Universidad Zaragoza

Trabajo Fin de Grado

Anonimización de datos en documentos médicos
Anonymization of health data and documents

Autor

Marta Morales Sabroso

Directores

Sergio Ilarri Artigas

Carlos Tellería Orriols

ESCUELA DE INGENIERÍA Y ARQUITECTURA

2022

TRABAJOS DE FIN DE GRADO / FIN DE MÁSTER

Escuela de Ingeniería y Arquitectura Universidad Zaragoza

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD

(Este documento debe acompañar al Trabajo Fin de Grado (TFG)/Trabajo Fin de Máster (TFM) cuando sea depositado para su evaluación).

D./Dª. _____,

con nº de DNI _____ en aplicación de lo dispuesto en el art. 14 (Derechos de autor) del Acuerdo de 11 de septiembre de 2014, del Consejo de Gobierno, por el que se aprueba el Reglamento de los TFG y TFM de la Universidad de Zaragoza,

Declaro que el presente Trabajo de Fin de (Grado/Máster) _____, (Título del Trabajo)

_____

_____

_____

_____

_____

_____,

es de mi autoría y es original, no habiéndose utilizado fuente sin ser citada debidamente.

Zaragoza, _____

Fdo: _____

AGRADECIMIENTOS

En primer lugar, me gustaría agradecer la labor de Sergio Ilarri y Carlos Tellería, tutores de este proyecto, por su implicación en él, su ayuda proporcionada en todo momento y sus indicaciones para elaborar la mejor de las versiones del mismo.

Agradecer a mi familia, por estar siempre ahí, apoyarme y guiarme en cada paso que doy. Ellos saben el esfuerzo que he invertido para llegar hasta aquí.

Agredecer también a mis amigos más cercanos por estar siempre de manera incondicional. Más concretamente a Pablo García y Alejandro Magallón, compañeros y amigos, por no soltarme nunca de la mano.

También, quería hacer especial mención a Borja Melendo, por su apoyo diario, comprensión, cariño y paciencia en el tiempo en el que he desarrollado este proyecto.

Finalmente, este trabajo se ha desarrollado como parte del proyecto de I+D+i PID2020-113037RB-I00, financiado por MCIN/AEI/ 10.13039/501100011033. Además del proyecto previo (proyecto NEAT-AMBIENCE), se agradece también el apoyo del Departamento de Ciencia, Universidad y Sociedad del Conocimiento del Gobierno de Aragón (Gobierno de Aragón: referencia de grupo T64_20R, grupo COSMOS).

Anonimización de datos en documentos médicos

RESUMEN

El tratamiento de la información de los ciudadanos cada vez tiene más relevancia en el ámbito de la investigación. Mucha de la información que se proporciona a las empresas requiere ser tratada para un procesamiento posterior. Además, suele contener datos sensibles desde la perspectiva del derecho a la intimidad. Es por esto que la anonimización de datos en documentos está cogiendo más importancia. Este proceso se caracteriza por la eliminación de información sensible con el objetivo de poder manejar información despersonalizada sin limitaciones, pero sin perder la información necesaria a analizar. En el caso del ámbito de la salud, la utilización de mecanismos de anonimización facilita el análisis de la documentación, ya que la mayoría de los informes clínicos almacenados poseen información identificativa de los pacientes.

Esta aplicación se ha desarrollado con el propósito de que los profesionales sanitarios puedan trabajar con informes médicos anonimizados a un cierto nivel, siendo ellos los que eligan el nivel de sensibilidad o de anonimización que desean aplicar a sus expedientes médicos. Este proyecto forma parte del marco del proyecto de investigación NEAT-AMBIENCE, y más concretamente el caso de uso de salud, que aborda la gestión de datos para ayudar a los ciudadanos en su vida cotidiana.

El sistema desarrollado consiste en un sistema de anonimización de datos en documentos médicos, donde el usuario carga los documentos a procesar junto a la configuración de anonimización que desea aplicarles. En este caso, el usuario podrá elegir entre distintos tipos de atributos y/o nivel de sensibilidad de los datos. Esto permite al usuario realizar diversas combinaciones y eliminar los datos que desee. Para el correcto funcionamiento del sistema se han incluido librerías que permiten realizar el procesamiento del lenguaje natural, ya que se trata de documentos no estructurados. Además, el sistema ofrece una técnica de clasificación de documentos, donde puede evaluar los ficheros originales junto con los que ha obtenido anonimizados y comparar sus resultados. Este sistema ha sido probado con informes reales proporcionados por el *Instituto Aragonés de Ciencias de la Salud (IACS)*.

Índice

Capítulo 1

Introducción y objetivos

En este capítulo se va a hablar de la anonimización de datos en documentos médicos, ya que es el tema a abordar en este proyecto. Este proceso está siendo muy demandado en el ámbito de la investigación, debido a su gran utilidad en investigaciones posteriores. En este caso, se van a procesar documetos clínicos escritos en castellano. Esto incrementa la dificultad, pues la mayoría de los avances obtenidos han sido con textos escritos en inglés. En la Sección 1.1 se introduce el contexto de la anonimización de datos en documentos médicos en la actualidad y la motivación del proyecto. En la Sección 1.2 se mencionan los objetivos y el alcance del proyecto. Finalmente, en la Sección 1.3 se explica la organización del proyecto y el contenido de esta memoria.

1.1. Anonimización de datos en documentos médicos y motivación del proyecto

En la actualidad, nuestros datos personales son requeridos por una gran cantidad de empresas y organismos que necesitan dicha información para obtener estadísticas, gestionar redes sociales, entidades bancarias, asuntos judiciales e, incluso, informes médicos. En todos estos casos, las entidades deben garantizar la integridad y la privacidad de los datos de sus clientes y evitar la posibilidad de que cualquiera de esos datos puedan estar al alcance de alguien que no tenga el permiso de su lectura o manejo.

Como bien se explica en el proyecto *NEAT-AMBIENCE* [1], a día de hoy, se requiere desarrollar la gestión de datos para ayudar a los ciudadanos en su vida cotidiana. Añadiendo, además, el tema del *Big Data*, ya que las personas cada vez usan más las tecnologías y la información cada vez está más digitalizada. Esto supone que se sumen más datos a la red y se haga más difícil su extracción, filtrado, manejo y gestión. Por esta razón, este trabajo se plantea en el marco del proyecto de

investigación *NEAT-AMBIENCE* en el caso de uso de salud.

En este proyecto se va a abordar la anonimización de datos en documentos médicos, ya que se considera que los informes clínicos deben ser tratados puramente desde el ámbito clínico, sin necesidad de tener que conocer ciertos datos personales de los pacientes. En este caso, se trata de informes clínicos redactados por los propios médicos. Se trata, por tanto, de lenguaje natural escrito en castellano.

Es cierto que la anonimización de datos en documentos médicos está incrementando el interés de los expertos, aunque los mayores avances obtenidos se han llevado a cabo en documentos escritos en inglés. Esta es otra de las motivaciones que llevan a la realización de este proyecto. Se utilizarán herramientas de análisis semántico y sintáctico, como *Spacy*, para probar cómo se comportan en textos en castellano, además del uso de diccionarios y expresiones regulares que permitirán el reconocimiento de los datos sensibles que aparecen en los informes.

1.2. Objetivos y alcance del proyecto

El objetivo de este proyecto es implementar un sistema que permita anonimizar documentos médicos en castellano escogiendo los atributos que se desean eliminar de los mismos. Para esto, se ha realizado un análisis de los documentos proporcionados por el *Instituto Aragonés de Ciencias de la Salud (IACS)*. Se desea conocer cuáles son los datos sensibles que se pueden encontrar en un informe médico, cuál es su nivel de sensibilidad y su nivel de importancia clínica.

Es por este motivo que la aplicación a implementar contará con dos tipos de selecciones ortogonales que permitirá al usuario escoger el conjunto de atributos que desea eliminar de sus documentos (información personal, información de contexto, información sobre su estilo de vida y/o información sobre eventos clínicos) y/o el nivel de anonimización que desee aplicar a cada conjunto de atributos (bajo, medio, alto o ninguno). La aplicación permitirá el análisis de ficheros de texto (en formato *".txt"*) o bien ficheros comprimidos (en formato *".zip"*), cuyo contenido sean ficheros *".txt"*.

Del mismo modo, la aplicación constará de una parte de minería de datos, donde se aplicará una técnica de evaluación denominada *Cross Validation* [2]. El usuario podrá cargar tanto el fichero originial como el anonimizado, obtenido en el apartado anterior, con el objetivo de comprobar que el proceso de anonimización ha funcionado

razonablemente bien y, que en ese proceso, no se pierde información clínica relevante. En este caso, el usuario podrá elegir el clasificador, el número de "folds" y la categoría o clase sobre la que aplicar la técnica de evaluación.

1.3. Organización del proyecto y contenido de la memoria

El desarrollo del proyecto se ha dividido en varias fases. La primera fue la detección y análisis de datos sensibles en documentos médicos. El siguiente paso fue el diseño del primer prototipo de la aplicación y su posterior implementación. En esta fase se encuentra la formación en las herramientas empleadas, como *Spacy*, utilizada para realizar el análisis sintáctico y semántico de las frases.

El siguiente paso fue el planteamiento de la parte de minería de datos, la cual se centra en la realización de una validación cruzada a partir de los documentos tanto anonimizados como sin anonimizar. Durante el desarrollo de cada una de las partes se han ido realizando pruebas y revisiones constantes con el objetivo de solventar errores.

La memoria está estructurada en las siguientes partes. En el Capítulo 2 se hace una introducción a la minería de textos médicos (escritos en castellano o no). Además, se narra el proceso de análisis de los datos clínicos proporcionados, su clasificación y su sensibilidad. En el Capítulo 3 se presenta el diseño y planteamiento de la aplicación junto con las tecnologías utilizadas para el proceso de detección de los atributos dentro de los textos escritos en lenguaje natural. Además, se habla sobre la arquitectura y el software de la aplicación. Más concretamente, se presentan los requisitos de la aplicación y su implementación. A continuación, en el Capítulo 4 se muestran la evaluación experimental realizada en cada una de las partes y las conclusiones obtenidas al respecto. En el Capítulo 5 se habla de los recursos y de las diferentes herramientas que se han utilizado en el desarrollo de este sistema. Finalmente, en el Capítulo 6 se habla de las conclusiones obtenidas, así como del trabajo futuro.

En la memoria también hay una serie de anexos que complementan la información de los capítulos anteriores. En el Anexo A se explica cómo se ha realizado la clasificación de los atributos y se muestran las distintas combinaciones de atributos que se pueden encontrar en los textos y que aumentan su sensibilidad cuando aparecen de manera conjunta, así como la sensibilidad conjunta de los atributos. En el Anexo

B se muestra el diseño de la aplicación; se incluye el prototipado principal, el mapa de navegación y la GUI de la aplicación. En el Anexo C se encuentran los manuales del prototipo, más concretamente el manual de instalación y el de usuario. Finalmente. en el Anexo D, se encuentran documentados los resultados detallados de las pruebas descritas en el Capítulo 4.

Capítulo 2

Estudio previo y análisis del proyecto

En este capítulo se va a hablar del estudio previo realizado a la implementación de la herramienta. Es de vital importancia conocer los proyectos y avances que existen en el ámbito de la anonimización de datos, más concretamente en documentos médicos escritos en castellano. El propósito de este capítulo es contextualizar el proyecto y resaltar aquellos análisis que se han realizado antes de implementar la herramienta.

En la Sección 2.1 se narra el estudio del estado del arte en minería de textos médicos. En la Sección 2.2 se concreta el estudio del estado del arte anterior respecto al castellano. En la Sección 2.3 se narra el estudio del estado del arte de anonimización de textos médicos en castellano. En la Sección 2.4 se explica cómo se han detectado los datos sensibles en los documentos médicos, así como los grupos en los que se han agrupado. En la Sección 2.5 se presenta la importancia de los atributos combinados. A continuación, en la Sección 2.6 se narra la necesidad de conocer la sensibilidad y la importancia clínica de los datos, tanto de manera independiente como de manera conjunta con otros atributos. Finalmente, en la Sección 2.7 se habla de la sensibilidad de los atributos.

2.1. Estado del arte de la minería de textos médicos

Las historias clínicas son una gran fuente de datos personales rica en información clínica. Estos documentos recogen información que podría conducir a la mejora de calidad de la asistencia sanitaria, a alcanzar acontecimientos históricos de investigación, reducciones de costes sanitarios, así como de la reducción de errores médicos. Sin embargo, el uso del lenguaje natural en ellos hace complicada la extracción de la información y su posterior procesamiento [3].

A pesar de ello, existen múltiples herramientas de procesamiento de lenguaje natural que ayudan a que esta tarea sea mucho más sencilla. En algunos estudios como [4], se señala que algunas herramientas y métodos de NLP son accesibles de forma gratuita y sencillos de usar. Sin embargo, estás herramientas se enfrentar a diversas complicaciones a la hora de aplicarlos al ámbito sanitario [5]. Al tratarse de textos escritos en lenguaje natural, se debe tener en cuenta la existencia de sinónimos, acrónimos, siglas, abreviaturas, etc., que pueden utilizarse para referirse a un mismo término. Esto ha llevado a la creación de sistemas de organización del conocimiento y ontologías [6].

A lo largo de los años, se han desarrollado múltiples estudios que han ayudado a construir avances en este campo. Es el caso de una iniciativa de investigación que se llevó a cabo en la Clínica Venderbit de Nueva York [7]. El objetivo era determinar si un programa de procesamiento del lenguaje natural podía codificar automáticamente la información sobre el estado funcional de acuerdo con los requisitos de la Clasificación Internacional del Funcionamiento, la Discapacidad y la Salud (CIF). De hecho, los investigadores ampliaron el proyecto existente para codificar los resúmenes de alta de rehabilitación. Además, posteriormente, un estudio realizado por la Universidad de Utah utilizó una versión modificada de este proyecto con el objetivo de extraer datos relacionados con los acontecimientos adversos relacionados con la colocación de catéteres venosos centrales. También es el caso de [8], donde se desarrolla y valida una escala de ictus prehospitalaria para predecir la oclusión arterial grande.

Por otro lado, se han realizado estudios sobre historias clínicas como es el caso de [9]. Se trata de unas herramientas que han analizado las historias clínicas del servicio de urgencias de un hospital con el fin de obtener conclusiones relacionadas con los servicios que se ofrecían. Se descubrió que había quejas similares que se trataban de manera diferente según el médico de guardia.

Se han producido múltiples avances en el ámbito sanitario respecto al procesamiento del lenguaje natural y abarcando otras especialidades como *Oncología* [10], donde a partir del crecimiento de la aplicación de NLP, se va creando una estructura para avanzar en el tratamiento del cáncer. En *Radiografía* [11] la mineríia de textos permite la automatización de diversas tareas en dicho campo. Finalmente, en *Geriatría* [12], donde se utilizó un modelo de CRF *(Conditional Random Fields)* con el objetivo de identificar los diferentes síndromes geriátricos a partir del texto obtenido de los

diferentes pacientes.

Además, también se han realizado estudios sobre la privacidad y confidencialidad de los datos clínicos en datos médicos textuales. Respecto a algunos estudios realizados, se ha tratado el reconocimiento de entidades médicas utilizando redes neuronales profundas en comparación con las técnicas actuales de vanguardia [13]. Además, estos estudios han comprobado la eficacia del marco propuesto y han destacado que mejora la recuperación de los datos, la precisión y la utilidad de los datos de documentos anonimizados hasta en un 13,79 %.

Además, no solo hay estudios que aplican técnicas sino evaluaciones sobre proyectos realizados. Este es el caso de [14], donde se evalúa si la traducción automática ha logrado una calidad lo suficientemente alta como para traducir títulos de PubMed para pacientes.

Todas estas investigaciones muestran los grandes avances que supone el procesamiento del lenguaje natural en el ámbito sanitario, así como las distintas aplicaciones que permite desarrollar en el sector.

2.2. Estado del arte de la minería de textos médicos en castellano

Como se ha comentado en la Sección 2.1, las herramientas de procesamiento del lenguaje natural se enfrentan a múlitples adversidades que complican su correcto funcionamiento. Anteriormente se ha hablado de sinónimos, acrónimos, siglas, etc. Sin embargo, por encima de ello está el lenguaje en el que está escrito el texto que se va a procesar. A día de hoy, existen proyectos realizados que abarcan el procesamiento del lenguaje natural en texto escritos en otro idioma que no es el inglés como es el caso de [15]. Es por esto, que la dificultad de este proyecto aumenta al tratarse de textos en castellano.

Se han desarrollado aplicaciones enfocadas al aprendizaje automático para el reconocimiento semántico y la normalización de términos clínicos, cuyos resultados han sido muy positivos. Es el caso de [16], donde se crean unos vectores de características para identificar términos equivalentes en los diferentes textos que se procesan. Otro ejemplo sería el de [17], donde se presenta una herramienta que trata de determinar las proposiciones negativas en textos clínicos en español.

Además, al tratarse de textos de ámbito clínico, muchas de las nomenclaturas o abreviaturas que se emplean en inglés no se corresponden con las que se utilizan en castellano. Hay abierta una enorme línea de investigación en el campo de procesamiento del lenguaje natural para textos en castellano. Un proyecto que abarca estos contenidos es el descrito en [18], el cual resume la configuración, los datos y los resultados sobre anonimización de documentos médicos en español de la pista MEDDOCAN *(Medical Document Anonymization)*.

El proyecto *NEAT-AMBIENCE* aborda la gestión de datos para ayudar a los ciudadanos en su vida cotidiana. La personalización de los datos adquiere una relevante importancia para proporcionar a cada ciudadano los datos que realmente necesita en cada momento. El proyecto que se desarrolla en este Trabajo Final de Grado forma parte del marco del proyecto de investigación de *NEAT-AMBIENCE* dentro del caso de uso de salud.

Además, existen proyectos dedicados a la minería de textos en castellano, [19]. El objetivo principal de este trabajo era el diseño e implementación de una aplicación que fuese capaz de devolver extractos de guías clínicas con la información que mejor resuelva una consulta concreta realizada por un médico. Además, la aplicación etiqueta de manera automática los términos médicos, lo cual se realiza en un tiempo mucho más reducido que si se realizase de forma manual.

2.3. Estado del arte de anonimización de textos médicos en castellano

La anonimización de textos médicos se está viendo impulsada en los últimos años debido al gran impacto que supone manejar documentos clínicos anonimizados. Como se ha comentado en las secciones anteriores, hay mucho desarrollo y aplicación en este sector, aunque una de las dificultades a las que se enfrenta este proceso es al tratamiento de textos en castellano.

Si se hace especial énfasis en proyectos cuyo objetivo es la anonimización de datos en documentos médicos en textos en castellano, se puede destacar [20]. Este forma parte del proyecto i2b2 *(Informatics for Integrating Biology to the Bedside)*, cuyos autores organizaron un desafío de procesamiento del lenguaje natural sobre

la eliminación automática de información de salud privada del alta médica. Otro proyecto a destacar es [21], donde se aborda la tarea de detectar y clasificar información sanitaria protegida a partir de datos españoles como un problema de etiquetado de secuencias y se investigan diferentes métodos de incrustación en una red neuronal.

Al margen del concurso MEDDOCAN, existen otros proyectos basados en anonimización de textos médicos como [22], donde se utilizan herramientas de procesamiento de lenguaje natural proporcionadas por el marco MEDTAG: un léxico semántico especializado en medicina y un conjunto de herramientas para el etiquetado de sentido de la palabra y morfosintáctico. Otro caso es el de [23], donde se realizó una búsqueda bibliográfica sistemática utilizando palabras clave de deidentificación, anonimización, depuración de datos y depuración de texto.

Por otro lado, existen múltiples proyectos que aplican la anonimización de datos mediante aprendizaje automático. Este es el caso de [24], que presenta un enfoque iterativo de reconocimiento de entidad nombrada basado en aprendizaje automático diseñado para su uso en documentos semiestructurados como registros de alta. También es el caso de [25], que presenta un sistema de anonimización automatizado para informes clínicos escritos en español. Se evalúan y comparan tres métodos diferentes. El primer método está basado en reglas, el segundo utiliza aprendizaje automático y el tercero es un método híbrido entre los dos primeros.

Estos son solo uno pocos ejemplos de la gran cantidad de proyectos e investigaciones que se están llevando a cabo en el ámbito clínico. El objetivo de este proyecto es realizar un proceso de anonimización de datos en documentos médicos escritos en español a partir del procesamiento del lenguaje natural que mantenga información relevante, pero protegiendo la privacidad del paciente. Además, será el usuario el que decida qué datos se eliminan en cada momento.

2.4. Análisis de los datos clínicos proporcionados

En primer lugar, se tuvo que realizar un análisis de los datos clínicos proporcionados para saber qué datos resultaban de interés para su posterior anonimización. En este caso, al tratarse de informes clínicos, se contaba con la aparición de atributos como el nombre, los apellidos, el domicilio, la edad y la fecha de nacimiento del paciente. Sin embargo, hay muchos otros datos que pueden resultar sensibles y proporcionar

información personal sobre los pacientes.

Tras realizar un intensivo análisis, se detectaron veinticuatro atributos sensibles que proporcionaban, en mayor o menor medida, información sobre el paciente. Estos atributos se pueden ver en la Tabla 2.1.

Atributo
Nombre
Apellidos
DNI
Pasaporte
Número Seguridad Social
Número de colegiado
Teléfono
Correo electrónico
Dirección
Género
Etnia
Edad
Lugar
País
Ciudad
Región
Hospital y/o centro de salud
Fechas de nacimiento y/o fallecimientos
Fechas de eventos clínicos
Familiar
Trabajo
Deporte
Historia personal
Hábitos

Tabla 2.1: Atributos encontrados en los documentos médicos

Dado que era un número considerablemente alto, se decidió dividirlos en cuatro grupos identificativos. Estos se muestran en la Tabla 2.2:

– El primer grupo recoge aquellos atributos que representan la información personal de los pacientes. En este grupo se recoge el nombre, los apellidos, el DNI, el pasaporte, la etnia, el número de teléfono, la dirección del domicilio, el correo electrónico, la fecha de nacimiento, el número de la seguridad social, el género y el número de colegiado, en el caso de que se trate de un médico. Este grupo se ha denominado *información personal*.

– En el segundo grupo se recogen aquellos datos que no identifican al paciente de manera directa, a diferencia de los atributos del grupo anterior, sino que se añade información que permite relacionar al usuario con su entorno. Este es el caso de información sobre familiares, su trabajo, ciudades, lugares, municipios o incluso el país al que pertenecen. Este grupo se ha denominado *información de contexto*.

– Respecto al tercer grupo, se recogen aquellos datos que informan sobre el estilo de vida que tiene el paciente. En este caso, se incluyen los deportes y los hábitos que practica. Dentro de los hábitos considerados están el alcohol, el tabaco y el trabajo sedentario. Este grupo se ha denominado *información de estilo de vida*.

– Finalmente, en el cuarto y último grupo se incluye la información clínica del paciente. Es decir, fechas de ingreso, fechas de alta, citaciones, hospitalizaciones, urgencias, hospitales y centros médicos donde han sido atendidos. Este grupo se ha denominado *información clínica*.

Grupo	**Atributos**
Información personal	Nombre, apellidos, edad, etnia, dni, teléfono, domicilio, correo electrónico, fecha de nacimiento, fecha de fallecimiento, número de la Seguridad Social, número de colegiado y género.
Información de contexto	Familiares, trabajo, lugares, países, ciudades y municipios.
Información sobre estilo de vida	Deportes y hábitos.
Información sobre eventos clínicos	Hospitalización, fecha de ingreso, fecha de urgencias y fecha de alta.

Tabla 2.2: Clasificación en grupos de los datos sensibles encontrados

La idea de este proyecto es poder eliminar de un conjunto de documentos toda la información identificativa y la personal que no sea necesaria para el caso de uso concreto en el que se van a utilizar los informes clínicos, y esa decisión depende del caso concreto. Por esta razón, en ciertas situaciones puede ser interesante eliminar, por ejemplo, información de contexto, y otras veces será necesario preservarla, y habrá que garantizar la privacidad de los pacientes en base a otros criterios y mecanismos.

2.5. Clasificación de los datos sensibles

Tras haber identificado los distintos datos sensibles en los documentos clínicos, se realizó una categorización de los mismos para poder agruparlos según la sensibilidad

o la importancia clínica que posean. Para ello, se les otorgó un nivel de sensibilidad y un nivel de importancia clínica. Esta clasificación se puede ver en la Tabla A.1. Esta clasificación se verá reflejada en la aplicación, ya que el usuario podrá elegir el nivel del anonimización que desea aplicar a los documentos. Los datos de "mayor sensibilidad" se relacionan con el "menor nivel de anonimización", ya que los datos de "mayor sensibilidad" son los que más comprometen la identidad y privacidad de los pacientes, y, por tanto, los que habrá que eliminar incluso cuando elijamos el menor nivel de anonimización. Del mismo modo, los datos "menos sensibles" solo serán eliminados si se selecciona el "mayor nivel de anonimización", que será el que más información relevante elimine. Los datos con "mayor sensibilidad" son recogidos por todos los niveles de anonimización, mientras que los de "menor sensibilidad" son recogidos únicamente por el "mayor nivel de anonimización".

La sensibilidad se reflejará en el nivel de anonimización que debe recibir el texto. En este caso, la sensibilidad puede ser alta, media o baja. Será alta cuando haya que anonimizar el texto siempre, media cuando el texto puede necesitar ser anonimizado en función del contexto, pero no siempre; y baja cuando no parece problemático mantener el texto.

En el caso de la utilidad, hace referencia a cómo de útiles son los datos desde el punto de vista médico. Por eso, su valor será alto cuando sea relevante desde la perspectiva médica, media cuando sea potencialmente relevante, parcialmente relevante cuando es relevante parte del texto, pero no todo; y baja cuando no parezca relevante.

2.6. Análisis de datos combinados

Dado que los datos que se utilizan para la realización de este proyecto son redactados por los propios médicos, es importante considerar la situación de que aparezcan varios atributos en un mismo informe. Es de especial consideración que dichos atributos tengan sensibilidad media o baja, ya que de manera conjunta puedan aumentar la sensibilidad global a alta o media, respectivamente. Es por ello, que se ha realizado un análisis de datos combinados.

Este análisis no considera aquellos atributos que, de manera independiente, ya poseen una sensibilidad alta. Se han combinado aquellos atributos de sensibilidad media

o baja para determinar si de manera conjunta aumentan la sensibilidad. Esto se puede ver reflejado en la Tabla A.2 que se encuentra en el Anexo A.2.

2.7. Sensibilidad de atributos

En la Sección 2.4 se ha explicado que a cada atributo se le ha otorgado un nivel de sensibilidad. Esto se realiza con el fin de poder clasificar los atributos en distintos niveles de anonimización. En el caso del nivel bajo se recogerán todos aquellos atributos que presenten una sensibilidad alta, ya que deben ser eliminados siempre. Respecto al nivel medio, se recogen aquellos atributos que pertenecen al grupo de anonimización bajo y todos los que tengan sensibilidad media. Finalmente, en el nivel alto, se recogerán todos los atributos que se han nombrado en la Sección 2.4.

A continuación, se presenta la Tabla 2.3, la cual muestra los distintos atributos que se han identificado en la colección de datos proporcionada, clasificados según la sensibilidad que tienen.

Patrón	**Sensibilidad**
Nombre y apellidos	Alta
DNI/Pasaporte	Alta
Número SS	Baja
Número de colegiado	Baja
Teléfono	Alta
Correo electrónico	Alta
Dirección	Alta
Género	Media
Etnia	Alta
Edad	Media
Lugar	Media
País	Media
Ciudad	Media
Región	Media
Hospital y/o centro de salud	Baja
Fechas de nacimiento y/o fallecimientos	Alta
Fechas de eventos clínicos	Baja
Familiar	Media
Trabajo	Media
Deporte	Media
Historia personal	Media
Hábitos	Baja

Tabla 2.3: Clasificación de atributos por sensibilidad

Se ha realizado esta clasificación por atributos para poder realizar el apartado de niveles de anonimización de la aplicación. Es decir, en la aplicación, el usuario puede determinar el tipo de anonimización que desea aplicar. En este caso, el usuario puede elegir entre cuatro opciones: "Don't apply", "Low", "Medium" y "High".

La opción "Don't apply" permite forzar al sistema para que no aplique ningún nivel de anonimización a los datos, aplicando únicamente anonimización sobre el conjunto de datos seleccionado en el apartado anterior. Es decir, al seleccionar esta opción sólo se aplicará anonimización sobre los datos personales, de contexto, de estilo de vida o eventos clínicos que el usuario haya seleccionado. En el caso de que seleccione la opción "Low", se eliminarán aquellos atributos que estén categorizados en la Tabla 2.3 con sensibilidad "Alta". En el caso de que seleccione la opción "Medium", se eliminarán aquellos atributos que en la Tabla 2.3 tengan sensibilidad "Alta" y/o "Media". Finalmente, si el usuario selecciona la opción "High", se eliminarán todos los atributos, incluidos los que aparecen en la Tabla 2.3 con sensibilidad "Baja".

Por otro lado, como se ha explicado en la Sección 2.5, hay atributos que aumentan su sensibilidad cuando aparecen de manera conjunta con otros atributos. Por este motivo, se ha establecido una sensibilidad conjunta a cada par de atributos combinados. Esta sensibilidad se ha tenido en cuenta en los grupos de anonimización anteriormente descritos.

En la Tabla A.3 se presenta el nivel de sensibilidad que muestran los atributos de manera conjunta. Esta tabla se encuentra en el Anexo A.3 de este documento.

Capítulo 3

Diseño, planteamiento, arquitectura y software de la aplicación

En este capítulo se va a mostrar el diseño realizado de la aplicación, así como su prototipado inicial. Se verá cómo se planteó inicialmente y cómo ha sido su implementación final. En la Sección 3.1 se explica el diseño y el prototipado inicial de la aplicación. En la Sección 3.2 se narra el procedimiento seguido para la detección de las palabras a anonimizar, así como las técnicas más utilizadas. En la Sección 3.3 se enumeran los requisitos que tiene la aplicación. En la Sección 3.4 se explican las técnicas de clasificación utilizadas en la parte de minería de datos. En la Sección 3.5 se explica cómo se ha realizado el procesamiento de ficheros de texto con WEKA. Por otro lado, en la Sección 3.6 se habla de las distintas versiones que se han realizado de la aplicación.

3.1. Diseño y prototipado inicial

Para el desarrollo de este proyecto se ha planteado el diseño e implementación de una aplicación de escritorio desarrollada en el lenguaje de programación *Java*. La aplicación se ha diseñado en varios idiomas: inglés y español, y consta de varias pestañas. La primera de ellas se denomina *Data* y permite al usuario configurar el proceso de anonimización que se va a llevar a cabo. Esta pestaña permitirá al usuario seleccionar el archivo o la carpeta de archivos que desea someter al proceso de anonimización.

Del mismo modo, el usuario podrá elegir cómo quiere aplicar dicha anonimización. Como se ha explicado en la Sección 2.4, los datos han sido agrupados en cuatro grupos diferentes: *información personal, información de contexto, información sobre estilo de vida e información sobre eventos clínicos*. El usuario podrá seleccionar una o varias opciones según la información que desee eliminar de sus documentos.

Por otro lado, como se ha explicado en la Sección 2.5, los datos han sido clasificados por nivel de sensibilidad. En este caso, los atributos han quedado agrupados en tres grupos diferenciados según si su sensibilidad es alta, media o baja. El usuario podrá elegir el nivel de sensibilidad que desea aplicar al proceso de anonimización de sus datos. En este caso, el usuario tiene la posibilidad de no aplicar ningún nivel de sensibilidad. Estas opciones son ortogonales a las explicadas anteriormente. Por tanto, el usuario puede aplicar anonimización a uno o varios grupos de atributos y añadir o no sensibilidad al proceso. Una vez que el usuario haya seleccionado las opciones que desee y haya finalizado el proceso de anonimización, podrá ver el archivo o la carpeta de archivos anonimizada.

La siguiente ventana se ha denominado *Mining* y se corresponde con la parte de minería de datos. Se ha implementado un interfaz para ejecutar distintos algoritmos de clasificación de textos mediante validación cruzada, con el objetivo de comparar documentos anonimizados y sin anonimizar con distintas características, y así evaluar las métricas de rendimiento del algoritmo de anonimización, tanto en la eliminación de datos identificativos de pacientes, como en la no eliminación de información clínica relevante. Esta funcionalidad se ha utilizado con este fin en la segunda parte del presente proyecto.

En esta pestaña el usuario podrá cargar ambos ficheros, seleccionar el tipo de clasificador a aplicar, el número de "folds" y la clase o categoría sobre la que aplicar la validación cruzada. Como en todas las aplicaciones realizadas, se ha realizado un prototipo inicial que se encuentra descrito en el Anexo B.

3.2. Detección de atributos en textos de lenguaje natural

Una vez se han detectado los distintos atributos que hay que eliminar en los textos, se deben localizar dentro de los textos que el usuario quiere anonimizar. Para ello, se han aplicado diferentes técnicas.

En primer lugar, los textos serán analizados línea por línea y se eliminarán aquellas palabras denominadas *empty words*, es decir, aquellas palabras o morfemas que no tienen significado léxico y que funcionan como un vínculo o marcador gramatical, más

que como un contenido.

En este caso, se ha utilizado la librería de procesamiento de lenguaje natural denominada *Spacy* https://spacy.io/. Esta librería permite realizar análisis semánticos en oraciones. Sin embargo, como esta librería está desarrollada, principalmente, para el lenguaje inglés y los textos con los que se ha tratado en este proyecto están escritos en castellano, se ha tenido que realizar un análisis sintáctico de las oraciones para asegurar que ambos análisis se estaban realizando correctamente. Además, dado que en el lenguaje natural se pueden producir ciertas ambigüedades, este análisis sintáctico ha ayudado a solventarlas. Un ejemplo sería el caso de que hay ciudades o municipios españoles que tienen el mismo nombre que personas.

Respecto a las técnicas utilizadas, se pueden destacar las expresiones regulares, los diccionarios de palabras y ficheros *.csv*, cuyo contenido ha sido comparado con cada línea del texto proporcionado por el usuario. Respecto a los ficheros *.csv*, guardan nombres y apellidos de personas que residen en Aragón, nombres y abreviaturas de hospitales y centros médicos pertenecientes a la comunidad de Aragón, así como países, familiares o municipios que pueden resultar de gran interés en los textos a procesar. Por otro lado, los diccionarios contienen información relacionada con bebidas alcohólicas, gentilicios y etnias, trabajos y profesiones, barrios de Zaragoza y deportes. Además, se consta de un diccionario que contiene términos o abreviaturas utilizadas en el ámbito clínico que no deben ser eliminados.

El proceso de anonimización de datos ha consistido en la detección, comparación y eliminación de atributos. En primer lugar, se debe destacar que el texto se procesa línea a línea, realizando un preprocesando de texto. La parte de detección comienza con las expresiones regulares seguido de la búsqueda en diccionarios y finalizando con la búsqueda en los ficheros *.csv*. Si alguno de los términos que aparecen en la línea a procesar coincide con alguno de los que constan en las expresiones regulares, ficheros o diccionarios, son eliminados de la oración.

En el caso de nombres y apellidos de los pacientes, se ha optado por realizar un análisis un poco más concreto, ya que se han detectado ciertas dificultades en su procesamiento. Cuando se detecta un nombre y/o apellidos, se compara con el contenido del fichero *.csv*. Si coinciden los resultados, se pasa a realizar un análisis sintáctico de la oración con el objetivo de garantizar que se trata de personas y no de lugares o enfermedades. Esto se ha realizado por situaciones concretas que se dan

cuando una persona se apellida del mismo modo que se llama un país o un municipio o, incluso, que una enfermedad. Por ejemplo, aquellas personas que se apellidan *Cáncer* o aquellas personas que se llaman *Borja*, como el municipio aragonés. Las líneas anonimizadas son escritas en un nuevo fichero como resultado del proceso.

En la Figura B.20 se muestra el diagrama de clases que representa el proceso de anonimización de los documentos.

3.3. Análisis de requerimientos de la aplicación

La aplicación que se ha desarrollado en este proyecto tenía el propósito de permitir al usuario poder anonimizar uno o un conjunto de ficheros según un grado de anonimización o respecto a un conjunto de atributos. Además, consta de un parte de minería de datos, en la que se aplica una técnica de evaluación denominada *Cross Validation*, donde se evalúan tanto los ficheros originales como los anonimizados con el fin de comparar sus resultados.

En las Figuras 3.1 y 3.2 se muestran los requisitos funcionales de la parte de anonimización de documentos y de minería de datos, respectivamente. En la Figura 3.3 se muestran los requisitos no funcionales de la aplicación.

Requisitos	**Descripción**
RF1	Importar el archivo en formato".txt" o la carpeta en formato ".zip" que desea anonimizar.
RF2	Seleccionar uno o varios conjuntos de atributos sobre los que aplicar el proceso de anonimización: *"Personal information", "Context information", "Style life" y/o "Clinic events".*
RF3	El usuario podrá visualizar una tabla informativa en la que se presenten los distintos atributos que se recoge en cada uno de los grupos nombrados anteriormente.
RF4	Seleccionar el nivel de anonimización a aplicar en el proceso: *"Don't apply", "Low", "Medium" o "High".*
RF5	Descargar el fichero o la carpeta ".zip" con los documentos anonimizados.

Tabla 3.1: Requisitos funcionales de la parte de anonimización

Requisitos	Descripción
RF1	Importar ficheros ".txt" sobre los que aplicar la evaluación.
RF2	Seleccionar el clasificador a utilizar en la validación cruzada.
RF3	Seleccionar el número de "folds" a aplicar en la validación cruzada.
RF4	Seleccionar la clase o categoría.
RF5	Visualizar los resultados obtenidos de la evaluación.

Tabla 3.2: Requisitos funcionales de la parte de minería de datos

Requisitos	Descripción
RNF1	El usuario tendrá la posibilidad de ejecutar la aplicación dentro de un ordenador con Sistema Operativo Windows o Linux.
RNF2	Se deberá ofrecer una interfaz amigable e intuitiva donde el usuario podrá ejecutar las acciones de la aplicación de manera sencilla.
RNF3	La aplicación deberá conectar con la API de Weka para fines de minería de datos.
RNF4	La aplicación deberá conectar con la herramienta Spacy para fines de procesamiento del lenguaje natural.
RNF5	La aplicación deberá ser un archivo ejecutable .jar que facilite a los usuarios el uso de la herramienta.
RNF6	La aplicación deberá ser fácil de utilizar.

Tabla 3.3: Requisitos no funcionales de la aplicación

Se considera oportuno reflejar un esquema general de cómo se va a plantear la aplicación final, ya que consta de dos partes diferenciadas. En la Figura 3.1, se muestra un diagrama de clases global de la aplicación. En el Anexo B.3 y en la Sección 3.5, se muestran con mayor detalle los diagramas de clases correspondientes a cada una de las partes implementadas.

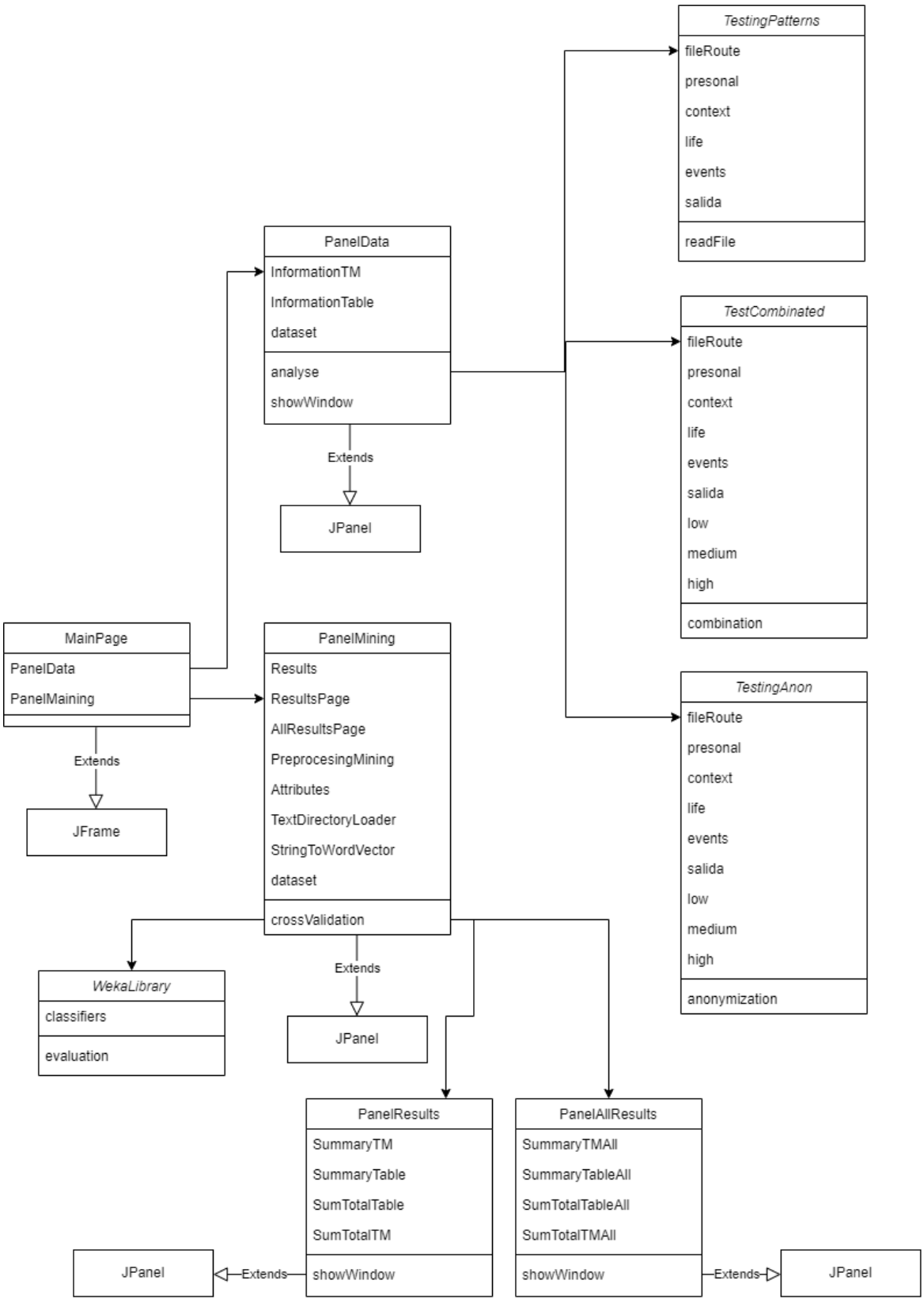


Figura 3.1: Diagrama de clases global de la aplicación

3.4. Técnicas de clasificación

Para este proyecto se habilitaron cuatro técnicas de clasificación. El objetivo es la comparación de los resultados obtenidos de clasificar un fichero sin anonimizar respecto del mismo tras haber pasado el proceso de anonimización, así como la comparación de conjuntos de informes de distinta índole, tanto anonimizados como sin anonimizar, para comprobar que el proceso de anonimización no destruye la capacidad del clasificador de distinguir entre estos documentos y, por tanto, que la información clínica relevante no ha sido eliminada.

Las técnicas de clasificación se aplicarán sobre la técnica de evaluación *"Validación cruzada" o "Cross Validation"* con un valor de "folds" determinado por el usuario. Las técnicas de clasificación que se ofrecen son accesibles desde la librería de *Weka*, la cual ha sido integrada durante el desarrollo de la aplicación. Las técnicas de clasificación se explican en mayor detalle en la Tabla 3.4.

Clasificador	**Descripción**
One Rule - OneR	Genera una regla para cada predictor en los datos, luego selecciona la regla con el error total más pequeño como su "regla única".
ZeroR	Predice sobre la clase o categoría principal. Es útil para determinar una base de performance sobre la cual medir los demas métodos de clasificación.
Naive Bayes	Se basa en el teorema de Bayes donde asume la independencia entre predictores. No posee parámetros estimativos iterativos complicados lo cual lo vuelve particularmente útil para datasets grandes.
Sequential Minimal Optimization - SMO	Implementa el algoritmo de optimización mínima secuencial de John Platt para entrenar un clasificador de vectores de soporte. Reemplaza globalmente todos los valores faltantes y transforma los atributos nominales en binarios. También normaliza todos los atributos por defecto.

Tabla 3.4: Algoritmos de clasificación considerados

Tras realizar la evaluación se obtienen una serie de métricas que se especifican en la Tabla 3.5.

Métrica	Descripción
Mean absolute error	Medida de errores entre observaciones pareadas que expresan el mismo fenómeno.
Root mean square error	Desviación estándar de los residuos (errores de predicción). Los residuos son una medida de qué tan lejos están los puntos de datos de la línea de regresión.
TP rate	Conjunto de datos clasificados como positivos siendo positivos realmente.
FP rate	Conjunto de datos clasificados como positivos cuando realmente son negativos.
TN rate	Conjunto de datos clasificados como negativos siendo negativos realmente.
FN rate	Conjunto de datos clasificados como negativos cuando realmente son positivos.
Precision	Predicciones de clase positivas que realmente pertenecen a la clase positiva.
Recall	Predicciones de clase positivas realizadas a partir de todos los ejemplos positivos del conjunto de datos.
F-measure	Proporciona una puntuación única que equilibra las preocupaciones de precision y recall en un solo número.

Tabla 3.5: Métricas de rendimiento calculadas

3.5. Procesamiento de ficheros de texto con WEKA

Se ha comentado en la Sección 3.4 que se va a utilizar una técnica de evaluación denominada *"Cross Validation"*. Esta técnica se va a integrar en la implementación mediante *Weka*. Normalmente, esta técnica se realiza a través de ficheros en formato *".arff"*, que es el formato de archivo de relación de atributos usado por Weka. Los archivos ARFF tienen formato de texto plano en ASCII y se diferencian en dos partes: la cabecera y los datos. Sin embargo, los datos que se manejan en este proyecto son ficheros de texto en formato *".txt"*, lo que supone una dificultad añadida.

Para solventar este problema se ha preprocesado el fichero con el fin de obtener un vector de palabras. En el preprocesado se han eliminado las denominadas *"empty words"*. Además, se ha hecho uso de la clase *StringToWordVector* de *Weka*, la cual realiza una transformación con la frecuencia de aparición de cada una de las palabras contenidas en el vector anterior. De esta manera, los atributos de cada instancia serán la frecuencia de cada una de las palabras que aparecen en el documento de texto.

Se puede ver el diagrama de clases de la parte de minería de datos en la Figura 3.2.

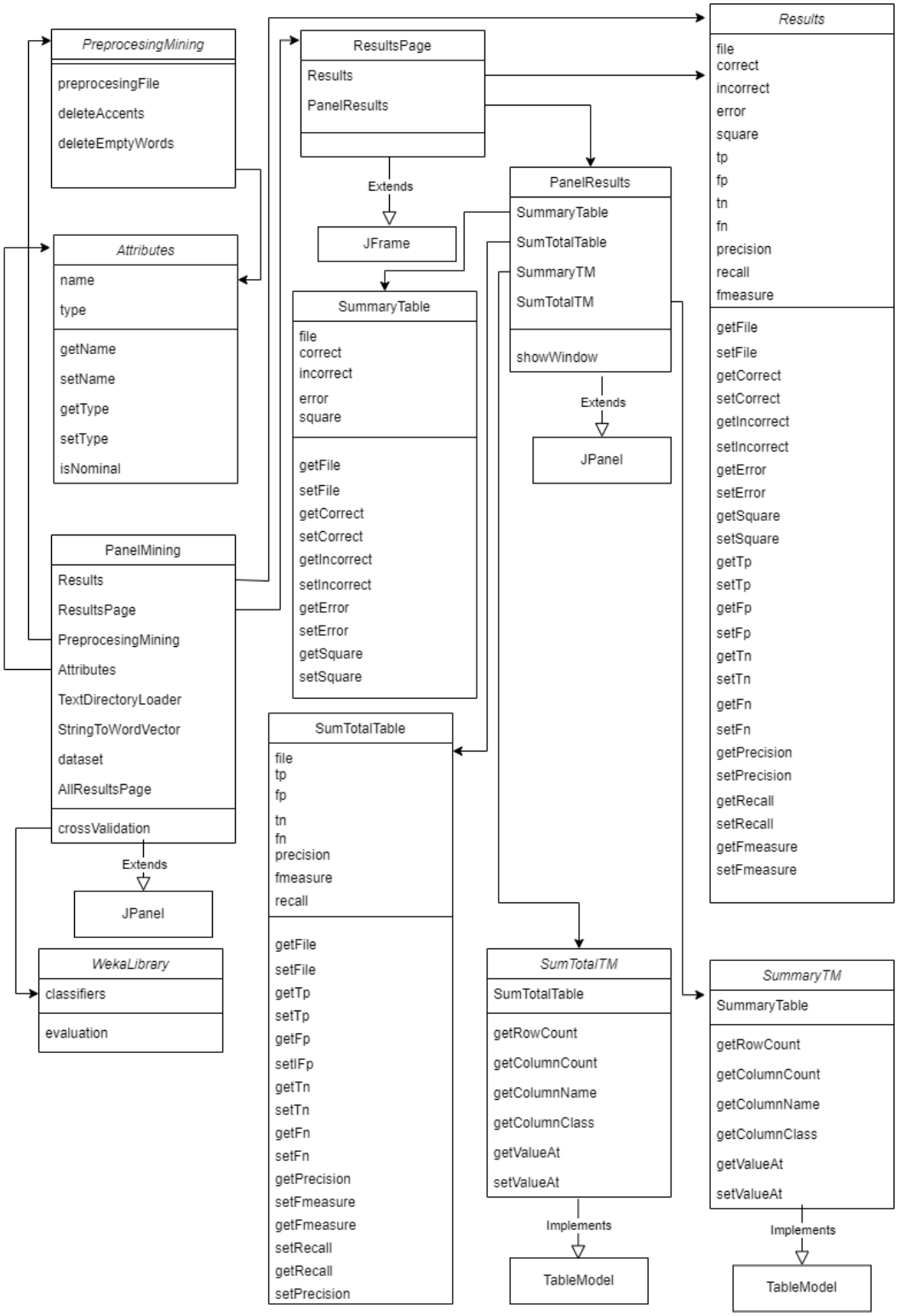


Figura 3.2: Diagrama de clases de la implementación de minería de datos

3.6. Versiones de la aplicación

Como se ha comentado a lo largo de todo este documento, en este proyecto se ha desarrollado una aplicación con interfaz de usuario, donde se configurarán las características del proceso de anonimización y el de evaluación de minería de datos. El diseño de esta aplicación se encuentra descrito en el Anexo B, donde se muestra el prototipado inicial de la aplicación (Anexo B.1) junto con su correspondiente mapa de navegación, (Anexo B.2).

Por otro lado, se ha realizado una segunda versión de la aplicación, la cual carece de interfaz de usuario. Esta versión se realizó para poder probar de forma sencilla la aplicación con lotes de documentos reales proporcionados por el *Instituto Aragonés de Ciencias de la Salud (IACS)*. Se trata de un fichero *".jar"* que se debe ejecutar a través de consola de comandos. Esta versión consta de un fichero de configuración donde se especifican las características del proceso de anonimización.

En la Figura 3.3, se muestra un ejemplo de fichero de configuración:

```
Write an X on those attributes you want to anonymize:

-Personal information:
-Context information:
-Style life information:
-Events information:

Write an X on the level of anonymization you want to apply: (You only can choose one of these four options. You must choose one.)
-Do not apply:
-Low level:
-Medium level:
-High level:

Write the path to the file which you want to get anonymized in the next line:

Write the path to the results folder in the next line:

Write the path to the collections folder in the next line:
```

Figura 3.3: Ejemplo de fichero de configuración

La aplicación sirve tanto para el Sistema Operativo Windows, como para Linux. En el Anexo C se encuentran los manuales de uso de dicha aplicación. Más concretamente, en el Anexo C.1 se encuentra el manual de configuración del entorno para la posterior ejecución y en el Anexo C.2 se encuentra el manual de uso de la aplicación.

Capítulo 4

Evaluación experimental

En este capítulo se explican las pruebas realizadas para comprobar el correcto funcionamiento de la herramienta. Principalmente se han realizado tres pruebas diferenciadas para comprobar la calidad del proceso de anonimización y la no pérdida de datos. Además, se detalla cómo se han realizado cada una de las pruebas y cómo se han obtenidos los datos con los que se ha realizado la evaluación experimental.

En la Sección 4.1 se comenta cómo se han obtenido los datos proporcionados y su procesado. En la Sección 4.2 se explican y documentan cada una de las pruebas realizadas en el proceso de anonimización. En la Sección 4.3 se documentan las pruebas de rendimiento realizadas. En la Sección 4.4 se narran y documentan las pruebas realizadas en la parte de minería de datos. Finalmente, en la Sección 4.5 se narran las conclusiones obtenidas sobre el proceso de pruebas realizado.

4.1. Datos proporcionados

Se han realizado una serie de pruebas para evaluar el comportamiento del sistema. La primera clasificación que se va a realizar en esta parte del proyecto consiste en diferenciar entre datos personales y no personales. Como bien se ha explicado a lo largo de este documento, los datos proporcionados por el *Instituto Aragonés de Ciencias de la Salud (IACS)*, son informes clínicos sintéticos que contienen información identificativa de pacientes. Es por esto que esos documentos serán los documentos personales. Este conjunto de datos conformaba un fichero en formato ".txt" de más de cuatro millones de líneas que recogen información sobre informes clínicos de pacientes que residen o reciben atención sanitaria en la Comunidad Autónoma de Aragón. Dado que se debe preservar la privacidad de los datos y que se trata de datos sensibles, para poder recibir los datos, el IACS tuvo que dividir los informes clínicos por líneas y mezclarlas todas ellas entre sí para poder cumplir este objetivo. Es justamente esa

la finalidad de este proyecto, la eliminación de la información sensible.

Dado que el conjunto de datos tiene una gran cantidad de información, se han seleccionado 3047 líneas del documento, creando así un subconjunto de datos que contiene datos personales de los pacientes. Las líneas que conforman este documento no han sido seleccionadas de manera aleatoria, ya que se necesitaba que constasen de ciertos datos personales para que se consideren documentos personales. Las líneas seleccionadas o bien contienen información personal de los pacientes como nombre, apellidos, edad, domicilio, DNI, género, historia personal, etc. o bien constan de eventos clínicos como nombres de hospitales y centros médicos, fechas de ingreso, fechas de alta, fechas de urgencias, etc.

Por otro lado, los documentos no personales son diferentes a los utilizados en este proyecto. El conjunto de datos proporcionado fue el utilizado en el Trabajo Final de Máster [19], cuyo autor es Carlos Sánchez Coronas. Este conjunto de datos ha sido proporcionado por los directores de este proyecto, Sergio Ilarri y Carlos Tellería, ya que también fueron directores de este Trabajo Final de Máster. Este conjunto de datos contiene recomendaciones de guías clínicas. Por tanto, su contenido no contiene información personal y se puede considerar como "no personal", aunque sí contiene información y terminología clínica.

Finalmente, para la realización de la clasificación de los documentos relacionados con ictus cerebrales u otras enfermedades, se han utilizado dos conjuntos de datos sintéticos proporcionados por el IACS, y se utilizan exclusivamente para esta evaluación. El primer conjunto de datos consta de información relacionada con ictus cerebrales, mientras que el segundo consta de información relacionada con otras enfermedades distintas al ictus cerebral.

4.2. Pruebas realizadas en el proceso de anonimización

En este caso, las pruebas han consistido en someter a cada uno de los ficheros a las diversas combinaciones de anonimización que la aplicación ofrece. De esta forma, se puede distinguir tres tipos de pruebas. La primera de ellas hace referencia a las distintas combinaciones de grupos de atributos que ofrece el sistema: "Personal information", "Context information", "Style life" y "Clinic events". Se hará referencia

a ellas como *Pruebas relacionadas con conjuntos de atributos* y se explican en la Sección 4.2.1. La segunda de ellas hace referencia a los niveles de anonimización: "Don't apply", "Low level", "Medium level" y "High level". Estas se denominan *Pruebas relacionadas con el nivel de anonimización* y se encuentran documentandas en el Sección 4.2.2. Finalmente, la tercera prueba consiste en la combinación de las anteriores y se demoninarán *Pruebas combinadas* y aparece en la Sección 4.2.3.

Las pruebas van a consistir en evaluar los valores de **True Positives**, **False positives** y **False negatives** de cada uno de los ficheros generados con cada una de las combinaciones que se lleven a cabo. Los **TP** hacen referencia a aquellos atributos que han sido eliminados y debían ser eliminados. Los **FP** son aquellos atributos que han sido eliminados y no deberían haberlo sido. Con esta métrica se podrá evaluar la pérdida de información. Finalmente, los **FN** son aquellos que se mantienen en el texto a pesar de que se deberían haber eliminado. Con esta métrica se podrá evaluar el riesgo de privacidad, pues cuanto más próximo sea al valor 0, mayor es la privacidad que se garantiza en el proceso.

Se calcularán métricas como el *Recall* y la calidad del algoritmo en su conjunto. Esta última se calculará mediante el cociente obtenido de dividir los **TP** entre los mal clasificados (**FP y FN**).

4.2.1. Pruebas relacionadas con conjuntos de atributos

Se van a realizar un total de 15 pruebas, cada una de las cuales representa una combinación distinta de los atributos que ofrece la aplicación. Los resultados obtenidos se muestran en el Anexo D.1.1.

Tras realizar las pruebas, se puede comentar que la mayoría de los datos que pertenecen a FP son datos que carecen de significado en el texto y que, aunque los elimina cuando no debería eliminarlos, no provocan que el texto pierda significado. En el caso de la métrica *Recall*, lo que sucede es que la mayoría de las pruebas dan resultados superiores al 60 %, el cual es un valor muy positivo para el proceso de anonimización.

Por lo que respecta a la calidad del algoritmo en su conjunto, todos los valores obtenidos son iguales o superiores a 1. Esto quiere decir que el algoritmo detecta más palabras correctamente de las que elimina o mantiene por error. Los resultados

obtenidos se muestran en la Figura 4.1.

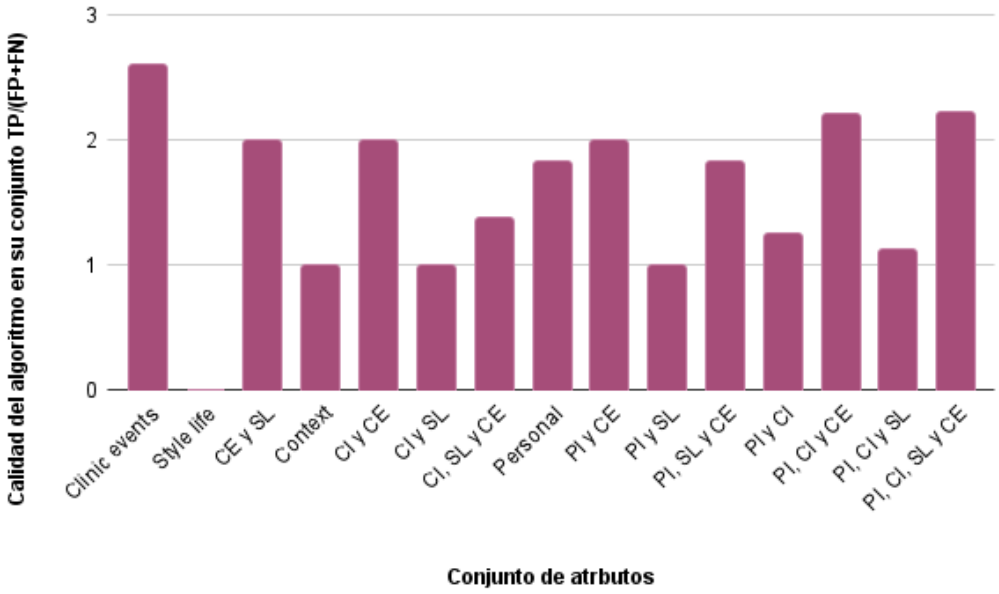


Figura 4.1: Gráfica con los diferentes resultados de calidad del algoritmo en pruebas relaciondas con conjuntos de atributos

Como conclusión general, se puede decir que el proceso de pruebas realizado es iterativo y que se toman como referencia para mejorar la aplicación final. De esta forma, se puede modificar el código para intentar disminuir la tasa de FN. Respecto a los valores obtenidos de FP, se destaca que son palabras vacías, es decir, que carecen de un significado muy relevante para los textos que se están tratando y que, aunque no deberían eliminarse, no tienen un efecto negativo sobre el contenido de los textos finales.

4.2.2. Pruebas relacionadas con el nivel de anonimización

En este caso, se van a realizar únicamente tres pruebas, cada una correspondiente a un nivel de anonimización: *"Low level", "Medium level" y "High level"*. Dado que la función de la opción *"Don't apply"* es no aplicar nivel de anonimización, no se considera oportuno en este tipo de pruebas. Los resultados obtenidos se muestran en el Anexo D.1.2.

El modo *Low level* elimina todos los datos sensibles que tiene que eliminar y no deja ninguno en el texto final. Esto justifica que el valor de la métrica *Recall* sea 1. Sucede lo mismo en el caso del nivel *Medium* y en el nivel *High*. Sin embargo, *el valor*

de la calidad del algoritmo en su conjunto no es superior a 1 en el caso del nivel *Low*, lo que implica que elimina o mantiene más información de la que debería en mayor proporción a la que detecta de manera correcta. Como se trata de pruebas iterativas, ha servido para modificar el código y reducir aquellos valores que aparecen como **FP** o **FN**.

Como conclusión general, se puede destacar que todos los niveles de anonimización detectan la información sensible que tienen que eliminar y la eliminan. Sin embargo, en todas las pruebas realizadas el número de FP no es muy pequeño. Es cierto, que los niveles de anonimización eliminan ciertas palabras que no deberían eliminar, pero estas son irrelevantes para el contenido del documento.

4.2.3. Pruebas combinadas

En este caso, las pruebas que se han realizado implican la combinación de las pruebas realizadas en la Sección 4.2.1 con las de la Sección 4.2.2. Como bien se ha comentado en la sección anterior, no se van a realizar pruebas con la opción *"Don't apply"*, ya que son las realizadas en la Sección *6.2.1.* Además, tampoco se consideran relevantes las pruebas con la opción *"High level"*, ya que solo con esta opción ya elimina toda la información sensible. Añadir atributos a este nivel de anonimización no tendrá efectos en la anonimización.

Se han realizado 15 pruebas siguiendo las distintas combinaciones posibles añadiendo el nivel de anonimización *"Low level"*. Los resultados obtenidos muestran en el Anexo D.1.3.

En estas pruebas se ha podido ver que la mayoría de valores de FN es muy pequeño. De hecho, en ninguna de las 15 pruebas supera el valor 5, en un intervalo entre 0 y 70. Además, dado que en las pruebas relacionadas con el nivel de anonimización este valor era prácticamente 0, se puede deducir que los fallos de estos FN se realizan en el apartado de los atributos a seleccionar. Es decir, en *"Personal information", "Context information", "Style life" y/o "Clinic events"*. Como bien se ha comentado en la Sección 4.2.2, las pruebas son iterativas y sirven para solventar errores.

Respecto al valor de los FP, sucede algo similar a lo que sucedía en las pruebas relacionadas con el nivel de anonimización. Los valores que son eliminados no afectan al contenido clínico del texto.

Finalmente, respecto a la calidad del algoritmo en su conjunto, los valores obtenidos en las pruebas suelen superar el valor de 1, lo que implica que se detectan y eliminan más atributos correctamente de los que se eliminan o mantienen por error. Sin embargo, hay algunos casos en los que el valor de esta métrica es inferior a 1. Estos casos son aquellos en los que no se marca la opción de *"Personal information"*. Estos resultados se pueden ver en la Figura 4.2. Dado que en estas pruebas estaba seleccionada la opción de *"Low level"*, se llega a la conclusión de que este nivel de anonimización elimina más información personal de la que debe. Como son pruebas iterativas, se han corregido aquellos errores que se han detectado para obtener una versión mejorada de la herramienta.

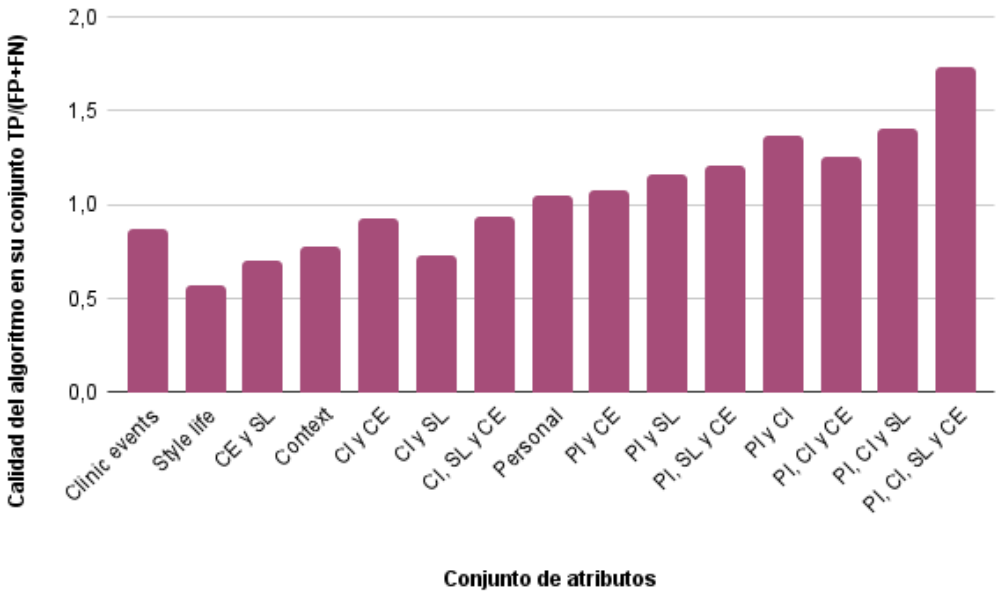


Figura 4.2: Gráfica con los diferentes resultados de calidad del algoritmo en pruebas combinadas

No se considera relevante realizar esta misma batería de pruebas para el nivel de anonimización *Medium level*, ya que es un nivel más restrictivo que el *Low level* y los resultados ya han sido buenos. El hecho de realizar las pruebas para este nivel no aportaría gran información a la evaluación realizada.

4.3. Pruebas de rendimiento

Estas pruebas pretenden reflejar cómo de costoso es el proceso de anonimización según las opciones que se escojan para el mismo. Las pruebas han sido realizadas

sobre un equipo con un procesador *intel core i7* con 16GB de RAM y 512GB SSD. Se debe destacar que es muy influyente el contenido de los documentos que se somenten a dicho proceso de anonimización. Si el fichero es largo, es posible que tarde más. Sin embargo, si el fichero no contiene mucha información sensible, aunque sea más largo, seguramente tarde menos que aquellos que sean cortos, pero contengan una gran cantidad de información sensible a eliminar.

Algunos de los resultados obtenidos más relevantes se muestran en la Figura 4.3. Estos resultados demuestran que el tiempo de ejecución es mayor cuanto mayor es el número de atributos que se desea eliminar. Es decir, cuando se selecciona una única opción de anonimización el tiempo de ejecución puede oscilar entre uno y cinco minutos. Esto depende también del conjunto de atributos que se deseen eliminar y la cantidad de datos sensibles que haya en el documento que se está procesando. En este caso, aquellos procesos de anonimización que constan de la opción de *"Personal information"* tardan más que el resto de las combinaciones que no lo involucran. Esto puede deberse a que esta opción considera muchos atributos a eliminar (trece atributos) a diferencia del resto que únicamente constan de dos, cuatro o seis atributos.

Las pruebas que más han tardado en ejecutarse son las que combinan nivel de anonimización con conjunto de atributos, ya que debe realizar más análisis que cuando solo se selecciona uno de ellos. La prueba más larga ha sido de 9 minutos.

4.4. Pruebas realizadas en minería de datos

En esta sección se van a explicar cada una de las pruebas que se van a realizar. Se puede diferenciar entre dos tipos de prueba. La primera de ellas va a consistir en evaluar la clasificación de los documentos en personales y no personales. Así, se podrá evaluar la calidad de la anonimización. Estas pruebas se muestran en la Sección 4.4.1.

Por otro lado, se realizará un segundo tipo de prueba que consistirá en la evaluación de la clasificación de los documentos en aquellos que están relacionados con ictus cerebrales y aquellos que no. Con esta segunda prueba se podrá evaluar la calidad de la no pérdida de información tras realizar la anonimización de los documentos. Estas pruebas se muestran en la Sección 4.4.2.

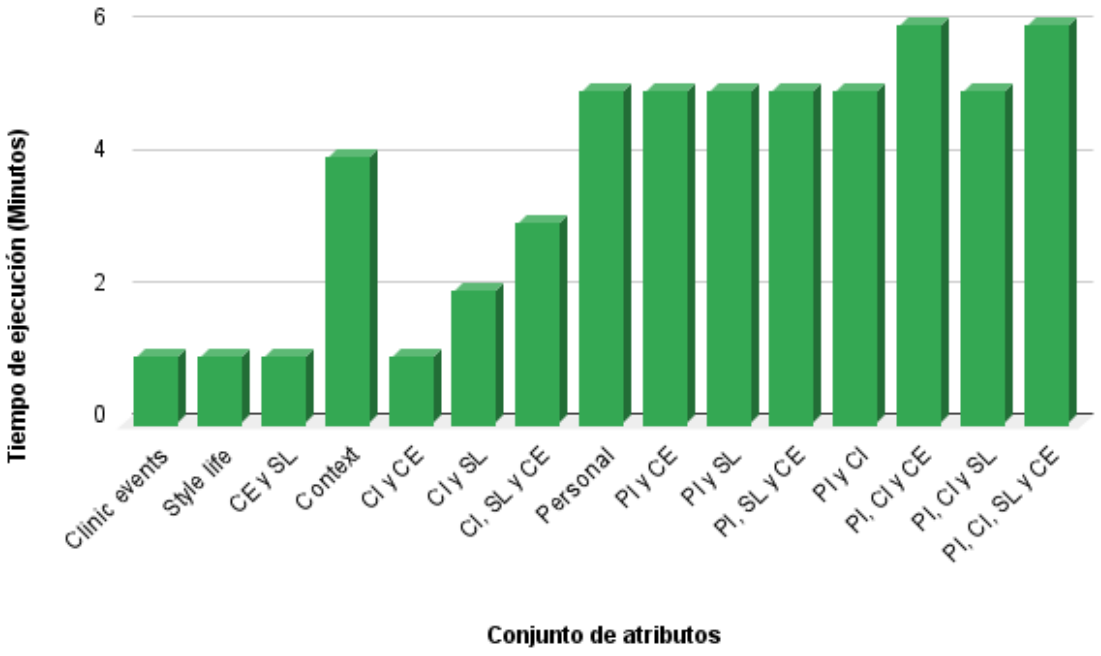


Figura 4.3: Gráfica de tiempos de ejecución obtenidos en las pruebas relacionadas con conjuntos de atributos

4.4.1. Clasificación entre documentos personales y no personales

Para la realización de esta prueba, se han tomado los informes médicos utilizados a lo largo de este proyecto y los correspondientes a guías clínicas [19]. Se ha realizado una primera evaluación con los documentos originales y, posteriormente, se ha anonimizado el conjunto de documentos personales respecto a información personal y eventos clínicos. Una vez se ha realizado esta anonimización, se ha realizado una segunda evaluación. El objetivo es comparar los resultados de ambas evaluaciones para comprobar la calidad de la anonimización. Es decir, lo que se pretende es ver si al aplicar la anonimización a los documentos personales estos pasan a ser indistinguibles de los documentos no personales para el clasificador, lo cual sería bueno, ya que el principal objetivo es eliminar información personal.

El proceso de minería de datos llevado a cabo ha consistido en lo siguiente. Se maneja un conjunto de documentos que contienen información personal de los pacientes y otro conjunto de datos que contiene guías clínicas, las cuales no se consideran documentos personales. En una carpeta del equipo se han creado dos carpetas: la carpeta *true*, donde se almacenan los documentos relacionados con datos personales; y la carpeta *false*, donde se almacenan los documentos relacionados con las

guías clínicas. Así se consigue realizar el etiquetado de los documentos de forma rápida.

A continuación, se procede a realizar la anonimización de los documentos que contienen datos personales. Una vez anonimizados, se organizarán en carpetas de la misma manera que los datos originales. En una carpeta, distinta a la anterior, se crearán las carpetas *true* y *false*. De esta manera, se obtendrán etiquetados los documentos para la evaluación de la clasificación con los datos anonimizados.

La interfaz de la aplicación permite al usuario cargar estos documentos. Para ello, el usuario debe cargar la carpeta padre en la que se encuentran las carpetas *true* y *false*. Realizará el mismo procedimiento para la carpeta que contiene los documentos anonimizados. La herramienta permite configurar la evaluación que se va a realizar mediante la técnica *Cross Validation*. El usuario seleccionará el clasificador, el número de *folds* y la clase que desea aplicar al proceso. En este caso, la clase a aplicar será el etiquetado que se ha realizado con las carpetas anteriores. *True* hace referencia a los documentos personales y *false* hace referencia a los datos no personales.

Este proceso se ha repetido cuatro veces, una por cada uno de los clasificadores que ofrece el sistema: *OneR, ZeroR, Naive Bayes y SMO*. Los resultados obtenidos se pueden ver en el Anexo D.

Tras realizar las pruebas, se ha obtenido que el clasificador *SMO* es el que mejor resultados ha obtenido respecto a los demás. Ha decrementado el número de *TP* y ha incrementado el número de *FP*. Es decir, ha cumplido con el objetivo de las pruebas: la evaluación realizada con los datos anonimizados ha clasificado un número inferior de documentos personales y un número más elevado de documentos no personales. Además, los resultados obtenidos de *Precision* y *Recall* son del 98 %, lo cual es muy beneficioso para el proceso de anonimización. Con estas pruebas se ha evaluado la calidad del proceso de anonimización.

4.4.2. Clasificación entre documentos relacionados con ictus cerebrales y otros documentos

En este caso, se va a realizar la clasificación de un dataset para clasificarlos como documentos relacionados con ictus cerebrales o como documentos relacionados con otros tipos de diagnósticos. Con esta prueba se pretende evaluar la calidad de no pérdida de información al realizar la anonimización de los documentos. Es

decir, se pretende ver si el clasificador se comporta igual a pesar de haber aplicado anonimización, lo que revelaría que no hay pérdida significativa de información útil desde la perspectiva de esa clasificación, al tiempo que se protege la información sensible.

La evaluación que se va a llevar a cabo consiste en realizar la evaluación de los ficheros proporcionados con los distintos clasificadores que la herramienta ofrece y obtener sus resultados. Del mismo modo, se realizará este proceso con los documentos anonimizados por la propia herramienta. La anonimización que se ha aplicado es *"Personal information" y "Clinic events"*. Cuando se obtengan ambos resultados, se podrán comparar para ver si la anonimización provoca pérdidas de información relevante. El objetivo de esta evaluación es que los resultados de la clasificación con documentos anonimizados y no anonimizados sean lo más similares posible.

Para realizar una correcta evaluación se debe crear una carpeta en la que se alojen los documentos relacionados con ictus cerebrales y otras enfermedades. Como se ha hecho en la Sección 4.4.2, se deben crear dos carpetas dentro de la carpeta padre. Una de ellas se denominará *true* y la otra *false*. En la primera carpeta se ubicarán los documentos relacionados con ictus cerebrales y en la segunda de ellas se encontrarán aquellos relacionados con otras enfermedades. Con esto se consigue realizar el etiquetado de los documentos, ya que el objetivo es realizar la clasificación de los mismos entre los que están relacionados con ictus y los que no. Luego, se realizará la anonimización de los datos, en esta ocasión, sobre todos los datos que se manejan (relacionados con ictus o no). A continuación, se realizará la misma organización de carpetas que con los datos originales para obtener el etiquetado de los datos. Finalmente, el usuario configurará la técnica de *Cross Validation* que se va a llevar a cabo. En este caso, se trata de una evaluación de 10 *"folds"* repetida cuatro veces, una por cada uno de los clasificadores que ofrece la herramienta. Los resultados obtenidos se pueden ver en el Anexo D.3.2.

Tras analizar los resultados, se debe destacar que el clasificador *ZeroR* ha clasificado los documentos exactamente igual con los datos anonimizados que con los datos originales. Sin embargo, los resultados obtenidos de *Precision* y *Recall* son bastante más bajos en comparación con los obtenidos con otros clasificadores. Es por esto que el clasificador que mejores resultados ha obtenido es *SMO*. Tanto con los datos originales como con los anonimizados ha obtenido un 99 % de *Precision* y *Recall*. Además, los resultados obtenidos con los datos originales varían únicamente en un

0.05 % respecto de los obtenidos con los anonimizados. Estos resultados determinan que el proceso de anonimización mantiene la calidad de no pérdida de información en este contexto.

4.5. Conclusiones obtenidas

La realización de esta aplicación no ha sido fácil, ya que el contenido de los textos viene escrito en castellano. En el caso de las pruebas realizadas con los documentos personales y no personales, las palabras que no son eliminadas y deberían eliminarse, se deben a errores de escritura, ya que los textos que se manejan para esta parte del proyecto son escritos por los propios profesionales sanitarios. Esto implica que se produzcan fallos ortográficos que limitan la herramienta.

Del mismo modo, las palabras que son eliminadas de más, pueden deberse a aquellas que ocupan distinto lugar, sintáctico o semántico, dependiendo del contexto de la información. Es decir, hay personas que se apellidan igual que nombres de municipios aragoneses o apellidos que se corresponden con enfermedades como, por ejemplo, la palabra "cáncer".

Es cierto que tras haber realizado las pruebas, se puede ver que el proceso de anonimización elimina la mayoría de las palabras sensibles que debe anonimizar, lo cual es una evaluación muy positiva sobre la herramienta. Dado que se trata de datos sensibles asociados a pacientes reales, es mejor que se elimine información de más a que quede mucha información sensible en los textos, ya que hay que garantizar la privacidad de las personas.

Además, me gustaría resaltar que la mayoría de las palabras que pertenecen al grupo de FP son palabras que no limitan la información clínica del paciente. Esto permite al personal sanitario leer el informe sin dificultades a pesar de que elimine alguna palabra de más.

Respecto a las pruebas realizadas con los documentos relacionados con los ictus cerebrales y otras enfermedades, las pruebas muestran unos resultados muy positivos sobre el proceso de anonimización, ya que la evaluación realizada con los datos originales es muy similar a la obtenida con los datos anonimizados. Se podría decir que, tras realizar las pruebas, la herramienta mantiene la calidad de la no pérdida de

datos, un aspecto muy positivo del sistema.

Finalmente, respecto a las pruebas de rendimiento, se puede ver que las pruebas que tardan más son aquellas que detectan más datos sensibles a eliminar. Es decir, la duración de ejecución de cada una de las pruebas dependerá de la longitud de los ficheros que se procesen y de la cantidad de datos sensibles que haya en los mismos. En el caso de los ficheros se han utilizado para la evaluación experimental tienen entre 168 y 314 palabras en total, aunque se debían eliminar entre 30 y 60 atributos.

Capítulo 5

Recursos y herramientas

En este capítulo se muestran las diferentes herramientas y recursos que se utilizaron para la realización de este proyecto de investigación. En la Sección 5.1 se nombran las herramientas utilizadas para el diseño de la aplicación. En la Sección 5.2 se enumeran los lenguajes de programación utilizados, así como las diferentes librerías y entornos de desarrollo que se han utilizado en la implementación de la aplicación. Finalmente, en la Sección 5.3 se cuenta cómo se ha llevado a cabo el control de versiones.

5.1. Diseño

Dado que se ha creado una aplicación para el desarrollo de este proyecto de investigación, se ha decidido realizar el diseño de la misma con una herramienta que permite la visualización de un primer prototipado de las distintas pantallas que van a formar parte de esta aplicación.

- **Balsamiq Mockups** [**3**]. Se trata de una herramienta de maquetación que permite realizar diseños para aplicaciones web, móviles y de escritorio de forma rápida, sencilla, eficaz y muy visual. Con esta herramienta también se puede concretar la navegabilidad de la aplicación. Esta herramienta de maquetas fue lanzada al mercado en 2008 por Balsamiq Studios, un proveedor de software independiente fundado ese mismo año por Peldi Guilizzoni.

5.2. Desarrollo

En esta sección se enumeran los lenguajes de programación que se han utilizado para la implementación de la aplicación, las librerías y el entorno de desarrollo.

Lenguajes de programación

- **Java 11**. Versión: 11.0.11. Lenguaje de programación utilizado para el desarrollo de la aplicación.

- **Python 3**. Versión: 3.10.5. Lenguaje de programación utilizado para las funcionalidades proporcionadas por la librería *Spacy* (https://spacy.io/), la cual fue programada para *Python*.

Librerías

- **Spacy** (https://spacy.io/). Librería de software para procesamiento de lenguaje natural desarrollado por Matt Honnibal y programado en lenguaje Python, aunque es adaptable para otros lenguajes. Es apta para castellano.

- **Jython** (https://www.jython.org/). Versión: 2.7.2. Jython es un lenguaje de programación de alto nivel, dinámico y orientado a objetos basado en Python e implementado íntegramente en Java. Es el sucesor de JPython. Jython al igual que Python es un proyecto de software libre. Se ha utilizado para poder utilizar la librería *Spacy* con el lenguaje de programación *Java*.

- **Ngrams** (https://github.com/DanielJohnBenton/Ngrams.java). Se trata de una librería que permite formar n-gramas de tamaño n, al pasarle como parámetro un conjunto de palabras.

- **Weka** (https://waikato.github.io/weka-wiki/downloading_weka/). Versión 3.7.0. Librería de software para el aprendizaje automático y la minería de datos escrito en Java y desarrollado en la Universidad de Waikato, Nueva Zelanda. Es software libre distribuido bajo la licencia GNU-GPL. Se ha utilizado para realizar la técnica de evaluación *"Validación cruzada"* y las diferentes técnicas de clasificación comentadas en la Sección 3.4.

- **Maven** (https://maven.apache.org/). Versión 3.6.3. herramienta de software para la gestión y construcción de proyectos *Java*. Es similar en funcionalidad a *Apache Ant*, pero tiene un modelo de configuración de construcción más simple, basado en un formato XML. Permite generar un fichero *.jar* de manera automática a partir de un proyecto *Java*.

– **Oracle VM VirtualBox** (https://www.virtualbox.org/). Versión 7.0.2. Se trata de un software de virtualización para arquitecturas x86/amd64. Ha sido utilizada para crear una máquina virtual con el Sistema Operativo *Ubuntu* configurada para la rápida ejecución de la herramienta diseñada.

Entorno de Desarrollo Integrado (IDE)

Para el desarrollo de la aplicación se hizo uso de un IDE denominado ***Eclipse*** (https://www.eclipse.org/ide/) con la versión 2022-03 (4.23.0). En este entorno de desarrollo ha sido posible realizar tanto la implementación de la interfaz gráfica de la aplicación, como las funciones que permiten desarrollar las distintas funcionalidades que tiene la misma.

Por otro lado, se ha utilizado un editor de código fuente denominado ***Visual Studio Code*** (https://code.visualstudio.com/), con la versión 1.69.0, para la realización de los scripts realizados en el lenguaje *Python* y su respectiva correcta ejecución.

5.3. Control de versiones

Para el control de versiones se decidió trabajar con ***Google Drive***, ya que el trabajo es individual y que permite ver la actividad y controlar las versiones del proyecto.

Por otro lado, se ha utilizado la herramienta de **GitHub** para que los tutores de este proyecto puedan realizar el seguimiento del mismo a través de un repositorio del que todos formamos parte.

Capítulo 6

Conclusiones y trabajo futuro

En este capítulo se van a desarrollar las conclusiones obtenidas del trabajo, así como las dificultades encontradas y los retos que se han superado a lo largo del desarrollo de este proyecto. Además, se habla del posible trabajo futuro.

En la Sección 6.1 se habla del trabajo realizado y de la conclusión personal del proyecto. En la Sección 6.2 se destacan las conclusiones obtenidas del proyecto. Para finalizar, en la Sección 6.3 se habla del trabajo futuro.

6.1. Trabajo realizado y conclusión personal

En mi opinión, la realización de este proyecto no ha sido algo sencillo. Es cierto que hay múltiples artículos y proyectos que hablan y trabajan sobre la anonimización de datos en documentos médicos, pero el hecho de trabajar con documentos en castellano ha complicado mucho el proyecto. Además, los datos propocionados por el *Instituto Aragonés de Ciencias de la Salud (IACS)* eran informes médicos en formato *".txt"*. Es decir, además de tratarse de textos en castellano, se trata de textos no estructurados escritos por los propios profesionales del sector sanitario.

Estos mismos problemas se han dado en la parte de minería de datos, ya que la técnica de evaluación, *Validación cruzada*, se realiza con Weka mediante ficheros en formarto ARFF. Sin embargo, en este proyecto, se trabajaba con ficheros en formato *".txt"*, lo que volvió a complicar la implementación de la aplicación. Sin embargo, a pesar de las dificultades encontradas y superadas a lo largo del desarrollo del proyecto, creo que se han alcanzado los objetivos que se marcaron al principio del mismo y de manera exitosa.

Este ha sido el primer proyecto de investigación que he realizado. Es un proyecto

de gran extensión que he comenzado desde cero y desconociendo múltiples tecnologías utilizadas. Es por esto que he aprendido muchas cosas nuevas, como la utilización y manejo de herramientas de procesamiento de lenguaje natural como es *Spacy*. Además, he programado una pequeña parte del proyecto en el lenguaje de programación *Python*, el cual apenas había utilizado nunca y he aprendido a configurar una máquina virtual desde cero.

Finalmente, me gustaría resaltar lo importante que es llevar una buena organización en un proyecto de unas dimensiones tan elevadas, ya que son muchas las tareas las que hay que realizar prácticamente en paralelo.

En la Tabla 6.1 se pueden ver las horas dedicadas a cada tarea desempeñada en el proyecto.

Tareas	**Descripción**	**Horas**
Estudio previo	Estudio de las herramientas necesarias para realizar el proyecto	**15h**
Análisis de datos	Análisis de los datos proporcionados por el *Instituto Aragonés de Ciencias de la Salud (IACS)* con el objetivo de encontrar los distintos tipos de datos sensibles que podían aparecer en los informes médicos	**24h**
Desarrollo Data	Implementación del proceso de anonimización de los datos en los documentos, así como la interfaz de esta parte de la aplicación	**74h**
Configuración de MV	Configuración del entorno en una máquina virtual Linux para poder ejecutar la versión de consola de comandos de la aplicación	**20h**
Estudio minería de datos	Estudio de artículos y proyectos basados en anonimización de de datos en documentos médicos aplicando distintas técnicas de minería de datos	**7h**
Desarrollo Minería	Implementación del proceso de evaluación de *Cross Validation*	**40h**
Pruebas	Realización de una batería de pruebas para comprobar el funcionamiento de la aplicación	**60h**
Memoria	Redacción de esta memoria, así como el diseño de diagramas y prototipos	**111h**
Total	Horas totales invertidas en el proyecto	**351h**

Tabla 6.1: Horas totales invertidas

En la Figura 6.1 se puede ver el diagrama de Gantt que refleja el calendario en el que se ha realizado este proyecto.

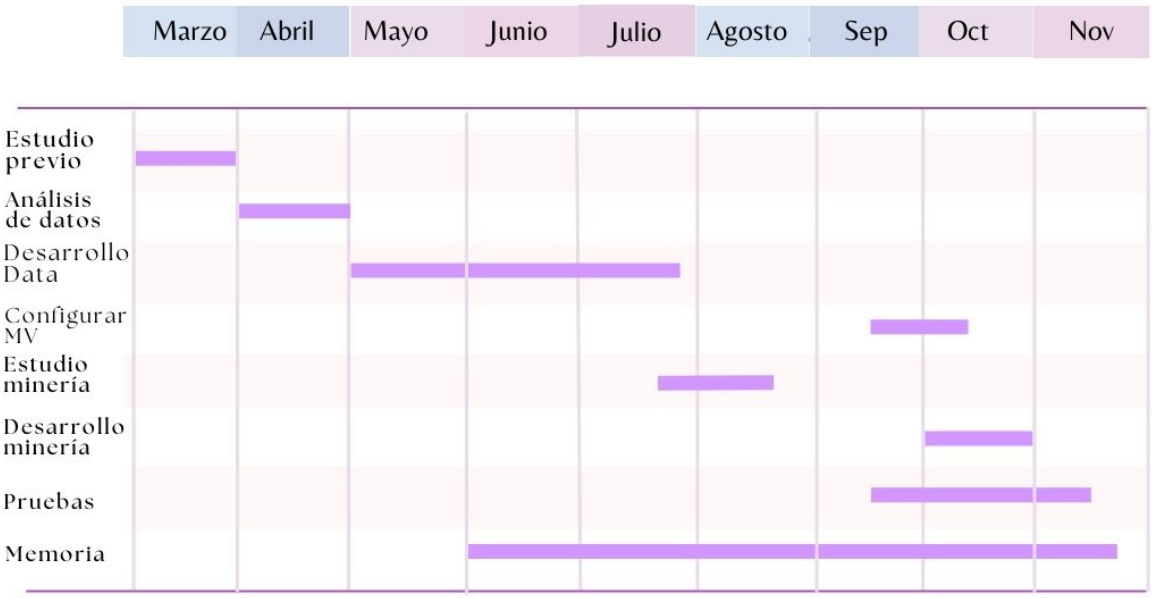


Figura 6.1: Diagrama de Gantt

6.2. Conclusiones del proyecto

El proyecto tenía como objetivo principal desarrollar una herramienta que permitiese evaluar fácilmente y a gran escala la aplicación de diversas estrategias, considerando la efectividad de la anonimización y su posible impacto en otras tareas de procesamiento posteriores. Además, se puntualizaba que los documentos fuesen escritos en castellano. Como resultados globales del proyecto se pueden destacar los siguientes:

- La herramienta permite al usuario seleccionar el tipo de anonimización que desea aplicar. Se trata de dos selectores ortogonales que permiten al usuario elegir atributos a anonimizar y/o niveles de anonimización a aplicar.
- La herramienta permite al usuario cargar documentos en formato ".txt" o carpetas comprimidas ".zip", cuyo contenido son ficheros en formato ".txt".
- La herramienta incluye de una versión que se ejecuta por consola de comandos que permite introducir un gran número de documentos de forma sensible.

- La herramienta consta de una parte de minería de datos donde se aplica una técnica de clasificación que permite evaluar fácilmente las diversas estrategias de anonimización.

- El usuario puede personalizar el proceso de evaluación escogiendo el clasificador a aplicar, el número de "folds" y la clase o categoría sobre la que aplicar la evaluación.

- La herramienta procesa textos en castellano no estructurados.

- Se han utilizado herramientas de procesamiento del lenguaje natural, como es el caso de *Spacy* (https://spacy.io/).

- Se han utilizado librerías de minería de datos como *Weka*.

- Se han utilizado diversos lenguajes de programación como *Java* y *Python* y se han utilizado diversas librerías que han ayudado al desarrollo de la aplicación como es el caso de *Jython* y *Ngrams*.

Además, considero que este trabajo inicial y el proyecto desarrollado puede ser útil para la investigación en gestión de datos en entornos de salud, en particular en el contexto del proyecto NEAT-AMBIENCE (PID2020-113037RB-I00, financiado por MCIN/AEI/ 10.13039/501100011033) en el que se enmarca este trabajo, e incluso en el futuro para otros trabajos desarrollados por investigadores de dicho proyecto.

6.3. Trabajo futuro

La arquitectura implementada durante el proyecto podrá seguir siendo mejorada en el futuro. Algunas de las posibles mejoras a realizar podrían ser las siguientes:

- Añadir nuevas técnicas de anonimización al sistema, más allá de la eliminación de datos, como la seudoanonimización o la generalización.

- Permitir al usuario realizar una configuración de la anonimización más concreta. Por ejemplo, que pueda seleccionar concretamente cuáles son los atributos que desea anonimizar, ya que ahora están divididos en cuatro únicos grupos.

- Añadir la posibilidad de trabajar con ficheros de otras extensiones a las ya permitidas.

- Añadir otros clasificadores.

– Añadir soporte para otras tareas de minería de datos como podrían ser las reglas de asociación.

Se cree que la arquitectura implementada podría seguir mejorándose y adaptándose en mayor medida a las necesidades de los usuarios. Más concretamente, al sector sanitario, ya que los datos con los que se ha trabajado forman parte de dicho sector y han sido proporcionados por el *Instituto Aragonés de Ciencias de la Salud (IACS)*. Del mismo modo, este trabajo se ha planteado en el contexto del proyecto NEAT-AMBIENCE, donde la herramienta y el estudio realizado puede seguir ampliándose y mejorándose. En adición, los directores del TFG analizarán la posibilidad de preparar un trabajo de investigación para su evaluación y potencial presentación en algún congreso, determinando previamente las adaptaciones o mejoras necesarias a lo realizado en el contexto de este TFG.

Referencias

[1] Sergio Ilarri. Rafael Tolosana. Ramón Hermoso. Raquel Trillo. Rafael del Hoyo. "NEAT-AMBIENCE". En: *webdiis.unizar. University of Zaragoza* (2021).

[2] Daniel Berrar. "Cross-validation." En: *Data Science Laboratory, Tokyo Institute of Technology 2-12-1-S3-70 Ookayama, Meguro-ku, Tokyo 152-8550, Japan* (2009).

[3] Névéol A. Zweigenbaum P. "Clinical natural langua- ge processing in 2015: leveraging the variety of texts of clinical interest." En: *Yearbook of Medical Informatics, 25(01):234–239* (2016).

[4] Névéol A. Zweigenbaum P. "Expanding the Diversity of Texts and Applications: Findings from the Section on Clinical Natural Language Processing of the International Medical Informatics Association Yearbook." En: *Yearbook of Medical Informatics, 27(01):193–198.* (2018).

[5] Marrero M. Sánchez-Cuadrado S. Urbano J. Morato J. y J.A. Moreiro. "Sistemas de recuperación de información adaptados a ldominio biomédico." En: *El profesional de la información, 19(3):246–25* (2009).

[6] Zhou G. Zhang J. Su J. Shen D.and Tan C. "Recognizing names in biomedical texts: a machine learning approach." En: *Bioinformatics, 20(7):1178–1190* (2004).

[7] R. Kukafka M. E. Bales A. Burkhardt y C. Friedman. "Human and automated coding of rehabilitation discharge summaries according to the international classification of functioning, disability, and health." En: *J Am Med Inform Assoc., vol. 13, pp. 508-523.* (2006).

[8] F. Gonzalo-Martín C. Millan M. Costumero R. Lopez y Menasalvas. "An approach to detect negation on medical documents in Spanish". En: *In International Conference on Brain Informatics and Health, pages 366–375. Springer.* (2014).

[9] P. Cerrito y J. C. Cerrito. "Data and text mining the electronic medical record to improve care and to lower costs." En: *SUGI 31 San Francisco, CA.* (2005).

[10] Yim W. Yetisgen M. Harris W. P. y Kwan S. W. "Natural language processing in oncology: a review." En: *JAMA oncology, 2(6):797–804* (2016).

[11] Pons E. Braun L. M. Hunink M. M. y Kors J. A. "Natural language processing in radiology: a systematic review." En: *Radiology, 279(2):329–343* (2016).

[12] Chen T. Dredze M. Weiner J. P. Hernandez L. Kimura J. y Kharrazi H. "Extraction of Geriatric Syndromes From Electronic Health Record Clinical Notes: Assessment of Statistical Natural Language Processing Methods." En: *JMIR Medical Informatics, 7(1):e13039* (2019).

[13] Syed Atif Moqurrab. Adeel Anjum. Abid Khan. Mansoor Ahmed. Awais Ahmad y Gwanggil Jeon. "Deep-Confidentiality: An IoT-Enabled Privacy-Preserving Framework for Unstructured Big Biomedical Data." En: *ACM Trans. Internet Technol. 22, 2, Article 42, 21 pages.* (2022).

[14] Wu C. Xia F. Deleger L. y Solti I. "Statistical machine translation for biomedical text: are we there yet?" En: *In AMIA Annual Symposium Proceedings, volume 2011, page 1290. American Medical Informatics Association* (2011).

[15] G. Névéol A. Dalianis H. Velupillai S. Savova y Zweigenbaum P. (2018a). "Clinical natural language processing in languages other than English: opportunities and challenges." En: *Journal of biomedical semantics, 9(1):12.* (2018).

[16] Castano J. Gambarte M. L. Park H. J. Williams M. d. P. A. Perez D. Campos F. Luna D. Benitez S. Berinsky H. y Zanetti S. "A machine learning approach to clinical terms normalization." En: *In 15th Workshop on Biomedical Natural Language Processing, pages 1–11.* (2016).

[17] Costumero R. Lopez F. Gonzalo Martín C. Millan M. y Menasalvas E. "An approach to detect negation on medical documents in Spanish." En: *In International Conference on Brain Informatics and Health, pages 366–375. Springer.* (2014).

[18] Marimon M. Gonzalez-Agirre A. Intxaurrondo A. Rodrguez H. Lopez Martin J. Villegas M. y Krallinger M. "Automatic de-identification of medical texts in Spanish: the MEDDOCAN track, corpus, guidelines, methods and evaluation of results." En: *In Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings (CEURWS.org), Bilbao, Spain* (2019).

[19] Carlos Sánchez. "Aplicación de Técnicas de Minería de Textos para Apoyar la Búsqueda de Información en Contextos Médicos en Español." En: *Repositorio de la Universidad de Zaragoza – Zaguan* (2019).

[20] Uzuner O. Luo Y. Szolovits P. "Evaluating the State-of-the-Art in Automatic De-identification." En: *Journal of the American Medical Informatics Association 14(5), 550–563.* (2007).

[21] Kushida CA. Nichols DA. Jadrnicek R. Miller R. Walsh JK. Griffin K. "Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies." En: *Med Care.50 Suppl(Suppl):S82-101. doi: 10.1097/MLR.0b013e3182585355. PMID: 22692265; PMCID: PMC6502465.* (2012).

[22] Ruch P. Baud RH. Rassinoux AM. Bouillon P. Robert G. "Medical document anonymization with a semantic lexicon." En: *Proc AMIA Symp. 2000:729-33. PMID: 11079980; PMCID: PMC2244050.* ().

[23] Kushida CA. Nichols DA. Jadrnicek R. Miller R. Walsh JK. Griffin K. "Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies." En: *Med Care. 50 Suppl(Suppl):S82-101. doi: 10.1097/MLR.0b013e3182585355. PMID: 22692265; PMCID: PMC6502465.* (2012).

[24] György Szarvas. Richárd Farkas. Róbert Busa-Fekete. "State-of-the-art Anonymization of Medical Records Using an Iterative Machine Learning Framework." En: *Journal of the American Medical Informatics Association, Volume 14, Issue 5, Pages 574–580.* (2007).

[25] Pilar López-Ubeda. Manuel C. Díaz-Galiano. L. Alfonso Urena-López y M. Teresa Martín-Valdivia. "Anonymization of Clinical Reports in Spanish: a Hybrid Method Based on Machine Learning and Rules." En: *IberLEF* (2019).

Lista de Figuras

Lista de Tablas

Anexos

Anexos A

Análisis y clasificación de los datos

En este anexo se va a explicar el análisis realizado sobre los datos de los documentos proporcionados. En el Anexo A.1 se narra la clasificación de los atributos. En el Anexo A.2 se muestran las combinaciones de atributos que aumentan la sensibilidad de manera conjunta. Finalmente, en el Anexo A.3 se muestra la sensibilidad conjunta de los atributos.

A.1. Clasificación de los atributos

A continuación, se muestra la Tabla A.1, que muestra las distintas entidades que se han encontrado en la colección de datos de los documentos médicos a tratar. Se han agrupado por patrones y se les ha otorgado un nivel de sensibilidad y de utilidad clínica.

La sensibilidad se verá reflejada en el nivel de anonimización que debería recibir texto. En este caso, la sensibilidad puede ser alta, media o baja. Será alta cuando haya que anonimizar el texto siempre, media cuando el texto puede necesitar ser anonimizado en función del contexto, pero no siempre; y baja cuando no parece problemático mantener el texto.

En el caso de la utilidad, hace referencia a cómo de útiles son los datos desde el punto de vista médico. Su valor será alto cuando sea relevante desde la perspectiva médica, media cuando sea potencialmente relevante, parcialmente relevante cuando es relevante parte del texto, pero no todo; y baja cuando no parezca relevante.

Por otro lado, cabe destacar que en la columna "Texto" se presentan diferentes ejemplos que aparecen en los documentos de historias clínicas. Cuando una entidad tiene más de uno es porque una misma entidad se ha visto representada en el documento de diferentes formas y se considera relevante para la clasificación. Algunos

datos aparecen reflejados con el patrón **XXXXX** por motivos de privacidad de los pacientes.

Patrón	**Texto**	**Sensibilidad**	**Utilidad clínica**
Nombre	nace mujer **Laura**,O+, peso 2760	Alta	Baja
	ENDOCRINOLOGIA C **DRA.GRACIA** (CME "GRANDE COVIÁN").	Baja	Alta
	Ainhoa además tuvo unas décimas	Alta	Baja
	Hola **Elisabet**: La información última es que...	Alta	Baja
	Acude la madre sin **Nicolas**	Alta	Parcialmente relevante
	Niña **Alicia** 2640, 2400 al alta.	Alta	Baja
	Acude su esposa- **Mª Jesus XXXXX.**	Alta	Baja
	Varon 3720gr **(Alexandro)**.	Alta	Baja
	Comenta su padre que **Josue**	Alta	Baja
	Refiere que su padre vive sólo, tiene 4 hijos, **(José María, Fcco Javier, Miguel Angel y Santiago)**.	Alta	Baja
Nombre y apellidos	Consulta Cirugia general 21-6-19 :21/06/2019 10:52 **JOSE LUIS**.	Alta	Baja
	Informado por: **JUAN JOSE XXXXX XXXXX**.	Alta	Media
	El/la paciente **TANIA, XXXXX XXXXX**,	Alta	Baja
	NEUMOLOGIA A **XXXXX XXXXX, LUISA MARGARITA** 17/03/2020	Alta	Baja
	11/12/2019 10:44 **MARIA JOSEFA XXXXX XXXXX** Anotación AS: TSH 1.63, T4L 0.77	Alta	Baja
	Os remito a **Natali XXXXX XXXXX**, es una niña de X años, que en el colegio un compañero le ha retorcido la muñeca y desde entonces nota dolor.	Alta	Baja
	- 19/08/2019 - 22:30 — **XXXXX XXXXX, OLGA** — comentado con adjunto de Urgencias (Dr Lahoz) que refiere que ...	Alta	Baja
	Saludos, **Marcos XXXXX**	Alta	Media

	Informado por: **JUAN JOSE XXXXX XXXXX** **Primer apellido: XXXXX Segundo apellido: XXXXX Nombre: Mª ISABEL** R: Dr. Enrique XXXXX XXXXX	Alta Alta Alto	Media Baja Baja
DNI/Pasaporte	DNI/NIE/TR/Pasaporte: **17154811P**	Alta	Baja
Número SS	El/la paciente TANIA, XXXXX XXXXX, con nº de SS **50/00707593/92** ha sido atendido/a	Baja	Baja
Número de colegiado	Zaragoza, 14/12/2016 Fdo Dr/a.LETICIA XXXXX XXXXX Medicina Nuclear Colegiado nº **16892**	Bajo	Bajo
Teléfono	TRABAJADORA SOCIAL **976667809**, ESTA DE VACACIONES LA PSIQUIATRA HASTA EL MARTES. **Tfno: 976576184**	Alta Alta	Baja Baja
Correo electrónico	Envío información **XXXXXX@gmail**.	Alta	Baja
Dirección	Acude por agresión por parte de una mujer conocida esta tarde a las 17:30 en la **C/ Villalpando Alonso**.	Alta	Baja
Género	**Sexo: MUJER** **Sexo: Varón**.	Media Media	Alta Alta
Etnia	**Paciente de raza negra,** de complexión atlética, adoptado hace 13 años, sin conocer antecedentes genéticos, presentó infestación activa de Malaria, tratada en su momento.	Alta	Alta
Edad	JOSE LUIS XXXX XXXX **52 años.** Un episodio a los **34 años** **Paciente de residencia** que acude por presentar cuadros presincopales con hipotensión en las últimas dos semanas sin pérdida de conciencia. **Paciete de 71 años**, con antecedentes poliartrosis y osteoporosis, malformacion cerebelosa de Dandy Walker, HTA	Media Media Media Media	Media Media Media Media

	Mujer de 16 años que acude por DOLOR DENTARIO Y DE ENCÍA ANTECEDENTES PERSONALES	Media	Media
	Niño de 10 años que consulta por sospecha de hipoacusia detectada en ámbito escolar y familiar durante el curso pasado.	Media	Media
	52 a. Fumador de purillos hasta hace 1 mes.	Media	Media
	Paciente de 23 años de edad .	Media	Media
	Hoy en **rev de 6 años**, talla por debajo de P3.	Media	Media
	paciente de 45a con dolor intenso e impotencia funcional en ambos codos que auemtna con el movimeinto.	Media	Media
	Edad de la madre: 30 años.	Media	Media
Lugar	Refiere que está pendiente de trasplante de riñon y pancreas en **Barcelona**.	Baja	Alta
	Paciente que ingresa en el **Hospital Nuestra Señora de Gracia de Zaragoza** de forma programada por CMA por presentar ciatalgia secundaria a anterolistesis L4-L5..	Baja	Media
	Fiebre de 38,5 de 3 horas de evolucion, **detectado en guardería**.	Media	Media
	Colegio Público de Valdespartera.	Media	Media
	Pte de estudio psicopedagogico en **su colegio, Marie Curie**, le dan apoyo.	Media	Media
	Trae informe de **Institut Català de la Salut**.	Baja	Media
	Llaman de **la residencia de los Pueyos,** hay problema de la diespensación con depakine 500, si tienen de 200 (problema de desabastecimiento).	Baja	Media
País	Quieren ir de viaje en unas semanas a **Rumania,** les doy recomendaciones de evitar el viaje (no está protegida de sarampion para viajar a un pais endémico).	Media	Alta
	Se van unos dias a **Marruecos.**	Media	Media

	Procedente de **Colombia**, lleva aquí 3 años.	Alto	Alta
	en Méjico.	Media	Media
	Es de Nigeria y viajó alli en agosto.	Alto	Alta
	Refiere que **viene de Nicaragüa** por reagrupamiento familiar.	Media	Media
	Procede de Rumania, refiere que en España ha cotizado 4 años y 7 meses, y en **Rumania** 23 años.	Alta	Alta
	En Inglaterra tras dolro intneso en región escapular izda, se h aDx Herpes zoster, no ha habido erupción .No traumatismo previo, el dolro lo tien localizado enregión cervical, suprescapular, mejora tumbada, duerme bien.	Baja	Alta
Ciudad	**En Asturias.**	Media	Media
	de su madre diagnosticada de Alzheimer que **vive en Alcañiz**.	Media	Media
	La tenían que haber revisado a los 6 meses, pero me cuenta que ya no hay especialista en **Calatayud y ahora vive en Zaragoza**	Media	Alta
	Provincia: **(ZARAGOZA)**.	Baja	Baja
Región	fam: La anciana vive rotando cada 6 meses entre **Andalucia y Aragon.**	Media	Media
Hospital y/o centro médico	**H. Royo Villanova.**	Baja	Alta
Fechas personales	**Fecha nacimiento: 18/04/1961 (58 años)**	Alta	Baja
	Fecha de Nacimiento: 26/04/2007.	Alta	Baja
	Padre **fallecido en 2003**	Alta	Baja
	madre fallecida **dia 22-6-16**	Alta	Baja
Fechas de eventos	En caso de que se le requiera el documento, le digo que podremos realizarlo con **fecha 25/01/2018**	Baja	Alta
	Lo ven en cirugia el **13 de agosto**.	Baja	Alta
	Nos comenta el paciente que **en el año 1986** sufrió un IAM y que rechazó coronariografia. Visto por urologo privado que pauta Nitrofurantoina y Urorec con mejoria importante tras iniciar este nuevo tratamiento desde hoy.	Baja	Alta

	2ª dosis de Hepatits A, **en enero de 2020**.	Baja	Alta
	Desde el**12.05.2017** en ttº con Amoxi.: 9.5 ml/8h.	Baja	Alta
	Fecha de INGRESO : 28/07/2016	Baja	Alta
	Fecha de ALTA : 28/07/2016	Baja	Alta
Familiar	**Acude su esposo** con informe de alta: adenocarcinoma de endometrio, histerectomia total con anexectomia bilateral con estuido intraoperatirio de invaision miometral negativo.(18-03-08) Pendien te de muestreo ganglional de iliaca interna y exerna bilateral	Media	Baja
	Hablo con la hija y le doy glucómetro para controlar cifras y ver si está siendo efectivo el tto puesto en hospital ADO	Media	Baja
	Padre, madre y hermano se realizan test rápido que es negativo **Padre y hermano** asintomático **Madre** síntomas dia 11 inespecíficos ahora asintomática Plan aislamiento madre del resto de la familia.	Bajo	Alta
	Terreno alérgico familiar: **Padre** reacción adversa a penicilina.	Bajo	Alta
	LLeva ya casi un año sin tomar heipram y ha llevado un año muy malo por problemas familiares, **hija 15 años** con ideación suicida en una ocasión , **marido en paro** de larga duración, **fallecimiento de su padre**...	Media	Parcialmente relevante
	Se encuentra mejor, algo nervioso por la espirometría dentro de una semana, más animado, han venido **unos familiares de Valencia** y ha estado entretenido.	Bajo	Parcialmente relevante
	trascribo visita 21/10/2021 11:33 MARIA ANGELES XXX XXX Anotación Psicología Refiere que **su "tío"(padre de su prima)** le ha vuelto a contactar para felicitarle el cumpleaños.	Bajo	Parcialmente relevante

	la niña pide pecho cada 4 horas , la madre la intenta poner antes pero no quiere , toma los 2 pechos 15 minutos en cadfa pecho.	Baja	Media
	aBUELOS PATERNO tiroides,**abuela** paterna HTA,DMID,ARRTIMIA LINFOMA.	Baja	Alta
	Acude Con su hijo de 17 años quien ha estado ingresado en centro de Lérdia 4 meses pasra desintoxicación.	Media	Alta
	Viene su hija (Fca).Demanda información , de solicitud de discapacidad/minusvalía.	Media	Baja
	Vive con un hno.	Media	Parcialmente relevante
	Su ex-marido es Covid + Hasta ahora una vecina le llevaba la compra pero ahora no tiene a nadie.	Media	Alta
	pers/fam: **Matrimonio**.	Media	Media
	Divorcio de los padres hace 2 años, los padres no tienen ninguna relación, la custodia la tiene la madre.	Media	Media
	Divorciado.	Media	Media
	Casada.	Media	Media
	tiene 4 hijos.	Media	Media
	Acude con una de sus 3 hijas, Ana, con quien está ahora hace unos días, por inf.	Alta	Parcialmente relevante
	Viuda hace unos 5 años , vive con un **hijo de 32 años** que ahora no trabaja y otra hija independiente.	Media	Baja
	Soltera, vive con su hermana.	Media	Media
Trabajo	**trabajadora de residencia**, le realizan el día 4/5 test rapido +(aislamiento) y pcr+ dias 7/5 (picarral).	Media	Alta
	Dejo volante a la **cuidadora** para que venga a citarla de nuevo.	Media	Parcialmente relevante
	Ayer en revisión se vió disminución de agudeza visual, hablé con la **madre (Enfermera)** estaba agobiada, remití como preferente.	Media	Parcialmente relevante

	econ: **Paciente pensionista** (SOVI), marido pensionista (1.3-1.4) P: Información ayudas individuales: Gestion DFA- no se ha publicado la convocatoria.	Media	Media
	Era administrativa de la empresa.	Media	Media
	Trabaja en la lavanderia de una residencia y levanta pesos de forma habitual.	Media	Media
	Trabajo: soldador.	Media	Media
	Enfermera de guardia de vigilancia.	Media	Media
	La paciente es teleoperadora y no sabe si podrá ir mañana a trabajar si necesita baja deberá citarse con su médico.	Media	Media
	Trabaja en SXXI, en noviembre de 2019 hubo un escape de particulas de aluminio.	Media	Alta
Deporte	**Practica baloncesto**.	Media	Media
	Me cuenta **un entrenamiento de rugby** y es duro: varias tandas de carreras, despues tablas de fuerza y abdominales, despues de nuevo carreras, en esta tercera parte es en la que else noto mal.	Media	Media
	Hace patinaje.	Media	Media
Historia personal	Adoptado hace 13 años	Media	Media

Tabla A.1: Datos sensibles encontrados

A.2. Combinaciones de atributos que aumentan la sensibilidad

A continuación, se presenta la Tabla A.2, con las diferentes combinaciones de las entidades anteriores, que pueden potenciar la sensibilidad de la información de manera conjunta.

La tercera columna representa si la combinación de las entidades aumenta o no la sensibilidad de los datos. En el caso de que el valor de esta casilla sea “Parcialmente”, implica que la combinación de los datos puede aumentar la sensibilidad de los datos en ciertas ocasiones, pero no siempre. Esto se debe a que no se sabe el contexto en el que se tratan los datos. Por ejemplo, si se trata de un pueblo pequeño o una ciudad enorme.

Combinación	**Texto**	**Aumenta la sensibilidad**
Etnia y deporte	* Adulto de raza negra que practica atletismo	Parcialmente
Etnia y edad	* Paciente de raza negra de 38 años	Parcialmente
Etnia y lugar	*Paciente de raza negra residente en Paracuellos	Sí
Etnia y país	*Paciente de raza negra procedente de Senegal	Parcialmente
Etnia y ciudad	*Paciente de raza negra que vive en Zaragoza	Parcialmente
Etnia y región	*Paciente de raza negra nacido en Aragón	Parcialmente
Etnia y familiar	*Niño de raza negra que viene acompañado de su madre	Parcialmente
Etnia y trabajo	*Paciente de raza negra que trabaja en el aeropuerto	Sí
Edad y lugar	*Niño de 2a que viene de la guardería	Parcialmente
Edad y ciudad	*Paciente de 35a que reside en Calatayud	Sí
Edad y país	*Paciente de 25 años procedente de Italia	Parcialmente
Edad y región	*Paciente de 63a de Andalucía	Parcialmente
Edad y trabajo	*Paciente de 46 años, soldador en una empresa	Sí
Edad y familiar	*15 años de edad, viene acompañada de su madre **hijo de 15 años** vive con un **hijo de 32 años**	Parcialmente Sí Sí
Trabajo y fecha	*03/05/2022: Policia con una muñena fracturada	Parcialmente
Trabajo y lugar	*Constructor en un colegio de Zaragoza	Parcialmente
Trabajo y ciudad	*Residente en Zaragoza, pero trabaja en Épila como fontanero	Sí
Trabajo y país	Bombero nacido en Francia	Sí
Trabajo y región	Policía nacional en la provincia de Zaragoza	Sí
Trabajo y deporte	*Instructor de zumba que juega a fútbol los fines de semana	Parcialmente
Deporte y lugar	Necesita it: **Federación Aragonesa de Futbol**, desde hoy 02/08. Le vieron **en la mutua de futbol**, no Rad le pusieron vendaje y estuvo un mes de reposo .	Parcialmente
Deporte y ciudad	*Juega a fútbol en Madrid	Parcialmente

Deporte y fecha	me comenta que es contacto porque **el dia 20/11** fueron en autobus su marido y su hijo a un **partido de futbol**, y hay varios positivos (estuvieron sin mascarilla en ocasiones).	Parcialmente
Lugar y familiar	**4 hijos fuera de Zaragoza**.	Sí
Ciudad y familiar	Vive en Calatayud con su hija Claudia	Sí
País y familiar	Procede de Portugal, viene con su hermana	Parcialmente
Región y familiar	Viene de Andalucía con su sobrino	Parcialmente
Ciudad y fecha	**Zaragoza, 20 de Diciembre de 2017**	No
Familiar y nombre	pers/fam: **Vive con su marido (Vicente Vergara)**.	Sí
Familiar y trabajo	**Padre: Taxista.**	Parcialmente
Familiar y deporte	Hablo **con su madre** por dolor en talon tras haber empezado **con entrenamientos de futbol**, es diestro.	Parcialmente
Número de SS y lugar	* Colegiado nº **16892**. Hospital Clínico	No
Número de SS y país	* nº de SS **50/00707593/92**, España	No
Número de SS y ciudad	* nº de SS **50/00707593/92**, Zaragoza	No
Número de SS y región	*nº de SS **50/00707593/92**, Aragón	No
Número de colegiado y lugar	*Colegiado nº **16892**. Hospital Clínico	No
Número de colegiado y país	*Colegiado nº **16892**. España	No
Número de colegiado y ciudad	*Colegiado nº **16892**, ZARAGOZA	No
Número de colegiado y región	*Colegiado nº **16892**, Cataluña	No
Hospital y lugar	*Paciente ingresado en Hospital Clínico, vive en Montecanal	Parcialmente
Hospital y ciudad	*Paciente ingresado en Hospital Clínico, vive en Madrid	Parcialmente
Hospital y país	*Paciente ingresado en Hospital Clínico, procedente de Italia	Parcialmente

Hospital y región	*Paciente ingresado en Hospital Clínico, viene de Asturias	Parcialmente
Hospital y edad	*Paciente de 31 años, ingresado en H. Royo	Parcialmente
Hospital y familiar	*Ingresado en H.Clinico, viene con su madre	No
Hospital y etnia	*Paciente de raza negra ingresado en H.Clínico	No
Género y edad	*Mujer de 23 años	No
Género y ciudad	*Hombre procedente de Valencia	Sí
Género y lugar	*Mujer del barrio de Delicias	Sí
Género y trabajo	*Mujer que trabaja como camarera en un bar	Sí
Género y país	*Varón procedente de Países Bajos	Sí
Género y región	*Varón residente en País Vasco	Sí
Género y deporte	*Hombre que practica fútbol	Parcialmente
Género y etnia	*Mujer de raza negra	No

Tabla A.2: Combinaciones que aumentan la sensibilidad de los datos

Dado que el nombre, los apellidos, el DNI, el pasaporte, la dirección de un domicilio, el teléfono y el correo electrónico se consideran altamente sensibles por sí solos, no se han tenido en cuenta para la realización de esta tabla.

Nota: Los campos que se encuentran en la columna de "Texto" con un *, es porque son ejemplos ficticios. Es decir, no aparecen en las historias clínicas disponibles, ya que están desordenados por cuestiones de privacidad. Sin embargo, al ver el historial médico completo de un paciente real pueden darse dichas situaciones.

A.3. Sensibilidad conjunta de atributos

A continuación, se presenta la Tabla A.3, donde se muestra el nivel de sensibilidad que muestran los atributos de manera conjunta.

Combinación	**Sensibilidad por separado**	**Aumenta la sensibilidad**	**Sensibilidad conjunta**
Edad y lugar	Media + Media	Parcialmente	Media
Edad y ciudad	Media + Media	Sí	Media
Edad y país	Media + Media	Parcialmente	Media
Edad y región	Media + Media	Parcialmente	Media
Edad y trabajo	Media + Media	Sí	Alta

Edad y familiar	Media + Media	Sí	Alta
Trabajo y fecha clínica	Media + Baja	Parcialmente	Media
Trabajo y lugar	Media + Media	Parcialmente	Media
Trabajo y ciudad	Media + Media	Sí	Alta
Trabajo y país	Media + Media	Sí	Media
Trabajo y región	Media + Media	Sí	Media
Trabajo y deporte	Media + Media	Parcialmente	Media
Deporte y lugar	Media + Media	Parcialmente	Media
Deporte y ciudad	Media + Media	Parcialmente	Media
Deporte y fecha clínica	Media + Baja	Parcialmente	Media
Lugar y familiar	Media + Media	Sí	Alta
Ciudad y familiar	Media + Media	Sí	Alta
País y familiar	Media + Media	Parcialmente	Media
Región y familiar	Media + Media	Parcialmente	Media
Familiar y trabajo	Media + Media	Parcialmente	Media
Familiar y deporte	Media + Media	Parcialmente	Media
Hospital y lugar	Baja + Media	Parcialmente	Media
Hospital y ciudad	Baja + Media	Parcialmente	Media
Hospital y país	Baja + Media	Parcialmente	Media
Hospital y región	Baja + Media	Parcialmente	Media
Hospital y edad	Baja + Media	Parcialmente	Media
Género y ciudad	Media + Media	Sí	Alta
Género y lugar	Media + Media	Sí	Alta
Género y trabajo	Media + Media	Sí	Alta
Género y país	Media + Media	Sí	Alta
Género y región	Media + Media	Sí	Alta
Género y deporte	Media + Media	Parcialmente	Media
Hospital y familiar	Baja + Media	No	**No afecta**
Género y edad	Media + Media	No	**No afecta**
Número SS y nº de colegiado	Baja + Baja	No	**No afecta**
Número de SS y género	Baja + Baja	No	**No afecta**
Número de SS y edad	Baja + Baja	No	**No afecta**
Número de SS y lugar	Baja + Media	No	**No afecta**
Número de SS y país	Baja + Media	No	**No afecta**
Número de SS y ciudad	Baja + Media	No	**No afecta**
Número de SS y región	Baja + Media	No	**No afecta**
Número de SS y hospital	Baja + Baja	No	**No afecta**
Número de SS y trabajo	Baja + Media	No	**No afecta**

Número de SS y deporte	Baja + Media	No	**No afecta**
Número de SS y historia personal	Media + Baja	No	**No afecta**
Número de colegiado y género	Baja + Baja	No	**No afecta**
Número de colegiado y edad	Baja + Baja	No	**No afecta**
Número de colegiado y lugar	Baja + Media	No	**No afecta**
Número de colegiado y país	Baja + Media	No	**No afecta**
Número de colegiado y ciudad	Baja + Media	No	**No afecta**
Número de colegiado y región	Baja + Media	No	**No afecta**
Número de colegiado y hospital	Baja + Baja	No	**No afecta**
Número de colegiado y trabajo	Baja + Media	No	**No afecta**
Número de colegiado y deporte	Baja + Media	No	**No afecta**
Número de colegiado y historia personal	Media + Baja	No	**No afecta**
Ciudad y fecha	Media + Baja	No	**No afecta**
Hospital y trabajo	Baja + Media	No	**No afecta**

Tabla A.3: Combinaciones que aumentan la sensibilidad de los datos y sensibilidad conjunta

Se debe destacar que para la realización de la Tabla A.3 no se han tenido en cuenta aquellas combinaciones en las que uno de los dos atributos tiene sensibilidad alta, ya que si se aplica anonimización dicho atributo se eliminará seguro.

Anexos B

Diseño de la aplicación

En este anexo se presenta el prototipado inicial de la herramienta, así como el diseño final de la misma. En el Anexo B.1 se muestra el prototipado inicial de la aplicación. En el Anexo B.2 se muestran los mapas de navegación correspondientes a cada parte de la aplicación. En el Anexo B.3, se muestra el diagrama de clases de la implementación llevada a cabo en el proceso de anonimización. En el Anexo B.4, se muestra el diseño e implementación de la herramienta.

B.1. Prototipado principal

A continuación se muestra un pequeño prototipo inicial de la aplicación. La aplicación consta de una pestaña denominada "Data" orientada al proceso de anonimización de datos y una pestaña denominada *"Mining"* enfocada al proceso de minería de datos. También consta de una ventana de resultados.

En la Figura B.1, se puede apreciar la página principal de la aplicación. En ella se puede ver que el usuario puede importar una carpeta con ficheros, los cuales se pretenden anonimizar. Además, el usuario puede elegir entre cuatro opciones de anonimizado, las cuales se explican en el cuadro central que aparece en pantalla. El usuario puede elegir más de una opción si lo desea.

En cuanto al desplegable del nivel de anonimización, se puede ver en la Figura B.2, que consta de cuatro niveles de anonimización: "Don't apply", "Low", "Medium" y "High". Cuando el usuario seleccione la opción "Don't apply" no se aplicará ningún nivel de anonimización. Por el contrario, cuando seleccione una de las otras tres opciones, se aplicará un nivel bajo, medio o alto de anonimización, respectivamante. Esto se puede ver en la Figura B.2.

Se puede observar que hay un botón denominado “Download folder” en la parte inferior de la pantalla que se encuentra deshabilitado. Este botón permitirá la descarga del documento anonimizado. Se habilitará cuando el usuario haya seleccionado la configuración del proceso.

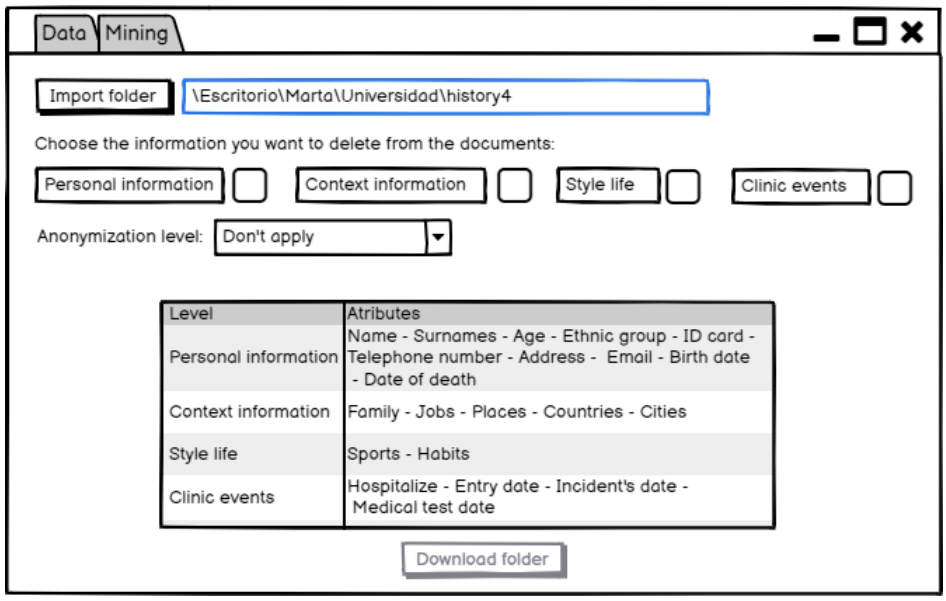


Figura B.1: Página principal de la aplicación

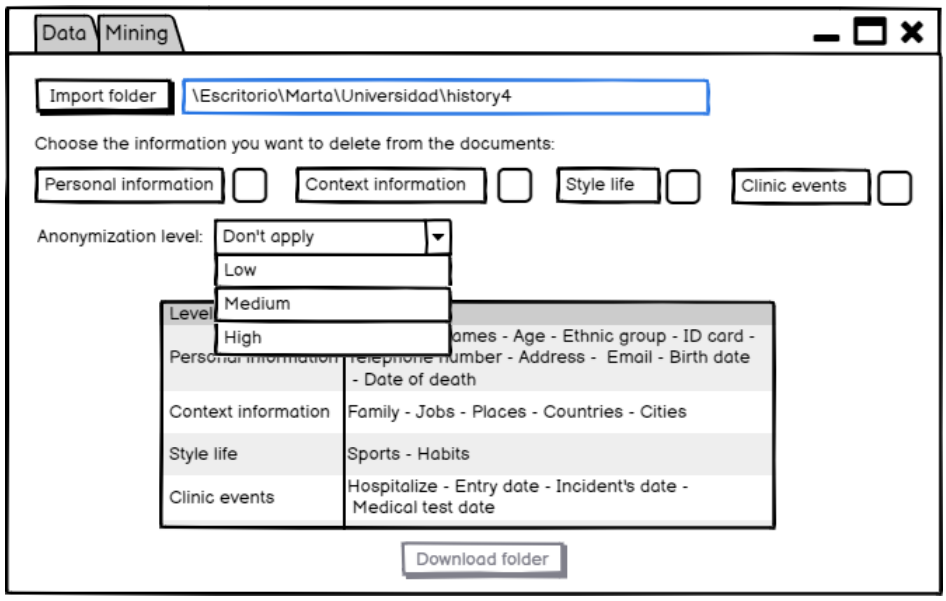


Figura B.2: Distintos niveles de anonimización

En la figura B.3, se puede ver lo que sucede cuando el usuario hace clic sobre el botón “Import folder” que aparece en la esquina superior izquierda. El usuario puede seleccionar la carpeta que desee, moviéndose por los directorios.

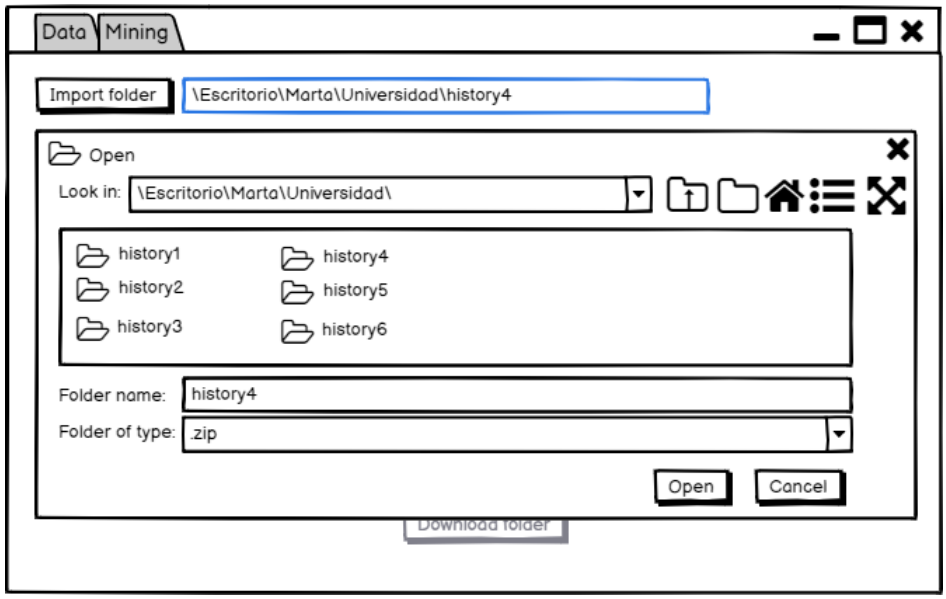


Figura B.3: Importar documentos

Finalmente, en la figura B.4, se puede ver que el usuario ha seleccionado varias opciones de las ofrecidas. Dado que ya ha seleccionado alguna de las opciones, el botón "Download folder", de la parte inferior, pasa a estar disponible para el usuario. Si el usuario pulsa en él, se descargarán los ficheros anonimizados en un fichero comprimido.

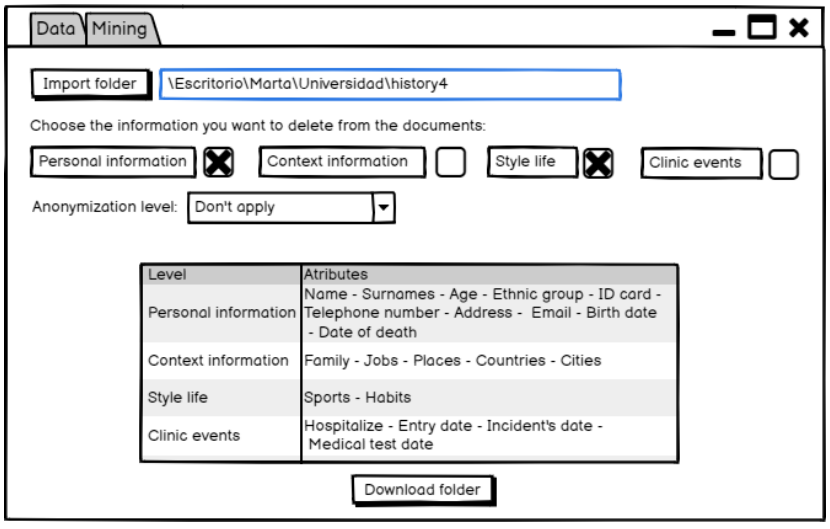


Figura B.4: Selección de opciones y descarga de archivo

En la Figura B.5, se presenta la ventana denominada *Mining*, que recoge la parte de minería de datos. La técnica de minería de datos que se va a llevar a cabo en este proyecto es la denominada *Validación Cruzada*. Esta validación se aplicará tanto al dataset original como al anonimizado.

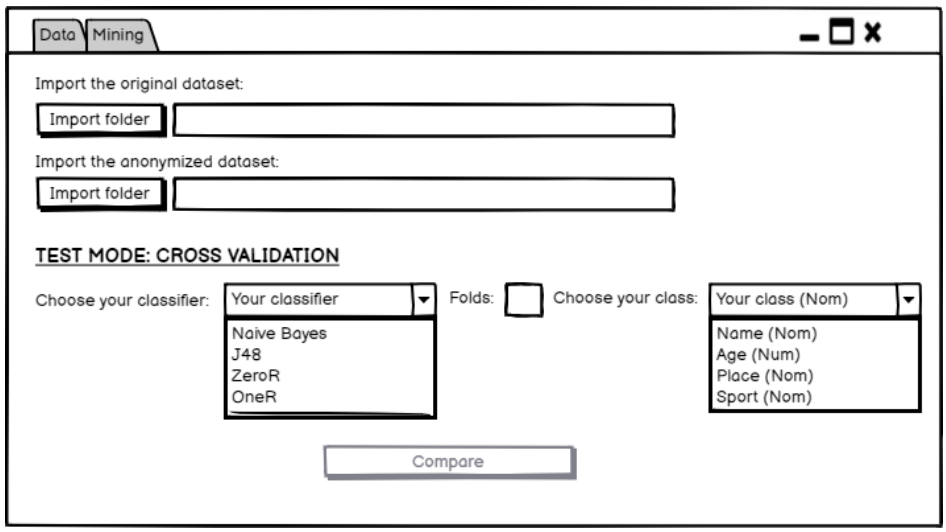


Figura B.5: Página principal de la ventana "Mining"

El usuario puede seleccionar los ficheros a procesar a través de los botones *"Import folder"* como se muestra en la Figura B.6.

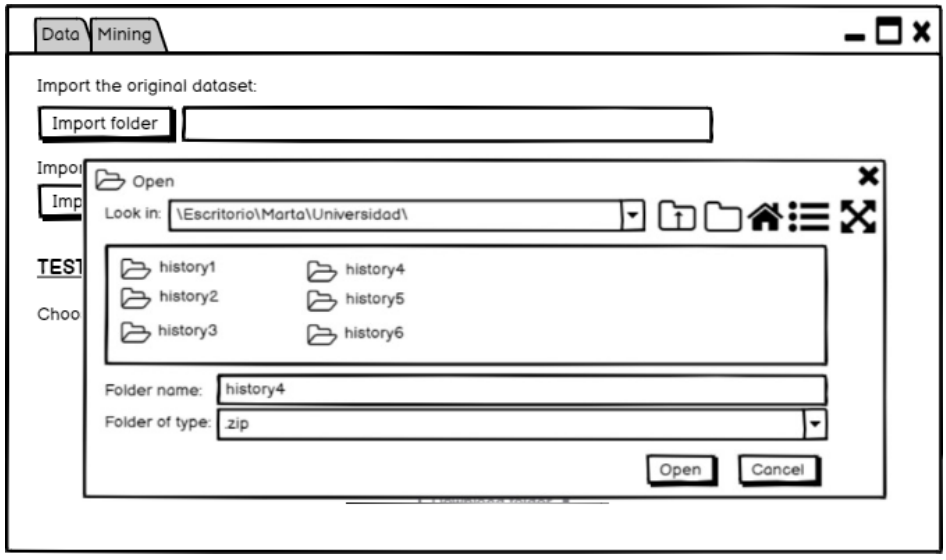


Figura B.6: Importar ficheros

Una vez el usuario haya importado los ficheros, debe proceder a configurar la validación cruzada. En este caso, el usuario debe elegir el clasificador que desea aplicar: Naive Bayes, SMO, ZeroR o OneR. Además, debe indicar el número de *"folds"* con los que desea aplicar la validación cruzada, así como la clase/categoría. Esto se puede ver en la Figura B.7.

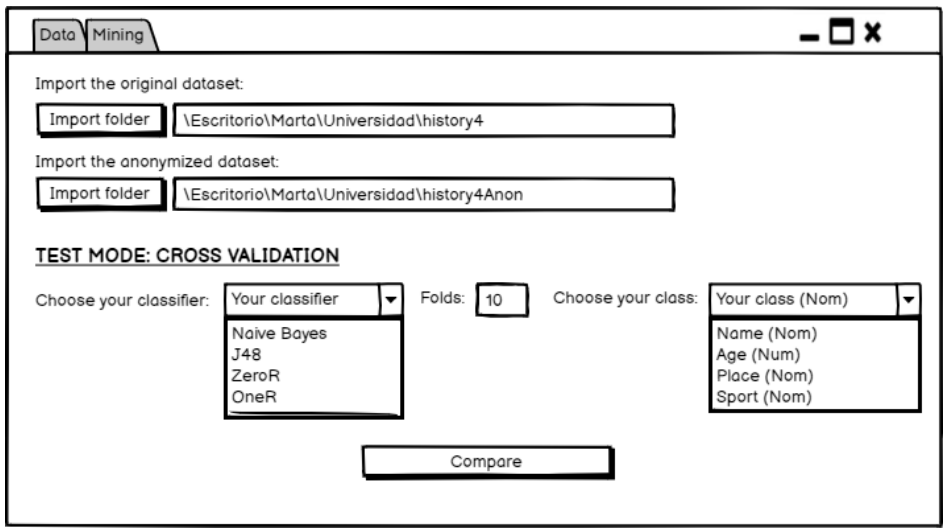


Figura B.7: Configuración de la validación cruzada

Una vez que el usuario haya importado los dos ficheros y haya configurado el proceso de validación cruzada, el botón *"Compare"* se habilitará y el usuario podrá hacer click sobre él. Esto se puede observar en la Figura B.7. Cuando lo haga, se abrirá la ventana denominada *Results* con los resultados obtenidos. Los resultados que mostrará serán: *True Positives, False Positives, True Negatives, False Negatives, Precision, Recall, F-measure, Correctly classified, Incorrectly classified, Mean absolut error y Root mean square error*. Esto se puede ver en la Figura B.8.

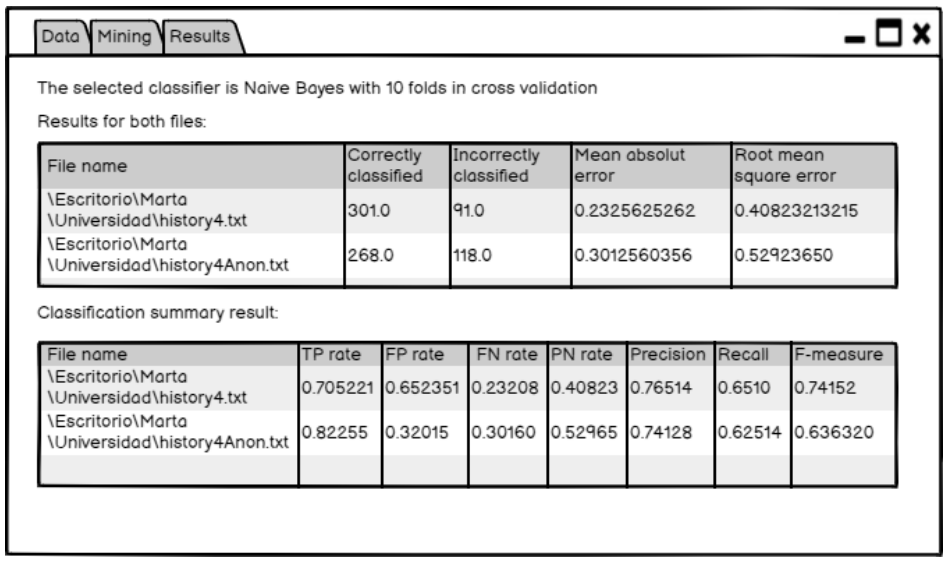


Figura B.8: Página principal de la aplicación

En las Figuras B.9 - B.17 se muestran las distintas pantallas que tiene la aplicación final.

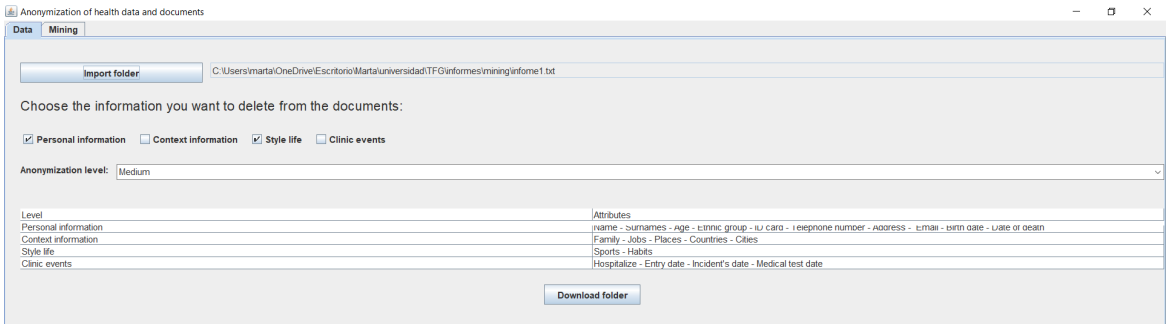


Figura B.9: Página principal de la aplicación

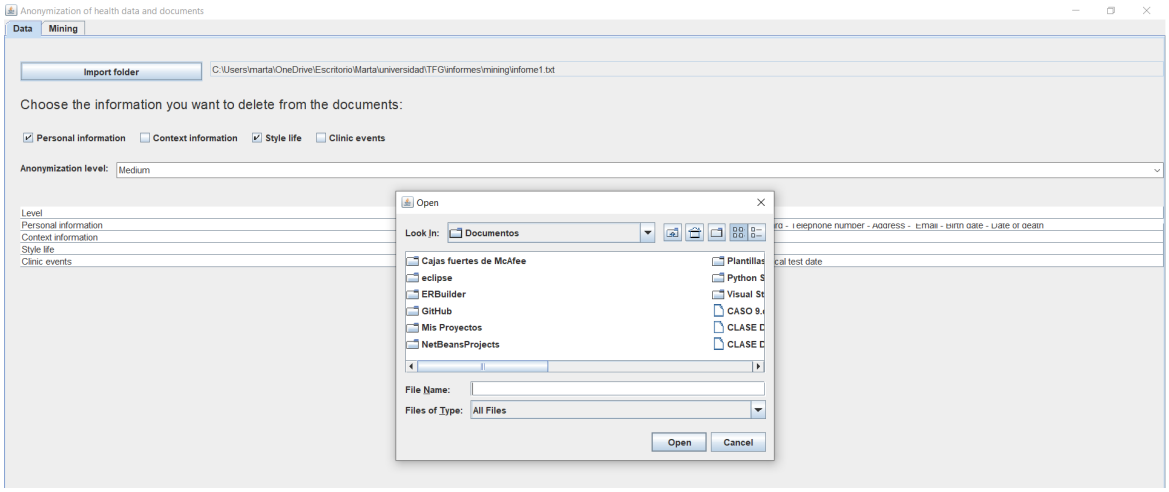


Figura B.10: Página que aparece al importar un archivo

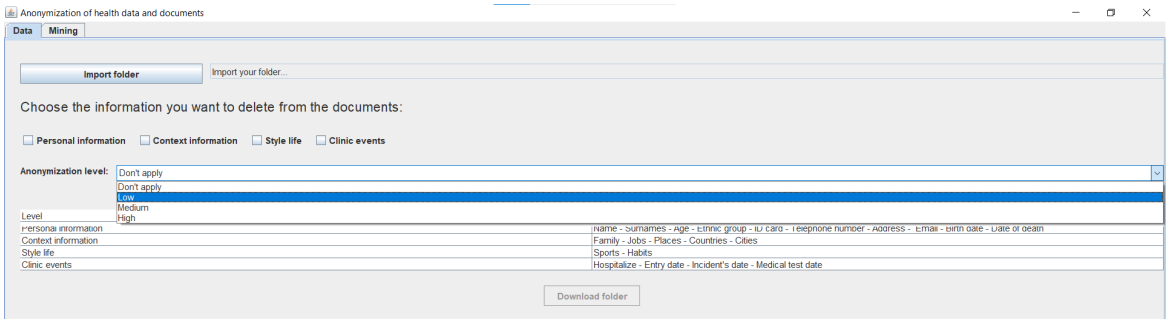


Figura B.11: Selector de nivel de anonimización

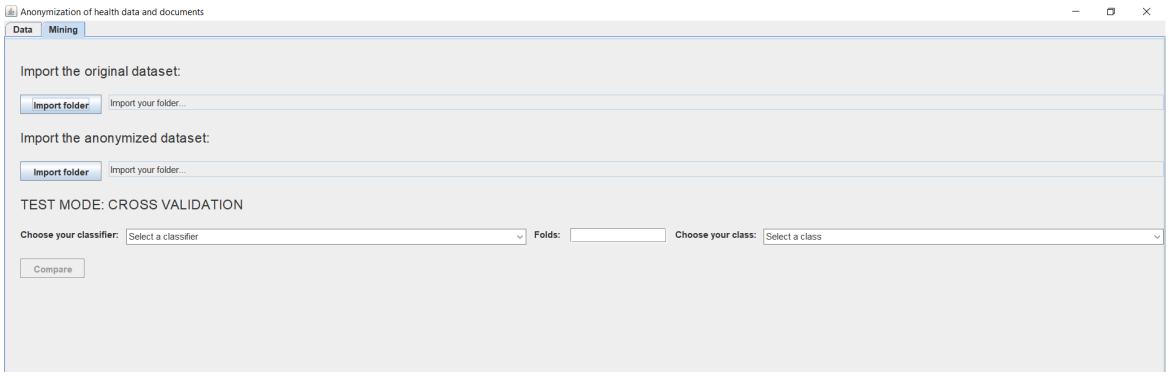


Figura B.12: Página principal de ventana *Mining*

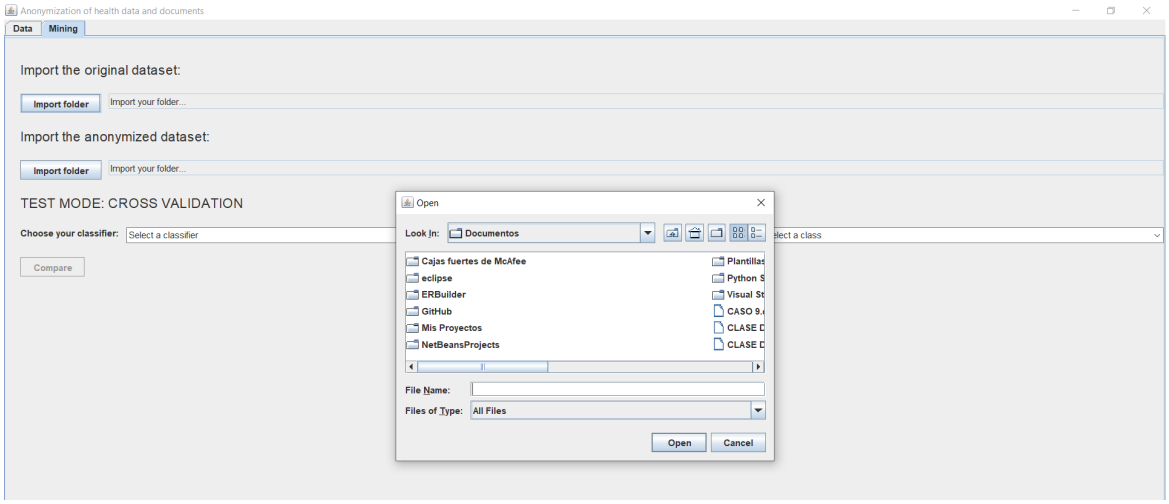


Figura B.13: Página que aparece al importar los documentos en la ventana de *Mining*

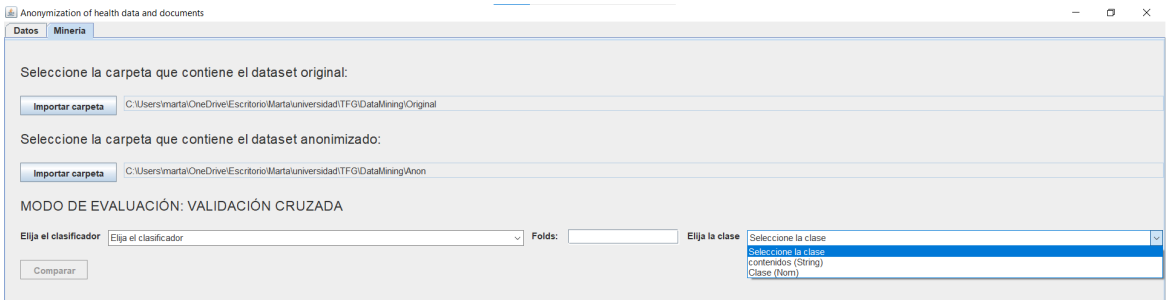


Figura B.14: Seleccionar la clase

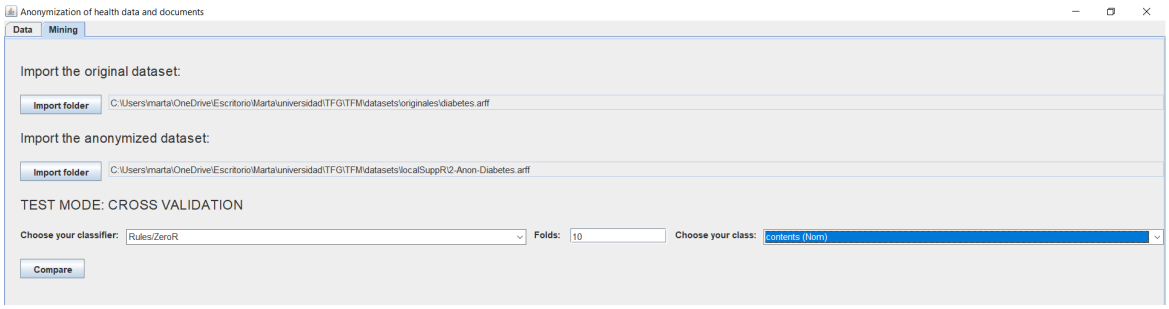


Figura B.15: Botón de *Compare* activado

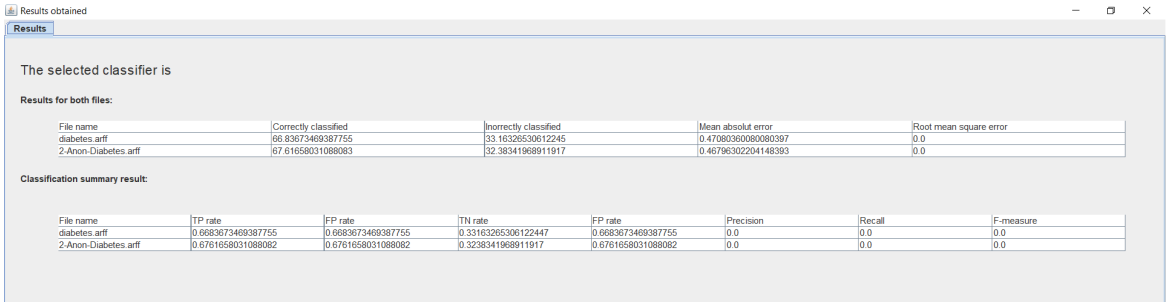
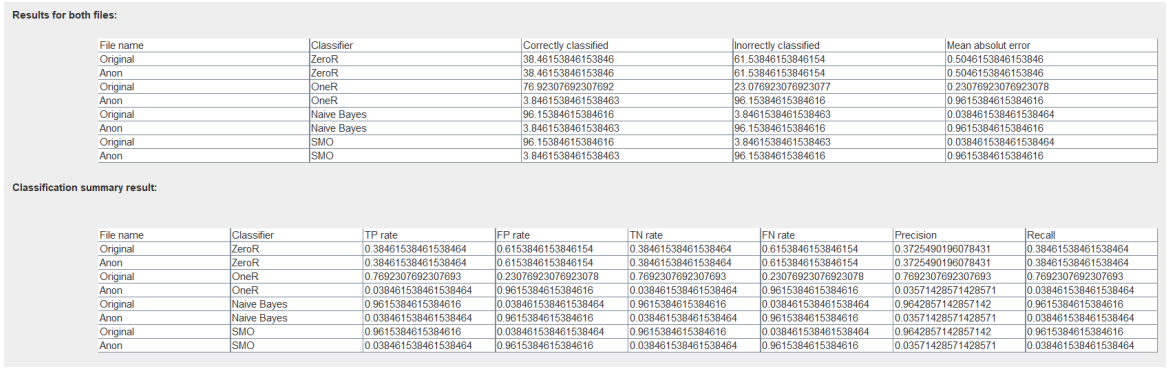


Figura B.16: Página final con los resultados obtenidos en la evaluación



Results for both files:

File name	Classifier	Correctly classified	Inorrectly classified	Mean absolut error
Original	ZeroR	38.46153846153846	61.53846153846154	0.5046153846153846
Anon	ZeroR	38.46153846153846	61.53846153846154	0.5046153846153846
Original	OneR	76.92307692307692	23.076923076923077	0.2307692307692308
Anon	OneR	3.8461538461538463	96.15384615384616	0.9615384615384616
Original	Naive Bayes	96.15384615384616	3.8461538461538463	0.038461538461538464
Anon	Naive Bayes	3.8461538461538463	96.15384615384616	0.9615384615384616
Original	SMO	96.15384615384616	3.8461538461538463	0.038461538461538464
Anon	SMO	3.8461538461538463	96.15384615384616	0.9615384615384616

Classification summary result:

File name	Classifier	TP rate	FP rate	TN rate	FN rate	Precision	Recall
Original	ZeroR	0.38461538461538464	0.6153846153846154	0.38461538461538464	0.6153846153846154	0.3725490196078431	0.38461538461538464
Anon	ZeroR	0.38461538461538464	0.6153846153846154	0.38461538461538464	0.6153846153846154	0.3725490196078431	0.38461538461538464
Original	OneR	0.7692307692307693	0.23076923076923078	0.7692307692307693	0.23076923076923078	0.7692307692307693	0.7692307692307693
Anon	OneR	0.038461538461538464	0.9615384615384616	0.038461538461538464	0.9615384615384616	0.03571428571428571	0.038461538461538464
Original	Naive Bayes	0.9615384615384616	0.038461538461538464	0.9615384615384616	0.038461538461538464	0.9642857142857142	0.9615384615384616
Anon	Naive Bayes	0.038461538461538464	0.9615384615384616	0.038461538461538464	0.9615384615384616	0.03571428571428571	0.038461538461538464
Original	SMO	0.9615384615384616	0.038461538461538464	0.9615384615384616	0.038461538461538464	0.9642857142857142	0.9615384615384616
Anon	SMO	0.038461538461538464	0.9615384615384616	0.038461538461538464	0.9615384615384616	0.03571428571428571	0.038461538461538464

Figura B.17: Página final con los resultados obtenidos en la evaluación con todos los clasificadores

B.2. Mapa de navegación

La Figura B.18 muestra el mapa de navegación de la ventana denominada *Data*, la especializada en realizar la configuración para la posterior anonimización de los documentos. La ventana principal es la que está marcada con una flecha de color azul. Desde ella, el usuario puede cargar los documentos a anonimizar, seleccionar los atributos a eliminar y elegir el nivel de anonimización a aplicar en el proceso.

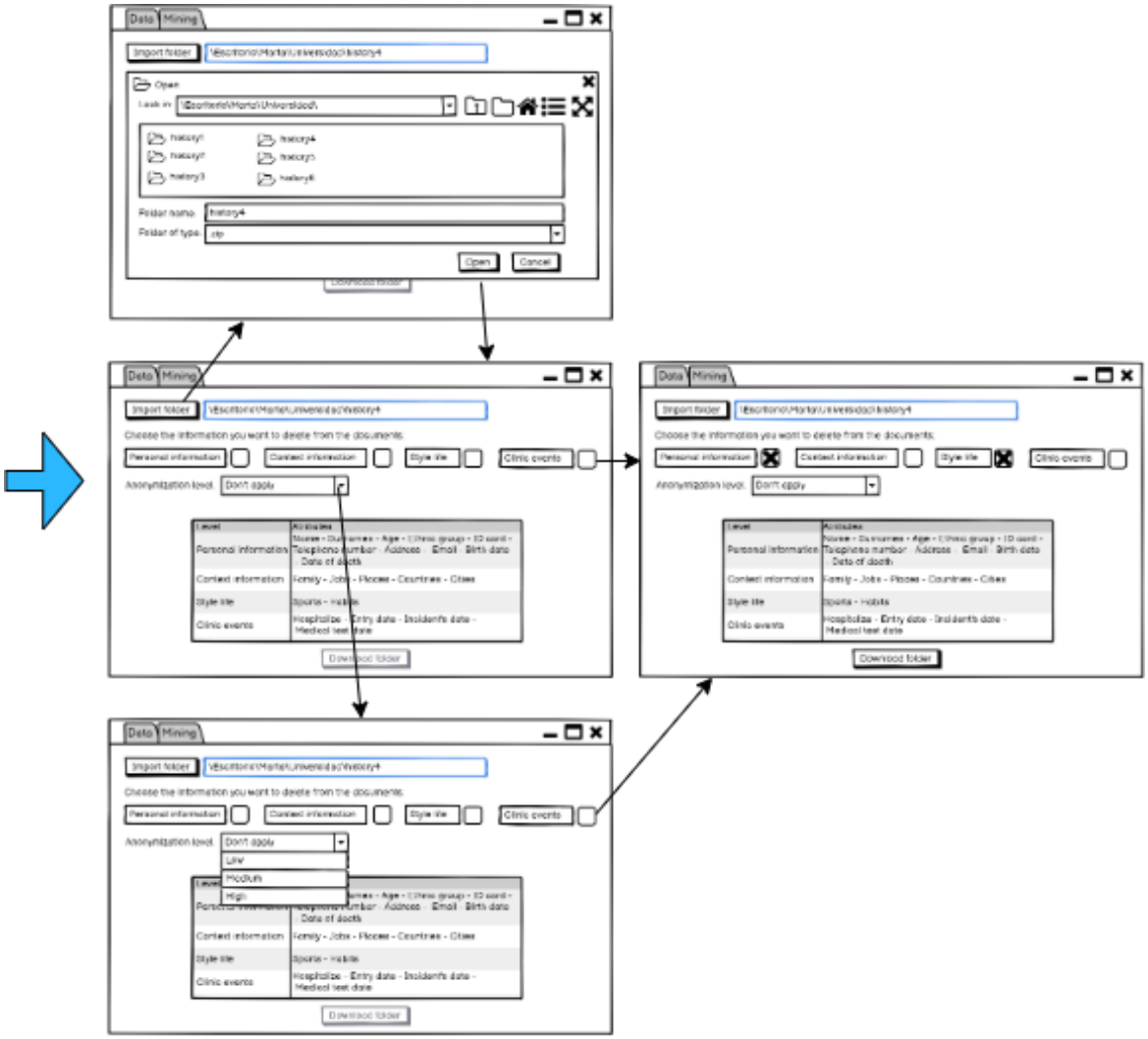


Figura B.18: Mapa de navegación de la ventana denominada *Data*

En la parte superior aparece la ventana denominada *Mining*. La página principal está señalada con una flecha de color azul. Desde esta ventana, el usuario puede cargar tanto el fichero original como el anominizado, seleccionar el selector a utilizar en la validación cruzada, el número de "folds" y la clase o categoría sobre la que aplicar la evaluación. Una vez el usuario haya seleccionado todo esto, podrá acceder a la ventana de resultados que aparecerá cuando pulse el botón *"Compare"*. Esto se puede ver en la Figura B.19.

Además, el usuario puede volver a las pestañas anteriores siempre que lo desee. Pulsando sobre la pestaña superior denominada *Data* puede volver a la parte de anonimización de datos. Del mismo modo, si pulsa sobre la pestaña superior denominada *Mining*, puede volver a la ventana dedicada a minería de datos.

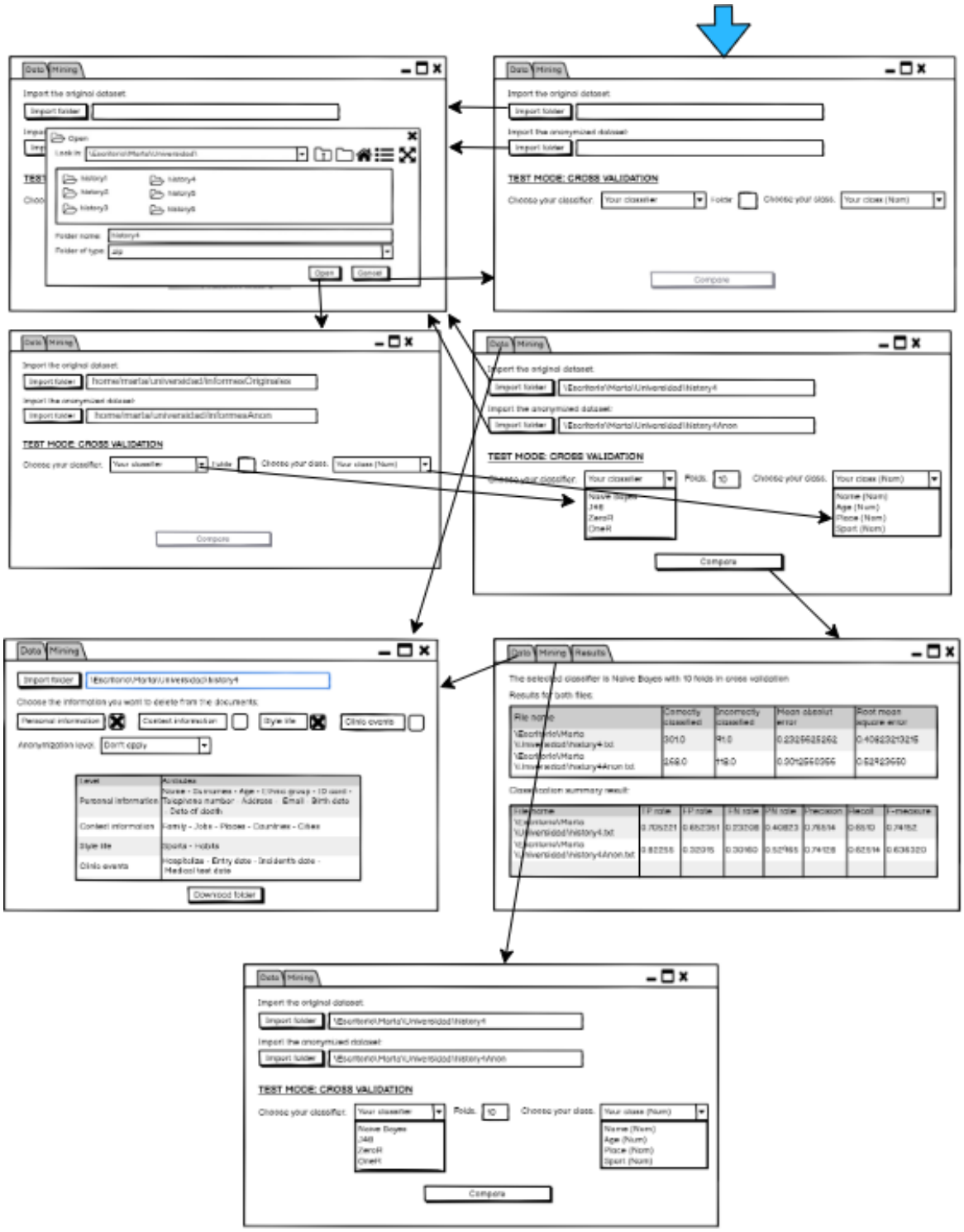


Figura B.19: Mapa de navegación de la ventana denominada *Mining*

B.3. Diagrama de clases

En la Figura B.20 se muestra el diagrama de clases que representa el proceso de anonimización de los documentos. La clase *PanelData* es la clase principal que gestiona el proceso de anonimización. Su método *analyse* es el que gestiona qué opciones de anonimización se han seleccionado. Si únicamente se seleccionan las opciones de conjuntos de atributos: *Personal information, Context information, Style life y Clinic events*, será la clase *TestingPatterns* la que se encargará de realizar la anonimización. Si por el contrario, solo se selecciona la opción de nivel de anonimización, será la clase *TestingAnon* la que realice la anonimización. Como tercera y última opción, si se seleccionan ambas opciones, la anonimización la realizará la clase *TestingComninated*.

La clase *InformationTypes* gestiona los conjuntos de atributos. Del mismo modo, la clase *AnonLevels* gestiona el nivel de anonimización seleccionado. Ambas dos utilizarán la clase *ContainsPattern* para la detección y eliminación de datos a través de expresiones regulares. También usarán la clase *BrowseInDictionaries* para realizar búsquedas de atributos en los diccionarios implementados. Por último, la clase *ReadFile* se encarga de la detección y eliminación de atributos que aparecen en los ficheros *.csv*.

Por otro lado, el análisis sintáctico se realizará a través de la clase *JavaRunSubject*. Tanto esta clase como *JavaRunCommand* hacen uso de la librería *Spacy* para el procesamiento del lenguaje natural.

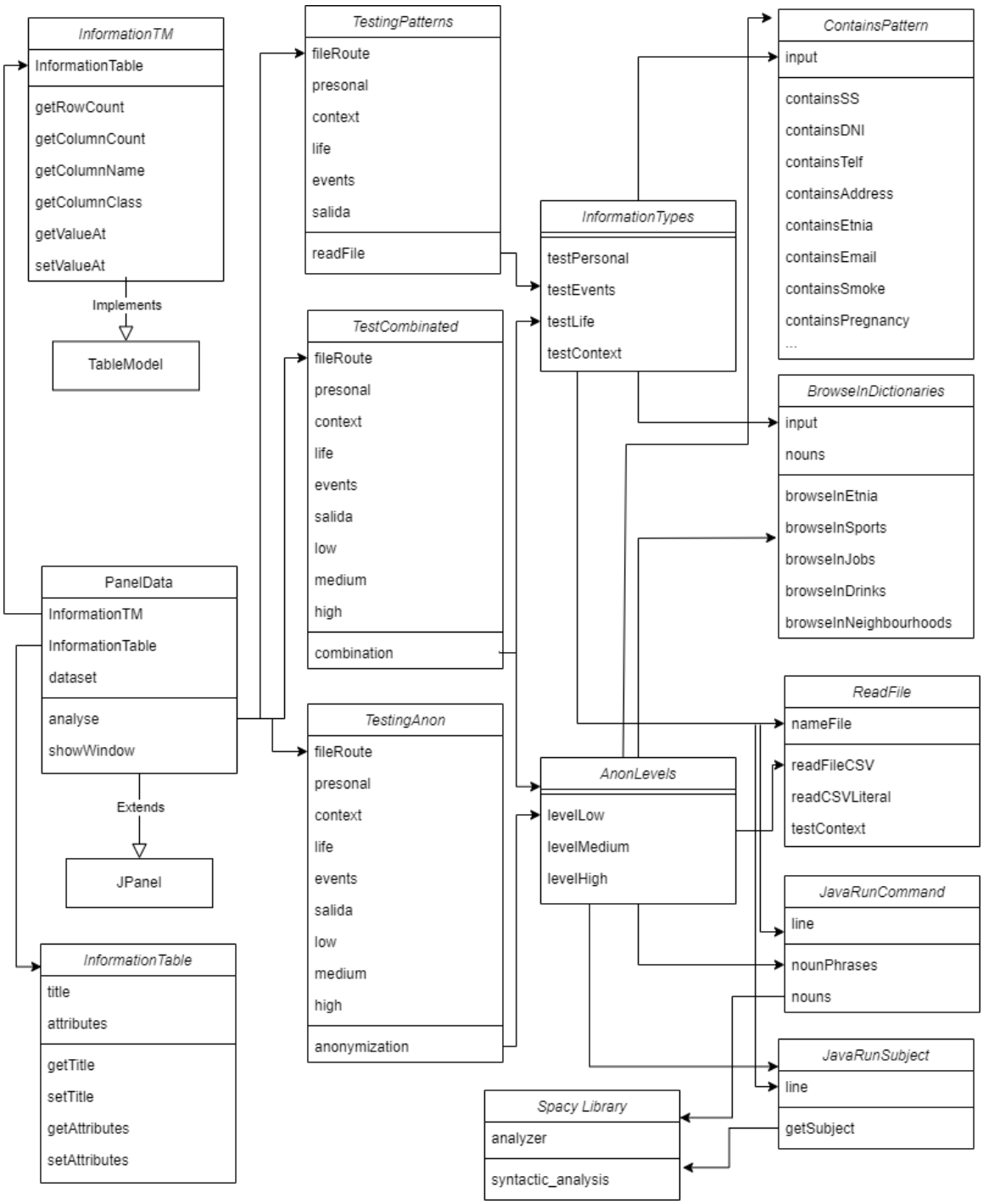


Figura B.20: Diagrama de clases del proceso de anonimización de datos

B.4. Diseño e implementación de la aplicación

El diseño de la interfaz de la aplicación se realizó con la herramienta *Balsamiq Cloud*. En este prototipo inicial se puede ver que la aplicación consta de dos ventanas diferenciadas. La primera de ellas se denomina *"Data"* y en ella se encuentran todos los requisitos que están relacionados con el proceso de anonimación de los textos. Es decir, en ella se puede importar la carpeta a anonimizar, seleccionar el nivel de anonimización y el conjunto de atributos sobre los que aplicar la anonimización, así como descargar la carpeta con los ficheros anonimizados.

Esta aplicación ha sido implementada en el lenguaje de programación *Java*. La aplicación consta de unas clases principales que realizan las funcionalidades principales:

- **PanelData**: es la clase que se encarga de mostrar la interfaz de la ventana denominada *"Data"* y, a su vez, de gestionar los botones de la misma.

- **TestingPatterns**: esta clase se ejecuta en el caso de que el usuario haya seleccionado la opción *"Don't apply"* como nivel de anonimización. Por tanto, la anonimización se aplicará sobre el conjunto de atributos que el usuario haya seleccionado: información personal, información de contexto, estilo de vida y/o eventos clínicos.

- **TestCombinated**: esta clase se ejecuta en el caso de que el usuario haya seleccionado tanto algún conjunto de atributos sobre los que aplicar anonimización como un nivel de anonimización concreto.

- **TestingAnon**: esta clase se ejecuta en el caso de que el usuario haya seleccionado únicamente un nivel de anonimización ("High", "Medium" o "Low") sin seleccionar un conjunto de atributos.

- **AttributesCombinatedLow**: esta clase detecta aquellos atributos que aumentan su sensibilidad cuando aparecen de manera conjunta con otros atributos. En este caso, detecta los que se clasifican con sensibilidad alta.

- **AttributesCombinatedMedium**: esta clase detecta aquellos atributos que aumentan su sensibilidad cuando aparecen de manera conjunta con otros atributos. En este caso, detecta los que se clasifican con sensibilidad media.

- **InformationTypes**: esta clase contiene los diferentes métodos necesarios a ejecutar en el caso de que haya que eliminar grupos de atributos.

- **Preprocesing**: esta clase realiza el preprocesado de las oraciones, eliminado las denominadas *"empty words"*, los espacios en blanco, las tildes, los signos de puntuación, etc.

- **ContainsPattern**: es aquella clase que contiene los métodos correspondientes a las diferentes expresiones regulares utilizadas.

- **BrowseInDictionaries**: esta clase contiene los métodos pertinentes para realizar búsquedas en los diccionarios de la aplicación.

- **ReadFile**: esta clase lee ficheros ".csv" y compara su contenido con el que está siendo procesado.

- **JavaRunCommand**, **JavaRunLocations**, **JavaRunNames** y **JavaRunSurnames**: estas clases se utilizan cuando se necesita hacer uso de la herramienta de procesamiento de lenguaje natural *Spacy* para detectar algún tipo de palabra concreta. Estas clases ejecutan un fichero ".py" que realiza el análisis requerido por *Spacy* y recoge la salida para procesarla.

- **JavaRunSubject**: esta clase es similar a las anteriores. Sin embargo, mientras que las otras realizan un análisis semántico, esta función realiza un análisis sintáctico de la oración que se está procesando.

Se ha comentado que la aplicación se ha implementado en el lenguaje de programación *Java*, pero la herramienta de procesamiento de lenguaje natural *Spacy* trabaja con el lenguaje de programación *Python*. Para poder utilizarla dentro de la aplicación *Java* se ha utilizado la librería *Jython*.

Por otro lado, la aplicación consta de una segunda ventana denominada *"Mining"*, en la cual se encuentran los requisitos relacionados con minería de datos: cargar los ficheros, seleccionar el clasificador, el número de "folds" y la clase o categoría y visualizar los resultados obtenidos de la evaluación.

Las clases más relevantes en este caso son las siguientes:

- **PanelMining**: muestra la ventana de *"Mining"* y realiza la validación cruzada.

- **PanelResults**: muestra la ventana con los resultados obtenidos en la evaluación.

- **Results**: guarda los resultados obtenidos mediante la técnica de evaluación.

Anexos C

Manual de instalación y de usuario

En este anexo se encuentran los manuales de instalación y de usuario de la aplicación. También se muestra cómo generar un fichero *.jar* del proyecto. En el Anexo C.1 se muestra el manual de instalación de la aplicación. Dicha aplicación es una versión especial para ejecutar sobre consola de comandos. En el Anexo C.2 se muestran los distintos manuales de usuario de la aplicación. En el Anexo C.3 se explica cómo generar un fichero .jar de la aplicación. Finalmente, en el Anexo C.4 se explica cómo se han llevado a cabo las pruebas automatizadas de la parte de minería de datos.

C.1. Manual de instalación

Este manual está especificado para el sistema operativo Ubuntu con versión 20.04.3 LTS. Junto con este manual se encuentran las carpetas y ficheros que se van a necesitar para la completa configuración del entorno de ejecución.

En primer lugar, se va a comenzar con la instalación de las herramientas, lenguajes de programación y librerías necesarias para ejecutar correctamente la aplicación.

Por un lado, dado que se trata de una aplicación implementada en el lenguaje de programación *Java*, se requiere de su disposición en el equipo. Del mismo modo, la aplicación hace uso de una serie de ficheros implementados en el lenguaje de programación *Python*. A continuación, se presentan las versiones que se requieren de cada uno de ellos.

- **Java 11**. Versión: 11.0.11. o superior.
- **Python 3**. Versión: 3.9.14. o superior.

A continuación, se muestran los comandos necesarios para realizar la instalación del lenguaje de programación *Java*.

```
sudo apt update
sudo apt install default-jre
```

Es posible que al instalar este lenguaje de programación por primera vez, aparezcan ciertos errores. Uno de ellos puede deberse a las dos PPAs *(Personal Packages Archives)*. Para poder solventarlo con éxito se deben eliminar las dos PPAs y ejecutar de nuevo los comandos mostrados anteriormente. Eliminación de las PPAs:

```
sudo apt-add-repository -r ppa:gnome3-team/gnome3
sudo apt-add-repository -r ppa:philip.scott/spice-up-daily
sudo apt update
```

Una vez el lenguaje de programación *Java* ha sido instalado correctamente, se puede proceder a la instalación del lenguaje *Python*. En este caso se instalará *Python3*. A continuación, se muestran los comandos necesarios para la instalación de este lenguaje.

```
sudo apt update
sudo apt install software-properties-common
sudo add-apt-repository ppa:deadsnakes/ppa
sudo apt install python3.9
python3.9 --version
```

Algunos de los ficheros que se han programado en *Python* hacen uso de una librería denominada *Pattern*. Esta librería puede estar incluida con la instalación anterior de *Python3* o no. Para su instalación hay que ejecutar lo siguiente:

```
pip install pattern3
```

A continuación, se debe instalar una de las librerías que se van a utilizar. Para instalar la librería de *Jython* se debe hacer lo siguiente:

```
sudo apt install jython
```

Finalmente, la aplicación hace uso de una herramienta de procesamiento del lenguaje natural denominada *Spacy*. Esta herramienta debe ser instalada correctamente. A continuación, se muestran los comandos necesarios para la instalación de *Spacy* para el sistema operativo *Linux*, en lenguaje castellano a través de pipes.

```
pip install -U pip setuptools wheel
pip install -U spacy
python -m spacy download es_core_news_sm
```

La instalación de *Spacy* puede realizarse para otros sistemas operativos mediante pipes, conda o "from source". Los pasos a seguir para este procedimiento vienen explicados en su página oficial: https://spacy.io/usage. Es de vital importancia que el lenguaje seleccionado sea "Spanish".

Una vez terminado el proceso de instalación de las herramientas necesarias para realizar la ejecución, se puede proceder a la configuración del entorno.

Dentro del fichero ".zip" donde se hallaba este manual también se pueden encontrar una carpeta denominada *"Pruebas"*. Esta carpeta contiene todas las carpetas y ficheros necesarios para realizar la ejecución, así como el fichero ejecutable de la aplicación.

En el interior de la carpeta *"Pruebas"* podemos encontrar los siguientes archivos/carpetas:

- **Carpeta *"Colecciones"*:** en ella se ubican las colecciones de datos que la aplicación consulta durante su ejecución.
- **Carpeta *"Resultados"*:** en ella aparecerán los ficheros anonimizados cuando termine el proceso de anonimización.
- **Archivo *"Configuration.txt"*:** fichero de configuración de aplicación. En él se podrá configurar el nivel de anonimización que se quiere aplicar y/o seleccionar el conjunto de atributos que se desean anonimizar del fichero introducido. Además, habrá que indicar en él las rutas de los ficheros a anonimizar, así como las rutas de las carpetas anteriormente descritas.
- **Ficheros *.py*:** son los ficheros implementados en el lenguaje de programación *Python* que la aplicación requiere para su correcto funcionamiento. Sus nombres son los siguientes: *"analyzer.py"*, *"analyze_names.py"*, *"analyze_names_surnames.py"*, *"personas.py"* y *"syntactic_analysis.py"*.
- **Carpeta *"Informes"*:** en esta carpeta puedes incluir, opcionalmente, los ficheros a anonimizar con la aplicación. Los ficheros pueden ser de extensión *"txt"* o *".zip"*. El programa detectará automáticamente de cuál de ellos se trata.
- **Ejecutable *App.jar*:** se trata de la aplicación a ejecutar en formato jar.

Siguiendo la descripción que se acaba de describir, el entorno debe quedar parecido a lo que se muestra en la Figura C.1.

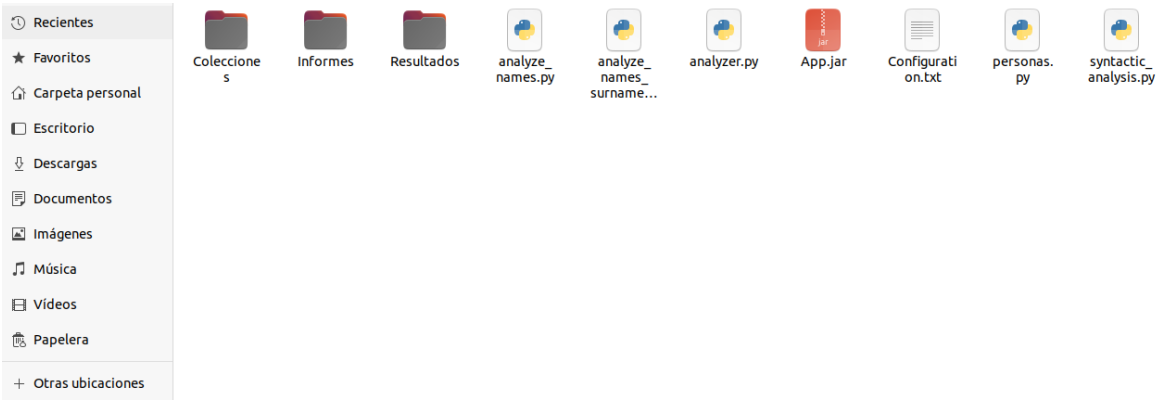


Figura C.1: Entorno final

Una vez llegado a este punto, el entorno ya está correctamente configurado para poder ejecutar la aplicación. También se ha construido una máquina virtual con el sistema operativo Ubuntu con versión 20.04.3 LTS que reproduce un entorno completo de ejecución sin necesidad de seguir los pasos anteriores.

C.2. Manual de usuario

C.2.1. Versión con fichero de configuración

En el caso de que se vaya a ejecutar la aplicación desde la máquina virtual proporcionada, se debe descargar e importar en *Oracle VM VirtualBox*. La descarga puede llevar unos minutos.

Una vez esté descargada e importada en *Oracle VM VirtualBox*, se debe acceder a ella mediante el usuario *"usuario"* y la contraseña *"usuario"*. Dentro se podrá ver el escritorio. En la carpeta personal hay una carpeta denominada *Pruebas*, donde se encuentran todos los ficheros y carpetas necesarias para el correcto funcionamiento de la aplicación, así como la aplicación en cuestión. En el Anexo C.1 se explica el contenido de dicha carpeta.

El fichero de configuración debe especificar las opciones correspondientes al tipo de anonimización que se desea aplicar. En la Figura C.2, se muestra el contenido del fichero de configuración.

Como se puede ver en la Figura C.2, el primer grupo a seleccionar hace referencia al conjunto de atributos que se desean anonimizar: "Personal information", "Context information", "Style life" y/o "Events information". En este caso, se debe marcar con

```
Write an X on those attributes you want to anonymize:

-Personal information:
-Context information:
-Style life information:
-Events information:

Write an X on the level of anonymization you want to apply: (You only can choose one of these four options. You must choose one.)
-Do not apply:
-Low level:
-Medium level:
-High level:

Write the path to the file which you want to get anonymized in the next line:

Write the path to the results folder in the next line:

Write the path to the collections folder in the next line:
```

Figura C.2: Ejemplo de fichero de configuración

una X aquellos atributos que se deseen anonimizar. Pueden ser uno o más conjuntos de atributos.

Para que quede claro qué atributos recoge cada grupo, se presenta la Tabla C.1.

Grupo	**Atributos**
Información personal	Nombre, apellidos, edad, etnia, dni, teléfono, domicilio, correo electrónico, fecha de nacimiento, fecha de fallecimiento, número de la Seguridad Social, número de colegiado y género.
Información de contexto	Familiares, trabajo, lugares, países, ciudades y municipios.
Información sobre estilo de vida	Deportes y hábitos.
Información sobre eventos clínicos	Hospitalización, fecha de ingreso, fecha de urgencias y fecha de alta.

Tabla C.1: Clasificación en grupos de los datos encontrados

Respecto al segundo grupo que presenta el fichero de configuración, en él se selecciona el nivel de anonimización que se desea aplicar: “Low level”, “Medium level”, “High level” o “Don’t apply”. A diferencia del caso anterior, solo se puede seleccionar uno de estos niveles y, en el caso de que no se desee aplicar ninguno, se debe seleccionar la opción “Don’t apply”.

Respecto a los atributos que recoge cada grupo de esta sección:

- **Don’t apply:** no se aplica ningún nivel de anonimización.
- **Low level:** se recogen aquellos atributos que tienen sensibilidad alta y aquellos que tienen otro nivel de sensibilidad inferior, pero que aparecen junto a otros que hacen que su sensibilidad se incremente a alta.

– **Medium level:** se recogen aquellos atributos que tienen sensibilidad alta o media y aquellos que tienen otro nivel de sensibilidad inferior, pero que aparecen junto a otros que hacen que su sensibilidad se incremente a media o alta.

– **High level:** se recogen todos los atributos, incluidos los que tienen sensibilidad baja.

A continuación de estos grupos, se encuentran los campos a rellenar sobre la información que se le debe proporcionar a la aplicación:

– **Ficheros a anonimizar:** se debe indicar la ruta del fichero a anonimizar. Opcionalmente se pueden incluir en la carpeta *"Informes"* que venía incluida en el fichero ".zip" que se ha descargado.

– **Carpeta de resultados:** se debe indicar la ruta de la carpeta *"Resultados"*, donde se obtendrán los ficheros anonimizados tras la ejecución. Es recomendable que esta carpeta esté siempre vacía antes de lanzar la aplicación. Si se quieren guardar en otra carpeta, se puede poner la ruta de la carpeta que se desee.

– **Carpeta de colecciones:** se debe indicar la ruta de la carpeta *"Colecciones"* que venía incluida en el fichero ".zip" que se ha descargado.

Se debe recalcar que dependiendo del sistema operativo en el que se vaya a ejecutar la aplicación, la ruta debe tener un formato u otro. Es decir, si el sistema operativo que se ejecuta es *Windows*, la barra que separa las carpetas debe ser doble: "\". Por el contrario, si se trata de los sistemas operativos *Linux* o *MacOS*, la barra debe ser una sola, pero invertida: "/". Esto se muestra en las Figuras C.3 y C.4.

Windows:

```
Write the path to the file which you want to get anonymized in the next line:
C:\Users\marta\OneDrive\Escritorio\Marta\universidad\TFG\comprimir\comprimir.zip
Write the path to the results folder in the next line:
C:\Users\marta\OneDrive\Escritorio\Marta\universidad\TFG\Entrada
Write the path to the collections folder in the next line:
C:\Users\marta\OneDrive\Escritorio\Marta\universidad\TFG\Colecciones
```

Figura C.3: Ejemplo de ruta para Windows

Linux/MacOS:

Una vez escrito todo esto, el fichero de configuración *"Configuration.txt"* debe tener una apariencia similar a la mostrada en la Figura C.5.

```
La ruta del fichero introducido es la siguiente: /home/usuario/Pruebas/Informes/prueba2.txt
La ruta de la carpeta de salida es la siguiente: /home/usuario/Pruebas/Resultados
La ruta de la carpeta de colecciones es la siguiente: /home/usuario/Pruebas/Colecciones
```

Figura C.4: Ejemplo de ruta para Linux

```
Write an X on those attributes you want to anonymize:

-Personal information: X
-Context information:
-Style life information:
-Events information:

Write an X on the level of anonymization you want to apply:
(You only can choose one of these four options. You must choose one.)
-Do not apply: X
-Low level:
-Medium level:
-High level:

Write the path to the file which you want to get anonymized in the next line:
C:\Users\marta\OneDrive\Escritorio\Marta\universidad\TFG\comprimir\comprimir.zip
Write the path to the results folder in the next line:
C:\Users\marta\OneDrive\Escritorio\Marta\universidad\TFG\Entrada
Write the path to the collections folder in the next line:
C:\Users\marta\OneDrive\Escritorio\Marta\universidad\TFG\Colecciones
```

Figura C.5: Fichero de configuración completado

Ahora ya está todo listo para poder ejecutar la aplicación. Se le debe pasar como argumento la ruta del fichero de configuración *"Configuration.txt"*. Los comandos necesarios para lanzarla son los siguientes:

```
1 java -jar App.jar /home/usuario/Pruebas/Configuration.txt
```

Cuando la aplicación haya terminado de ejecutarse, mostrará por pantalla lo que se muestra en la Figura C.6. Como muestra la figura, los ficheros anonimizados se encuentran en la carpeta que haya indicado en el fichero de configuración como *"carpeta de resultados"*. Aún así, el programa informa cuál es la ruta que ha recibido como *"carpeta de resultados"*.

```
usuario@ubuntu-20:~/Pruebas$ java -jar App.jar /home/usuario/Pruebas/Configuration.txt
Este es un fichero de configuración
Se ha seleccionado la opción: -Personal information: X
Se ha seleccionado la opción: -Do not apply: X
La ruta del fichero introducido es la siguiente: /home/usuario/Pruebas/Informes/prueba2.txt
La ruta de la carpeta de salida es la siguiente: /home/usuario/Pruebas/Resultados
La ruta de la carpeta de colecciones es la siguiente: /home/usuario/Pruebas/Colecciones
Se trata de un fichero TXT
Su fichero /home/usuario/Pruebas/Informes/prueba2.txt se está procesando
POR FAVOR, ESPERE UNOS SEGUNDOS
Se ha terminado de anonimizar la línea 1
Número total de líneas anonimizadas: 1
PROCESO DE ANONIMIZACIÓN TERMINADO. ENCONTRARÁ SUS ARCHIVOS EN EL DIRECTORIO /home/usuario/Pruebas/Resultados
usuario@ubuntu-20:~/Pruebas$
```

Figura C.6: Proceso de ejecución terminado

C.2.2. Manual de usuario de la versión con GUI

Como se ha comentado a lo largo del proyecto, la herramienta consta de dos partes diferenciadas: proceso de anonimización y minería de datos. El proceso de anonimización se encuentra en la ventana de *"Data"*. En esta ventana se pueden cargar los documentos a procesar. Se puede tratar de un único fichero en formato *".txt"* o un conjunto de ficheros de texto comprimidos en un fichero *".zip"*.

A continuación, el usuario debe seleccionar el o los conjuntos de atributos que deseea eliminar. Estos atributos se agrupan en cuatro grupos: *"Personal information", "Context information", "Style life" y "Clinic events"*. El usuario puede seleccionar uno, varios o ninguno de los grupos. Los atributos que recoge cada grupo vienen especificados en la Tabla C.2.

Grupo	**Atributos**
Información personal	Nombre, apellidos, edad, etnia, dni, teléfono, domicilio, correo electrónico, fecha de nacimiento, fecha de fallecimiento, número de la Seguridad Social, número de colegiado y género.
Información de contexto	Familiares, trabajo, lugares, países, ciudades y municipios.
Información sobre estilo de vida	Deportes y hábitos.
Información sobre eventos clínicos	Hospitalización, fecha de ingreso, fecha de urgencias y fecha de alta.

Tabla C.2: Clasificación en grupos de los datos encontrados

El siguiente apartado que puede seleccionar el usuario es el nivel de anonimización, el cual es totalmente complementario al anterior. El sistema dispone de tres niveles de anonimización: *"Low level", "Medium level" y "High level"*. El nivel *"Low"* recoge aquellos atributos que tienen una sensibilidad alta, el nivel *"Medium"* recoge aquellos que tienen sensibilidad media o alta y, finalmente, el nivel *"High"* recoge todos los atributos, ya que recoge los clasificados con sensibilidad alta, media o baja. Además, el usuario puede seleccionar la opción de *"Don't apply"*. Esta opción hace que no se aplique ningún nivel de anonimización sobre el documento a procesar y, por tanto, solo eliminará los atributos recogidos en los grupos de atributos que haya seleccionado el usuario en el apartado anterior.

Tras completar la configuración del proceso de anonimización que se va a realizar, se pulsa el botón de *Download folder*. Cuando el proceso haya finalizado, podrá

encontrar en la carpeta *"Resultados"*, los documentos anonimizados.

Por otro lado, la parte de minería se encuentra en la ventana de *Mining*, donde se puede configurar la técnica de evaluación de *Cross Validation* que se va a llevar a cabo. Antes de explicar cómo se organiza la GUI de minería de datos, se debe resaltar cómo estructurar las carpetas para que la evaluación sea un éxito.

Partiendo del conjunto de datos que se quieren someter a la evaluación, se debe crear una carpeta por cada una de las etiquetas que se quieran plantear en la evaluación, de manera que en cada una de estas carpetas residan los documentos que pertenezcan a dicha etiqueta. Por ejemplo, si se desea realizar una clasificación entre documentos personales y no personales, se deben crean dos carpetas. Una carpeta denominada, por ejemplo, *True*, que guarde los documentos que sean de carácter personal y otra carpeta denominada, por ejemplo, *False*, que guarde los documentos no personales. Así se realiza el etiquetado. En otra carpeta distinta, se debe realizar el mismo procedimiento para el dataset anonimizado.

La herramienta permite cargar estas carpetas. Es por esto que se debe cargar la carpeta padre donde residen las subcarpetas que representan las etiquetas de la evaluación. Una vez subidas ambas carpetas a la herramienta, se debe realizar la configuración de la evaluación a realizar. En primer lugar, se debe seleccionar el clasificador, el número de *"folds"* y la clase. En el caso del clasificador, se puede aplicar *ZeroR, OneR, Naive Bayes y SMO*. Además, si seleccionas la opción de *"All classifiers"*, se realizará la evaluación con todos los clasificadores.

Una vez finalizada la configuración, se podrá hacer click en el botón de *"Compare"*. Cuando el proceso de evaluación finalice, aparecerá por pantalla una nueva ventana *"Results"* con los resultados obtenidos.

C.3. Generar fichero .jar

Esta versión de la aplicación es un proyecto *Maven* que permitirá generar un fichero *.jar* a través de un comando. Para ello, se debe acceder a la carpeta del proyecto *configurationProject*. Dentro de ella se debe ejecutar el siguiente comando:

```
mvn clean install package
```

Una vez se haya ejecutado, se debe acceder a la carpeta denominada *target*, donde se encontrará un fichero .jar denominado *"configurationProject-0.0.1-SNAPSHOT.jar"*. Este es el fichero final con el que se puede ejecutar la aplicación desarrollada.

C.4. Pruebas automatizadas

En esta sección se va a explicar cómo se han realizado las pruebas automatizadas de la parte de minería de datos. Se debe tener el conjunto de documentos original y el mismo conjunto de documentos anonimizado para realizar correctamente la evaluación.

Para realizar el etiquetado de los documentos, se debe crear una carpeta por cada una de las etiquetas que se deseen crear. Por ejemplo, en este caso se han creado dos carpetas: la carpeta *true*, donde se almacenan los documentos relacionados con datos personales; y la carpeta *false*, donde se almacenan los documentos relacionados con las guías clínicas. Este mismo procedimiento se debe realizar con los documentos anonimizados. Como resultado se debería tener una carpeta donde se guardan los datos originales en las subcarpeteras correspientes al etiquetado. Lo mismo en otra carpeta para los datos anonimizados.

A continuación, se procede a ejecutar la aplicación con la versión de GUI. Esta herramienta ofrece dos pestañas: *Data* y *Mining*. Se debe acceder a la pestaña de *Mining*. Esta pestaña se muestra en la Figura C.7. Esta pestaña permite al usuario cargar los datos con los que se va a realizar la evaluación. Para cargar correctamente los datos, se debe subir la carpeta padre que contiene las subcarpetas con el etiquetado. Esto se debe hacer tanto para los datos originales como para los anonimizados. A continuación, se debe configurar la evaluación. Dado que se va a realizar la evaluación automática, se debe seleccionar la opción *"All classifiers"* en el apartado de *"Choose your classifier"*. Esto se puede ver en la Figura C.8. También se indicará el número de *"folds"* y la clase a aplicar.

Al haber seleccionado la opción de *"All classifiers"*, se realizará la evaluación con los cuatro clasificadores que se ofrecen en la herramienta: *ZeroR, OneR, Naive Bayes y SMO*. Cuando este proceso finalice, se mostrarán por pantalla todos los resultados obtenidos en la evaluación con cada uno de los clasificadores y con cada uno de los conjuntos de datos. Esto se puede ver en la Figura C.9.

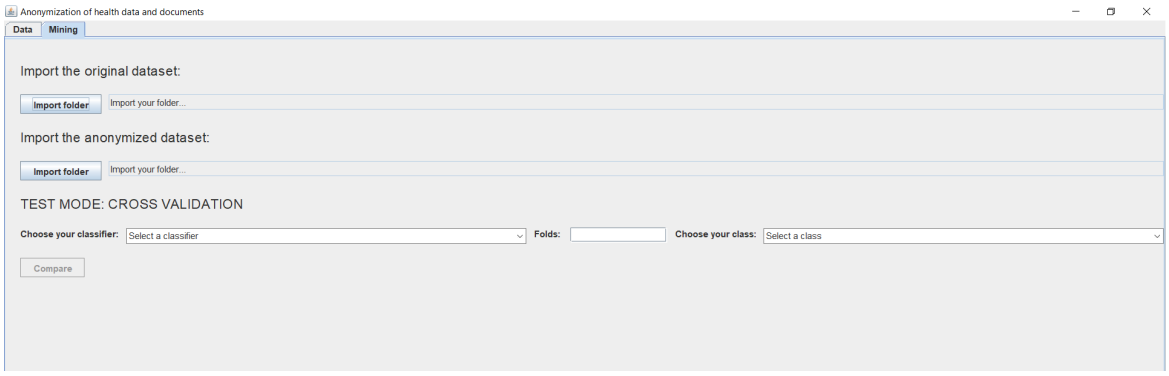


Figura C.7: Página principal de ventana *Mining*

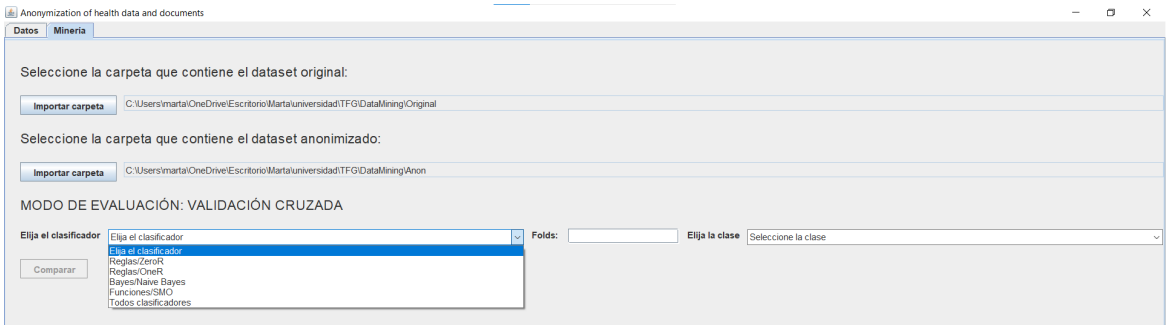


Figura C.8: Opción de "All classifiers"

Results for both files:

File name	Classifier	Correctly classified	Inorrectly classified	Mean absolut error
Original	ZeroR	38.46153846153846	61.53846153846154	0.5046153846153846
Anon	ZeroR	38.46153846153846	61.53846153846154	0.5046153846153846
Original	OneR	76.92307692307692	23.07692307692307	0.2307692307692308
Anon	OneR	3.8461538461538463	96.15384615384616	0.9615384615384616
Original	Naive Bayes	96.15384615384616	3.8461538461538463	0.038461538461538464
Anon	Naive Bayes	3.8461538461538463	96.15384615384616	0.9615384615384616
Original	SMO	96.15384615384616	3.8461538461538463	0.038461538461538464
Anon	SMO	3.8461538461538463	96.15384615384616	0.9615384615384616

Classification summary result:

File name	Classifier	TP rate	FP rate	TN rate	FN rate	Precision	Recall
Original	ZeroR	0.38461538461538464	0.6153846153846154	0.38461538461538464	0.6153846153846154	0.3725490196078431	0.38461538461538464
Anon	ZeroR	0.38461538461538464	0.6153846153846154	0.38461538461538464	0.6153846153846154	0.3725490196078431	0.38461538461538464
Original	OneR	0.7692307692307693	0.2307692307692308	0.7692307692307693	0.2307692307692308	0.7692307692307693	0.7692307692307693
Anon	OneR	0.038461538461538464	0.9615384615384616	0.038461538461538464	0.9615384615384616	0.03571428571428571	0.038461538461538464
Original	Naive Bayes	0.9615384615384616	0.038461538461538464	0.9615384615384616	0.038461538461538464	0.9642857142857142	0.9615384615384616
Anon	Naive Bayes	0.038461538461538464	0.9615384615384616	0.038461538461538464	0.9615384615384616	0.03571428571428571	0.038461538461538464
Original	SMO	0.9615384615384616	0.038461538461538464	0.9615384615384616	0.038461538461538464	0.9642857142857142	0.9615384615384616
Anon	SMO	0.038461538461538464	0.9615384615384616	0.038461538461538464	0.9615384615384616	0.03571428571428571	0.038461538461538464

Figura C.9: Resultados obtenidos con todos los clasificadores

Anexos D

Resultados obtenidos durante las pruebas

En este anexo se van a documentar las pruebas realizadas en este proyecto. Se van a mostrar tablas, gráficas y conclusiones más detalladas sobre los resultados obtenidos en la evaluación experimental del proyecto. En el Anexo D.1, se muestran los resultados obtenidos durante las pruebas relacionadas con el proceso de anonimización: aquellas relacionadas con conjuntos de atributos, sobre niveles de anonimización y combinadas. En el Anexo D.2, se muestran los resultados obtenidos en las pruebas de rendimiento. Finalmente, en el Anexo D.3, se muestran los resultados obtenidos en las pruebas realizadas en la parte de minería de datos: documentos personales y no personales, así como las relacionadas con ictus cerebrales y otras enfermedades.

D.1. Resultados obtenidos en las pruebas del proceso de anonimización

En este anexo se muestran los resultados obtenidos en las pruebas realizadas. En el Anexo D.1.1 se muestran los resultados obtenidos en las pruebas con conjuntos de atributos. En el Anexo D.1.2 se muestran los resultados obtenidos con las pruebas relacionadas con el nivel de anonimización. Finalmente, en el Anexo D.1.3 se muestran los resultados obtenidos con las pruebas combinadas.

D.1.1. Pruebas con conjuntos de atributos

Se van a realizar 15 pruebas, cada una de ellas representa una combinación distinta de los atributos que ofrece la aplicación. Los resultados obtenidos de las mismas figuran en la Tabla D.1.

Test	**TP**	**FP**	**FN**	**Recall**	**Algorithm quality**
Clinic events	13	0	5	0.72	2.6
Style life	1	0	0	1	ERROR
Style life y Clinic events	10	5	0	0.66	2
Context information	10	9	1	0.91	1
Context information y Clinic events	24	7	5	0.83	2
Context information y Style life	12	8	4	0.75	1
Context life, Style life y Clinic events	22	8	8	0.73	1.38
Personal information	31	12	15	0.67	1.82
Personal information y Clinic events	44	13	9	0.83	2
Personal information y Style life	28	19	9	0.76	1
Personal information, Style life y Clinic events	40	14	8	0.83	1.82
Personal information y Context information	35	19	9	0.80	1.25
Personal information, Context information y Clinic events	53	15	9	0.85	2.21
Personal information, Context information y Style life	35	19	12	0.74	1.13
Personal information, Context information, Style life y Clinic events	51	14	9	0.85	2.22

Tabla D.1: Resultados obtenidos de las pruebas relacionadas con conjuntos de atributos

Para ver con más claridad los resultados obtenido se va a mostrar la gráfica con los resultados de *Recall* en la Figura D.1. En ella se puede ver que todos alcanzan un nivel superior a 0.5, e incluso, superando el 0.75.

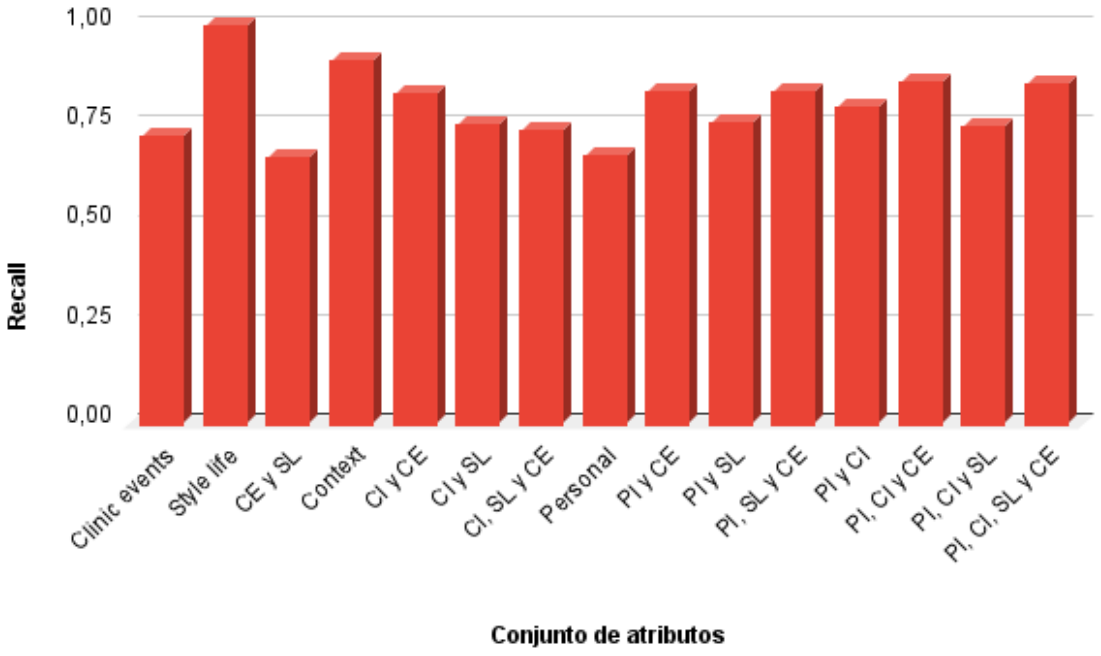


Figura D.1: Resultados de la métrica *Recall* obtenidos en las pruebas relacionadas con conjuntos de atributos

D.1.2. Pruebas relacionadas con el nivel de anonimización

Se van a realizar un total de tres pruebas, una para cada uno de los niveles de anonimización que ofrece la herramienta. Los resultados obtenidos se muestran en la Tabla D.2.

Test	**TP**	**FP**	**FN**	**Recall**	**Algorithm quality**
Low level	18	25	0	1	0.72
Medium level	31	19	0	1	1.63
High level	33	14	1	0.97	2.2

Tabla D.2: Resultados obtenidos en las pruebas relacionadas con el nivel de anonimización

D.1.3. Pruebas combinadas

Se han realizado 15 pruebas siguiendo las distintas combinaciones posibles que ofrece la herramienta añadiendo el nivel de anonimización *"Low level"*. Los resultados obtenidos se pueden ver en la Tabla D.3.

Test	**TP**	**FP**	**FN**	**Recall**	**Algorithm quality**
Clinic events	24	28	0	1	0.86

Style life	14	23	2	0.88	0.56
Style life y Clinic events	18	27	2	0.90	0.69
Context information	20	22	4	0.83	0.77
Context information y Clinic events	23	20	5	0.82	0.92
Context information y Style life	21	25	4	0.84	0.72
Context information, Style life y Clinic events	25	24	3	0.89	0.93
Personal information	28	27	0	1	1.04
Personal information y Clinic events	29	25	2	0.94	1.07
Personal information y Style life	31	27	0	1	1.15
Personal information, Style life y Clinic events	30	24	1	0.94	1.2
Personal information y Context information	34	24	1	0.97	1.36
Personal information, Context information y Clinic events	35	26	2	0.97	1.25
Personal information, Context information y Style life	35	24	1	0.97	1.4
Personal information, Context information y Style life	43	24	1	0.98	1.72

Tabla D.3: Resultados obtenidos de las pruebas combinadas

Para ver con más claridad los resultados obtenido se van a mostrar las gráficas con los resultados de *Recall* en la Figura D.2. En este caso se puede ver que todas las pruebas realizadas tienen un alto valor de *Recall*. De hecho, la mayoría de ellas rozan el valor máximo 1.

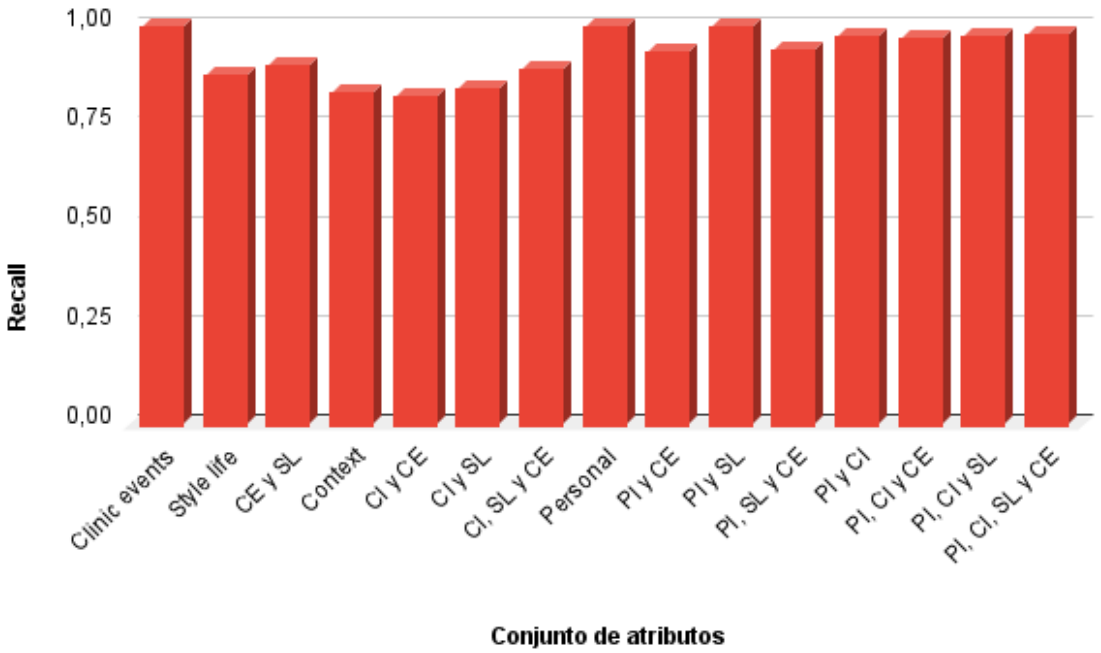


Figura D.2: Resultados de la métrica *Recall* obtenidos en las pruebas combinadas

D.2. Resultados obtenidos en las pruebas de rendimiento

Estas pruebas pretenden reflejar cómo de costoso es el proceso de anonimización según las opciones que se escojan para el mismo. Se debe destacar que es muy influyente el contenido de los documentos que se someten a dicho proceso de anonimización. Si el fichero no contiene mucha información sensible, aunque sea más largo, seguramente tarde menos que aquellos que sean cortos, pero contengan una gran cantidad de información sensible a eliminar.

D.2.1. Pruebas relacionadas con conjuntos de atributos

En este caso, se han realizar las 15 pruebas realizadas en la Sección 4.2.1, con su fichero correspondiente. Los resultados obtenidos se muestran en la Tabla D.4. Para

visualizar mejor los datos se ofrece la gráfica que aparece en la Figura D.3.

Test	**Execution time**
Clinic events	1 min
Style life	1 min
Style life y Clinic events	1 min
Context information	4 mins
Context information y Clinic events	1 min
Context information y Style life	2 mins
Context information, Style life y Clinic events	3 mins
Personal information	5 mins
Personal information y Clinic events	5 mins
Personal information y Style life	5 mins
Personal information, Style life y Clinic events	5 mins
Personal information y Context information	5 mins
Personal information, Context information y Clinic events	6 mins
Personal information, Context information y Style life	5 mins
Personal information, Context information, Style life y Clinic events	6 mins

Tabla D.4: Tiempos de ejecución obtenidos de las pruebas relacionadas con conjuntos de atributos

Se puede observar en la Tabla D.4 que la duración de cada una de las pruebas no es excesivamente largas, ya que el fichero contenía bastante información sensible a pesar de constar únicamente de 7 líneas. Sin embargo, es cierto que las pruebas que tienen una duración más larga son aquellas que utilizan la opción de *"Personal information"*. Esto se debe a que el sistema debe realizar muchas más comprobaciones que en el resto de casos, además de consultar varios ficheros en formato *".csv"*.

Es por este mismo motivo que hay cuatro pruebas que tardan un minuto aproximadamente en realizar la anonimización. Estos casos son los que implican realizar búsquedas en diccionarios o aplicar expresiones regulares.

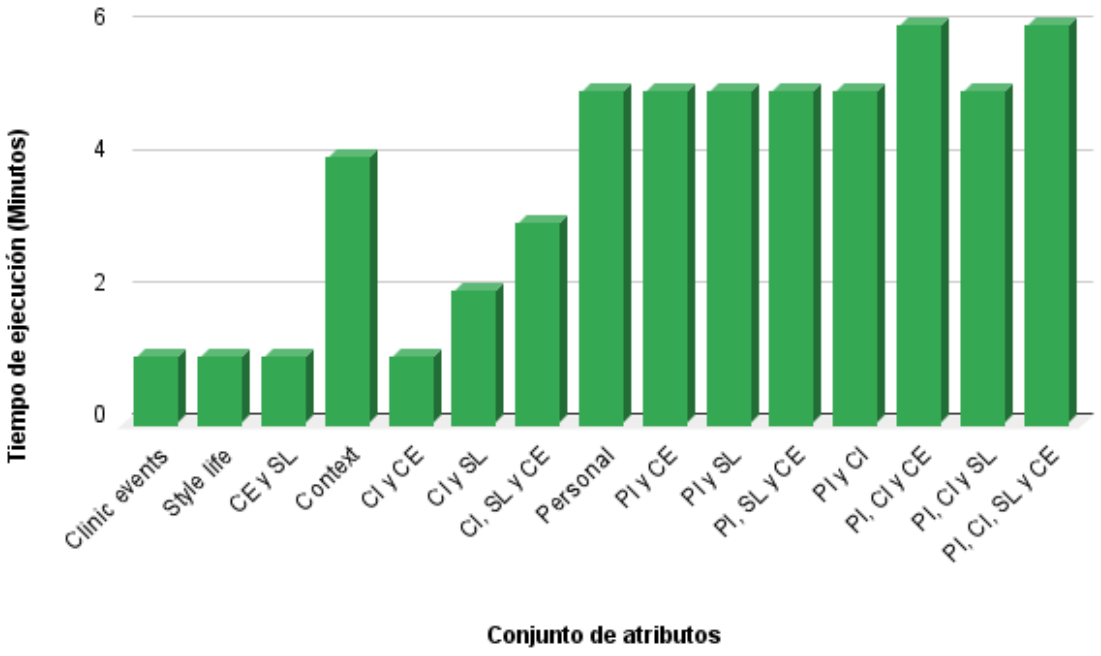


Figura D.3: Gráfica de tiempos de ejecución obtenidos en las pruebas relacionadas con conjuntos de atributos

D.2.2. Pruebas relacionadas con el nivel de anonimización

En este caso, se han realizado tres pruebas, las correspondientes a la Sección 4.2.2. En este caso, la prueba se ha realizado para un fichero de texto que contiene información sensible para cada uno de los niveles de anonimización que se ofrece en el sistema. Es decir, en las tres pruebas debe eliminar cierta información sensible. Los tiempos de ejecución obtenidos se muestran en la Tabla D.5.

Anonymization level	**Execution time**
Low level	5 mins
Medium level	9 mins
High level	8 mins

Tabla D.5: Tiempos de ejecución obtenidos de las pruebas relacionadas con el nivel de anonimización

D.2.3. Pruebas combinadas

En este caso, se han realizar las 15 pruebas realizadas en la Sección 4.2.3, con su fichero correspondiente. Los resultados obtenidos se muestran en la Tabla D.6. Para visualizar mejor los datos se ofrece la gráfica que aparece en la Figura D.4.

Test	Execution time
Clinic events	6 mins
Style life	6 mins
Style life y Clinic events	5 mins
Context information	4 mins
Context information y Clinic events	8 mins
Context information y Style life	8 mins
Context information, Style life y Clinic events	7 mins
Personal information	8 mins
Personal information y Clinic events	7 mins
Personal information y Style life	8 mins
Personal information, Style life y Clinic events	8 mins
Personal information y Context information	7 mins
Personal information, Context information y Clinic events	7 mins
Personal information, Context information y Style life	6 mins
Personal information, Context information, Style life y Clinic events	9 mins

Tabla D.6: Tiempos de ejecución obtenidos de las pruebas combinadas

En este caso, las pruebas realizadas son las que más tiempo lleva realizarlas. Esto es evidente, ya que se aplica un combinación de anonimización entre las pruebas realizadas en las Sección 4.2.1 y las realizadas en la Sección 4.2.2. Aún así, se sigue viendo que las pruebas con menor duración son las cuatro primeras. Las que tienen que realizar búsquedas menos costosas o las que constan de menos atributos a identificar.

D.3. Resultados obtenidos en las pruebas de minería de datos

En el Anexo D.3.1 se muestran los resultados obtenidos en las pruebas relacionadas con los documentos personales y no personales. En el Anexo D.3.2 se muestran los resultados obtenidos en las pruebas relacionadas con documentos de ictus cerebrales y

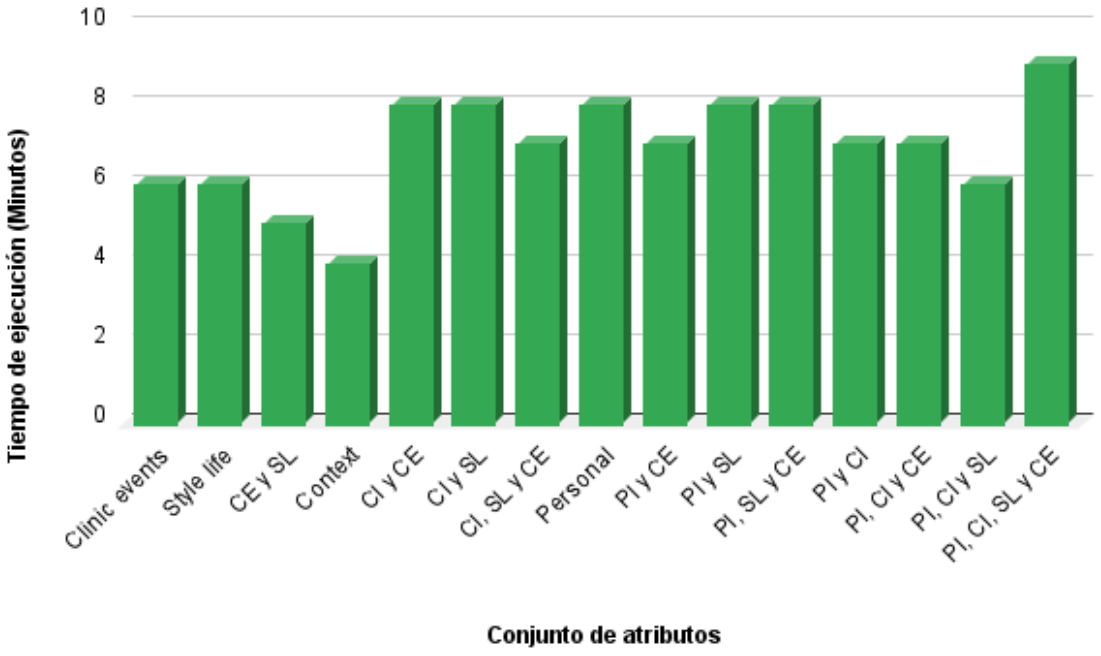


Figura D.4: Gráfica de tiempos de ejecución obtenidos en las pruebas combinadas

otras enfermedades.

D.3.1. Pruebas relacionadas con documentos personales y no personales

En este Anexo se pueden ver los resulatos obtenidos tras realizar las diferentes pruebas relacionadas con los documentos personales y no personales. Los resultados obtenido se pueden ver en las Tablas D.8 y D.7, junto con las gráficas de las Figuras D.5 y D.6.

Dataset original				
Classifier	**Correctly Classified**	**Incorrectly Classified**	**Mean absolute error**	**Root mean square error**
OneR	68.83	31.17	0.21	0.46
ZeroR	56.59	43.41	0.35	0.42
Naive Bayes	92.02	7.97	0.05	0.21
SMO	99.18	0.81	0.22	0.27
Dataset anonimizado				
Classifier	**Correctly Classified**	**Incorrectly Classified**	**Mean absolute error**	**Root mean square error**
OneR	67.81	32.18	0.32	0.57

ZeroR	50.83	49.17	0.50	0.50
Naive Bayes	92.79	7.20	0.07	0.24
SMO	97.94	2.05	0.02	0.14

Tabla D.7: Resultados obtenidos de la clasificación de documentos personales y no personales I

Original dataset						
Classifier	**TP**	**FP**	**TN**	**FN**	**Precision**	**Recall**
OneR	0.69	0.40	0.60	0.40	0.75	0.69
ZeroR	0.57	0.57	0.43	0.57	0.32	0.57
Naive Bayes	0.92	0.06	0.94	0.06	0.93	0.92
SMO	0.99	0.01	0.99	0.01	0.99	0.99
Anonymized dataset						
Classifier	**TP**	**FP**	**TN**	**FN**	**Precision**	**Recall**
OneR	0.68	0.33	0.67	0.33	0.69	0.68
ZeroR	0.51	0.51	0.49	0.51	0.26	0.51
Naive Bayes	0.93	0.07	0.93	0.07	0.93	0.93
SMO	0.98	0.02	0.98	0.02	0.98	0.98

Tabla D.8: Resultados obtenidos de la clasificación de documentos personales y no personales II

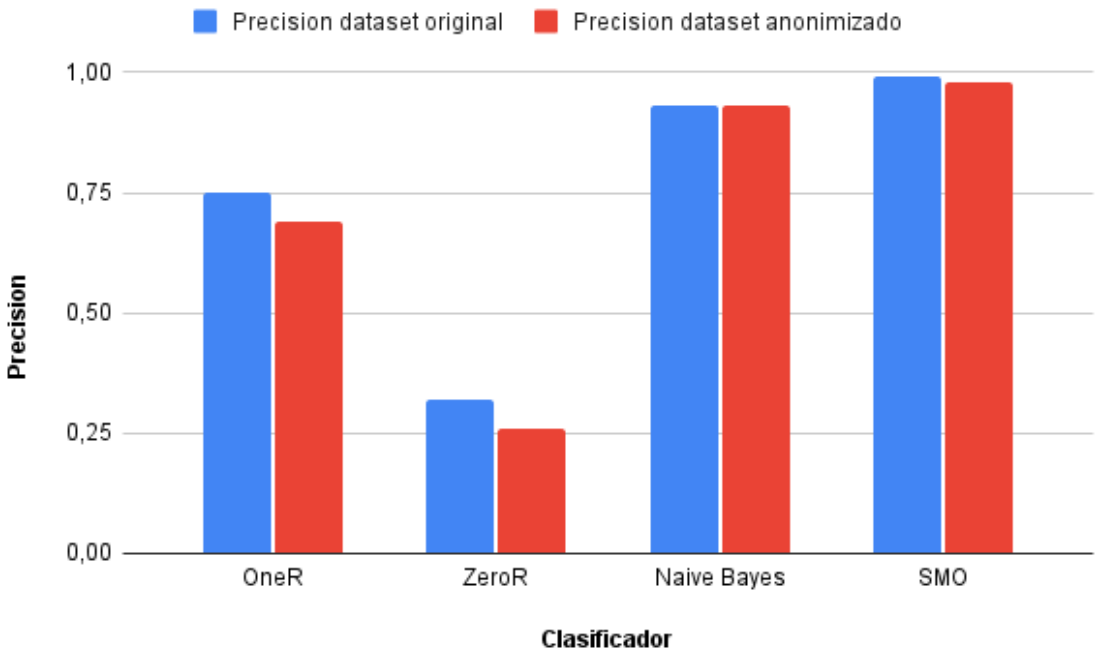


Figura D.5: Resultados de la métrica *Precision* obtenidos en la clasificación de documentos personales y no personales

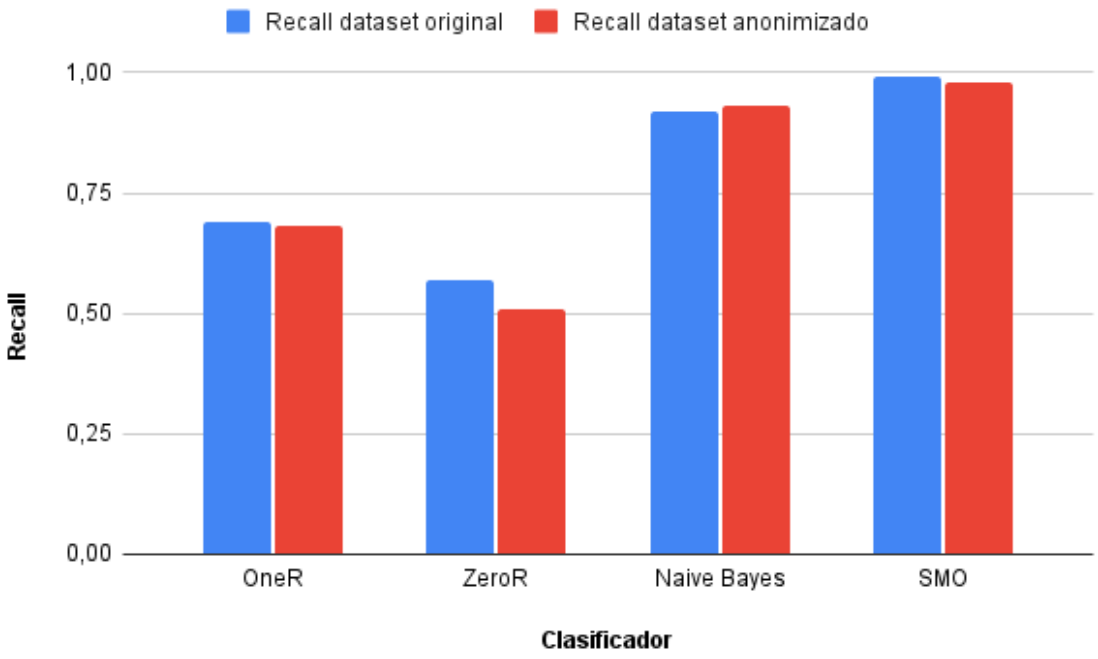


Figura D.6: Resultados de la métrica *Recall* obtenidos en la clasificación de documentos personales y no personales

Clasificador OneR

En este caso, se puede ver que la diferencia de correctos clasificados entre ambos datasets es de un 1 %, lo que supone que la clasificación ha sido muy parecida. Sin embargo, si se observa la Tabla D.8, se puede observar que el clasificador ha clasificado el 9 % de los datos peor en el caso del dataset anonimizado, aunque su *precision* sea un 6 % más alta que en el caso del dataset original.

Como se puede ver en la Tabla D.8, el valor de FP no aumenta en el caso del conjunto de datos anonimizado. De hecho, el mismo valor que ha disminuido en FP es el que ha aumentado en TP. Del mismo modo, el valor de TP ha disminuido un 1 % y ese mismo valor ha aumentado en FN. Se puede decir que la anonimización ha hecho que el clasificador clasifique más documentos como personales cuando en verdad no lo son.

Clasificador ZeroR

Como se puede ver en la Tabla D.7, la diferencia de clasificados correctamente e incorrectamente entre ambos dataset es de un 6 %, siendo el dataset anonimizado el

que clasifica un 6 % menos de datos correctamente. Además, tiene un error medio de un 8 % superior al del dataset original.

Como se puede observar en la Tabla D.8, el clasificador *ZeroR* clasifica un 6 % más de datos en TP y FP en el dataset original que en el anonimizado. Es por eso que el valor de la métrica *precision* es un 6,2 % más baja en el dataset anonimizado.

Estos resultados muestran que el clasificador ha detectado un número menor de TP, lo cual es uno de los objetivos de la evaluación. En este caso, el clasificador detecta que hay un 6 % de los datos que dejan de ser personales, pero lo que aumenta no es el FP sino el FN y el TN. Además, también disminuye en un 6 % el valor de FP, lo que supone que hay documentos no personales que deja de considerarlos como tal y los clasifica o bien como FN o como TN.

Clasificador Naive Bayes

En el caso del clasificador *Naive Bayes*, se puede ver en la Tabla D.7 que el valor de correctos clasificados es muy parecido con ambos datasets. Sin embargo, el número de clasificados incorrectamente disminuye un 0.77 % en el caso del dataset anonimizado. Es por este motivo que el valor de error en el caso del dataset anonimizado es un 0.05 % inferior al del original.

En este caso, el valor de TP es prácticamente igual en ambos casos, como se ha comentado anteriormente. Sin embargo, el valor de FP es mayor en el caso del dataset anonimizado. Del mismo modo, los valores de FN y TN disminuyen en el caso del dataset anonimizado. Esto nos hace llegar a la conclusión de que el proceso de anonimización ha producido que algunos de los casos que estaban clasificados incorrectamente hayan sido clasificados correctamente como no personales. Esto es un factor positivo para el proceso de anonimización.

Clasificador SMO

Respecto al clasificador SMO, se muestra en la Tabla D.7 que con el dataset original clasifica un 1,6 % de los datos más como correctos que con el dataset anonimizado. Sin embargo, el valor obtenido del error, es menor en el caso del dataset anonimizado que en el original, un 13,2 %.

Si se hace referencia a la Tabla D.8, se puede ver que el clasificador clasifica un 1,7 % de datos más como datos personales en el caso del dataset original a diferencia del dataset anonimizado. Sin embargo, la tasa de FP es más elevada en el caso del anonimizado.

La tasa de TP ha disminuido un 1,7 % del dataset original al anonimizado. Esto quiere decir que 104 documentos que antes se clasificaban como personales, ahora no lo son. De hecho, el valor de FP aumenta en un 1,2 % con el dataset anonimizado. Por tanto, tras la anonimización, el número de documentos no personales aumenta. Esto implica que el proceso de anonimización ha sido positivo.

D.3.2. Pruebas relacionadas con documentos sobre ictus cerebrales y otras enfermedades

En este Anexo se pueden ver los resulatos obtenidos tras realizar las diferentes pruebas realizadas con documentos relacionados con ictus cerebrales y otras enfermedades. Los resultados obtenido se pueden ver en las Tablas D.9 y D.10 junto con las gráficas que aparecen en las Figuras D.7 y D.8.

Original dataset				
Classifier	**Correctly Classified**	**Incorrectly Classified**	**Mean absolute error**	**Root mean square error**
OneR	95.00	5.00	0.05	0.22
ZeroR	50.00	50.00	0.50	0.50
Naive Bayes	96.50	3.5	0.03	0.19
SMO	98.50	1.5	0.02	0.12
Anonymized dataset				
Classifier	**Correct Classified**	**Incorrect Classified**	**Mean absolute error**	**Root mean square error**
OneR	87.00	13.00	0.13	0.36
ZeroR	50.00	50.00	0.50	0.50
Naive Bayes	94.50	5.50	0.06	0.23
SMO	99.00	1.00	0.01	0.1

Tabla D.9: Resultados obtenidos de la evaluación de documentos relacionados con ictus y otras enfermedades I

Original dataset						
Classifier	**TP**	**FP**	**TN**	**FN**	**Precision**	**Recall**
OneR	0.95	0.05	0.95	0.05	0.95	0.95
ZeroR	0.50	0.50	0.50	0.50	0.25	0.25
Naive Bayes	0.97	0.04	0.97	0.04	0.97	0.97
SMO	0.99	0.02	0.99	0.02	0.99	0.99
Anonymized dataset						
Classifier	**TP**	**FP**	**TN**	**FN**	**Precision**	**Recall**
OneR	0.87	0.13	0.87	0.13	0.87	0.87
ZeroR	0.50	0.50	0.50	0.50	0.25	0.50
Naive Bayes	0.95	0.06	0.95	0.06	0.95	0.95
SMO	0.99	0.01	0.99	0.01	0.99	0.99

Tabla D.10: Resultados obtenidos de la evaluación de documentos relacionados con ictus y otras enfermedades II

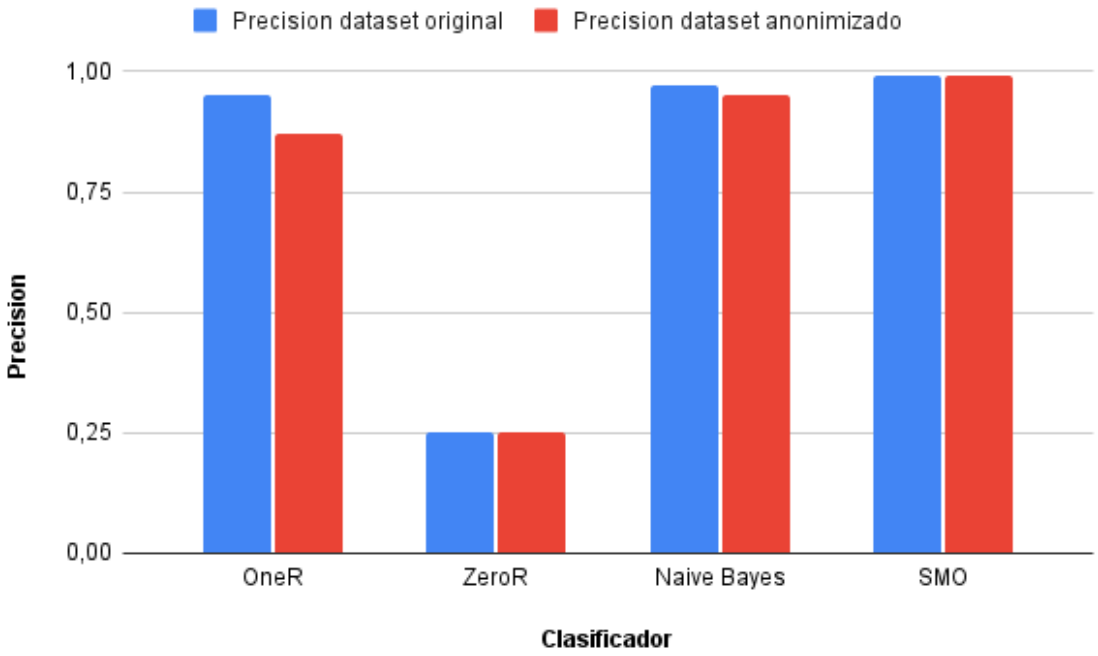


Figura D.7: Resultados de la métrica *Precision* obtenidos en la evaluación de documentos relacionados con ictus y otras enfermedades

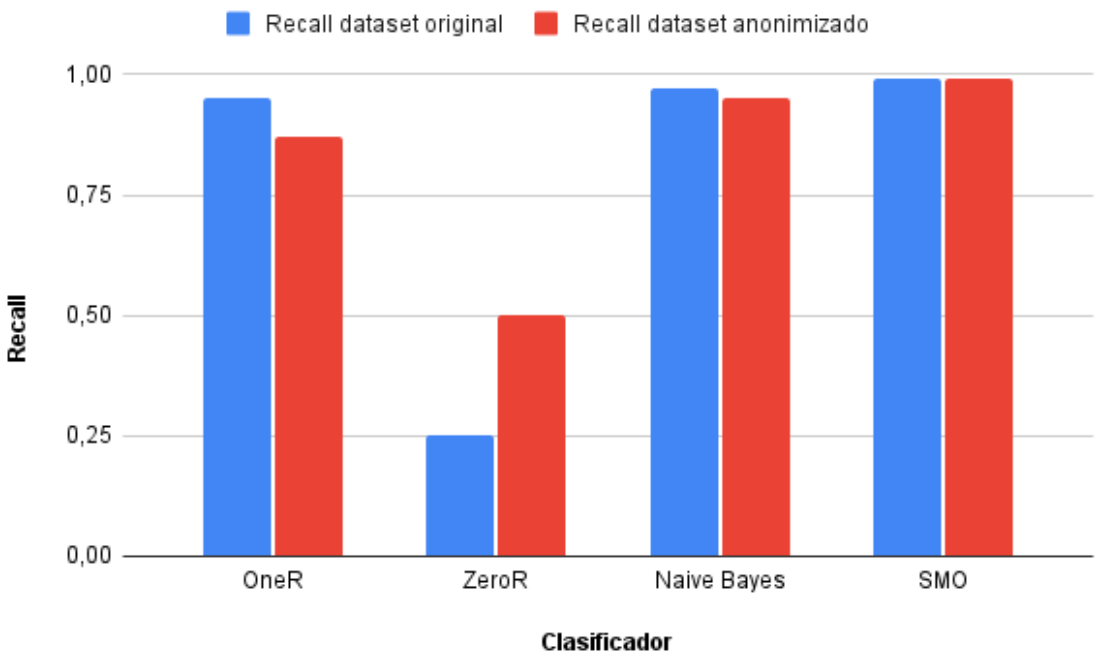


Figura D.8: Resultados de la métrica *Recall* obtenidos en la evaluación de documentos relacionados con ictus y otras enfermedades

Clasificador ZeroR

En el caso del clasificador *ZeroR*, se puede apreciar que los valores obtenidos con los dataset originales son los mismos que los obtenidos con el dataset anonimizado. Esto quiere decir que el proceso de anonimización no elimina información importante para la clasificación. Es decir, el clasificador ha realizado la misma clasificación con los datos originales y con los datos anonimizados, lo que implica que la anonimización no afecta de manera negativa al contenido de los informes en este contexto.

Clasificador OneR

Con el clasificador *OneR*, se pueden apreciar ciertos cambios a diferencia del caso anterior. En este caso, la diferencia es mínima, ya que se trata de un 8 % de diferencia en los datos clasificados. Es decir, el clasificador ha clasificado un 8 % de los datos incorrectamente con los datos anonimizados respecto de los originales. Dado que lo que se quiere observar, es la calidad de no pérdida de información, se puede concluir que un 8 % de los datos es un número lo suficientemente pequeño como para determinar que el proceso de anonimización no afecta negativamente al contenido de los informes. En otras palabras, el proceso de anonimización elimina información, pero no

tanto como para determinar que afecta a la semántica del contenido de los documentos.

Clasificador Naive Bayes

En el caso del clasificador *Naive Bayes*, también hay cierta diferencia en la clasificación con el dataset anonimizado respecto al original. Sin embargo, la diferencia es del 2 %. Esto supone que el proceso de anonimización no elimina información de forma masiva que haga perder la semántica del contenido del documento original. El hecho de que la diferencia entre ambas clasificaciones sea tan pequeño permite determinar que la calidad de la no pérdida de información es positiva.

Clasificador SMO

En el caso del clasificador *SMO*, la diferencia entre ambas clasificaciones es de un 0,05 %. Al igual que en los casos anteriores, la diferencia que existe entre la evaluación con los datos originales y los anonimizados es prácticamente mínima. Por tanto, se puede concluir que el proceso de anonimización no ha eliminado información de forma masiva, manteniendo la semántica del contenido del conjunto de datos original.

Como conclusiones generales, se puede observar en las Tablas D.9 y D.10, que la diferencia que existe entre la evaluación realizada con los conjuntos de datos originales y los anonimizados es como máximo de un 8 %. Dado que esta cantidad no es muy elevada, se puede determinar que la clasificación es similar y que el proceso de anonimización no afecta de manera negativa al contenido de los datos.