UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Guía de anonimización y pseudonimización de datos para personal sanitario

Trabajo Fin de Máster

Máster Universitario en Ciberseguridad y Ciberinteligencia

AUTOR/A: Fort Palau, César Martín

Tutor/a: Oltra Gutiérrez, Juan Vicente

CURSO ACADÉMICO: 2022/2023

Resumen

La creciente utilización de datos médicos facilitada por el avance de las tecnologías de la información y las comunicaciones unida a la evolución de la legislación europea en materia de protección de datos personales coloca a quien quiera extraer y procesar dichos datos en una situación en la que debe buscar la forma de utilizar datos potencialmente sensibles de una forma segura respetando la privacidad de los pacientes manteniendo la utilidad de la información. El objetivo principal del presente trabajo de final de máster es el diseño de un marco teórico que contenga las técnicas de anonimización y seudonimización más conocidas, y como objetivo secundario, la creación de una guía que permita al investigador sanitario la selección y aplicación de la técnica de anonimización o seudonimización más adecuada según sus necesidades.

Palabras clave: anonimización, seudonimización, ciencia de datos, protección de datos personales, datos de salud

Abstract

The increasing use of medical data eased by the advances in information and communication technologies together with the evolution of European legislation on the protection of personal data puts those who want to extract and process such data in a situation in which they must find a way to use potentially sensitive data in a safe way while respecting the privacy of patients and maintaining the usefulness of the information. The main objective of this master's thesis is the design of a theoretical framework containing the best known anonymization and pseudonymization techniques, and as a secondary objective, the creation of a guide that allows healthcare researcher to select and apply the most appropriate anonymization or pseudonymization technique according to his or her needs.

Keywords: anonymization, anonymization, pseudonymization, data science, personal data protection, healthcare data

Resum

La creixent utilització de dades mèdiques facilitada per l'avanç de les tecnologies de la informació i les comunicacions unida a l'evolució de la legislació europea en matèria de protecció de dades personals posa a qui vulga extraure i processar aquestes dades en una situació en la qual ha de buscar la manera d'utilitzar dades potencialment sensibles d'una forma segura respectant la privacitat dels pacients mantenint la utilitat de la informació. L'objectiu principal del present projecte de final de màster és el disseny d'un marc teòric que continga les tècniques d'anonimització i seudonimiztació més conegudes, i com a objectiu secundari, la creació d'una guia que permeta a l'investigador sanitari la selecció i aplicació de la tècnica d'anonimització o seudonimització més adequada segons les seues necessitats.

Paraules clau: anonimització, peudonimització, ciència de dades, protecció de dades personals, dades de salut

Agradecimientos

Este trabajo ha sido posible gracias a la guía y los consejos de mi director, el profesor D. Juan Vicente Oltra Gutiérrez. Empezando por que me animó a cursar este máster, y después siguiendo sus indicaciones a la hora de elaborar este trabajo pero aun así dejándome buscar mi propio camino, he sido capaz de plasmar la vaga idea de la que partía en un principio, cuando quería aprovechar los conocimientos recibidos para ayudar en la medida de mis posibilidades a médicos y enfermeras a tratar de una forma segura una información que probablemente les haya costado mucho esfuerzo obtener, cuidando uno de los derechos fundamentales de los pacientes como es su protección en relación con el tratamiento de sus datos personales.

Quisiera también agradecerle al profesor D. Emilio Pedro Vivancos Rubio el haberme aportado los conocimientos en materia de anonimización y seudonimización de los que he partido para el desarrollo de este trabajo, enclavados dentro de la asignatura Criptología y Seguridad de los Datos, aprovechando para disculparme por la intensidad que suelo aportar al aprendizaje cuando lo estoy disfrutando.

Pero sobretodo, no hubiera sido capaz de llevar a cabo esta nueva locura que ha sido estudiar este máster sin el apoyo y la paciencia de Ángela, mi eterna compañera de vida, y también de nuestra hija Julia, a la que he privado de demasiados juegos y atenciones.

A todos y cada uno de ellos les estoy inmensamente agradecido por haberme ayudado a lograr esta nueva meta que tuve la osadía de plantearme.

Sumario

Glosario....7

Listado de Acrónimos....9

Índice de figuras....10

Índice de tablas....11

1. Introducción....13

1.1. Presentación....13

1.2. Motivación....13

1.3. Objetivos del trabajo....13

1.4. Estructura del trabajo....14

2. Descripción del problema....14

3. Plan de trabajo....16

4. Contexto tecnológico....20

4.1. Conceptos básicos....20

4.1.1. Tipos de atributos....20

4.1.2. Datos personales y datos disociados....20

4.1.3. Anonimización frente a seudonimización....20

4.2. Terminologías médicas....21

4.2.1. CIE-11....21

4.2.2. SNOMED CT....22

4.3. Técnicas de seudonimización....25

4.3.1. Contador....25

4.3.2. Números aleatorios....26

4.3.3. Función resumen....26

4.3.4. Cifrado simétrico....27

4.3.5. Cifrado asimétrico....28

4.4. Técnicas de anonimización....31

4.4.1. Supresión de atributos....31

4.4.2. Supresión de registros....32

4.4.3. Enmascaramiento de cadenas....33

4.4.4. Generalización....33

4.4.5. K-anonimidad....35

4.5. Consideraciones adicionales sobre la seguridad de la información....37

4.5.1. Finalidad....37
4.5.2. Consentimiento....37
4.5.3. Riesgo del uso de datos anonimizados o seudonimizados....38
5. Solución propuesta....39
5.1. Propósito....39
5.2. Datos de entrada....39
5.3. Información de salida....40
5.4. Asignación del método más adecuado....40
5.5. Contenido de la guía....42
6. Conclusiones. Líneas futuras....43
6.1. Conclusiones....43
6.2. Líneas futuras....43
Bibliografía y referencias....45
ANEXO I. Guía de anonimización y pseudonimización de datos para personal sanitario 47
0. Introducción....47
1. Identificadores, seudoidentificadores, datos sensible y no sensibles....47
2. Anonimización y seudonimización....48
3. Comenzando....48
4. Selección de la técnica....49
5. Técnicas....51
6. Consideraciones adicionales....58
ANEXO II. Objetivos de desarrollo sostenible....59
Reflexión sobre la relación del TFM con los ODS y con el/los ODS más relacionados....59

Glosario

Afectado o interesado: Según el [RD 1720/2007], es la persona física titular de los datos que sean objeto del tratamiento.

Anonimizar: Según el diccionario de la [RAE 2023], es expresar un dato relativo a entidades o personas eliminando la referencia a su identidad.

Atributos cuasi-identificadores: Según [Casas Roma y Romero Tris 2017], los cuasi-identificadores son un conjunto de atributos que potencialmente podrían identificar a un individuo. Son atributos que por sí mismos no permiten identificar a una persona, pero que junto con otros sí lo permitirían, como la fecha de nacimiento, la dirección postal, o el teléfono.

Atributos identificadores: Según [Casas Roma y Romero Tris 2017], son un conjunto de atributos que permiten identificar de forma explícita a un individuo; por ejemplo, el nombre y los apellidos, el DNI o el número de la seguridad social.

Atributos sensibles y no sensibles: Según [Casas Roma y Romero Tris 2017], los atributos sensibles son aquellos que presentan información específica y sensible de un individuo en concreto. Ejemplos de datos sensibles podrían ser las enfermedades que padece una persona, su salario o sus creencias religiosas. Por extensión, los atributos no sensibles son aquellos que no presentan información específica y sensible de un individuo en concreto. Ejemplos de datos no sensibles podrían ser si tiene o no una mascota, o su nivel de estudios.

Autenticación: Según el [RD 1720/2007], es el procedimiento de comprobación de la identidad de un usuario.

Cesión o comunicación de datos: Según el [RD 1720/2007], es un tratamiento de datos que supone su revelación a una persona distinta del interesado.

Consentimiento del interesado: Es, según el [RD 1720/2007], toda manifestación de voluntad, libre, inequívoca, específica e informada, mediante la que el interesado consienta el tratamiento de datos personales que le conciernen.

Contraseña: Es, según el [RD 1720/2007], información confidencial, frecuentemente constituida por una cadena de caracteres, que puede ser usada en la autenticación de un usuario o en el acceso a un recurso.

Datos disociados: Según el [RD 1720/2007] son aquellos que no permiten la identificación de un afectado o interesado.

Datos personales: Según el [RGPD 2017] (Reglamento General de Protección de Datos), son toda información sobre una persona física identificada o identificable («el interesado»); se considerará persona física identificable toda persona cuya identidad pueda determinarse, directa o indirectamente, en particular mediante un identificador, como por ejemplo un nombre, un número de identificación, datos de localización, un identificador en línea o uno o varios elementos propios de la identidad física, fisiológica, genética, psíquica, económica, cultural o social de dicha persona.

Destinatario o cesionario: Es, según el [RD 1720/2007], la persona física o jurídica, pública o privada u órgano administrativo, al que se revelen los datos. Podrán ser también destinatarios los entes sin personalidad jurídica que actúen en el tráfico como sujetos diferenciados.

Identificación: Es, según el [RD 1720/2007], el procedimiento de reconocimiento de la identidad de un usuario.

Persona identificable: Según el [RD 1720/2007], es toda persona cuya identidad pueda determinarse, directa o indirectamente, mediante cualquier información referida a su identidad física, fisiológica, psíquica, económica, cultural o social. Una persona física no se considerará identificable si dicha identificación requiere plazos o actividades desproporcionados.

Procedimiento de disociación: Es, según el [RD 1720/2007], todo tratamiento de datos personales que permita la obtención de datos disociados.

Recurso: Es, según el [RD 1720/2007], cualquier parte componente de un sistema de información.

Reidentificación: Es la determinación, con un alto nivel de confianza, de la identidad de un individuo descrito por un registro específico.

Seudonimización o pseudonimización de datos: Es, según, el [RGPD 2017], el proceso por el cual se transforman datos personales, de forna que no sea posible identificar a un interesado sin utilizar información adicional, siempre que dicha información adicional 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.

Sistema de información: Es un conjunto de ficheros, tratamientos, programas, soportes y en su caso, equipos empleados para el tratamiento de datos de carácter personal.

Tercero: Según el [RD 1720/2007], es la persona física o jurídica, pública o privada u órgano administrativo distinta del afectado o interesado, del responsable del tratamiento, del responsable del fichero, del encargado del tratamiento y de las personas autorizadas para tratar los datos bajo la autoridad directa del responsable del tratamiento o del encargado del tratamiento. Podrán ser también terceros los entes sin personalidad jurídica que actúen en el tráfico como sujetos diferenciados.

Transferencia internacional de datos: Es, según el [RD 1720/2007], un tratamiento de datos que supone una transmisión de los mismos fuera del territorio del Espacio Económico Europeo, bien constituya una cesión o comunicación de datos, bien tenga por objeto la realización de un tratamiento de datos por cuenta del responsable del fichero establecido en territorio español.

Tratamiento de datos personales: Un tratamiento de datos personales es según, el [RGPD 2017], cualquier operación o conjunto de operaciones realizadas sobre datos personales o conjuntos de datos personales, ya sea por procedimientos automatizados o no, como la recogida, registro, organización, estructuración, conservación, adaptación o modificación, extracción, consulta, utilización, comunicación por transmisión, difusión o cualquier otra forma de habilitación de acceso, cotejo o interconexión, limitación, supresión o destrucción.

Usuario: Es una persona o un proceso que está autorizado a acceder a datos o recursos.

Utilidad de los datos: La utilidad es la propiedad de los datos que indica lo apropiados y eficaces que son para el tratamiento que se desee hacer sobre ellos en un uso particular.

Listado de Acrónimos

AEPD: Agencia Española de Protección de Datos

ATC: Anatomical, Therapeutic, Chemical classification system

CIE: Clasificación Internacional de Enfermedades

DPD: Delegado de Protección de Datos

EIPD: Evaluación de Impacto en Protección de datos

LOPDGDD: Ley Orgánica de Protección de Datos y Garantía de los Derechos Digitales

MD5: Message-Digest Algorithm

ODS: Objetivos de Desarrollo Sostenible

PGP: Pretty Good Privacy

RD: Real Decreto

RGPD: Reglamento General de Protección de Datos

SHA: Secure Hash Algorithm

SCTID: SNOMED CT Identifier

SNOMED CT: Systematized Nomenclature of Medicine – Clinical Terms

WBS: Work Breakdown Structure

Índice de figuras

Fig. 1: División en tareas (WBS) del trabajo....16
Fig. 2: Diagrama de Gantt. Primera semana....17
Fig. 3: Diagrama de Gantt. Semanas segunda y tercera....17
Fig. 4: Diagrama de Gantt. Semanas cuarta y quinta....18
Fig. 5: Diagrama de Gantt. Semanas sexta y séptima....18
Fig. 6: Diagrama de Gantt. Octava semana....19
Fig. 7: SNOMED CT. Detalles del cólera....24
Fig. 8: SNOMED CT. Diagrama del cólera....24
Fig. 9: Generación del par de claves....28
Fig. 10: Contenido del almacén de claves....29
Fig. 11: Cifrado de los datos....29

Índice de tablas

Tabla 1: CIE-11. Primer nivel....22
Tabla 2: CIE 11. Codificación del cólera....22
Tabla 3: SNOMED CT. Primer nivel....23
Tabla 4: Contador. Datos originales....25
Tabla 5: Seudonimización por contador....25
Tabla 6: Seudonimización por números aleatorios....26
Tabla 7: Seudonimización por función resumen....27
Tabla 8: Seudonimización por cifrado simétrico....27
Tabla 9: Datos cifrados con cifrado asimétrico....30
Tabla 10: Supresión de atributos. Datos originales....31
Tabla 11: Supresión del atributo Apellidos....32
Tabla 12: Supresión de registros....32
Tabla 13: Enmascaramiento de cadenas....33
Tabla 14: Generalización. Datos originales....34
Tabla 15: Obtención de información asociada....34
Tabla 16: Generalización....35
Tabla 17: Generalización alternativa....35
Tabla 18: K-anonimidad. Datos originales....36
Tabla 19: Clases de equivalencia....36
Tabla 20: Combinaciones según las variables de entrada....41
Tabla 21: Ejemplo de tabla....49
Tabla 22: Ejemplo de foco....49
Tabla 23: Selección de la técnica....50
Tabla 24: Seudonimización por contador....51
Tabla 25: Seudonimización por números aleatorios....52
Tabla 26: Seudonimización por funciones resumen....53

Tabla 27: Seudonimización por cifrado simétrico.....53
Tabla 28: Anonimización por supresión de columnas.....54
Tabla 29: Anonimización por supresión de filas.....54
Tabla 30: Anonimización por enmascaramiento de cadenas.....55
Tabla 31: Generalización. Datos originales.....55
Tabla 32: Anonimización por generalización con SNOMED CT.....56
Tabla 33: Anonimización por generalización con CIE-11.....57
Tabla 34: K-anonimidad. Datos originales.....57
Tabla 35: Aplicación de la K-anonimidad.....57
Tabla 36: Relación con los Objetivos de Desarrollo Sostenible (ODS).....59

1. Introducción

1.1. Presentación

Me llevo dedicando a la informática sanitaria desde el principio de mi carrera profesional, ya en un lejano 1997, y en este tiempo he llevado a cabo proyectos de todo tipo para diferentes hospitales y clínicas, y también para distintas especialidades médicas.

Cuando se publicó la primera ley de protección de datos en España, en el año 1999, consciente de la relevancia que tenía en el sector sanitario, empecé a desarrollar tareas de implantación y auditoría de la por entonces novedosa ley, pero sobretodo con el impulso y relevancia que la protección de datos personales ha adquirido en los últimos años, esta disciplina ha llegado a ocupar buena parte de mi vida profesional, siendo uno de los primeros delegados de protección de datos certificados por la Agencia Española de Protección de Datos, y ejerciendo como tal actualmente para un hospital y varias clínicas en la ciudad de Valencia.

1.2. Motivación

Para llevar a cabo eficazmente las tareas que demanda un panorama legislativo y técnico que va evolucionado día a día, también es necesario conocer de primera mano la parte técnica que subyace, y en este contexto encaja perfectamente el programa de estudios del Máster Universitario en Ciberseguridad y Ciberinteligencia de la Universidad Politécnica de Valencia. En él, se introduce en las tareas de proteger los datos identificativos de las personas utilizando técnicas de anonimización y seudonimización, y si se profundiza es posible dar respuestas segura y fiable a muchas de las cuestiones que plantean los médicos, enfermeras y estudiantes sobre cómo tratar de forma adecuada sus datos para sus estudios, tesis, trabajos de final de grado, etc.

Este trabajo responde a esa necesidad de aplicar los conocimientos adquiridos en un área que, si bien puede no ser tan emocionante como otras estudiadas en el máster, es muy relevante para el día a día de todas las personas, ya que antes o después todos vamos a tener alguna dolencia, vamos acudir a un centro sanitario, y seguramente todos deseamos que se trate nuestra información de la forma más segura posible.

1.3. Objetivos del trabajo

El objetivo de este trabajo es la creación de un marco teórico en el que estarán contempladas las técnicas de anonimización y seudonimización más conocidas, de forma que partiendo de unas condiciones iniciales ofrecerá indicaciones sobre qué técnica sería la más apropiada y cómo llevar a cabo la anonimización o seudonimización, en su caso. Estas condiciones iniciales son:

- El conjunto de datos a tratar.
- El objetivo: obtener datos estadísticos, buscar registros anómalos...

- La información que tenga especial interés dentro del conjunto de datos.
- Las potenciales transferencias de datos que se pretendan llevar a cabo.
- Si se pretende reutilizar los datos, una vez extraídos y procesados.

Esto se plasmará, como objetivo secundario, en una guía de cómo debe proceder el facultativo para llevar a cabo el proceso de una forma procedimentada y eficaz.

Como líneas futuras de trabajo se propone la validación de la guía aquí desarrollada en un uso real por profesionales de la salud, la creación de una herramienta software que implemente la guía aquí desarrollada, y finalmente la creación de una versión más avanzada de esta misma guía.

1.4. Estructura del trabajo

El presente trabajo se estructura en 5 bloques:

- En la primera parte, *Descripción del problema*, se detallará el problema que se desea abordar.
- En la segunda parte, *Plan de trabajo*, se detalla, en forma de proyecto, cómo se ha llevado a cabo la planificación de este trabajo.
- En la tercera parte, *Contexto tecnológico*, se hace un repaso al estado de arte de la materia, describiendo las técnicas que se emplean habitualmente.
- En la cuarta, *Solución propuesta*, se detalla la propuesta aportada.
- En la quinta parte, *Conclusiones. Líneas futuras*, se analizan los resultados obtenidos y se describe cómo podría ser la ulterior evolución de este trabajo de final de máster.

2. Descripción del problema

Cuando una persona quiere hacer uso de una fuente de datos para una investigación científica, para un uso publicitario, o simplemente para un uso personal, se encuentra con que debe solicitar esos datos al titular de los mismos, y luego se llega a la disyuntiva de qué debe hacer para tratar esos datos de forma segura y además cumplir con la legislación vigente en materia de protección de datos personales.

La cuestión subsiguiente a la que se enfrenta es cómo hacerlo, teniendo en cuenta además que si se transforma demasiado la información para preservar la privacidad puede llegarse a un punto en el que la información deje de ser útil para el fin perseguido.

Nos hemos referido solo a una persona, pero muchas veces la situación es más compleja, ya que intervienen más participantes. En un centro sanitario, por ejemplo, la petición por parte de un equipo de investigación típicamente le llegará al área de informática o al director médico, que probablemente consulten con el delegado de protección de datos esperando sus indicaciones. En otro tipo de organizaciones pueden intervenir, según los datos requeridos, el área de recursos

humanos, el de prevención de riesgos laborales, etc., con la misma pregunta básica: ¿Cual es el método que debo seguir?

En el presente trabajo se intentará dar respuesta a esa cuestión, recurriendo a los métodos de anonimización y seudonimización más conocidos, y proponiendo un método que puedan seguir todos aquellos que deseen llevar a cabo un tratamiento de datos personales ocultando o eliminando parte o toda la información que permita identificar a los individuos cuyos datos son tratados.

Se hace especial énfasis en que los resultados de este trabajo deben ser fácilmente interpretables por alguien que no sea experto en ciberseguridad ni en protección de datos personales, de forma que pueda resultarle útil para su trabajo diario.

3. Plan de trabajo

El desarrollo del presente trabajo se ha estructurado en las siguientes fases:

1. Estudio de las principales técnicas de anonimización y seudonimización.
2. Estudio de las guías y documentos sobre anonimización proporcionados por la Agencia Española de Protección de Datos, en adelante la AEPD.
3. Diseño de la solución.
4. Implementación de la solución.
5. Redacción de la presente memoria.

Se ha establecido la fecha de inicio del del trabajo en el 3 de Julio de 2023, y la división en trabajos (WBS, Work Breakdown Structure) ha sido la siguiente:

WBS	Nombre	Inicio	Fin	Trabajo	Duración	Desperdicio	Coste	Asignado a	% Completado
1	Inicio del proyecto	jul 3	jul 3	N/D	N/D		0	Cesar Fort	100
2	Estudio previo y planificación	jul 3	jul 7	5d	5d		0	Cesar Fort	100
3	**Estudio de técnicas de anonimización**	**jul 10**	**jul 19**	**8d**	**8d**		**0**	Cesar Fort	100
3.1	Contador	jul 10	jul 10	1d	1d		0		100
3.2	Nümeros Aleatorios	jul 11	jul 11	1d	1d		0		100
3.3	Función resumen	jul 12	jul 13	2d	2d		0		100
3.4	Cifrado simétrico	jul 14	jul 17	2d	2d		0		100
3.5	Cifrado asimetrico	jul 18	jul 19	2d	2d		0		100
4	**Estudio de técnicas de seudonimización**	**jul 20**	**jul 28**	**7d**	**7d**		**0**	Cesar Fort	100
4.1	Supresión de atributos	jul 20	jul 20	1d	1d		0		100
4.2	Supresión de registros	jul 21	jul 21	1d	1d		0		100
4.3	Enmascaramiento de cadenas	jul 24	jul 24	1d	1d		0		100
4.4	Generalización	jul 25	jul 26	2d	2d		0		100
4.5	K-anonimidad	jul 27	jul 28	2d	2d		0		100
5	Diseño de la solución	jul 31	ago 2	3d	3d		0	Cesar Fort	100
6	Implementación de la solución	ago 3	ago 8	4d	4d		0	Cesar Fort	100
7	Redacción de la memoria	ago 9	sep 1	18d	18d		0	Cesar Fort	100

Fig. 1: División en tareas (WBS) del trabajo

Asimismo, el cronograma del trabajo, expresado mediante un diagrama de Gantt es tal y como se muestra en las sucesivas tres páginas:

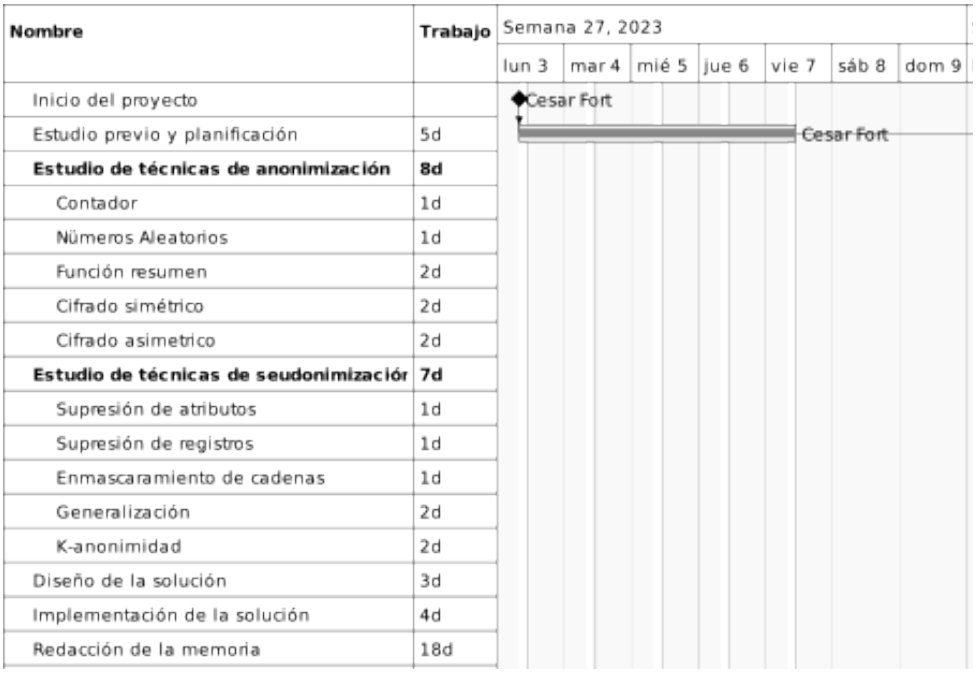

Fig. 2: Diagrama de Gantt. Primera semana

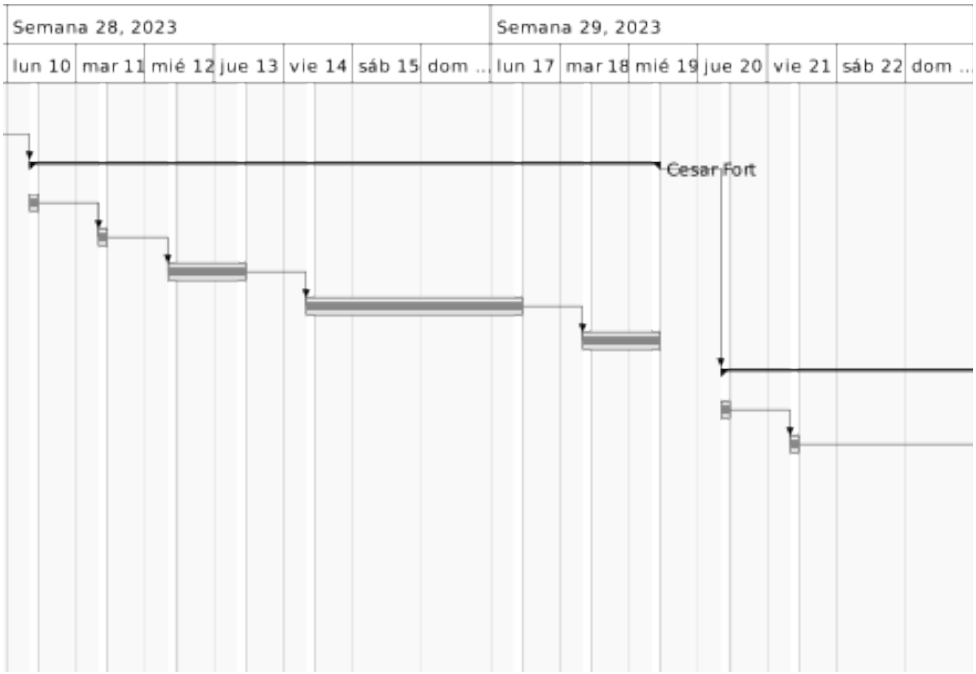

Fig. 3: Diagrama de Gantt. Semanas segunda y tercera

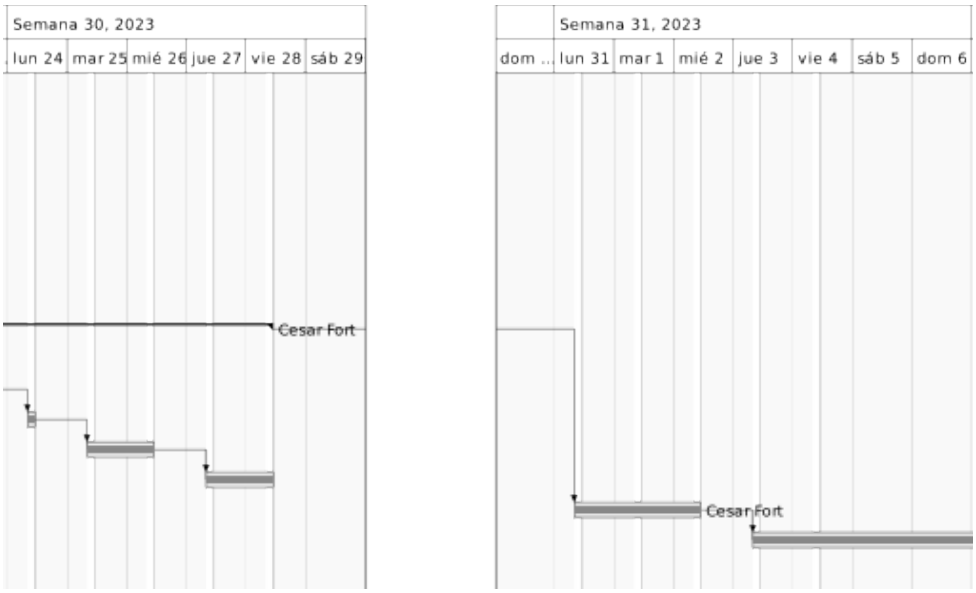

Fig. 4: Diagrama de Gantt. Semanas cuarta y quinta

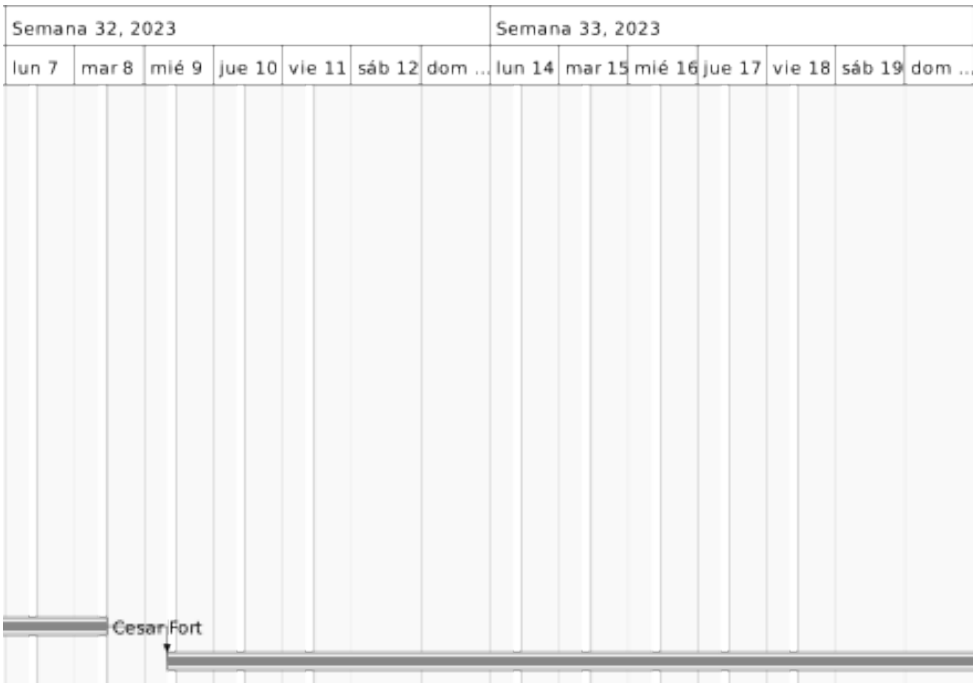

Fig. 5: Diagrama de Gantt. Semanas sexta y séptima

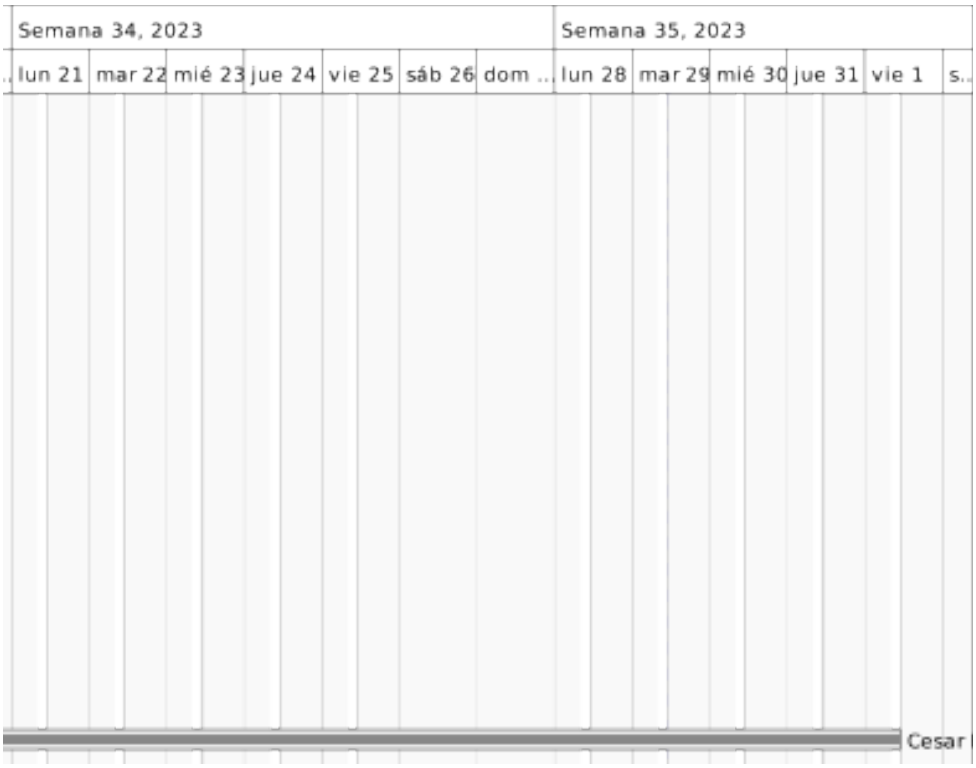

Fig. 6: Diagrama de Gantt. Octava semana

4. Contexto tecnológico

Existen numerosas técnicas para manipular información separando los datos identificativos que contiene, pero para abordarlas primero es necesario introducir una serie de conceptos fundamentales.

4.1. Conceptos básicos

4.1.1. Tipos de atributos

En este trabajo se van a tratar conceptos que se referirán a distintos tipos de atributos, y es importante conocer las diferencias entre ellos:

Según [Casas Roma y Romero Tris 2017], desde el punto de vista de la privacidad, los atributos de un conjunto de datos se dividen en cuatro clases según el tipo de información que contienen:

- Los **identificadores**, que son atributos aislados o grupos de ellos que permiten identificar de forma explícita y unívoca a un individuo.
- Los **cuasi-identificadores** son atributos o conjuntos de ellos que, si bien por sí mismos no contienen información suficiente para identificar explícitamente a un individuo, junto con otros sí que podrían contribuir a hacerlo.
- Los atributos **sensibles** son los que contienen información personal y privada de un individuo.
- Los atributos **no sensibles** serían los que no se pueden englobar en ninguna de las tres categorías mencionadas previamente.

4.1.2. Datos personales y datos disociados

Los **datos personales** son datos en los que es posible identificar a una persona. Si en estos datos se extrae la información identificativa, se convierten en **datos disociados**, que aun pudiendo contener información sensible, como datos médicos o sobre creencias religiosas, no es posible atribuirlos a ninguna persona en concreto.

4.1.3. Anonimización frente a seudonimización

La anonimización es el proceso por en cual se extraen los datos identificativos de una persona de un conjunto de datos, e idealmente es irreversible.

Por el contrario, la seudonimización es un proceso por el cual se sustituyen los datos identificativos por otros, de forma que con estos nuevos datos procesados no es posible identificar a una persona en concreto, pero con más información (que debe figurar por separado), sí debería ser posible reidentificar a un individuo.

4.2. Terminologías médicas

Para poder almacenar y procesar datos médicos de una forma consistente y estandarizada se utilizan las terminologías médicas.

Existen distintas terminologías aceptadas internacionalmente, con distintos enfoques, como por ejemplo:

- El código ATC, acrónimo de Anatomical, Therapeutic, Chemical classification system que es un índice de clasificación de sustancias farmacológicas y medicamentos.
- La CIE-11, acrónimo de Clasificación Internacional de Enfermedades, que es una compilación ordenada de enfermedades.
- SNOMED CT, acrónimo de Systematized Nomenclature of Medicine – Clinical Terms, que es toda una terminología médica que permite no solo la codificación estandarizada de conceptos, sino también la generación de expresiones complejas para expresar propiedades y relaciones entre ellos.

Como el presente trabajo está enfocado a médicos, enfermeras a investigadores de la salud, profundizaremos un poco en la CIE-11 u SNOMED CT, que son los más utilizados.

4.2.1. CIE-11

La CIE es una clasificación que, según [OMS 2022], proporciona un lenguaje común que permite a los profesionales de la salud compartir información estandarizada en todo el mundo. La undécima revisión contiene unos 17000 códigos únicos y más de 120000 términos codificables, y ahora es totalmente digital.

Se trata de una taxonomía por sistemas (nervioso, digestivo, etc.), y se puede acceder online a la versión en castellano mediante el enlace: [OMS 2023].

Tiene una estructura en árbol, y el primer nivel es el siguiente:

CIE-11 para estadísticas de mortalidad y morbilidad (Versión : 01/2023)
01. Algunas enfermedades infecciosas o parasitarias
02 Neoplasias
03 Enfermedades de la sangre o de los órganos hematopoyéticos
04 Enfermedades del sistema inmunitario
05 Enfermedades endocrinas, nutricionales o metabólicas
06 Trastornos mentales, del comportamiento y del neurodesarrollo
07 Trastornos del sueño y la vigilia
08 Enfermedades del sistema nervioso
09 Enfermedades del sistema visual
10 Enfermedades del oído o de la apófisis mastoides
11 Enfermedades del sistema circulatorio

12 Enfermedades del sistema respiratorio
13 Enfermedades del sistema digestivo
14 Enfermedades de la piel
15 Enfermedades del sistema musculo esquelético o del tejido conjuntivo
16 Enfermedades del sistema genitourinario
17 Condiciones relacionadas con la salud sexual
18 Embarazo, parto o puerperio
19 Algunas afecciones que se originan en el período perinatal
20 Anomalías del desarrollo
21 Síntomas, signos o hallazgos clínicos no clasificados en otra parte
22 Traumatismos, intoxicaciones u otras consecuencias de causas externas
23 Causas externas de morbilidad o mortalidad
24 Factores que influyen en el estado de salud o el contacto con los servicios de salud
25 Códigos para propósitos especiales
26 Capítulo complementario de condiciones de la medicina tradicional - Módulo I
V Sección suplementaria para la evaluación del funcionamiento
X Códigos de extensión

Tabla 1: CIE-11. Primer nivel

Por ejemplo, la enfermedad del cólera está identificada por el código 1A00, y el recorrido del árbol de enfermedades que lleva hasta ella es:

CIE-11 para estadísticas de mortalidad y morbilidad (Versión : 01/2023)
01 Algunas enfermedades infecciosas o parasitarias
Gastroenteritis o colitis de origen infeccioso
Infecciones intestinales bacterianas
1A00 Cólera

Tabla 2: CIE 11. Codificación del cólera

4.2.2. SNOMED CT

Según [Min. Sanidad 2023], *«SNOMED CT es la terminología clínica integral, multilingüe y codificada de mayor amplitud, precisión e importancia desarrollada en el mundo. SNOMED CT es, también, un producto terminológico que puede usarse para codificar, recuperar, comu-*

nicar y analizar datos clínicos permitiendo a los profesionales de la salud representar la información de forma adecuada, precisa e inequívoca.»

Se trata de un estándar internacional distribuido por SNOMED International, organización a la que pertenece España como país miembro.

SNOMED CT expresa conocimiento clínico por medio de conceptos, descripciones y relaciones. Tiene también una estructura jerárquica similar a la de CIE-11, pero con mucho más contenido, no estando solo limitado a las enfermedades, sino a otro tipo de conceptos como procedimientos quirúrgicos, antecedentes, etc..

En este caso, el primer nivel de la jerarquía es el siguiente:

Concepto de SNOMED CT	
	Ambiente o localización geográfica
	Calificador
	Componente del modelo de SNOMED CT
	Concepto especial
	Contexto social
	Elemento de registro
	Entidad observable
	Espécimen
	Estadificaciones y escalas
	Estructura corporal
	Evento
	Fuerza física
	Hallazgo clínico
	Objeto físico
	Organismo
	Procedimiento
	Producto biológico/farmacéutico
	Situación con contexto explicado
	Sustancia

Tabla 3: SNOMED CT. Primer nivel

En el caso del cólera, mediante el buscador de SNOMED CT disponible online en [SNOMED 2023-1], vemos que su código, que SNOMED CT denomina SCTID, acrónimo de SNOMED CT Identifier, es el 63650001, y tiene como ancestro dentro de la jerarquía a *Infección de intestino causada por Vibrio (trastorno)*, tal y como se muestra en la siguiente figura:

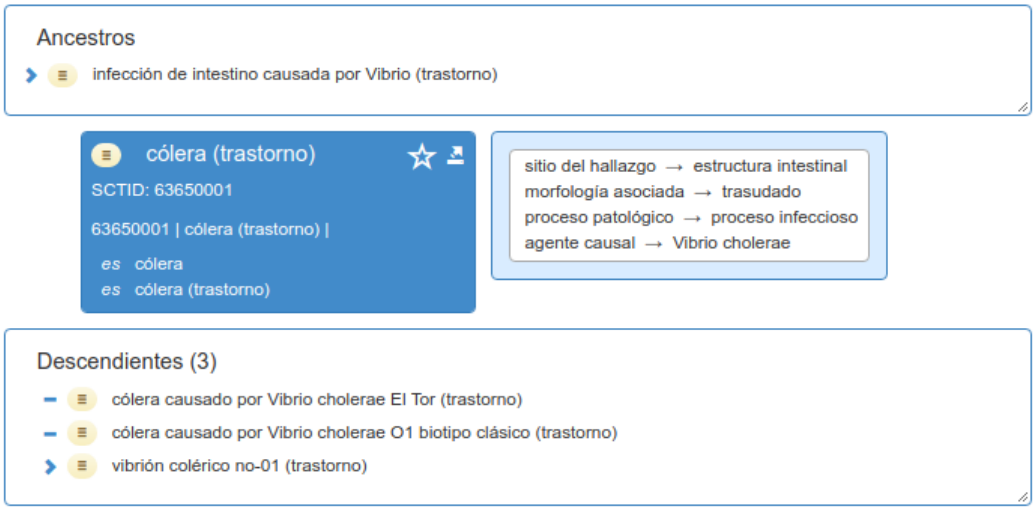

Fig. 7: SNOMED CT. Detalles del cólera

En la figura anterior podemos ver también que tiene tres descendientes, y que uno de ellos tiene a su vez más descendientes, con lo que ya podemos comprobar que esta clasificación tieen un grano más fino que la CIE-11.

SNOMED CT también permite expresar lo conceptos en forma de diagramas, donde se expresan las relaciones entre conceptos, como por ejemplo siguiendo con el caso del cólera:

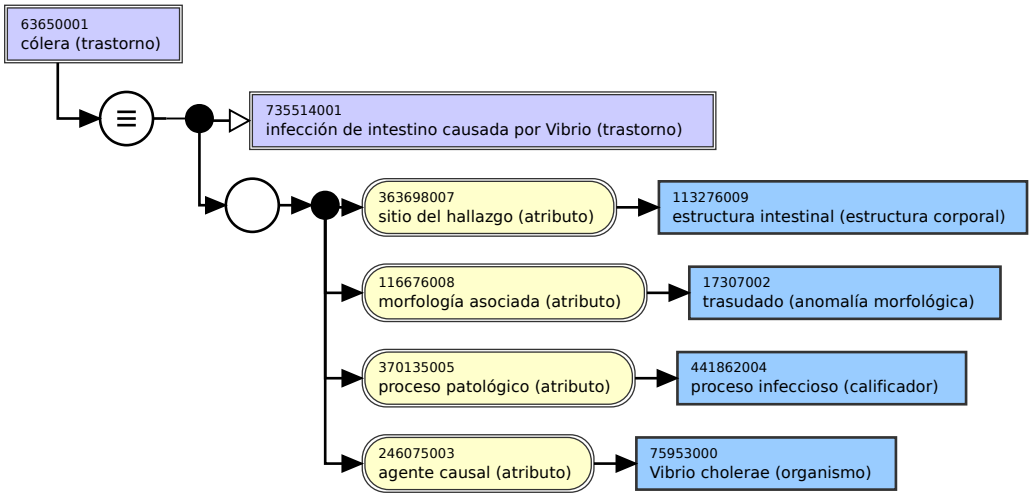

Fig. 8: SNOMED CT. Diagrama del cólera

Como se ha mencionado anteriormente, para representar entidades más complejas es posible utilizar expresiones, como por ejemplo, la siguiente expresión expresa un antecedente familiar de infarto cerebral:

```
281666001|antecedente familiar de trastorno|:246090004|hallazgo asociado|
=432504007 |infarto cerebral
```

Hay que añadir que hay disponible en [SNOMED 2023-2] una guía para llevar a cabo la correspondencia de SNOMED CT a la versión anterior de la CIE, la CIE 10, pero al revés no es posible establecer una correspondencia, ya que el ámbito que abarca SNOMED CT es mayor que el de la CIE en cualquiera de sus versiones.

4.3. Técnicas de seudonimización

Según [ENISA 2022], las técnicas más utilizadas para generar seudónimos a partir de datos fuente son las siguientes:

4.3.1. Contador

Se trata de una función monótona que comienza con un determinado valor que se va incrementando cuando se necesita un nuevo seudónimo.

Por ejemplo, sea esta tabla-ejemplo ya vista anteriormente en el apartado 4.4.4. Generalización:

ID	Nombre	C. P.	Uso (kW)
1	Oscar	46118	230
2	Alberto	46181	118
3	Vicente	28001	189
4	Enrique	28023	217
5	Jose Maria	28400	212

Tabla 4: Contador. Datos originales

Si quisiéramos seudonimizar el atributo *Nombre*, podemos sustituirlo por números a partir del 24 aumentando de tres en tres en orden creciente:

ID	Nombre (seudonimizado)	C. P.	Uso (kW)
1	24	46118	230
2	27	46181	118
3	30	28001	189
4	33	28023	217
5	36	28400	212

Tabla 5: Seudonimización por contador

En resumen, las características de esta técnica son:

- Muy simple y fácil de utilizar.
- Adecuada para trabajar con conjuntos de datos en los que sea importante no perder información de ningún registro.
- Adecuada para reutilizar los datos.
- Poco segura.
- Poco adecuada para llevar a cabo transferencias por su simplicidad.

4.3.2. Números aleatorios

Con esta técnica, cada vez que se necesita generar un nuevo seudónimo se extrae un valor aleatorio entre un límite mínimo y un límite máximo.

Por ejemplo, siguiendo el ejemplo del punto anterior, y tomando números aleatorios entre el 10 y el 500. el atributo *Nombre* quedaría anonimizado de la siguiente forma:

ID	Nombre (seudonimizado)	C. P.	Uso (kW)
1	463	46118	230
2	186	46181	118
3	252	28001	189
4	352	28023	217
5	112	28400	212

Tabla 6: Seudonimización por números aleatorios

En resumen, las características de esta técnica son:

- Relativamente simple y fácil de utilizar.
- Adecuada para trabajar con conjuntos de datos en los que sea importante no perder información de ningún registro.
- Adecuada para reutilizar los datos.
- Poco segura.
- Poco adecuada para llevar a cabo transferencias por su simplicidad.

4.3.3. Función resumen

Mediante esta técnica, cada vez que se necesita generar un nuevo seudónimo se utiliza una función matemática no reversible que transforma los datos personales de entrada en otros valores, numéricos o alfanuméricos, de longitud fija.

Algunas funciones resumen comúnmente utilizadas en criptografía son la función MD5 (Message-Digest Algorithm 5), la SHA-1 (Secure Hash Algoruthm-1), y la familia de algoritmos SHA-2: SHA-224, SHA-256, SHA-384 y SHA-512.

Siguiendo con el ejemplo anterior, si seudonimizamos el atributo *Nombre* utilizando la función SHA-1:

ID	Nombre (seudonimizado)	C. P.	Uso (kW)
1	07f24c146c9cd13d69fdc5ef719e97aec36f24fe	46118	230
2	3b3e55fdc7886baea165a854d080caf9808cac97	46181	118
3	7277f956e52a382f15ed35f8e1af3fc78663e9af	28001	189

4	9cdcdaec09c4ecdeb7df30db270618e77ec4eee5	28023	217
5	f377b6375dd4b3940bc67de1cc74c9fb616114df6	28400	212

Tabla 7: Seudonimización por función resumen

En resumen, las características de esta técnica son:

- Bastante simple, pero de utilización algo laboriosa.
- Adecuada para trabajar con conjuntos de datos en los que sea importante no perder información de ningún registro.
- Adecuada para reutilizar los datos.
- Seguridad intermedia.
- Adecuada para llevar a cabo transferencias por su seguridad.

4.3.4. Cifrado simétrico

En este caso se utiliza una función criptográfica bidireccional (reversible) que transforma los datos de entrada en valores que pueden volver a transformarse en su formato original utilizando la misma clave.

Los algoritmos usados en la criptografía simétrica son principalmente operaciones booleanas y de transposición, y es más eficiente que la criptografía asimétrica.

Por ejemplo, en el Cifrado César se utiliza como clave un número menor que el número de letras del alfabeto, y se sustituye cada una de las letras por la letra que resulta de desplazarse a la derecha el número de veces indicado en la clave. Si tomamos las 27 letras del alfabeto español y una clave = 5, siguiendo con el mismo ejemplo resultaría:

ID	Nombre	C. P.	Uso (kW)
1	Txhfw	46118	230
2	Fpgjwyt	46181	118
3	Anhjryj	28001	189
4	Jrwnvzj	28023	217
5	Ñtxj Qfwnf	28400	212

Tabla 8: Seudonimización por cifrado simétrico

En resumen, las características de esta técnica son:

- Algo compleja y laboriosa.
- Adecuada para trabajar con conjuntos de datos en los que sea importante no perder información de ningún registro.

- Adecuada para reutilizar los datos.
- Bastante segura.
- Adecuada para llevar a cabo transferencias por su seguridad.

4.3.5. Cifrado asimétrico

Con el cifrado simétrico, no existe una sola clave sino dos: una clave privada que se utiliza para cifrar, y una clave pública, que se utiliza para descifrar. Ambas claves son vinculadas mediante un algoritmo matemático, de forma que los datos cifrados con la clave pública solo pueden descifrarse con la clave privada. Existen múltiples implementaciones del esquema de clave pública y privada, como PGP, acrónimo de Pretty Good Privacy.

Siguiendo con al mismo ejemplo, para anonimizar el atributo *Nombre* lo primero será crear un par de claves. Se puede hacer mediante las distintas implementaciones de PGP.

Mediante el comando `gpg -gen-key` e introduciendo los datos requeridos, que son el nombre, correo electrónico, algoritmo, longitud de clave, expiración y contraseña, obtenemos las correspondientes claves privada y pública:

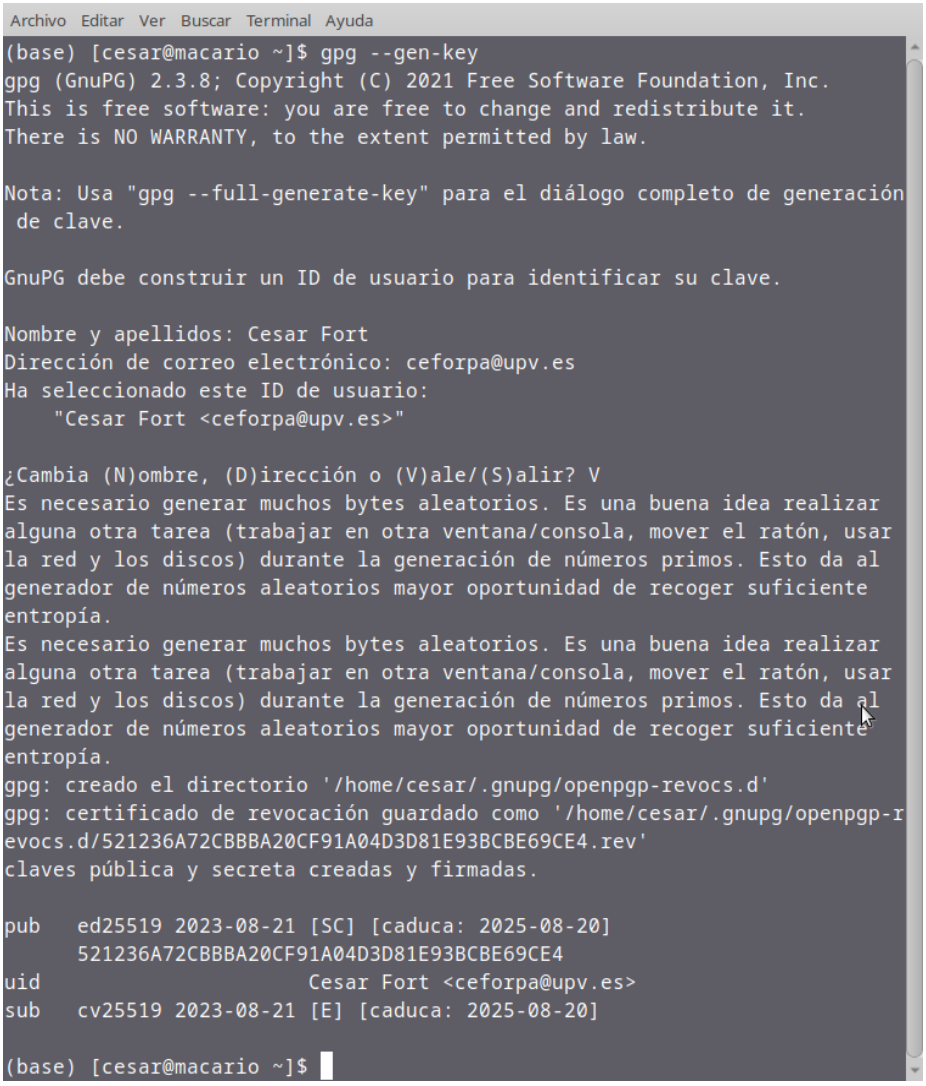

Fig. 9: Generación del par de claves

Con el comando `gpg -k` es posible ver el contenido del almacén de claves:

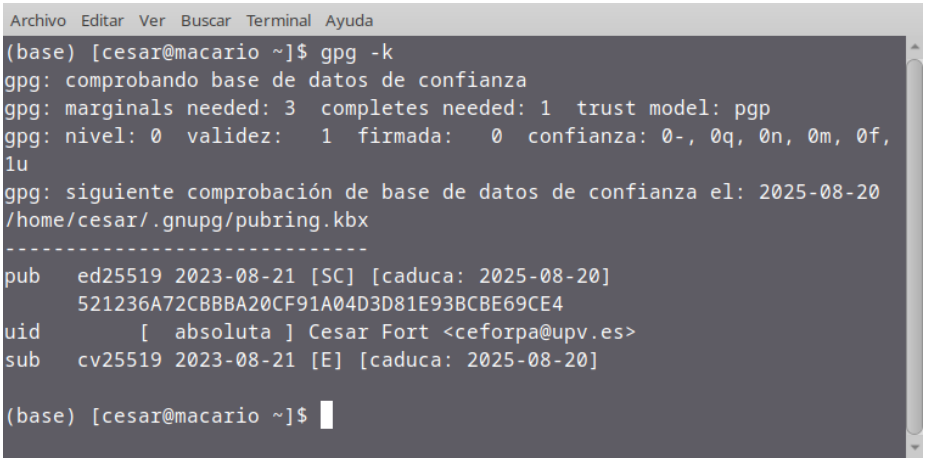

Fig. 10: Contenido del almacén de claves

Ahora, siguiendo con el mismo ejemplo, para cifrar cada uno de los elementos del atributo *Nombre*, utilizamos:

`echo "[texto]" | gpg --encrypt --armor -r [clave publica]`, de la siguiente forma:

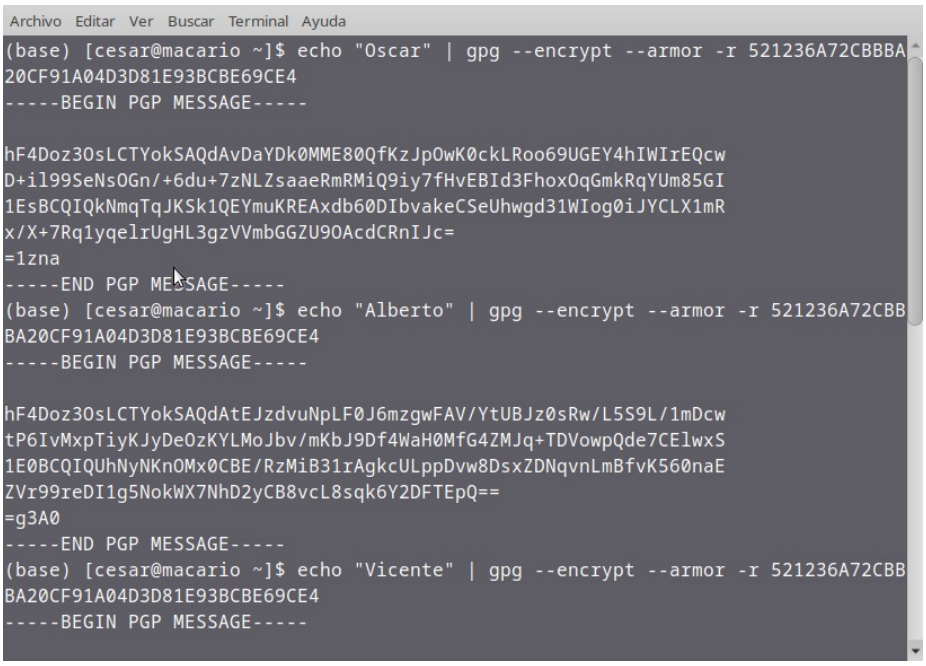

Fig. 11: Cifrado de los datos

Como resultado, se obtiene el cifrado de los valores del atributo *Nombre*:

ID	Nombre	C. P.	Uso(kW)
1	hF4Doz3OsLCTYokSAQdAvDaYDk0MME80QfKzJ-pOwK0ckLRoo69UGEY4hIWIrEQcw D+il99SeNsOGn/+6du+7zNLZsaaeRmRMiQ9iy7fHvE-	46118	230

	BId3FhoxOqGmkRqYUm85GI 1EsBCQIQkNmqTqJKSk1QEYmuKREAxdb60DIb- vakeCSeUhwgd31WIog0iJYCLX1mR x/+7Rq1yqelrUgHL3gzVVmbGGZU9OAcdCRnIJc= =1zna		
2	hF4Doz3OsLCTYokSAQdAtEJzdvuNpLF0J6mzgw- FAV/YtUBJz0sRw/L5S9L/1mDcw tP6IvMxpTiyKJyDeOzKYLMoJbv/mKbJ9Df4WaH0M- fG4ZMJq+TDVowpQde7CElwxS 1E0BCQIQUhNyNKnOMx0CBE/RzMiB31rAgkcUL- ppDvw8DsxZDNqvnLmBfvK560naE ZVr99reDI1g5NokWX7NhD2yCB8vcL8sqk6Y2DFTE- pQ===g3A0	46181	118
3	hF4Doz3OsLCTYokSAQdAwX7NSH+CP4ZoGRRYr- Y5wh5mfprq7SqGuae9wZ69BWGgw ep1vAovhOgR25VLSa93Qtedv2Ooe16xawMAbRKbY- gFl3SlYyepSKB2SIPg3kkYsE 1E0BCQIQwxc80QAXyxu6Uy6E7xzUMAqg0BZdma- xJ0biG+W3tmZ73pKpHHoCOWY6T YYxNLLJF32CXJrqqGvzeXKGlH+KgjoNklWLOa- YN5jA== =vfPd	28001	189
4	hF4Doz3OsLCTYokSAQdAHUNcM0eOQ+kWTSm- L0SBxuRDKNMx8/rJnFeOd1HeMV2Iw ObdxuipxsRGiro8LiCi2conv2Q3WEziBfhJhkw4I4yeI- L6dAJv/6aMgbfpjUI6jg 1E0BCQIQsKCfwUfWOASb7EdxlXko3f6yFVGx+tW- Q3Add6wp0m4z2C3g6neG5lOs0 CuCl1+oP2SdIckauptCwMmve3ACOI6vqgedhEQpySg ===Mlqd	28023	217
5	hF4Doz3OsLCTYokSAQdA5zoxcz8eRgpO+47clP2oe- yWxGgKhF5r5KRxyhDjIAjkw 0LehHat2u+Bnj9N6f4Sn0+k6M1Y7mp3qb9kZ/cRcQs- siar09qqp4WvhePVlS57Lz 1FABCQIQleRG10Q7akBm1AsSw6CglRcdcezmO2o- L36SzuWM3NXcNTxesDRSiY4Yf CDt6oa6M55MhcKOek0M08w9bIaLTRabEh0OaLTW/ RtJ7nQ===Qlff	28400	212

Tabla 9: Datos cifrados con cifrado asimétrico

En resumen, las características de esta técnica son:

- Compleja y laboriosa de utilizar.
- Adecuada para trabajar con conjuntos de datos en los que sea importante no perder información de ningún registro.
- Adecuada para reutilizar los datos.
- Bastante segura.
- Adecuada para llevar a cabo transferencias por su seguridad.

4.4. Técnicas de anonimización

Las técnicas más utilizadas para llevar a cabo la anonimización de datos son las siguientes:

4.4.1. Supresión de atributos

En [Murty 2019] se describe la supresión de atributos como la eliminación de parte de los atributos de un conjunto de datos cambiando sus valores por otros sin significado, como por ejemplo ***. Esta técnica puede llegar a producir que los datos sean inutilizables si no se eligen adecuadamente los atributos a eliminar.

Por ejemplo, sea el siguiente conjunto de datos:

ID	Nombre	Apellidos	Sueldo anual
1	Oscar	Fort	45298€
2	Alberto	Ruiz	65000€
3	Vicente	Sanz	46250€
4	Julia	Palau	82500€
5	Enrique	Perez	37860€
6	Jose Maria	Gomez	28990€
7	Angela	Del Villar	67930€
8	Joaquín	Izquierdo	37980€

Tabla 10: Supresión de atributos. Datos originales

Podría eliminarse el atributo *Apellidos* para que fuera más difícil la identificación de los interesados, de forma que quedaría:

ID	Nombre	Sueldo anual
1	Oscar	45298€
2	Alberto	125200€

3	Vicente	46250€
4	Julia	82500€
5	Enrique	37860€
6	Jose Maria	28990€
7	Angela	67930€
8	Joaquín	57980€

Tabla 11: Supresión del atributo Apellidos

En resumen, las características de esta técnica son:

- Muy simple y fácil de utilizar.
- Adecuada para trabajar con conjuntos de datos en los que se pueda prescindir de algún atributo.
- No adecuada para reutilizar los datos.
- Segura si se suprimen los atributos adecuados.
- Bastante adecuada para llevar a cabo transferencias por su seguridad.

4.4.2. Supresión de registros

En [Murty 2019], se describe de forma análoga la supresión de registros como la eliminación de parte de registros de un conjunto de datos, eliminándolos por completo, o cambiando sus valores por otro sin significado, como por ejemplo ***.

Siguiendo el ejemplo anterior partiendo de la Tabla 10: Supresión de atributos. Datos originales, podrían eliminarse los registros de sueldos anormalmente altos o bajos para reducir el riesgo de reidentificación:

ID	Nombre	Apellidos	Sueldo anual
1	Oscar	Fort	45298€
***	***	***	***
3	Vicente	Sanz	46250€
4	Julia	Palau	82500€
***	***	***	***
6	Jose Maria	Gomez	28990€
7	Angela	Del Villar	67930€
8	Joaquín	Izquierdo	57980€

Tabla 12: Supresión de registros

Esta técnica, igual que en el caso de la supresión de registros, puede producir el efecto indesea-do de que los datos sean inutilizables si no se eligen adecuadamente los registros a eliminar.

En resumen, las características de esta técnica son:

- Muy simple y fácil de utilizar.
- Adecuada para trabajar con conjuntos de datos en los que se pueda prescindir de ciertos registros.
- No adecuada para reutilizar todos los datos.
- Poco segura.
- Poco adecuada para llevar a cabo transferencias por su poca seguridad.

4.4.3. Enmascaramiento de cadenas

El enmascaramiento de cadenas consiste en ocultar una serie de caracteres, sustituyéndolos por asteriscos o similares. Es muy utilizado en las tarjetas de crédito, como por ejemplo:

N.º TARJETA	CADUCIDAD
**** **** **** 6052	04/24
**** **** **** 2002	11/23
**** **** **** 9512	10/24
**** **** **** 0341	12/25
**** **** **** 1467	12/23
**** **** **** 6052	02/26
**** **** **** 2444	05/24
**** **** **** 2002	09/26

Tabla 13: Enmascaramiento de cadenas

En resumen, las características de esta técnica son:

- Muy simple y fácil de utilizar.
- Adecuada para trabajar con conjuntos de datos en los que se pueda prescindir de partes de la información de de algún atributo.
- No adecuada para reutilizar todos los datos.
- Bastante segura si se elige adecuadamente la información que se enmascara.
- Adecuada para llevar a cabo transferencias por su seguridad.

4.4.4. Generalización

La generalización consiste en la sustitución para uno o más atributos, de ciertos valores por otros más generales.

Por ejemplo, en al conjunto de datos;

ID	Nombre	C. P.	Uso (kW)
1	Oscar	46118	230
2	Alberto	46181	118
3	Vicente	28001	145
4	Julia	22005	234
5	Enrique	28023	189
6	Jose Maria	28400	217
7	Angela	31670	212
8	Joaquín	25539	467

Tabla 14: Generalización. Datos originales

Mediante la información pública disponible sobre códigos postales se podría averiguar la población a la que pertenece cada usuario:

ID	Nombre	C. P.	Población	Uso (kW)
1	Oscar	46118	Paterna, Valencia	230
2	Alberto	46181	Serra, Valencia	118
3	Vicente	28001	Madrid, Madrid	145
4	Julia	22005	Pozuelo de Alarcón, Madrid	234
5	Enrique	28023	Madrid, Madrid	189
6	Jose Maria	28400	Collado Villalba, Madrid	217
7	Angela	28001	Madrid, Madrid	212
8	Joaquín	28036	Madrid, Madrid	467

Tabla 15: Obtención de información asociada

En esta tabla se podrían generalizar los códigos postales manteniendo las dos primeras cifras, de forma que se perdería precisión y tan solo sería posible identificar la provincia de los usuarios, de la siguiente forma:

ID	Nombre	C. P.	Población	Uso (kW)
1	Oscar	46000	Valencia	230
2	Alberto	46000	Valencia	118
3	Vicente	28000	Madrid	145
4	Julia	22000	Madrid	234

5	Enrique	28000	Madrid	189
6	Jose Maria	28000	Madrid	217
7	Angela	28000	Madrid	212
8	Joaquín	28000	Madrid	467

Tabla 16: Generalización

También se podría llevar a cabo una generalización por rangos de la variable Uso (kW):

ID	Nombre	C. P.	Uso (kW)
1	Oscar	46000	200-300
2	Alberto	46000	100-200
3	Vicente	28000	100-200
4	Julia	22000	200-300
5	Enrique	28000	100-200
6	Jose Maria	28000	200-300
7	Angela	31000	200-300
8	Joaquín	28000	400-500

Tabla 17: Generalización alternativa

En resumen, las características de esta técnica son:

- Medianamente simple y algo laboriosa de utilizar.
- No adecuada para reutilizar todos los datos.
- Bastante segura si se elige lleva a cabo la generalización sobre los registros adecuados.
- Bastante adecuada para llevar a cabo transferencias por su seguridad.

4.4.5. K-anonimidad

Formalmente, se define en [Sweeney 2022] la k-anonimidad como:

Sea $RT(A_1,...,A_n)$ una tabla y QI_{RT} el cuasi-identificador asociada a ella. Se dice que RT satisface la k-anonimidad si y solo si cada secuencia de valores en $RT[QI_{RT}]$ aparece con al menos k ocurrencias de $RT[QI_{RT}]$.

Se trata de una propiedad que permite asegurar, dado un valor k, que no es posible distinguir un individuo de otros k-1 individuos, ya que comparten el mismo valor en los cuasi-identificadores.

Por ejemplo, en el conjunto de datos:

ID	Edad	C. P.	Antecedentes tumorales
1	51	46118	Sí
2	52	46181	No
3	49	28001	No
4	35	22005	Sí
5	46	28023	Sí
6	43	28400	No
7	37	31670	No
8	49	25539	No

Tabla 18: K-anonimidad. Datos originales

Si aplicamos generalización sobre los atributos *Edad* y *Código Postal* sustituyendo los valores de Edad por intervalos de decena en decena, y los códigos postales como se ha hecho en la Tabla 16: Generalización para quedarnos solo con la provincia:

ID	Edad	C. P.	Antecedentes tumorales
1	51-60	46000	Sí
2	51-60	46000	No
3	41-50	28000	No
4	31-40	28000	Sí
5	41-50	28000	Sí
6	41-50	28000	No
7	31-40	28000	No
8	41-50	28000	No

Tabla 19: Clases de equivalencia

Así, se obtiene una 2-anonimidad con tres clases de equivalencia.

En resumen, las características de esta técnica son:

- Algo compleja.
- Adecuada para trabajar con conjuntos de datos en los que se pueda prescindir de información detallada de algunos registros.
- No adecuada para reutilizar todos los datos.
- Muy segura si se elige adecuadamente la información que se anonimiza.
- Bastante adecuada para llevar a cabo transferencias por su seguridad.

4.5. Consideraciones adicionales sobre la seguridad de la información

En el proceso de anonimización o seudonimización siembre debe tenerse en cuenta la legislación en protección de datos personales vigente. En España en la fecha de la presentación de este trabajo son especialmente relevantes el [RGPD 2017], el [RD 1720/2007], y la [LOPDGDD 2018] (Ley Orgánica de Protección de Datos y Garantía de los Derechos Digitales).

Es importante considerar los siguientes aspectos al llevar a cabo el proceso de anonimización o seudonimización:

4.5.1. Finalidad

Los datos deben siempre recogerse con una **finalidad**. El artículo 6 del [RGPD 2017] señala que la finalidad del tratamiento puede ser:

1. Fines científicos.
2. Fines históricos.
3. Fines estadísticos.
4. La ejecución de un contrato del cual el interesado es parte, o para la aplicación de medidas precontractuales.
5. El cumplimiento de una obligación legal por parte del responsable del tratamiento.
6. La protección de intereses vitales del interesado.
7. El cumplimiento de una misión de interés público.
8. La satisfacción de intereses legítimos perseguidos por el responsable del tratamiento o por un tercero, siempre que sobre estos intereses no prevalezcan los derechos y libertades del interesado, en particular cuando este sea un menor de edad.

Para llevar a cabo un tratamiento de datos, es necesario tener definida previamente la finalidad del tratamiento, y dicha finalidad debe ser compatible si es necesario o no el consentimiento según los supuestos contemplados por la ley.

Además, según [Europa 2014], el tratamiento que generan los datos anonimizados es un tratamiento de datos personales, que puede considerarse compatible con el fin original del tratamiento de datos personales del que proceden los datos. En otras palabras, en un tratamiento de datos en el que se lleva a cabo una seudonimización, se puede considerar que la finalidad de los datos anonimizados en la misma que la de los datos sin anonimizar.

4.5.2. Consentimiento

En la guía para pacientes y usuarios de la sanidad de la AEPD disponible en [AEPD 2019], la Agencia Española de Protección de Datos aclaró las dudas de interpretación del [RGPD 2017] y de la [LOPDGDD 2018] sobre si es necesario que un profesional sanitario o una clínica recaben el consentimiento de los pacientes para que les sea prestada asistencia médica, porque considera que la licitud del tratamiento no está basada en el consentimiento, sino que está basada en que el

tratamiento es necesario para la ejecución de un contrato en el que el interesado es parte, o para la aplicación a petición de este de medidas precontractuales.

Esto no es aplicable a los datos cuya finalidad sea otra que la propia prestación de la asistencia médica, con lo que para llevar a cabo otros tratamientos con sus datos, que perseguirían otros fines como la investigación biomédica, o incluso el posible envío de recomendaciones basándose en sus condiciones de salud, es necesario su consentimiento expreso, y además el paciente debe de ejercerlo de una forma libre y mediante una acción positiva.

En cuanto a los datos anonimizados, por sí mismos no son objeto del [RGPD 2017] ni la [LOPDGDD 2018], ya que no contienen datos personales que permitan identificar a un individuo, pero sí lo son en el tratamiento que los origina, con lo que se debe recabar en los supuestos exigidos por la ley.

4.5.3. Riesgo del uso de datos anonimizados o seudonimizados

Según [Europa 2014], hay que tener en cuenta los siguientes riesgos:

1. En los datos seudonimizados, como es posible reidentificar a los individuos, entran dentro del ámbito de aplicación de la legislación en protección de datos.
2. Aunque los datos estén anonimizados, se debe respetar la confidencialidad de las comunicaciones.

Esto significa que es necesario llevar a cabo para los datos seudonimizados los procedimientos impuestos por el [RGPD 2017], por la [LOPDGDD 2018] y por el [RD 1720/2007], entre ellos:

- Llevar a cabo un registro de actividades de tratamiento, que en el caso de los datos de salud es obligatorio.
- Llevar a cabo un análisis de riesgos.
- Llevar a cabo una Evaluación de Impacto en Privacidad (EIPD)
- Poner los medios para que los pacientes puedan ejercer sus derechos de acceso, limitación del tratamiento, oposición, rectificación, supresión, y no ser objeto de decisiones automatizadas.
- Disponer de un procedimiento de notificaciones de quiebras de seguridad.
- Garantizar la disponibilidad de la información mediante las copias de seguridad pertinentes.
- Cifrar las comunicaciones.
- Guardar un registro de accesos

5. Solución propuesta

5.1. Propósito

El problema considerado consiste en que el profesional sanitario, cuando tiene que generar datos para ser utilizados por sí mismo o por otros profesionales, no suele tener formación ni información sobre técnicas que le permitan reducir el riesgo de atribución. Además, con la promulgación de las últimas leyes y reglamentos en materia de protección de datos personales y la adaptación del entorno sanitario a dicha normativa, así como la información disponible sobre los cada vez más frecuentes ciberataques, el profesional del área de la salud es cada vez más consciente del peligro y los riesgos que conlleva el tratamiento de datos de un modo inseguro.

El facultativo que se encuentra en esta situación, si no trabaja como profesional independiente, tiene la opción de consultar sobre estos aspectos al Delegado de Protección de Datos (DPD) de su centro asistencial, ya que según el artículo 34.1 de la [LOPDGDD 2018], los centros sanitarios están obligados a contar con un DPD, pero un DPD no está obligado a estar certificado como tal, aunque la propia AEPD promueva la certificación, que está regulada por [AEPD 2019-2] con lo que en muchos casos la formación del DPD es deficiente y no contempla este tipo de técnicas, lo que lleva a que el profesional sanitario tampoco recibe una respuesta satisfactoria.

La Agencia Española de Protección de Datos suele salir al paso de este tipo de cuestiones, publicando guías y herramientas, como [AEPD 2016], pero la información disponible, en unos casos por ser demasiado técnica y en otros justo por lo contrario, no facilita suficientemente la tarea.

Llegados a este punto, dada la diversidad de técnicas que es posible utilizar para llevar a cabo la anonimización y la seudonimización de datos, sería interesante disponer de una herramienta que permitiera seleccionar de forma eficaz la más adecuada según nuestras necesidades, y que además estuviera enfocada a su uso por personal sanitario para facilitarles su labor.

5.2. Datos de entrada

Para los efectos de este trabajo, supondremos que el usuario dispone de:

1. Un **origen de datos**, por simplificar una tabla, con atributos identificadores, cuasi-identificadores, sensibles y/o no sensibles. En consonancia con el objetivo del presente trabajo, se asumirá que los datos sensibles son datos referidos a la salud del interesado.
2. Un **objetivo**, que puede ser uno de los siguientes:
    - Qué registros del origen de datos responden a unas determinadas características.
    - El resultado de aplicar estudios estadísticos, como el estudio de la distribución estadística, la búsqueda de valores como media, mediana, desviación típica, cuartiles, percentiles, etc..

3. Un **foco**, entendiendo como tal un atributo o conjunto de atributos que sean de interés para preservar la utilidad de los datos.
4. Si se van a llevar a cabo **transferencias** o no.
5. Si se pretende o no **reutilizar** los datos extraídos, por ejemplo, una vez ampliados o procesados por un tercero.

5.3. Información de salida

La salida del sistema propuesto debe ser:

1. El métodos más idóneo
2. Instrucciones de cómo llevar a cabo la anonimización o la seudonimización indicadas
3. Una serie de recomendaciones

5.4. Asignación del método más adecuado

Repasando los métodos considerados en el apartado 4. Contexto tecnológico, recordemos que eran los siguientes:

1. Técnicas de seudonimización
    1. Contador
    2. Números aleatorios
    3. Función resumen
    4. Cifrado simétrico
    5. Cifrado asimétrico
2. Técnicas de anonimización
    1. Supresión de atributos
    2. Supresión de registros
    3. Enmascaramiento de cadenas
    4. Generalización
    5. K-anonimidad

Y recordemos también las variables de entrada:

1. Origen de datos
2. Objetivo
3. Foco

4. Transferencias
5. Reutilización

Para llevar a cabo la selección de la técnica, se utilizarán el consentimiento, el objetivo, las transferencias y la reutilización.

Para seleccionar el método más adecuado, se valorarán las distintas posibilidades, intentando elegir aquellos que no exijan un esfuerzo desproporcionado dentro de los aplicables, y además teniendo en cuenta la legislación en protección de datos personales vigente, concretamente el [RGPD 2017], el [RD 1720/2007] y la [LOPDGDD 2018].

Según esto, las posibles alternativas son las siguientes:

Objetivo	Transferencias	Reutilización	Técnica/s adecuada/s
Cuales	No	No	• Enmascaramiento de cadenas
Cuales	No	Sí	• Contador • Números aleatorios
Cuales	Sí	No	• Supresión de atributos • Enmascaramiento de cadenas • Generalización
Cuales	Sí	Sí	• Función resumen • Cifrado simétrico • Cifrado asimétrico
Estudio estadístico	No	No	• Generalización • Números aleatorios • Supresión de registros
Estudio estadístico	No	Sí	No procede
Estudio estadístico	Sí	No	• Generalización • K-anonimidad • Supresión de registros
Estudio estadístico	Sí	Sí	No procede

Tabla 20: Combinaciones según las variables de entrada

En la anterior tabla se ha indicado *No procede* en algunas filas porque una vez transformado un conjunto de datos en una por medio de un estudio estadístico, normalmente no es posible llevar a cabo la operación inversa de obtener los registros originales.

5.5. Contenido de la guía

En el ANEXO I. Guía de anonimización y pseudonimización de datos para personal sanitario se muestra la guía que se estableció como objetivo secundario de este trabajo.

Teniendo en cuanta que la guía va enfocada a profesionales de la sanidad, en la guía se usará el término «paciente» como sinónimo de «interesado» o persona objeto del tratamiento de datos. Se intentará usar un lenguaje lo menos técnico posible a nivel informático, legal, o de ciencia de datos, para facilitar la compresión por parte del usuario a quien va dirigida. Así pues, entre otros, se usará el término «columna» en lugar de «campo» o «atributo», y «fila» en lugar de «registro».

6. Conclusiones. Líneas futuras

6.1. Conclusiones

Dijimos en 1.3. Objetivos del trabajo que el principal objetivo era la creación de un marco teórico en el que estarían contempladas las técnicas de anonimización y seudonimización más conocidas, de forma que partiendo de unas condiciones iniciales ofreciera indicaciones sobre qué técnica sería la más apropiada y cómo llevar a cabo la anonimización o seudonimización, en su caso.

En la propuesta ofrecida por el presente trabajo, que efectivamente incorpora dicho marco, se ha intentado acercar los conceptos aprendidos en el Máster Universitario en Ciberseguridad y Ciberinteligencia al día a día de los profesionales de la salud.

La guía que se incorpora como objetivo secundario en el ANEXO I. Guía de anonimización y pseudonimización de datos para personal sanitario ha sido diseñada sin demasiados tecnicismos, de una forma que resulte comprensible para el personal sanitario de forma que puedan incorporarla como una herramienta útil y eficaz a su trabajo diario.

Es precisamente por esto que, habiendo podido considerar otras técnicas más complejas, como por ejemplo la privacidad diferencial, se ha desechado la idea para no complicar demasiado el producto resultante y generar rechazo en el profesional que pretenda utilizarla.

La mayor dificultad de este trabajo ha sido poder precisar y concretar más en la sistematización de los datos de entrada, más en concreto de la fuente de datos, y ello ha sido porque es inherente al problema el hecho de que los datos puedan ser de cualquier procedencia, en cualquier formato, y contener información de cualquier tipo, de forma que se deja al usuario la tarea de organizar esos datos dentro de las categorías propuestas por el presente trabajo y decidir qué es lo que quiere obtener, aunque se le intenta guiar de una forma lo más amena posible para conseguirlo.

6.2. Líneas futuras

Como líneas futuras de trabajo, la primera de ellas sería validar si realmente es útil la guía propuesta, dándole un formato visual y poniéndola a disposición de los profesionales de la salud para que la usen, utilizando sus apreciaciones y comentarios como realimentación para mejorarla.

La segunda línea de trabajo sería el desarrollo de una herramienta software que implementara los conceptos y directrices propuestos en este trabajo, de forma que facilitara todavía más la tarea de anonimización y seudonimización y, sobretodo, la elección en el camino a tomar según las necesidades a intereses del usuario.

Una tercera línea de trabajo sería crear una versión de la guía más compleja destinada a usuarios acostumbrados a trabajar con conceptos estadísticos y de ciencia de datos, de forma que se pudieran incorporar conceptos no solo como la mencionada privacidad diferencial, sino también otros más novedosos como la aplicación de técnicas de *blockchain* para proteger los datos personales.

Para finalizar, la preparación de este proyecto ha sido de gran ayuda para repasar y profundizar en las tareas de anonimización y seudonimización, hasta tal punto que puedan ser trasladadas a otras personas sin conocimientos sobre estas materias, o como dijo Ralph Waldo Emerson:

«El educador es el hombre que hace que las cosas difíciles parezcan fáciles.»

Bibliografía y referencias

RD 1720/2007: España. Real Decreto 1720/2007, de 21 de diciembre, por el que se aprueba el Reglamento de desarrollo de la Ley Orgánica 15/1999, de 13 de diciembre, de protección de datos de carácter personal. *Boletín Oficial del Estado*, 19 de enero de 2008, núm. 17.

RAE 2023: Real Academia Española, 2023. *Diccionario de la lengua española | Edición del Tricentenario | RAE*. [Consulta: 16 julio 2023] . Disponible en: https://dle.rae.es/

Casas Roma y Romero Tris 2017: CASAS ROMA, Jordi, ROMERO TRIS, Cristina. *Privacidad y anonimización de datos*. Barcelona: Universitat Oberta de Catalunya. (UOC), 2017. ISBN: 9788491169383

RGPD 2017: Unión Europea. Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos. *Diario oficial de la Unión Europea*, 4 de mayo de 2016, núm. L 119.

OMS 2022: Organización Mundial de la Salud, 2022. *Publicación de la CIE-11 2022*. [Consulta: 24 julio 2023] . Disponible en: https://www.who.int/es/news/item/11-02-2022-icd-11-2022-release

OMS 2023: Organización Mundial de la Salud, 2023. *CIE-11*. [Consulta: 26 julio 2023] . Disponible en: https://icd.who.int/es

Min. Sanidad 2023: Ministerio de Sanidad de España, 2023. *SNOMED CT*. [Consulta: 29 julio 2023] . Disponible en: https://www.sanidad.gob.es/areas/saludDigital/interoperabilidadSemantica/factoriaRecursos/snomedCT/home.htm

SNOMED 2023-1: SNOMED International, 2023. *SNOMED CT Browser*. [Consulta: 20 julio 2023] . Disponible en: https://browser.ihtsdotools.org/?perspective=full&conceptId1=404684003&edition=MAIN/SNOMEDCT-ES/2023-04-30&release=&languages=es,en

SNOMED 2023-2: Snomed International, 2023. *SNOMED International Home*. [Consulta: 20 julio 2023] . Disponible en: https://confluence.ihtsdotools.org/

ENISA 2022: Agencia de la Unión Europea para la Ciberseguridad, 2022. *La adopción de técnicas de seudonimización. El caso del sector sanitario*. [consulta: 15 julio 2023] , ISBN: 978-92-9204-576-0. Disponible en: https://www.aepd.es/es/documento/tecnicas-seudonimizacion-sector-sanitario-enisa.pdf

Murty 2019: MURTY, Suntherasvaran et al. A Comparative Study of Data Anonymization Techniques. 2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing (HPSC), and IEEE Intl Conference on Intelligent Data and Security (IDS), 2019

Sweeney 2022: Sweeney, Latanya, "K-anonymity: a model for protecting privacy". *International Journal on Uncertainty*. 2002 , núm. 10 (5), p. 557-570

LOPDGDD 2018: . Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales. *Boletín Oficial del Estado*, 294, de 06/12/2018.

Europa 2014: Grupo de trabajo sobre protección de datos del Artículo 29, 2014. *Dictamen 05/2014 sobre técnicas de anonimización*. [consulta: 16 julio 2023] . Disponible en: https://www.aepd.es/es/documento/wp216-es.pdf

AEPD 2019: Agencia Española de Protección de Datos, 2019. *Guía para pacientes y usuarios de la sanidad*. [Consulta: 18 de julio 2023] . Disponible en: https://www.aepd.es/sites/default/files/2019-12/guia-pacientes-usuarios-sanidad.pdf

AEPD 2019-2: Agencia Española de Protección de Datos, 2019. *Esqauema de certificación de delegados de protección de datos de la Agencia Española de Protección de Datos (Esquema AEPD-DPD)*. [Consulta: 25 julio 2023] . Disponible en: https://www.aepd.es/es/documento/esquema-aepd-dpd.pdf

AEPD 2016: Agencia Española de Protección de Datos, 2016. *Orientaciones y garantías en los procedimientos de anonimización de datos personales*. [Consulta: 16 julio 2023] . Disponible en: https://www.aepd.es/es/documento/guia-orientaciones-procedimientos-anonimizacion.pdf

CodeBeautify 1: CodeBeautify, 2023. *MD5 Hash Generator*. [consulta: 18 julio 2023] . Disponible en: https://codebeautify.org/md5-hash-generator

CodeBeautify 2: CodeBeautify, 2023. *SHA-1 Hash Generator*. [consulta: 18 julio 2023] . Disponible en: https://codebeautify.org/sha1-hash-generator

CodeBeautify 3: CodeBeautify, 2023. *SHA-2 Hash Generator*. [consulta: 18 julio 2023] . Disponible en: https://codebeautify.org/sha2-hash-generator

CodeBeautify 4: CodeBeautify, 2023. *SHA224 Hash Generator*. [consulta: 18 julio 2023] . Disponible en: https://codebeautify.org/sha224-hash-generator

CodeBeautify 5: CodeBeautify, 2023. *SHA256 Hash Generator*. [consulta: 18 julio 2023] . Disponible en: https://codebeautify.org/sha256-hash-generator

CodeBeautify 6: CodeBeautify, 2023. *SHA384 Hash Generator*. [consulta: 18 julio 2023] . Disponible en: https://codebeautify.org/sha384-hash-generator

CodeBeautify 7: CodeBeautify, 2023. *SHA512 Hash Generator*. [consulta: 18 julio 2023] . Disponible en: https://codebeautify.org/sha512-hash-generator

SNOMED 2023: Snomed International, 2023. *SNOMED International Home*. [Consulta: 20 julio 2023] . Disponible en: https://confluence.ihtsdotools.org/

OMS 2023-2: Organización Mundial de la Salud, 2023. *CIE-11 para estadísticas de mortalidad y morbilidad (Versión : 01/2023)*. [Consulta: 22 julio 2023] . Disponible en: https://icd.who.int/browse11/l-m/es

ANEXO I. Guía de anonimización y pseudonimización de datos para personal sanitario

0. Introducción

En la sociedad actual se llevan a cabo multitud de tratamientos de datos, tratamientos que, gracias a las técnicas actuales de procesamiento, proporcionan grandes facilidades para la investigación médica; sin embargo, si no se protege de forma adecuada, el tratamiento de datos personales puede poner en riesgo la libertad de las personas por un uso indebido de sus datos personales.

Con esta guía se pretende facilitar a los médicos, estudiantes e investigadores de ciencias de la salud la tarea de utilizar sus datos de forma segura mediante técnicas que permiten reducir el riesgo de que una persona sea identificada, además de una forma fácil de comprender.

Esta guía está estructurada en los siguientes apartados:

1. **Identificadores, seudoidentificadores, datos sensible y no sensibles**. En este apartado se introducen estos conceptos básicos para llevar a cabo el proceso de anonimización o seudonimización.
2. **Anonimización y seudonimización**. Aquí se introducen estos dos conceptos, fundamentales en la tarea que se pretende llevar a cabo.
3. **Comenzando**. Se darán las directrices para empezar a trabajar.
4. **Selección de la técnica**. Se guiará para elegir la técnica más adecuada.
5. **Técnicas**. En este apartado se explican de una forma sencilla las técnicas que se utilizan para reducir o eliminar el riesgo de que una persona sea identificada.
6. **Consideraciones adicionales.** Se ofrecerán indicaciones que hay que tener en cuenta para llevar a cabo la tarea de anonimización o seudonimización.

1. Identificadores, seudoidentificadores, datos sensible y no sensibles

En todo conjunto de datos, que normalmente vendrán en forma de una o varias tablas, para que no se pueda distinguir a los pacientes asociados es necesario ocultar o eliminar la información que permita identificarlos.

Podemos catalogar las columnas de una tabla como:

1. **Identificadoras**: Son las columnas que permiten identificar directamente los pacientes, como el nombre y los apellidos, el DNI, el número de la seguridad social, etc.
2. **Cuasi-identificadoras**: Son las columnas que por sí mismo no dan información suficiente como para identificar a un paciente, pero que junto con otras sí lo permitirían, como la fecha de nacimiento, la dirección postal, o el teléfono.
3. **Sensibles**: En nuestro caso, cualquier dato médico perteneciente a la historia clínica o relacionado con la vida sexual.
4. **No sensibles**: El resto, como su lugar de nacimiento o su profesión.

2. Anonimización y seudonimización

Para que no sea posible, o por lo menos requiera un gran esfuerzo, identificar a un paciente dentro de una tabla, se utilizan dos estrategias, la **anonimización** y la **seudonimización**.

Veamos en qué consisten.

- La **anonimización** es el proceso de transformar los datos personales de forma que no sea posible identificar a un paciente. Para ello, transformaremos las columnas identificadoras y cuasi-identificadoras.
- La **seudonimización**, por otra parte, es el proceso por el cual se transforman los datos de forna que no sea posible identificar a un paciente sin utilizar información adicional, como por ejemplo una contraseña o una descripción del método que se ha seguido para «ocultar» la información.

3. Comenzando

Para llevar a cabo la operación que hará que se oculte la información que permita identificar a un paciente, es necesario disponer de:

1. Una lo varias **tablas** con datos de pacientes. Como se ha indicado en el apartado anterior, las columnas de esta tabla podrán ser identificadoras, cuasi-identificadoras, sensibles y/o no sensibles.

   Por ejemplo:

Id	Nombre	1er. Apellido	Presión arterial	Motivo visita
1	Oscar	Fort	80/122	Dolor de oídos
2	Alberto	Ruiz	76/130	Vértigo
3	Vicente	Sanz	79/119	Pérdida de audición
4	Julia	Palau	60/114	Revisión cirugía
5	Enrique	Pérez	85/129	Acúfenos
6	Jose Maria	Gómez	100/173	Cefaleas

7	Angela	Del Villar	65/140	Anosmia
8	Joaquín	Izquierdo	88/180	Ronquera

Tabla 21: Ejemplo de tabla

2. Un **foco**, que serán las columnas que sean necesarias para que los datos sean útiles. Por ejemplo, si disponemos de los apellidos, dirección y motivo de visita de los pacientes del último mes y queremos estudiar los pacientes que viven en Valencia y han acudido a la consulta por dolor de oídos, las columnas que tendremos que mantener serán sobretodo las que indiquen la población y el motivo de visita:

ID	Apellidos	Dirección	Motivo visita
1	Fort Palau	San Vicente, 45. Valencia	Dolor de oídos
2	Ruiz Tomás	Oronet, 24. Serra	Vértigo
3	Sanz Taberner	Ramón Llull 23. Valencia	Pérdida de audición
4	Palau Ferrer	Mayor, 3. Paterna	Revisión cirugía
5	Pérez Sanchís	217, 28. Paterna	Acúfenos
	Gómez Pascual	San Carlos, 3. Albal	Cefalea
	Del Villar Ferrer	Lebón, 36. Valencia	Anosmia
	Izquierdo Peris	La Safor, 24. Valencia	Ronquera

Tabla 22: Ejemplo de foco

4. Selección de la técnica

Para elegir la técnica, necesitaremos:

1. Un **objetivo**, que puede ser uno de los siguientes:
   a. **Qué filas de la tabla responden a unas determinadas características**. Por ejemplo, en una tabla con diagnósticos, qué filas corresponden al diagnóstico «triquinosis» o si están agrupadas por aparatos, cuales de ellas corresponden con enfermedades del aparato respiratorio.
   b. **El resultado de un estudio estadístico sobre la tabla**. Por ejemplo, qué distribución estadística siguen las de la muestra, un estudio por áreas geográficas de pacientes han tenido un accidente cerebrovascular, etc.
2. Si se van a llevar a cabo o no **transferencias** a otros médicos o centros sanitarios.
3. Si se pretende o no llevar a cabo una **reutilización** de los datos extraídos. Por ejemplo, podríamos querer pasar a otro centro de forma anonimizada los datos de pacientes que han tenido un tipo de alergia y que ese centro nos devolviera información ampliada.

La selección de la técnica se llevará a cabo según lo indicado en la siguiente tabla:

Objetivo	Transferencias	Reutilización	Técnica/s adecuada/s
Qué filas	No	No	• Anonimización por enmascaramiento de cadenas
Qué filas	No	Sí	• Seudonimización por contador • Seudonimización por números aleatorios
Qué filas	Sí	No	• Anonimización por supresión de columnas • Anonimización por enmascaramiento de cadenas • Anonimización por generalización
Qué filas	Sí	Sí	• Seudonimización por función resumen • Seudonimización por cifrado simétrico • Seudonimización por cifrado asimétrico
Estudio estadístico	No	No	• Anonimización por generalización • Seudonimización por números aleatorios • Anonimización por supresión de filas
Estudio estadístico	No	Sí	No procede
Estudio estadístico	Sí	No	• Anonimización por generalización • K-anonimidad • Anonimización por supresión de filas
Estudio estadístico	Sí	Sí	No procede

Tabla 23: Selección de la técnica

5. Técnicas

Las técnicas más utilizadas para llevar a cabo la anonimización o seudonimización son las siguientes:

1. Seudonimización por contador
2. Seudonimización por números aleatorios
3. Seudonimización por función resumen
4. Seudonimización por cifrado simétrico
5. Seudonimización por cifrado asimétrico
6. Anonimización por supresión de columnas
7. Anonimización por supresión de filas
8. Anonimización por enmascaramiento de cadenas
9. Anonimización por generalización
10. K-anonimidad

En detalle:

1. **Seudonimización por contador**. Se elije un valor al azar, y se incrementa cuando se necesita un nuevo seudónimo.

   Por ejemplo:

ID	Nombre	Nombre (seudonimizado)	Presión arterial
1	Oscar	5	80/122
2	Alberto	6	76/130
3	Vicente	7	79/119
4	Julia	8	60/114
5	Enrique	9	85/129
6	Jose Maria	10	100/173
7	Angela	11	65/140
8	Joaquín	12	88/180

Tabla 24: Seudonimización por contador

2. **Seudonimización por números aleatorios**. Se elige un valor al azar entre unos límites mínimo y máximo.

   Por ejemplo, eligiendo valores aleatorios entre el 1 y el 50:

ID	Nombre	Nombre (seudonimizado)	Presión arterial
1	Oscar	22	80/122
2	Alberto	2	76/130
3	Vicente	13	79/119
4	Julia	24	60/114
5	Enrique	13	85/129
6	Jose Maria	24	100/173
7	Angela	32	65/140
8	Joaquín	33	88/180

Tabla 25: Seudonimización por números aleatorios

3. **Seudonimización por función resumen**. Se utiliza una función matemática que transforma los datos personales de entrada en otros valores, numéricos o compuestos por letras y números, de longitud fija.

   Algunas funciones resumen muy utilizadas son: MD5, SHA-1, SHA-224, SHA-256, SHA-384, y SHA-512.

   Queda fuera del ámbito de esta guía la explicación en detalle de estas funciones resumen, pero existen programas para calcularlas, y también herramientas *online,* como [CodeBeautify 1], [CodeBeautify 2], [CodeBeautify 3], [CodeBeautify 4], [CodeBeautify 5], [CodeBeautify 6] y [CodeBeautify 7].

   Por ejemplo aplicando a la columna *Nombre* la función SHA-1:

ID	Nombre	Nombre (seudonimizado)	Presión arterial
1	Oscar	07f24c146c9cd13d69fdc5e-f719e97aec36f24fe	80/122
2	Alberto	3b3e55fdc7886baea165a854-d080caf9808cac97	76/130
3	Vicente	7277f956e52a382f15ed35f8e1a-f3fc78663e9af	79/119
4	Julia	e64d664b335757ab1b0ed70-dc6883a5f412be34b	60/114
5	Enrique	9cdcdaec09c4ecdeb7df30-db270618e77ec4eee5	85/129

6	Jose Maria	377b6375dd4b3940bc67de1cc74-c9fb616114df6	100/173
7	Angela	5e69b41340098b1812be6593-f81e97e6500e8d5b	65/140
8	Joaquín	063e723c004481488-d301e3607135ea393a08d14	88/180

Tabla 26: Seudonimización por funciones resumen

4. **Seudonimización por cifrado simétrico.** En este caso se utiliza una función para cifrar los datos con una contraseña, y se utiliza la misma contraseña para descifrarlos.

   Por ejemplo, se puede utilizar como clave un número del 1 al 27, y se sustituye cada una de las letras por la letra que resulta de desplazarse a la derecha el número de veces indicado en la clave.

   Con una clave = 5, siguiendo con el mismo ejemplo resultaría:

ID	Nombre	Nombre (seudonimizado)	Presión arterial
1	Oscar	Xblja	80/122
2	Alberto	Fpgjwyt	76/130
3	Vicente	Zmgiqxi	79/119
4	Julia	Ñzpnf	60/114
5	Enrique	Jrwnvzj	85/129
6	Jose Maria	Ñtxj Qfwnf	100/173
7	Angela	Frljpf	65/140
8	Joaquín	Ñtfvzír	88/180

Tabla 27: Seudonimización por cifrado simétrico

5. **Seudonimización por cifrado asimétrico**. Con el cifrado simétrico, no existe una sola clave sino dos: una clave privada que se utiliza para cifrar el texto, y una clave pública, que se utiliza para descifrarlo.

   Para hacerlo se utiliza un software específico, como PGP (Pretty Good Privacy), ya que ni Microsoft Windows ni Apple OS proporcionan por defecto ninguna aplicación para llevar a cabo esta tarea, y el funcionamiento es el siguiente:

   1. Se genera el par de claves: la clave pública y la privada
   2. Se cifran los textos con la clave privada
   3. Si es necesario revertir el cifrado, se descifran con la clave pública

6. **Anonimización por supresión de columnas.** Se trata de cambiar todos los valores de alguna de las columnas por otros sin significado, como por ejemplo ***. Es importante

que la o las columnas en las que se suprimas los atributos no sea ninguna de las del foco que se mencionó en el apartado 3. Comenzando.

Por ejemplo:

ID	Nombre	Nombre (anonimizado)	Presión arterial
1	Oscar	*****	80/122
2	Alberto	*****	76/130
3	Vicente	*****	79/119
4	Julia	*****	60/114
5	Enrique	*****	85/129
6	Jose Maria	*****	100/173
7	Angela	*****	65/140
8	Joaquín	*****	88/180

Tabla 28: Anonimización por supresión de columnas

7. **Anonimización por supresión de filas**. Se trata de eliminar parte de las filas de la tabla o de cambiar sus valores por otro sin significado, como ***.

   Por ejemplo:

ID	Nombre	Presión arterial
***	***	***
***	***	***
3	Vicente	79/119
4	Julia	60/114
***	***	***
6	Jose Maria	100/173
7	Angela	65/140
8	Joaquín	88/180

Tabla 29: Anonimización por supresión de filas

   Es importante utilizar esta técnica teniendo en cuenta las celdas que se eliminan. Por ejemplo, se puede utilizar para eliminar filas con valorea anormalmente altos o bajos.

8. **Anonimización por enmascaramiento de cadenas**. Consiste en ocultar una serie de letras o números, sustituyéndolos por aspas o similares. Es muy utilizado en las tarjetas de crédito, como por ejemplo:

N.º TARJETA	CADUCIDAD
**** **** **** 6052	04/24
**** **** **** 2002	11/23
**** **** **** 9512	10/24
**** **** **** 0341	12/25
**** **** **** 1467	12/23
**** **** **** 6052	02/26
**** **** **** 2444	05/24
**** **** **** 2002	09/26

Tabla 30: Anonimización por enmascaramiento de cadenas

9. **Anonimización por generalización**. Se trata de sustitución, para uno o más columnas, de ciertos valores por otros más generales.

   Por ejemplo, en la siguiente tabla, que utiliza SNOMED CT:

ID	Nombre	Diagnóstico	Edad (años)
1	Oscar	Hiperqueratosis (SCTID: 26996000)	51
2	Alberto	Hiperplasia linfoide (SCTID: 43961000)	52
3	Vicente	Edema agudo (SCTID: 40829002)	49
4	Julia	Colesteatoma (SCTID: 363668000)	35
5	Enrique	Quiste cutáneo (SCTID: 285302001)	46
6	Jose Maria	Otoesclerosis coclear (SCTID: 91108004)	43
7	Angela	Micosis dérmica (SCTID: 14560005)	37
8	Joaquín	Eccema de la cara (SCTID: 300924008)	49

Tabla 31: Generalización. Datos originales

   Se puede sustituir el diagnóstico por uno más general, y además generalizar la edad estableciendo rangos de edades.

   Esta generalización puede llevarse a cabo con ayuda del navegador de la terminología médica SNOMED CT disponible en [SNOMED 2023]:

ID	Nombre	Diagnóstico (generalizado)	Edad (años) (generalizada)
1	Oscar	Hiperplasia (SCTID: 76197007)	51-60
2	Alberto	Hiperplasia (SCTID: 76197007)	51-60
3	Vicente	Edema (SCTID: 79654002)	41-50
4	Julia	Queratosis (SCTID: 254666005)	31-40
5	Enrique	Quiste (SCTID: 441457006)	41-50
6	Jose Maria	Otoesclerosis (SCTID: 11543004)	41-50
7	Angela	Micosis superficial (SCTID: 276206000)	31-40
8	Joaquín	Eccema (SCTID: 43116000)	41-50

Tabla 32: Anonimización por generalización con SNOMED CT

Utilizando CIE-11 mediante el navegador disponible en [OMS 2023-2], la generalización quedaría de la siguiente forma:

ID	Nombre	Diagnóstico (generalizado)	Edad (años) (generalizada)
1	Oscar	ED51 Hiperplasia	51-60
2	Alberto	DB11.1 Hiperplasia del apéndice	51-60
3	Vicente	EE20 Síndrome de distensión cutánea aguda	41-50
4	Julia	Enfermedades del oído medio o de la apófisis mastoides	31-40
5	Enrique	Proliferaciones, neoplasias y quistes benignos de la piel	41-50
6	Jose Maria	Enfermedades del oído interno	41-50
7	Angela	1F28 Dermatofitosis	31-40
8	Joaquín	EA81 Dermatitis sebo-	41-50

		rreica y afecciones relacionadas	

Tabla 33: Anonimización por generalización con CIE-11

Como podemos comprobar, la generalización puede variar utilizando una terminología médica u otra, por lo que es un aspecto a tener en cuenta.

10. **K-anonimidad.** Se trata de una propiedad que permite asegurar, dado un valor k, que no es posible distinguir un paciente de otros k-1 pacientes, ya que comparten el mismo valor en los cuasi-identificadores.

    Por ejemplo, en el conjunto de datos:

ID	Peso	Teléfono	Antecedentes tumorales
1	95	963 303 582	Sí
2	78	963 378 643	No
3	74	963 293 183	No
4	67	911 786 532	Sí
5	65	911 765 234	Sí
6	69	911 423 879	No
7	92	963 245 356	No

Tabla 34: K-anonimidad. Datos originales

Podemos aplicar generalización sobre las columnas *Peso* y *Teléfono,* sustituyendo los valores de *Peso* por intervalos de decena en decena, y enmascarando los números de teléfono de forma que queden solo las tres cifras que corresponden al prefijo telefónico de la provincia:

ID	Peso	Teléfono	Antecedentes tumorales
1	91-100	963 *** ***	Sí
2	71-80	963 *** ***	No
3	71-80	963 *** ***	No
4	61-70	911 *** ***	Sí
5	61-70	911 *** ***	Sí
6	61-70	961 *** ***	No
7	91-100	963 *** ***	No

Tabla 35: Aplicación de la K-anonimidad

Así, se obtiene una **2-anonimidad** con **tres clases de equivalencia**.

6. Consideraciones adicionales

Debe tener en cuenta lo siguiente:

1. No es necesario el consentimiento expreso de los pacientes para prestarles asistencia médica, pero sí para tratar sus datos con fines científicos.
2. Si los datos que está tratando incluyen identificadores o cuasi-identificadores, debe contar con el consentimiento de los pacientes para tratarlos.
3. Por el contrario, si los datos de partida no contienen identificadores ni cuasi-identificadores, puede tratarlos libremente sin consentimiento.
4. Si los datos se obtuvieron los datos con un fin determinado, como es el tratamiento médico del paciente, no se pueden utilizar para otro fin, como la investigación biomédica, sin su consentimiento expreso.
5. Con los datos seudonimizados hay que tomar medidas de seguridad al igual que si no lo estuvieran, ya que es posible reidentificar a los pacientes, entre ellas:
    - Llevar a cabo un registro de actividades de tratamiento.
    - Llevar a cabo un análisis de riesgos en privacidad.
    - Llevar a cabo una Evaluación de Impacto en Protección de Datos (EIPD).
    - Poner los medios para que los pacientes puedan ejercer sus derechos de acceso, limitación del tratamiento, oposición, rectificación, supresión, y no ser objeto de decisiones automatizadas.
    - Disponer de un procedimiento de notificaciones de quiebras de seguridad.
    - Garantizar la disponibilidad de la información mediante las copias de seguridad pertinentes.
    - Cifrar las comunicaciones.
    - Guardar un registro de accesos
6. Aunque los datos estén anonimizados, se debe respetar del mismo modo la confidencialidad de las comunicaciones.

ANEXO II. Objetivos de desarrollo sostenible

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenibles	Alto	Medio	Bajo	No Procede
ODS 1. Fin de la pobreza				●
ODS 2. Hambre cero				●
ODS 3. Salud y bienestar	●			
ODS 4. Educación de calidad			●	
ODS 5. Igualdad de género				●
ODS 6. Agua limpia y saneamiento				●
ODS 7. Energía asequible y no contaminante				●
ODS 8. Trabajo decente y crecimiento económico				●
ODS 9. Industria, innovación e infraestructuras		●		
ODS 10. Reducción de las desigualdades				●
ODS 11. Ciudades y comunidades sostenibles				●
ODS 12. Producción y consumo responsables				●
ODS 13. Acción por el clima				●
ODS 14. Vida submarina				●
ODS 15. Vida de ecosistemas terrestres				●
ODS 16. Paz, justicia e instituciones sólidas		●		
ODS 17. Alianzas para lograr objetivos	●			

Tabla 36: Relación con los Objetivos de Desarrollo Sostenible (ODS)

Reflexión sobre la relación del TFM con los ODS y con el/los ODS más relacionados

La Agenda 2030 sobre el Desarrollo Sostenible, aprobada en 2015 por la Organización de las Naciones Unidas, es una oportunidad para que los países y sus sociedades lleven a cabo acciones con el que mejorar la vida de todas las personas.

La Agenda cuenta con 17 Objetivos de Desarrollo Sostenible enumerados en la Tabla 36: Relación con los Objetivos de Desarrollo Sostenible (ODS).

En este trabajo se aprecia una cierta relación con los siguientes objetivos:

- ODS 3. Salud y bienestar
- ODS 4. Educación de calidad
- ODS 9. Industria, innovación e infraestructuras
- ODS 16. Paz, justicia e instituciones sólidas
- ODS 17. Alianzas para lograr objetivos

En este trabajo tiene especial relevancia el **ODS 3. Salud y bienestar** ya que va destinado a tratar datos clínicos de forma segura, de forma que los profesionales puedan llevar a cabo estudios científicos sin temor a estar contraviniendo la legislación vigente en materia de datos personales, y por lo tanto facilitando ese tipo de estudios e investigaciones.

Tiene también una cierta relación con el **ODS 4. Educación de calidad**, ya que el fin de este trabajo es en buena parte didáctico y se pretende explicar de una forma clara, sencilla y con pocos tecnicismos las herramientas para dificultar la identificación de los individuos en los ficheros con datos personales.

En cuanto al **ODS 9. Industria, innovación e infraestructuras**, está relacionado con este trabajo porque se tratan materias y técnicas innovadoras aplicables inmediatamente a entornos reales, y que además puede poner en peligro la seguridad de las personas si no se aplican adecuadamente.

El **ODS 16. Paz, justicia e instituciones sólidas** tiene bastante relevancia en este trabajo porque en definitiva se trata de preservar la seguridad de los datos personales, reconocidos como datos fundamentales en el artículo 17 de la Constitución Española. Preservando esa seguridad, estamos preservando la igualdad entre las personas en cuanto al tratamiento de sus datos, previniendo posibles abusos o usos indebidos.

Para finalizar, el presente trabajo de final de máster está muy implicado en el **ODS 17. Alianzas para lograr objetivos**, porque parte de sus objetivos es facilitar el intercambio de datos clínicos dificultando o imposibilitando la identificación de los individuos. Ese intercambio de datos puede ser crucial para llevar a cabo investigaciones de forma segura por parte de distintos científicos o grupos de investigación que pueden llevar a importantes avances en asuntos tales como el tratamiento de enfermedades o en la mejora de la asistencia sanitaria.