamos a clasificar los atributos de nuestra información en identificadores, cuasi-identificadores y sensibles:
  - Los campos identificadores son: DNI, nombre y apellidos, número de la seguridad social y dirección
  - Los campos cuasi-identificadores son: edad, profesión y código postal
  - El campo especialmente protegido, es el que contiene la información sensible y en este caso es: diagnóstico positivo
2. Descarga de aules el archivo de datos: DataSet\_Personal.csv, se trata de una tabla de excel guardada como csv, formato que acepta Amnesia.

3. Nos conectamos a la URL: <https://amnesia.openaire.eu/amnesia/> y subimos el fichero que acabamos de descargar, arrastrando el fichero o pulsando Choose Dataset



4. Nos aparecerá la pantalla siguiente, en la que añadiremos “,” en la caja de texto **Delimiter\***. Pulsamos next

## Select the delimiters and the dataset type

## Dataset preview

**DNI,Nombre,NSS,Direccion,CP,Edad,Profesion,Positivo**

12348791,Andres Calamar,73648791,C. de Vicent Serra i Orvay 43,7800,22,Cientifico\_de\_datos,SI

45793584,Antonia Molina,45793224,Av. de les Corts Catalanes 184,8038,55,Programador\_Multi,NO

15935748,Rebeca Ridruejo,14444448,C. Riu Segre 19 ,43006,60,Ingeniero\_de\_datos,SI

...

## Delimiter \*

,

## Dataset type :

Simple table

Previous

Next

Cancel

5. En la siguiente pantalla deberemos eliminar los campos identificadores, desmarcándolos. Es decir, desmarcamos los campos dni, nombre, nss y dirección. Fíjate que Amnesia ha interpretado el tipo de cada campo, permite ahora en algunos casos cambiar el tipo. Los tipos permitidos en Amnesia son: strings, integers, floats y dates. Cambia el tipo del CP a string. Una vez seleccionados solo los campos cuasi-identificadores y especialmente protegido pulsamos sobre Finish



## Choose the attributes and their types

Select All ☒

Select the attributes to include in the output dataset and assign a type to each attribute.

DNI <input type="checkbox"/>	Nombre <input type="checkbox"/>	NSS <input type="checkbox"/>	Direccion <input type="checkbox"/>	CP <input checked="" type="checkbox"/>	Edad <input checked="" type="checkbox"/>	Profesion <input checked="" type="checkbox"/>	Positivo <input checked="" type="checkbox"/>
<input type="text" value="int"/>	<input type="text" value="string"/>	<input type="text" value="int"/>	<input type="text" value="string"/>	<input type="text" value="string"/>	<input type="text" value="int"/>	<input type="text" value="string"/>	<input type="text" value="string"/>
45793584	Antonia Molina	45793224	Av. de les Corts Catalanes 184	8038	55	Programador_Multi	NO
15935748	Rebeca Ridruejo	14444448	C. Riu Segre 19	43006	60	Ingeniero_de_datos	SI
36987412	Felipe Romero	36985858	C. Catedratico Ferre Vidiella 4	3005	31	Ingeniero_de_datos	SI
32147985	Ramon Garcia	12127985	Av. Navarro Reverter 15	12400	40	Analista_de_datos	SI
25897461	Susana Gonzalez	99977461	C. de Josefa Valcarcel 44	28027	38	Administrador_Sist	NO

Previous

Finish

Cancel

6. Ahora nos aparece el DataSet resultante de eliminar todos los campos identificadores. Fíjate que en los campos de tipo string aparece un botón de Masking. Al hacer clic en él, una ventana emergente muestra cada carácter de un valor aleatorio de la columna en un pequeño cuadro. En ese momento se puede establecer el carácter especial deseado para la máscara (por ejemplo, \*, &, ^, etc.). Y finalmente las posiciones donde el caracter elegido enmascarará el valor real. En este caso no utilizaremos este recurso

## Pre-Anonymization Process

Show  entries

CP	<input type="button" value="Masking"/>	Edad	Profesion	<input type="button" value="Masking"/>	Positivo	<input type="button" value="Masking"/>
7800		22	Cientifico_de_datos		SI	
8038		55	Programador_Multi		NO	
43006		60	Ingeniero_de_datos		SI	
3005		31	Ingeniero_de_datos		SI	
12400		40	Analista_de_datos		SI	
28027		38	Administrador_Sist		NO	
7825		59	Cientifico_de_datos		SI	
8039		29	Analista_de_datos		NO	
43011		35	Ingeniero_de_datos		NO	
3000		42	Programador_Web		SI	

Showing 1 to 10 of 55 entries

7. En esta misma pantalla podemos comprobar si se cumple una k-anonimización para una determinada combinación de los atributos cuasi-identificadores, pulsando sobre Check Anonymization. Veamos qué pasa con la combinación de los atributos cuasi-identificadores y con el objetivo de llegar al 2-anonimato.

### Check if dataset is Anonymous

Choose attributes :

☒ CP  
☒ Edad  
☒ Profesion  
☐ Positivo

Choose K :

2

Close


Show Anonymization

### Check if dataset is Anonymous

"CP,Edad,Profesion" is anonymous for k : 1.

Percentage of displayed dataset is : 100%

To produce a k = 2 suppress 100%



Close

Suppress

Apply

😞 No parece que tengamos anonimización, para conseguir un  $k=2$  sugiere eliminar el 100% de los datos. No parece razonable quedarnos sin dataset borrando todos los datos, así que pulsa sobre "Close"

8. Como con la técnica de borrado no conseguimos nuestro propósito, vamos a trabajar en la anonimización de los datos, utilizando jerarquías. Pulsamos en el botón de "Proceed to Hierarchies". Las jerarquías podemos definirlas manualmente nosotros en la web, subirlas o dejar que nos las genere automáticamente. Vamos a recurrir a esta última forma.
- En primer lugar vamos a generalizar el campo edad haciendo rangos de edad de 10 años, es decir de 20 a 30, de 30 a 40, de 40 a 50 y de 50 a 60. Pulsamos

sobre el botón Autogenerate Hierarchy y completamos las pantallas que nos aparecen de la siguiente manera:

#### Hierarchy Autogenerate

1. Choose Attribute and Hierarchy Type

2. Hierarchy Info

### Choose Attribute

On Attribute

Edad

Type

Range

VarType

Integer

Previous

Next

Cancel

Hemos elegido el atributo Edad que es entero y vamos a generalizar haciendo un rango, forma habitual cuando tenemos valores numéricos. Pulsamos Next y completamos como se muestra en la imagen:

#### Hierarchy Autogenerate

1. Choose Attribute and Hierarchy Type

2. Hierarchy Info

### Hierarchy Information

Step

10

Name

Edad\_hier

Domain start  
End limit

20-60

Fanout

10

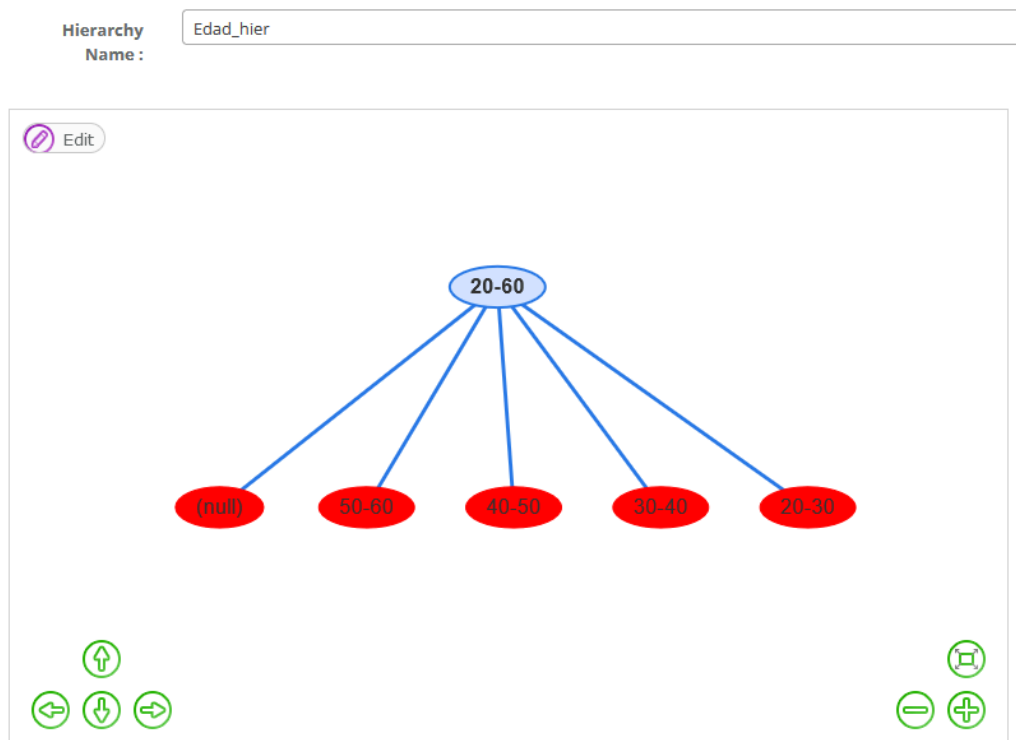
Previous

Finish

Cancel

El valor step es el tamaño del rango: en nuestro caso van a ser 10 años, el inicio y fin del dominio de los datos, en nuestro caso es 20 y 60 años respectivamente. El fallout, número de hijos por nodo lo dejamos por defecto a 10. Y pulsamos sobre Finish. Como se puede apreciar en la siguiente imagen nos ha salido una jerarquía de 2 niveles, en el nivel 1

tenemos las edades agrupadas en rangos de 10 y en el nivel 2 tendríamos el rango de edades total 20-60.



Vamos ahora a generalizar el campo profesión, generando su jerarquía. Pulsamos de nuevo en Autogenerate Hierarchy. Y completamos según las siguientes imágenes:

## Hierarchy Autogenerate

1. Choose Attribute and Hierarchy Type

2. Hierarchy Info

### Choose Attribute

On Attribute	Profesion	▼
Type	Group Based	▼
VarType	String	▼

Previous

Next

Cancel

## Hierarchy Autogenerate

1. Choose Attribute and Hierarchy Type

2. Hierarchy Info

### Hierarchy Information

Sorting	alphabetical	▼
Name	Profesion_hier	
Fanout	2	

Previous

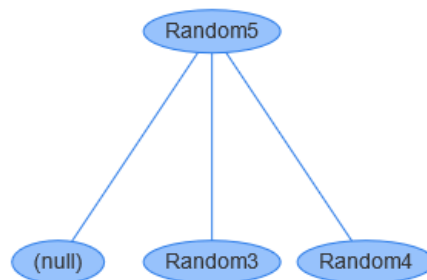
Finish

Cancel

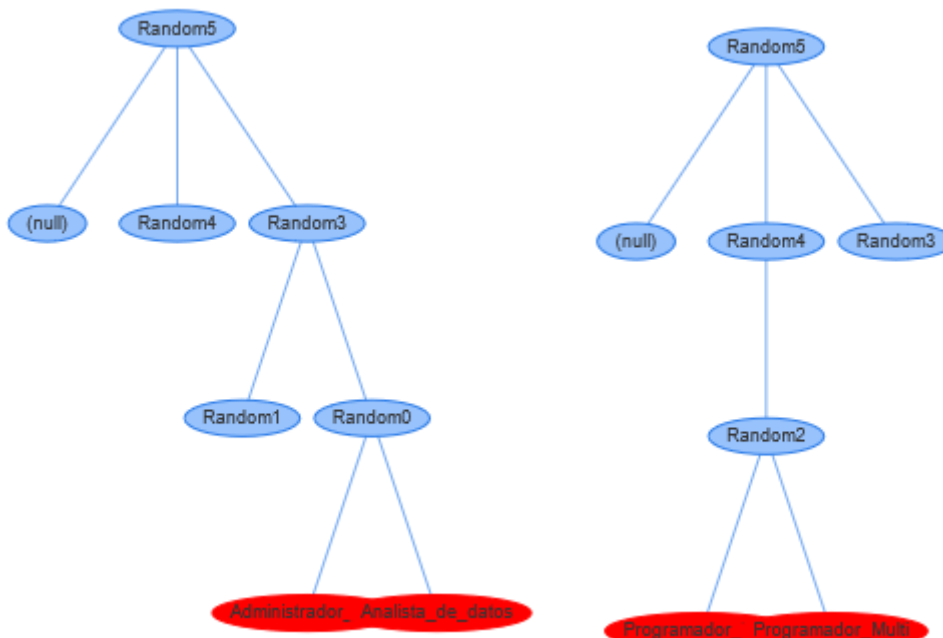
El tipo de jerarquía puede ser basada en agrupamiento o basada en máscara, como estamos en un campo no codificado donde podríamos aplicar máscara para generalizar, por ejemplo código postal elegimos Group Based. En cuanto a "Sorting" podemos hacer los grupos ordenando alfabéticamente las profesiones o aleatoriamente, en un caso real podríamos agrupar las profesiones por familia profesional y seguramente aportaría más

valor que las opciones automáticas que nos proporciona esta herramienta. Elegimos la alfabética, por ejemplo.

Respecto al campo “fanout”, indicar a modo general que cuando más bajo sea el valor numérico de este menor número de datos originales de la tabla en cuestión se agruparán bajo cada nodo, lo cual hará que inevitablemente aumente el número de nodos. Si elegimos el valor 2, eso quiere decir que cada 2 nodos hoja vamos a añadir un nodo de generalización. Lo vemos más claro analizando el resultado que nos aparece al pulsar Finish:



La figura no muestra todos los nodos, para ver que hay debajo de Random3 o Random4 tenemos que pulsar dos veces sobre estos nodos. Y 2 veces más en cada uno de los que van apareciendo. Observa que realmente ha agrupado de 2 en 2 las profesiones y que los grupos se han realizado por orden alfabético. Hemos conseguido una jerarquía de 4 niveles, nivel 0 cada profesión, nivel 1 los grupos Random 0, Random1 y Random2 que generalizan las profesiones de dos en dos y Nivel 2 agrupación de nuevo de dos en dos y así sucesivamente.



Por último nos queda crear la jerarquía del código postal. Procederemos de forma similar al caso de la edad, solo que en este caso el step lo pondremos a 1000.

1. Choose Attribute and Hierarchy Type

2. Hierarchy Info

### Choose Attribute

On Attribute

Type

VarType

### Hierarchy Information

Step

Name

Domain start

End limit

Fanout

9. Nos queda ahora aplicar estas jerarquías a nuestros datos. Desde la pantalla en la que estamos trabajando pulsamos sobre Proceed to Algorithms y vamos enlazando los campos con las correspondientes jerarquías. Como no tenemos muchos datos no podremos utilizar valores K altos, que es lo deseable. Pongamos 2, eso quiere decir que al menos existen 2 individuos para cualquier combinación de los atributos cuasi-identificadores. Y finalmente pulsamos Execute.

#### Bind Hierarchies with Attributes

Indicate with generalization hierarchy will be used for each dataset attribute. The same hierarchy can be used in multiple attributes. A hierarchy must be defined for each quasi identifier.

CP

cp\_hier

Edad

Edad\_hier

Profesion

Profesion\_hier

Positivo

#### Algorithm Parameters

Type:  
Parallel k-anonymization

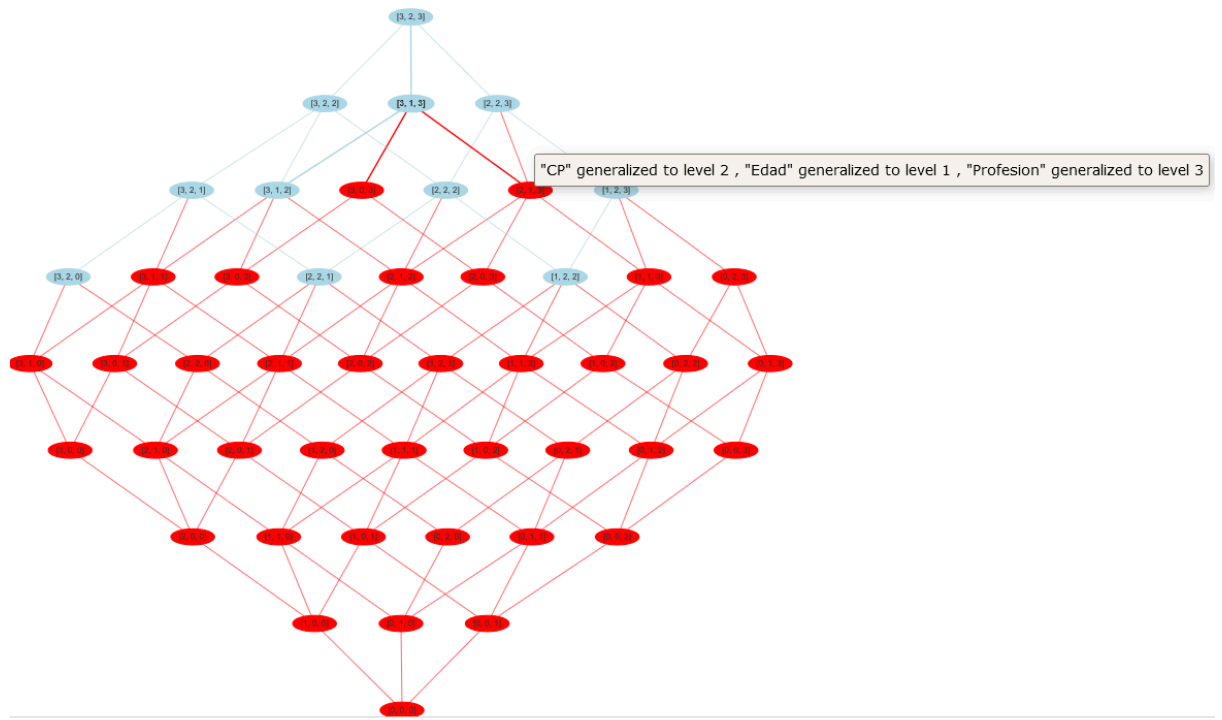
K:

-

2

+

Obtenemos el siguiente grafo



Donde los nodos azules están indicando soluciones seguras y los nodos rojos soluciones inseguras. Pasa por encima de los nodos el cursor para ver el nivel de generalización de los atributos. Selecciona un nodo para ver como quedarían los datos según esta anonimización y explorar sus estadísticas. Las soluciones inseguras pueden transformarse en seguras eliminando los datos que no las hacen seguras.

Si exploramos la opción 2(cp)1(edad)2(profesión), pulsando una vez sobre este nodo, obtenemos lo siguiente:

## Statistics and Anonymized menu

[Preview of the Anonymized dataset](#)  
[Show Statistics of the dataset with this solution](#)  
[Show Information Loss of this solution](#)

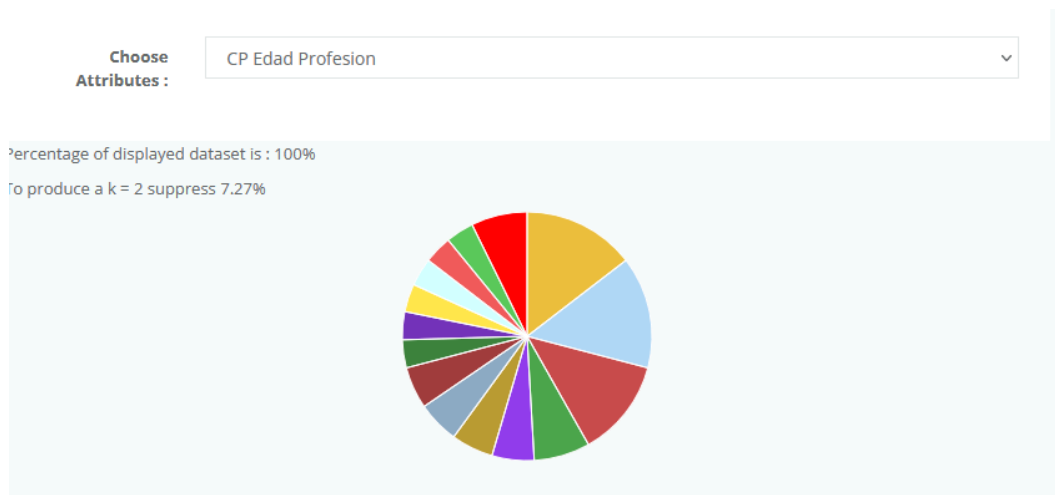
Close

Previsualizar los datos anonimizados. Observa como se ha generalizado cada columna al nivel correspondiente 2-1-2 respectivamente



CP	Edad	Profesion	Positivo
3000-13000	20-30	Random3	SI
3000-13000	50-60	Random4	NO
33000-43055	50-60	Random3	SI
3000-13000	30-40	Random3	SI

### Show statistics of the data set with this solution



Supongamos que nos podemos permitir perder el 7,27% de los datos y así tener un k=2 anonimización, vamos a darle a suprimir y aplicar a esos datos para así obtener nuestro data set 2-anonimizado. Ahora podemos salvar nuestro dataset resultante y trabajar con él. Observa que las filas a borrar aparecen en rojo.

## 6. Conclusiones

El deber del responsable del tratamiento es velar por la privacidad de los sujetos de los que trata datos. Algunas entidades consideran que suprimir o enmascarar los atributos de carácter identificador resulta suficiente para garantizar la anonimidad de los sujetos objeto de estudio, sin embargo, es posible que campos comunes presentes en diferentes fuentes de datos, convenientemente agrupados y cruzados, se conviertan en un atributo seudointificador que llegue a comprometer la privacidad de las personas.

Para ello, durante las fases de concepción y diseño de un tratamiento de datos de carácter personal, se ha de realizar un análisis del grado de fidelidad necesario en el resultado del tratamiento para determinar, de forma precisa, los márgenes adecuados de generalización y eliminación, dentro de límites razonables que impidan la distorsión de la realidad.

Igualmente, hay que hacer un análisis y correcto balance entre los riesgos para los derechos y libertades de los ciudadanos y los beneficios legítimos y para la sociedad que conlleva la realización de dicho tratamiento con un determinado grado de precisión.

Derivado de ambos análisis, es preciso alcanzar un equilibrio entre el beneficio que se obtendrá para la sociedad en la realización de un tratamiento con un grado de fidelidad determinado y el coste que dicho tratamiento implica para los derechos y libertades de los sujetos de los datos

## 7. Otros ejercicios

- a. En Aules hay una jerarquía diferente de profesiones, abre el fichero .txt para ver que sintaxis se utiliza en la definición de jerarquías. Importa esta jerarquía y aplícala a los datos.
- b. Edita la jerarquía de códigos postales, agrupando los valores por comunidades autónomas. En el caso de los datos de Madrid y las islas baleares vamos a imaginar que los códigos postales 28XXX son 38XXX y haremos una agrupación denominada islas. Limita en esta jerarquía los valores de fanout a 2. ¿Cuántos niveles te salen? Aplica esta nueva jerarquía a los datos.

## 8. Referencias

- a. [Nota técnica de la Unidad de Evaluación y Estudios Tecnológicos de la AEPD: La K-ANONIMIDAD como medida de la privacidad.](#)
- b. [Anonimización de datos en Amnesia: un caso práctico en investigación biomédica](#)
- c. [Documentación de Amnesia](#)
- d. [Video tutoriales de Amnesia](#)