



UNIDAD 8

PRINCIPIOS LEGALES

ANEXO -1 ANONIMIZACIÓN

1. CONCEPTOS



- **Anonimización:** consiste en la conversión de datos personales en datos que no se pueden utilizar para identificar a ningún individuo
- **Desidentificación:** consiste en la eliminación de identificadores (por ejemplo, nombre, dirección, número de documento nacional de identidad) que identifican directamente a un individuo
- **Reidentificación:** se refiere a la identificación de individuos a partir de un conjunto de datos que previamente fue desidentificado o anonimizado
- **Reversibilidad:** en el caso que la organización que aplica la anonimización conserve la capacidad de recrear el conjunto de datos original a partir de los datos anonimizados; el proceso de anonimización es “reversible”

1. CONCEPTOS



Importante:

- Ninguna técnica de anonimización podrá garantizar en términos absolutos la imposibilidad de la reidentificación, ya que existirá siempre un índice de probabilidad de reidentificación que debemos intentar atenuar mediante la correspondiente gestión de riesgos.
- Los datos anonimizados no se consideran datos personales y, por lo tanto, **no** se rigen por la normativa de protección de datos.
- La desidentificación se puede confundir con la anonimización, pero es solo el primer paso de la anonimización. Un conjunto de datos desidentificado puede volver a identificarse fácilmente cuando se combina con datos que son de acceso público o fácil.

1. CONCEPTOS

Ejemplo.

Datos personales → Desidentificación → Datos desidentificados

Registro de datos de Albert en SuperHungry

Nombre

Alberto Ruiz

Restaurante favorito

Restaurante Katong

Comida favorita

Combo de 3 piezas de pollo

Fecha de nacimiento

01/01/1990

Género

Hombre

Empresa

ABC Pte Ltd

Nombre

Alberto Ruiz

Restaurante favorito

Restaurante Katong

Comida favorita

Combo de 3 piezas de pollo

Fecha de nacimiento

01/01/1990

Género

Hombre

Empresa

ABC Pte Ltd

Combinando la información desidentificada con redes sociales podríamos seguramente identificar a Alberto.

1. CONCEPTOS

- **Identificadores directos:** atributos de datos exclusivos de un individuo y se pueden usar como atributos de datos clave para volver a identificar a un individuo
- **Identificadores indirectos o seudoidentificadores:** atributos de datos que no son exclusivos de un individuo, pero pueden volver a identificar a un individuo cuando se combinan con otra información
- **Atributos objetivo:** atributos de datos que contienen la utilidad principal del conjunto de datos. Este tipo de atributos de datos puede ser de naturaleza sensible y puede dar lugar a un alto potencial de efecto adverso para un individuo cuando se divulga

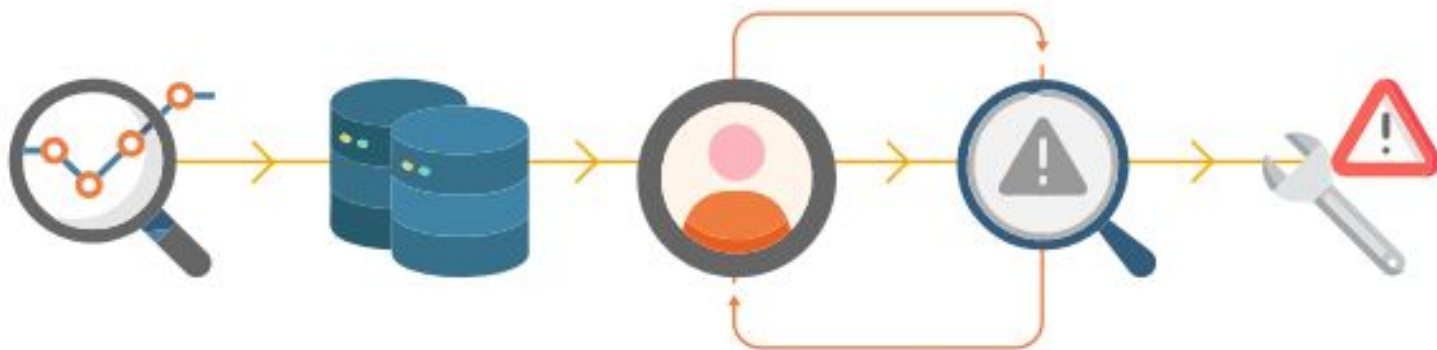
1. CONCEPTOS



Ejemplos
comunes en
un conjunto de
datos

Identificadores directos	Identificadores indirectos o seudoindentificadores	Atributos objetivo
<ul style="list-style-type: none">• Nombre• Dirección de correo electrónico• Número de teléfono móvil• Número DNI• Número de pasaporte• Número de cuenta• Número de certificado de nacimiento• Número de permiso de trabajo• Nombre de usuario de redes sociales	<ul style="list-style-type: none">• Edad• Género• Carrera• Fecha de nacimiento• Dirección• Código postal• Título del trabajo• Nombre de la empresa• Estado civil• Altura• Peso• Dirección de protocolo de Internet (IP)• Número de matrícula del vehículo• Número de bastidor del vehículo• Localización del Sistema de Posicionamiento Global (GPS)	<ul style="list-style-type: none">• Transacciones (por ejemplo, compras)• Salario• Calificación crediticia• Póliza de seguro• Diagnóstico médico• Estado de vacunación

2. Proceso de anonimización



Paso 1

Conoce tus
datos

Paso 2

Desidentifique
sus datos

Paso 3

Aplice
técnicas de
anonimización

Paso 4

Calcule su
riesgo

Paso 5

Gestione sus
riesgos

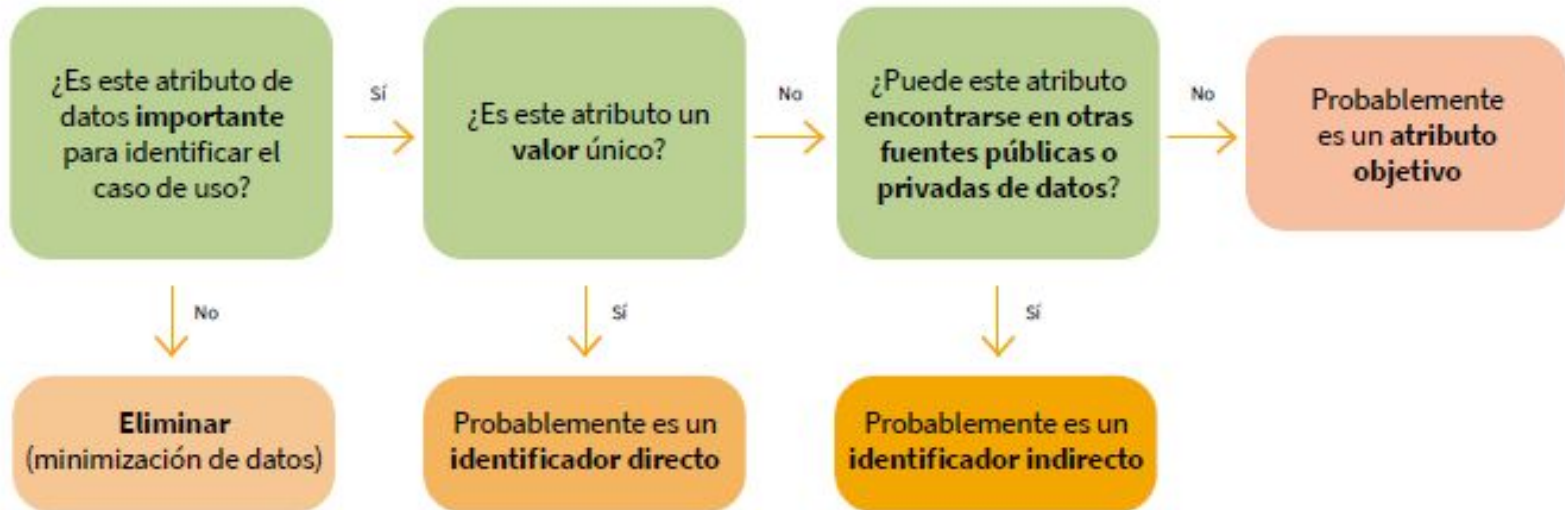
2.1 Conoce tus datos



- Un registro de datos personales se compone de atributos de datos que tienen diversos grados de identificabilidad y sensibilidad a un individuo, identificadores directos, indirectos y atributos objetivos.
- Cualquier atributo de datos que no sea necesario en el conjunto de datos resultante debe eliminarse como parte de la minimización de datos
- Generalmente los identificadores directos se eliminan
- Se modifican con técnicas de anonimización los identificadores indirectos
- Normalmente los atributos objetivos se mantienen sin cambios.

2.1 Conoce tus datos

- Diagrama de flujo para clasificar los atributos



2.1 Conoce tus datos

- Ejercicio 1: Clasifica los atributos de las siguiente fuente de datos
 - a)

Nº personal	Nombre	Género	Fecha de nacimiento	Código postal	Ocupación	Ingresos	Educación	Estado civil
39192	Sandra García	M	05/08/1975	570150	Científico de datos	53.000€	Master	Viuda
37030	Paula Martínez	H	14/12/1973	787589	Profesor universidad	52.000€	Doctorado	Casado
22722	Bernardo Sánchez	H	02/03/1985	408600	Investigador	47.000€	Doctorado	Divorciado
28760	Estefanía Gómez	M	27/03/1968	570150	Administrador	48.000€	Licenciado	Casada
13902	Javier Muñoz	H	25/06/1967	199588	Arquitecto	50.000€	Master	Soltero

2.1 Conoce tus datos

- Ejercicio 1: Clasifica los atributos de las siguiente fuente de datos
 - b)

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

2.2 Desidentifica tus datos

- Este paso es siempre parte del proceso de anonimización
- Consiste en eliminar todos los identificadores directos.
- Opcionalmente se asigna un seudónimo a cada registro si se desea conservar la capacidad de vincular el registro de datos desidentificados al registro original en un momento posterior.

Nombre	Token	Edad	Serie favorita
Alex	1234	25	The Big Bang Theory
Bosco	5678	54	Friends
Charlene	5432	42	Grey's Anatomy

Nombre	Token
Alex	1234
Bosco	5678
Charlene	5432

2.2 Desidentifica tus datos

- Ejercicio 2: Desidentifica tus datos

Nº personal	Nombre	Departamento	Género	Fecha de nacimiento	Fecha de incorporación	Tipo de jornada
39192	Sandra García	Investigación & Desarrollo	M	08/01/1971	02/03/1997	Media jornada
37030	Paula Martínez	Ingeniería	M	15/05/1976	08/03/2015	Jornada completa
22722	Bernardo Sánchez	Ingeniería	H	31/12/1973	30/07/1991	Jornada completa
28760	Estefanía Gómez	Ingeniería	M	24/12/1970	18/03/2010	Media jornada
13902	Javier Muñoz	Recursos Humanos	H	15/07/1973	28/05/2012	Media jornada

2.3 Aplicar técnicas de anonimización



- Las técnicas de anonimización se aplica a los identificadores indirectos.
- El objetivo es evitar que se puedan combinar fácilmente con otros conjuntos de datos que puedan contener información adicional para volver a identificar a las personas
- Técnicas:
 - Supresión de registros
 - Supresión de atributos
 - Enmascaramiento de caracteres
 - Generalización
 - Perturbación de datos
 - Seudonimización
 - Intercambio de datos

2.3 Aplicar técnicas de anonimización



- **Supresión de registros:**
 - La eliminación de un registro (es decir, una fila de datos, especialmente cuando dichos datos pueden contener valores de datos únicos que no se pueden anonimizar más)
 - La supresión de registros se utiliza para eliminar registros atípicos que son únicos o que no cumplen otros criterios, como la k-anonimidad. Los valores atípicos pueden conducir a una fácil reidentificación.

2.3 Aplicar técnicas de anonimización



- **Enmascaramiento de caracteres:**
 - la sustitución de algunos caracteres del valor de datos por un símbolo coherente (por ejemplo, * o x). Por ejemplo, enmascarar un código postal implicaría cambiarlo de “28029” a “28xxx”.
 - El enmascaramiento de caracteres se utiliza cuando el valor de los datos es una cadena de caracteres y ocultar parte de ella es suficiente para proporcionar el grado de anonimato requerido
 - No confundir con el caso de los enmascaramientos que se realiza para que los interesados puedan reconocer sus propios datos(Ej: enmascaramiento de los dígitos de la cuenta bancaria en las notificaciones del banco)

2.3 Aplicar técnicas de anonimización




- **Seudonimización:**
 - La seudonimización se refiere a la sustitución de datos de identificación por valores inventados.
 - Los seudónimos pueden ser irreversibles cuando los valores originales se eliminan
 - Pueden ser reversibles (por el propietario de los datos originales) cuando los valores originales se guardan de forma segura, pero se pueden recuperar y vincular al seudónimo en caso de que surja la necesidad

2.3 Aplicar técnicas de anonimización



- **Generalización:**
 - La reducción de la granularidad de los datos (por ejemplo, mediante la conversión de la edad de una persona en un rango de edad). Por ejemplo, generalizar la edad de una persona de “26 años” a “25-29 años”.
 - La generalización se utiliza para valores que pueden generalizarse y seguir siendo útiles para el propósito previsto
 - Considere la posibilidad de suprimir cualquier registro que aún se destaque después de la generalización
 - Un rango de datos que es demasiado grande puede implicar una pérdida significativa en la utilidad de datos, mientras que un rango de datos que es demasiado pequeño puede significar que los datos apenas se modifican
 - Si se utiliza la k-anonimidad, el valor k elegido también afectará al rango de datos. Tenga en cuenta que el primer y el último rango pueden ser un rango más grande para acomodar el número típicamente menor de registros en estos extremos

Ejercicio 3. Aplica las técnicas de generalización y enmascaramiento al siguiente conjunto de datos



Persona	Edad	Dirección
357703	24	Avenida de Madrid, 22 1ºB
233121	31	Calle Alcalá, 18 1ºB
938637	44	Calle Mayor, 27 3ºB
591493	29	Avenida de Madrid, 22 6ºC
202626	23	Calle Serrano, 40 3ºD
888948	75	Carretera de Stonehenge, 5
175878	28	Calle Serrano, 40 5ºA
312304	50	Calle Mayor, 27 1ºA
214025	30	Avenida de Madrid, 22 2ºA
271714	37	Alcalá, 18 1ºC
341338	22	Calle Serrano, 40 7ºA
529057	25	Calle Serrano, 40 2ºB
390438	39	Calle Alcalá, 18 4ºB

2.3 Aplicar técnicas de anonimización



- **Intercambio de datos:**
 - Es reorganizar los datos en el conjunto de datos de modo que los valores de los atributos individuales sigan representados en el conjunto de datos, pero generalmente no correspondan a los registros originales.
 - El intercambio se utiliza cuando no hay necesidad de analizar las relaciones entre los atributos a nivel de registro
 - Consiste en reasignar el valor de unos atributos a otros registros del conjunto de datos.

2.3 Aplicar técnicas de anonimización

- Ejemplo:

Antes de la anonimización				
Persona	Título del trabajo	Fecha de nacimiento	Tipo de membresía	Promedio de visitas por mes
A	Profesor universitario	03/01/1970	Plata	0
B	Vendedor	05/02/1972	Platino	5
C	Abogado	07/03/1985	Oro	2
D	Profesional de TI	10/04/1990	Plata	1
E	Enfermera	13/05/1995	Plata	2

Después de la anonimización*				
Persona	Título del trabajo	Fecha de nacimiento	Tipo de membresía	Promedio de visitas por mes
A	Abogado	10/04/1990	Plata	1
B	Enfermera	07/03/1985	Plata	2
C	Vendedor	13/05/1995	Platino	5
D	Profesional de TI	03/01/1970	Plata	2
E	Profesor universitario	05/02/1972	Oro	0

2.3 Aplicar técnicas de anonimización



- **Perturbación de datos:**
 - Modificación de los valores en los datos agregando “ruido” a los datos originales (por ejemplo, valores aleatorios +/- a los datos).
 - La perturbación de datos se utiliza para identificadores indirectos (normalmente números y fechas), que pueden ser potencialmente identificables cuando se combinan con otras fuentes de datos, pero los cambios leves en el valor son aceptables para el atributo
 - Entre las técnicas están la de redondeo y agregar ruido aleatorio
 - Cuando el cálculo se realiza sobre valores de atributos que se han perturbado antes, el valor resultante puede experimentar perturbaciones en una medida aún mayor

2.3 Aplicar técnicas de anonimización

- Ejemplo de redondeo:

Atributo	Anonimación técnica
Altura (en cm)	Redondeo base-5 (se elige 5, siendo algo proporcional al valor de altura típico de 120 a 190 cm).
Peso (en kg)	Redondeo base-3 (se elige 3, siendo algo proporcional al valor de peso típico de 40 a 100 kg).
Edad (en años)	Redondeo de base 3 (se elige 3, siendo algo proporcional al valor de edad típico de 10 a 100 años).

2.3 Aplicar técnicas de anonimización

- Ejemplo de redondeo:

Altura	Peso (kg)	Edad (años)
160	50	30
177	70	36
158	46	20
173	75	22
169	82	44

Altura	Peso (kg)	Edad (años)
160	51	30
175	69	36
160	45	18
175	75	21
170	81	42

2.3 Aplicar técnicas de anonimización



Ejercicio 5.

Piensa al menos dos técnicas de anonimización para los siguientes identificadores.
Muestra el resultado con ejemplos

- dni
- edad
- altura
- peso
- raza
- fecha de nacimiento
- dirección
- código postal

2.4 Calcula tu riesgo



- Una técnica para calcular el nivel de riesgo de reidentificación de un conjunto de datos es la **k-anonimidad**:
 - Se refiere al k número de registros idénticos que se pueden agrupar en un conjunto de datos
 - Solo se consideran los identificadores indirectos para su cálculo
 - Un valor de k-anonimidad más alto significa que existe un menor riesgo de reidentificación
 - Un valor de k-anonimidad de 1 significa que el registro es único
 - El umbral de la industria para el valor de k-anonimidad es de 3 ó 5
 - Si no se alcanza el umbral de la k-anonimidad hay que volver al paso anterior y seguir aplicando técnicas de anonimización

Ejercicio 4. Anonimiza el siguiente conjunto de datos hasta obtener k-anonimización con $k=5$



Conjunto de datos antes de la anonimización				
Número de serie	Edad	Género	Ocupación	Promedio de viajes por semana
1	21	Femenino	Oficial Asistente de Protección de Datos	15
2	38	Masculino	Consultor Líder de TI	2
3	25	Femenino	Banquero	8
4	34	Masculino	Administrador de bases de datos	3
5	30	Femenino	Director de Privacidad	1
6	29	Femenino	Delegado Regional de Protección de Datos	5
7	38	Masculino	Programador	3
8	32	Masculino	Analista de TI	4
9	25	Femenino	Delegado Adjunto de Protección de Datos	2
10	23	Femenino	Gerente, Oficina de DPO	11
11	31	Masculino	Diseñador UX	0

2.5 Gestión de los riesgos de reidentificación y divulgación



- Riesgos
 - **Reidentificación**: puede surgir cuando la anonimización no fue suficiente, la reidentificación mediante vinculación o la inversión del seudónimo (si el seudónimo se creó con un algoritmo fácilmente adivinable)
 - **Revelación de atributos**: consiste en determinar con alto nivel de confianza que un atributo descrito en el conjunto corresponde a un individuo aunque no se pueda distinguir el registro (ej: registros anónimos de clientes de un cirujano estético en particular que revela que todos sus clientes menores de 30 años se han sometido a un procedimiento en particular. Si se sabe que un individuo en particular tiene 28 años y es cliente de este cirujano, entonces sabemos que este individuo se ha sometido al procedimiento en particular)

2.5 Gestión de los riesgos de reidentificación y divulgación



- Riesgos
 - **Revelación de inferencias:** Hacer una inferencia, con un alto nivel de confianza, sobre un individuo, incluso si él o ella no está en el conjunto de datos por propiedades estadísticas del conjunto de datos. Por ejemplo, si un conjunto de datos publicado por un investigador médico revela que el 70% de las personas mayores de 75 años tienen una determinada afección médica, esta información podría inferirse sobre una persona que no está en el conjunto de datos.

2.5 Gestión de los riesgos de reidentificación y divulgación



- Medidas de protección
 - Control de acceso a nivel de aplicación: nivel mínimo de complejidad de la contraseña 12 caracteres alfanuméricos, mayúsculas, minúsculas, números, y caracteres especiales. Hacer revisiones periódicas de que las cuentas y derechos asignados sean correctos.
 - Protección de los equipos mediante contraseña, bloquear la pantalla después de un período de inactividad
 - Cifrar los datos, revisando que el método es reconocido por la industria como relevante y seguro
 - Cifrar las tablas de correspondencia de identidades y no compartirse
 - Desarrollar plan de gestión para responder a las incidencias del tipo pérdida de datos
 - Mantener un registro de todos los datos compartidos/desidentificados/anonimizados para garantizar que los datos compartidos combinado no den lugar a una nueva identificación.
 - Realizar periódicamente revisiones de reidentificación
 - Cumplir las políticas de confidencialidad
 - Eliminar los datos cuando no sean necesarios
 - Realizar periódicamente auditorías para garantizar el cumplimiento de todas las medidas.



Referencias