Universidad Zaragoza
1542

Trabajo Fin de Máster

ANONIMIZACIÓN PARA MINERÍA DE DATOS EN ENTORNOS MÉDICOS

ANONYMIZATION FOR DATA MINING IN MEDICAL ENVIRONMENTS

Autor

Verónica Irene Liñayo Vega

Director

Sergio Ilarri Artigas

Escuela de Ingeniería y Arquitectura - Universidad de Zaragoza

Zaragoza, Septiembre 2020

Resumen

Este trabajo es un estudio sobre la anonimización de datos en contextos médicos, partiendo del estado del arte y las diversas técnicas y herramientas disponibles para anonimizar datos hoy en día, en donde se seleccionan y evalúan conjuntos de datos médicos con el fin de comparar diversas técnicas y su impacto.

Para agilizar la evaluación sobre los datos se desarrolla una aplicación basada en Java, mediante la cual a través de la integración con la API de Weka (herramienta para minería de datos) y mediante el uso de herramientas externas como R, se genera una herramienta que permita la comparación entre diversas técnicas de anonimización seleccionadas, con lo cual se puede observar el impacto que dichas técnicas pueden tener sobre los conjuntos de datos estudiados.

Mediante la aplicación desarrollada en la elaboración de este trabajo es posible realizar un análisis en el cual se estudia la posible relación entre el impacto generado por la anonimización y las diversas técnicas aplicadas, permitiendo estudiar así el nivel de compromiso privacidad-precisión que se obtiene tras aplicar las técnicas de anonimización. De esta forma se demuestra que las diversas técnicas de anonimización estudiadas pueden tener mayor o menor impacto en el nivel de compromiso privacidad-precisión obtenido para los conjuntos de datos estudiados, y que dicho impacto depende de varios factores, como el tipo de anonimización aplicada, el tipo de clasificador usado, el conjunto de datos, los diversos tipos de atributos encontrados dentro del conjunto de datos, entre otros.

Se espera que el trabajo desarrollado sirva como base para futuros trabajos e investigaciones llevadas a cabo por grupos de investigación de la Universidad de Zaragoza, en particular el grupo COSMOS (Computer Science for Complex System Modelling). Además, podría servir también de aplicación en contextos docentes, como base de estudio en asignaturas como Manipulación y Análisis de Grandes Volúmenes de Datos del Máster Universitario en Ingeniería Informática.

Agradecimientos

A mi familia, por apoyarme siempre en todos y cada uno de mis pasos, por impulsarme a ser siempre una mejor versión de mí, por su amor incondicional, por su comprensión, por su paciencia, por escucharme y aconsejarme cuando más los he necesitado, por enseñarme a no darme por vencida ante las adversidades, y por todos los sacrificios que han realizado para ayudarme a estar hoy aquí.

A Iván Aldea Morales, por su cariño, comprensión, apoyo y paciencia durante este tiempo que he dedicado a la elaboración del Máster, por aguantar mis malos días e impulsarme a seguir, logrando siempre sacarme una sonrisa incluso en los días más difíciles.

A mi querida profesora y amiga Marlene Goncalves, quien fue mi tutor académico en el trabajo de grado, que con su apoyo, su guía, sus consejos y sugerencias me incentivaron a continuar mis estudios más allá del grado, que gracias a sus recomendaciones logré contactar con la Universidad de Zaragoza y el Prof. Javier Zarazaga, y por quien se despierta en mí cierto interés por el mundo de los datos hasta el punto de decidir realizar mi trabajo final en relación a esto.

Al Prof. Javier Zarazaga, por su cálida recepción en mi llegada a Zaragoza, por ser un apoyo invaluable en todo el proceso de admisión a estos estudios que hoy culmino, por su generosidad en guiarme y aconsejarme en mi paso por el Máster e incluso por ser un puente que me permitió conseguir mi primer trabajo aquí en España con el cual pude financiar mis estudios.

A mi tutor académico, Prof. Sergio Ilarri Artigas, a quien le agradezco por el tiempo dedicado para guiarme en la elaboración de este trabajo, por poner la barrera alta desde un principio, exigiéndome a cada paso, demostrándome que soy capaz de realizar un trabajo fuera de mi área de confort y siendo un apoyo constante aún ante las dificultades que se presentaron durante este inusual año.

Autoria

TRABAJOS DE FIN DE GRADO / FIN DE MÁSTER

Escuela de Ingeniería y Arquitectura Universidad Zaragoza

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD

(Este documento debe entregarse en la Secretaría de la EINA, dentro del plazo de depósito del TFG/TFM para su evaluación).

D./Dª. Verónica Irene Liñayo Vega ,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) Máster Universitario en Ingeniería Informática (Título del Trabajo)

Anonimización para minería de datos en entornos médicos

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

Zaragoza, Septiembre 2020

Fdo: Verónica Liñayo Vega

Tabla de Contenidos

Índice de Figuras

Índice de Tablas

CAPÍTULO I. Introducción

A continuación se presenta el capítulo de introducción al trabajo, en el cual se puede observar la motivación que da inicio a esta investigación, sus objetivos, la estructura y una breve introducción que permite dar contexto a los siguientes capítulos.

Antes de adentrarnos en el tema principal del trabajo de fin de máster es necesario abordar los conceptos principales que se verán a lo largo del mismo. Primordialmente hablaremos de temas como la anonimización, los conjuntos de datos, la privacidad, y la minería de datos.

Actualmente la Real Academia Española (RAE) define ***anonimizar*** como "*Expresar un dato relativo a entidades o personas, eliminando la referencia a su identidad"* [1]. A su vez la Agencia Española de Protección de Datos o AEPD define la anonimización como *"la ruptura de la cadena de identificación de las personas"* [2].

Donde un dato [3] puede considerarse según la RAE como:

- *"Información sobre algo concreto que permite su conocimiento exacto o sirve para deducir las consecuencias derivadas de un hecho."*
- *"Documento, testimonio, fundamento."*
- *"Información dispuesta de manera adecuada para su tratamiento por una computadora."*

De lo que se puede concluir que anonimizar datos consiste en eliminar de estos todas las referencias que permitan la identificación directa de una persona o entidad. La idea principal de anonimizar datos nace del objetivo de buscar proteger a las entidades y/o personas que puedan ser identificadas mediante los mismos, especialmente cuando se trabaja con datos sensibles como lo son los datos provenientes del área de Salud.

Dentro del campo de la minería de datos [4] también resulta importante la anonimización, partiendo de la idea de que el objetivo del área es tratar datos para un fin dado, y que consiste en la extracción de información partiendo de un conjunto de datos, transformando la información en una estructura comprensible que permita su uso posterior. Por lo tanto, dentro de la minería de datos se presenta la necesidad de mantener los datos protegidos a la vez que se aprovechan y transforman para un determinado fin.

Para la RAE, la protección de datos se define como *"Sistema legal que garantiza la confidencialidad de los datos personales en poder de las Administraciones públicas u otras organizaciones"* [5]. Y también define el procesamiento de datos como aquella *"Aplicación sistemática de una serie de operaciones sobre un conjunto de datos, generalmente por medio de máquinas, para explotar la información que estos datos representan"* [6].

Los conceptos que se mencionan anteriormente permiten dar sustento a la motivación de este trabajo y permiten definir los objetivos que se esperan alcanzar en el mismo, para así poder continuar con la presentación de la estructura del trabajo de investigación y dar inicio al **Capítulo II**.

I.1. Motivación

Hoy en día, nuestros datos personales son requeridos constantemente por múltiples organismos y empresas, desde entidades privadas que los solicitan para diversos fines como la gestión de perfiles en redes sociales, gestión de publicidad, o simplemente para proveer un servicio; hasta entidades públicas como hospitales, ayuntamientos, universidades, entre otros; que solicitan nuestros datos para fines diversos como por ejemplo llevar un registro y control de enfermedades de los pacientes, o las notas y datos académicos de los estudiantes, permitiendo a quienes proveen el servicio público poder identificar a un individuo dentro de la sociedad.

Esta necesidad constante que tienen dichos organismos y empresas de obtener nuestros datos representa una enorme responsabilidad, en donde dichas entidades deben asegurarse de resguardar la integridad de estos datos que se les han compartido, y evitar a toda costa la posibilidad de que un individuo quede expuesto partiendo de dicha información.

Hoy en día existen diversas formas de proteger los datos, desde bases de datos cifradas, hasta el uso de bases de datos distribuidas y técnicas de aislamiento de datos. Sin embargo, se hace difícil proteger aquellos datos que son de dominio público, en donde las empresas y organismos están obligadas a hacer públicos los datos con los que trabajan, o que por motivos de investigación o estadísticos deben compartir los datos con otras entidades que harán uso de estos para proveer a la población con información de interés, como por ejemplo estadísticas del uso del coche en la ciudad, o promedio de adolescentes embarazadas, entre otros.

Es por este motivo que la idea de la anonimización de datos toma fuerza, dado que permite eliminar o “disfrazar” aquellos datos de carácter público, que puedan llevar a la identificación de un individuo, permitiendo así publicar o compartir los mismos para su uso y estudio.

El problema está cuando se intentan estudiar y analizar estos datos anonimizados, debido a que al aplicar técnicas de minería sobre conjuntos de datos anonimizados se puede estar perdiendo cierta información de valor al momento de generar conclusiones que sirvan de uso y resulten de interés para la población. En consecuencia se podría estar generando resultados que no son del todo precisos o al menos que no son tan precisos como lo serían si se trabajara con datos sin anonimizar.

También se tiene la creciente problemática que ha causado la “nueva” normativa de protección de datos que está constantemente cambiando y que regula y restringe aún más el tipo de datos que puede ser de dominio público. De manera que nace una interrogante de interés, en donde se plantea el estudio del impacto que pueden ocasionar diversas técnicas de anonimización sobre los conjuntos de datos y cómo afecta dicho impacto en la minería de estos.

Esto ha dado lugar a un interés creciente por el concepto de anonimización, sus técnicas y el impacto de aplicarlas, en donde para propósitos de esta investigación se centra el interés principalmente en datos provenientes de entornos médicos, buscando así realizar un estudio sobre diversos conjuntos de datos provenientes del mundo de la salud como por ejemplo: enfermedades del corazón, diabetes, entre otros; en donde se

pueda aplicar diversas técnicas de minería y analizar los posibles resultados antes y después de anonimizar dichos datos, con el fin de evaluar hasta qué punto puede afectar la anonimización de datos a las conclusiones que se pueden obtener mediante la minería de los mismos.

I.2. Objetivos

En este trabajo el objetivo general es el de realizar un estudio sobre las diversas estrategias de anonimización y comparar las mismas mediante el uso de diversas técnicas de minería de datos con el fin de analizar el compromiso privacidad-precisión al usar dichas técnicas. De la misma forma se presentan como objetivos específicos:

- Estudiar el estado del arte de la anonimización.
- Estudiar la normativa actual sobre la protección de datos.
- Analizar y seleccionar diversos conjuntos de datos pertenecientes al ámbito de salud.
- Aplicar diversas estrategias de anonimización sobre los conjuntos de datos seleccionados.
- Aplicar diversas técnicas de minería de datos sobre los conjuntos de datos seleccionados.
- Comparar los resultados obtenidos de las estrategias y técnicas usadas con los conjuntos de datos seleccionados.
- Desarrollar una interfaz gráfica que facilite las comparaciones entre las técnicas de anonimización aplicadas a los conjuntos de datos.

Además de estos objetivos se propone como una posibilidad adicional considerar el hacer uso de otras herramientas externas distintas de Weka, en donde se puedan aplicar técnicas de anonimización, para así hacer uso de ellas y comparar las mismas con la herramienta generada. Ampliando así las posibilidades de trabajos futuros que puedan partir de esta investigación.

I.3. Estructura del trabajo

Este trabajo de investigación contiene una estructura diferenciada por capítulos en los cuales se presenta lo siguiente:

- **Capítulo I.** Introducción: actual capítulo donde se presenta al lector una idea general de la temática de este trabajo de investigación, junto a los objetivos planteados y la motivación de este trabajo.

- **Capítulo II.** Estado del arte: se presentan los diversos conceptos relacionados a la investigación, en donde se muestra al lector las actuales formas de clasificación de datos, técnicas de anonimización actualmente usadas, riesgos más comunes al anonimizar, herramientas comunes para anonimizar, información relacionada a la minería de datos y las herramientas más usadas hoy en día para la minería de datos, finalizando con la normativa asociada a la protección de datos incluyendo la forma de gestión de datos generales y de entornos médicos.

- **Capítulo III.** Recursos y Herramientas: en este capítulo se mencionan y explican las herramientas utilizadas para la elaboración de este trabajo de investigación, diferenciando las herramientas mediante 5 secciones, empezando por las herramientas usadas para la minería de datos, las utilizadas para la realización del diseño de la aplicación desarrollada, y continuando con las herramientas utilizadas para el desarrollo, gestión y control de versiones.

- **Capítulo IV.** Obtención y estudio de los conjuntos de datos: en este capítulo se presenta la metodología seguida en el trabajo para la obtención y análisis de los conjuntos de datos. Así como también se presentan los conjuntos de datos seleccionados para la investigación; en donde para cada uno de ellos se presentan las características, un contexto general del conjunto de datos, un resumen del contenido del mismo, y se muestra la realización del pre-procesado de dicho conjunto de datos.

- **Capítulo V.** Desarrollo de herramienta de apoyo a la evaluación de estrategias de anonimización sobre conjuntos de datos: en este capítulo se expone el desarrollo realizado para dar apoyo a la investigación en la fase de evaluación de conjuntos de datos, explicando el diseño, la metodología y los resultados finales obtenidos en el desarrollo de la herramienta.

- **Capítulo VI.** Resultados y análisis de resultados: en este capítulo se presentan los resultados obtenidos al aplicar las diversas técnicas de anonimización y minería de datos sobre los conjuntos de datos seleccionados en el capítulo III, y se realiza un análisis sobre dichos resultados, permitiendo generar conclusiones sobre lo observado.

- **Capítulo VII.** Conclusiones y trabajo futuro: se observa en este capítulo un resumen final del trabajo realizado y se presentan conclusiones personales del autor, junto con las dificultades encontradas, aprendizaje obtenido durante la elaboración del trabajo, el tiempo invertido y el trabajo futuro que se puede prever partiendo del trabajo realizado en esta investigación.

- **Anexo I.** Información complementaria al Capítulo II. Estado del arte: se añade un anexo en el cual se puede observar información complementaria o simplemente que pueda resultar de interés al lector en referencia al estado del arte presentado en el capítulo II.

- **Anexo II.** Desarrollo de la herramienta CPDA. Se presenta en este anexo los diseños realizados para la aplicación desarrollada en el trabajo, partiendo desde los diseños iniciales hasta el diseño final que dio base a la aplicación desarrollada. También se presenta en detalle la aplicación de la metodología usada para el desarrollo de la aplicación, explicando los diferentes *sprints* definidos, el tiempo estimado para cada uno de ellos, y la gestión realizada mediante la herramienta de Asana para hacer seguimiento del desarrollo y del tiempo dedicado a este.

- **Anexo III.** Manual de usuario de CPDA. Se presenta un manual de usuario mediante el uso de capturas de pantalla de la interfaz de la aplicación CPDA desarrollada para este trabajo, presentando las capturas según las secciones que se encuentran en la aplicación, como la sección de carga de datos, sección de minería, sección de resultados y la sección de gráficos.

- **Anexo IV.** Resultados y extractos de código. Se muestra en detalle la aplicación de las técnicas de anonimización y minería de datos estudiadas en la realización del trabajo, y es posible observar en ella los resultados obtenidos incluyendo la sección gráfica generada por la herramienta CPDA. Finalmente se puede encontrar ejemplos de los scripts usados para aplicar técnicas de anonimización mediante la herramienta externa R y análisis complementarios de los resultados obtenidos en la investigación.

Partiendo de la introducción y la estructura de trabajo anterior, es posible avanzar con el trabajo planteado, empezando con la presentación de técnicas más usadas para anonimizar datos, hasta llegar a los posibles impactos que estas técnicas y estrategias puedan acarrear.

CAPÍTULO II. Estado del arte

A continuación se presentan los diferentes conceptos relacionados con el trabajo de investigación, en donde sitúa en contexto al lector, permitiendo tener una idea general de la temática estudiada y su situación en el día de hoy, como la anonimización, herramientas y algoritmos usados hoy en día para anonimizar, riesgos encontrados, minería de datos y herramientas más usadas, así como otros temas asociados que son de interés para el trabajo.

II.1. Clasificación de datos

Los datos se clasifican según el área de saber en la que se encuentran o estudian, de esta forma el concepto de dato resulta sumamente impreciso, ya que este dependerá del área conforme a la cual se esté tratando, pero de forma general, un dato se corresponde con una porción de la realidad que se representa mediante el mismo y le da valor. De esa manera los datos pueden clasificarse como estadísticos, informáticos, personales, entre otros. En este trabajo se dará prioridad a los datos y los tipos de datos identificados por la Ley Orgánica de Protección de datos (LOPD) y el Reglamento General de Protección de Datos (RGPD), es decir, se habla de datos personales.
Para anonimizar los datos personales es importante la elección de qué anonimizar, porque una elección incorrecta de los atributos conlleva una desprotección de los datos, o una sobreprotección de los mismos, dificultando la extracción del conocimiento.
Para ayudar en los procesos de anonimización el *Dr. Khaled El Emam* realiza la siguiente clasificación:

- *Identificadores*: son aquellos campos de la base de datos que permiten identificar a una persona de forma única, como puede ser el DNI, nombres y apellidos, número de la seguridad social, número de la tarjeta sanitaria, teléfono móvil, entre otros [7].
- *Cuasi-identificadores*: aquellos datos personales que no son identificadores, y que no tienen un valor especial para nuestro análisis, pero que si bien por sí solos no revelan información, cuando se combinan entre ellos o con otras fuentes externas de datos pueden llegar a desvelar la identidad de una persona, algunos ejemplos de este tipo de datos serían: fecha de nacimiento, municipio de residencia, código postal, género, edad, entre otros [8].
- *No identificadores*: aquellos datos personales que no identifican por sí solos ni en combinación con otros atributos a un individuo, un ejemplo podría ser el tipo de calzado favorito de un individuo, el medio de transporte más frecuentado, el color de ojos, entre otros.

Los cuasi-identificadores son datos especialmente relevantes para nuestro análisis, ya que estos son los datos que se suelen alterar para proteger la privacidad de las personas. Es importante destacar también que los cuasi-identificadores no serán los mismos en todos los casos de uso, ya que dependen del análisis que queramos hacer posteriormente con los datos.
A la clasificación dada anteriormente, se le suma un tipo más dado por la LOPD y el RGPD, donde añaden el concepto de datos sensibles, los cuales no pueden ser

tratados sin previa confirmación del individuo, y que para cuyo tratamiento deben seguir una serie de pautas y requisitos indicados por la ley.

- *Datos sensibles*: se trata de datos personales que, por su naturaleza, son particularmente sensibles en relación con los derechos y las libertades fundamentales de un individuo, ya que el contexto de su tratamiento puede entrañar importantes riesgos para los derechos y las libertades fundamentales de los interesados. Es en este grupo de datos donde se sitúan los datos de salud, en los cuales se enfoca la investigación realizada, entre algunos ejemplos de datos sensibles se puede mencionar: religión, ideología política, datos genéticos, biométricos, de salud, datos referentes a la raza o etnia, entre otros.

En la actualidad existen más formas de clasificar datos personales además de los tipos mencionados en este capítulo, algunos de los cuales pueden verse reflejados en el **Anexo I**.

II.2. Técnicas de anonimización más usadas hoy

Existen diversas prácticas y técnicas de anonimización con diferentes grados de solidez. Esta sección aborda las principales técnicas usadas hoy en día en el ámbito de anonimización de datos.

II.2.1. Seudonimización

Hoy en día no es posible hablar de anonimización de datos sin antes hablar de la seudonimización, la cual para el RGPD se define como "*el tratamiento de datos personales de manera tal que ya no puedan atribuirse a un interesado sin utilizar información adicional, siempre que dicha información adicional figure por separado y esté sujeta a medidas técnicas y organizativas destinadas a garantizar que los datos personales no se atribuyan a una persona física identificada o identificable*." [9]. En palabras más simples, consiste en tratar los datos personales sin los datos identificativos del interesado, pero sin suprimir totalmente la vinculación entre dichos datos con aquellos que permiten identificar al individuo.

Un ejemplo sería la sustitución de los nombres de clientes por un código o por un identificador numérico, es decir, cambiar los datos personales por seudónimos.

La seudonimización reduce la vinculabilidad de un conjunto de datos con la identidad del interesado; se trata, por tanto, de una medida de seguridad útil, pero no es considerado por la AEPD como un método de anonimización. De todas formas se menciona junto al resto de técnicas ya que sigue siendo muy utilizada hoy día cómo forma de "proteger" los datos de los individuos [10].

Se puede encontrar la explicación detallada sobre la seudonimización en la sección de *seudonimización* del **Anexo I**, pero para fines de este trabajo solo se hace una breve mención dado que el objeto principal del estudio está en las técnicas de anonimización explicadas a continuación.

II.2.2. Generalización

Consiste en el reemplazo de datos por otros menos específicos, pero semánticamente consistentes. La generalización es la acción y efecto de generalizar,

que, según la RAE, es "*Abstraer lo que es común y esencial a muchas cosas, para formar un concepto general que las comprenda todas*" [11].

La generalización es conocida como uno de los tipos de anonimización debido a que a través de algoritmos y técnicas de generalización es posible reducir la precisión de los datos sin perder su utilidad, se puede observar un ejemplo sencillo en la Figura 1**¡Error! No se encuentra el origen de la referencia.**, en donde se generalizan dos atributos, el sexo y la edad. En un estudio en el que se desea analizar el año de nacimiento de una persona, se puede modificar el atributo que contiene la fecha de nacimiento (día, mes y año) a simplemente contener el año; de esta forma se reduce la probabilidad de identificar a los individuos del estudio, mientras que aún es posible analizar los datos para el objetivo pensado.

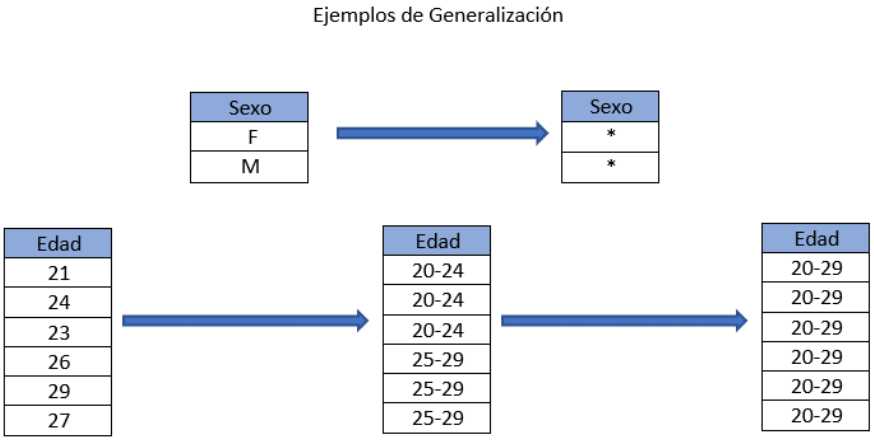


Figura 1 - Ejemplo de Generalización, adaptado de [12]

Existen distintas formas de generalizar datos, de las cuales se pueden mencionar entre las más usadas y comunes hoy en día las siguientes:

- Agregación: Consiste simplemente en agrupar los datos de las personas, buscando impedir que un individuo pueda ser identificado dentro del grupo, como en el ejemplo observado en la Figura 1, donde se pueden observar formas de generalización, en el caso del atributo de Edad, simplemente se genera una agrupación por edades separadas en un rango de 4 o 9 años según el nivel de agrupación que se realice, en el primer caso podría ser más sencillo para un atacante dar con una persona específica, debido a que tiene menos registros de datos en común con su objetivo, por ejemplo, si se busca a una persona de 22 años, ya se pueden descartar las últimas 3 filas de datos, y luego mediante otros atributos se podría intentar de identificar a la persona, claramente el segundo nivel de agrupación abarca muchas más filas por lo que dificulta un poco más la posibilidad de identificar a una persona, más no es del todo imposible si se tiene cierto conocimiento sobre la persona a identificar.

- K-anonimidad: Es una técnica que permite cuantificar y aplicar un determinado grado de anonimidad a la información de los sujetos que figuran en un determinado conjunto de datos. Para ello, se eliminan los datos identificadores y se aplican métodos que evitan que los datos cuasi-identificadores se puedan relacionar con los datos sensibles [13].

- Diversidad L: Es una forma de anonimización basada en grupos que se utiliza para preservar la privacidad en conjuntos de datos al reducir la granularidad de una representación de datos. El modelo de diversidad l es una extensión del modelo de anonimato k, pero en él se manejan algunas de las debilidades presentes en el modelo de k-anonimato donde las identidades protegidas al nivel de k-individuos no es equivalente a proteger los valores sensibles correspondientes que fueron generalizados o suprimidos, especialmente cuando los valores sensibles dentro de un grupo exhiben homogeneidad [14].

- Proximidad T: Busca generalizar los datos de forma que la distribución de cuasi identificadores en cada clase de equivalencia sea similar a la distribución de datos original. La idea de la proximidad t es que la distribución de datos confidenciales en todos los grupos no está demasiado lejos de la distribución en toda la población. Si dentro de un grupo de datos, aquellos datos considerados sensibles/privados no destacan entre el resto de datos sensibles, se frustra el ataque de homogeneidad y el ataque de conocimiento de fondo [15].

Queda a disposición del lector una sección dentro del **Anexo I** en donde se explica en mayor detalle cada una de estas técnicas mencionadas anteriormente, en donde se pueden encontrar referencias a los artículos en donde se definen dichas técnicas, así como visualizar ejemplos prácticos de los mismos.

II.2.3. Aleatorización

Consiste en la acción y efecto de aleatorizar, lo cual consiste según la RAE en, “Someter algo o a alguien a un proceso aleatorio” [16]. La aleatorización es una técnica usada en la anonimización de datos que consiste en aplicar una modificación aleatoria a los datos. Se conocen tres métodos principales de aleatorización, la inyección de ruido, las permutaciones y la privacidad diferencial, los cuales se explican detalladamente en del **Anexo I**.

II.2.4. Eliminación

La eliminación como su nombre bien lo indica consiste en eliminar ya sea un atributo o un registro que no se considera necesario para el estudio a realizar y/o que a su vez puede representar un riesgo de re-identificación, la eliminación se podría hacer parcial o total según sea el caso.

Podría considerarse el método más simple de todos los mencionados debido a que simplemente busca eliminar atributos o registros que permitan la identificación de una persona. Pero de igual forma tiene su dificultad, en especial cuando se quiere mantener la usabilidad de estos datos para posibles estudios e investigaciones, por lo que se debe proceder con cuidado al momento de aplicar la eliminación, y siempre es recomendable hacer uso de más técnicas de anonimización y no sólo la eliminación, debido a que aún eliminando cierto número de registros, los datos podrían estar expuestos a distintos ataques como los de homogeneidad, conocimiento de fondo, entre otros; y de esa forma un atacante podría lograr la re-identificación de un individuo.

II.3. Riesgos claves de la anonimización

Una vez presentadas las distintas técnicas de anonimización, es importante hablar de los riesgos que se deben identificar una vez anonimizados los datos, debido a que como se observó anteriormente, no hay técnica que sea 100% segura, siempre existe un mínimo de riesgo en que se pueda re-identificar algún individuo a través de los datos, por lo que es necesario saber cómo se mide ese nivel de riesgo a modo de buscar reducir el mismo en los datos anonimizados.

1. *Singularizar*: La posibilidad de aislar algunos o todos los registros que identifican a un individuo en el conjunto de datos
2. *Conectividad*: La capacidad de vincular al menos dos registros sobre el mismo sujeto de datos o un grupo de sujetos de datos (ya sea en la misma base de datos o en dos bases de datos diferentes)
3. *Inferencia*: La posibilidad de deducir, con probabilidad significativa, el valor de un atributo a partir de los valores de un conjunto de otros atributos.

Como se puede ver en la Figura 2 cada técnica tiene su propio conjunto de fortalezas y debilidades, hoy en día ninguna técnica es capaz de asegurar completamente la protección de los datos sensibles, pero sí son capaces de reducir el riesgo de identificación.

Técnicas	Singularizar	Conectividad	Inferencia
	¿Se mantiene el riesgo?		
Ruido	Si	Probablemente no	Probablemente no
Sustitución	Si	Si	Probablemente no
Agregación o K-anonymity	No	Si	Si
l-diversity	No	Si	Probablemente no
t-closeness	No	Probablemente no	Probablemente no
Privacidad Diferencial	Probablemente no	Probablemente no	Probablemente no

Figura 2 - Comparativa de técnicas de Anonimización basada en [17]

II.4. Herramientas para anonimizar datos

Algunas de las herramientas usadas hoy en día para la anonimización de datos son:

- *ARX Data Anonymization Tool* [18]: ARX es un software de código abierto para anonimizar datos personales confidenciales. Admite una amplia variedad de (1) modelos de privacidad y riesgo, (2) métodos para transformar datos, y (3) métodos para analizar la utilidad de los datos de salida. El software se ha utilizado en una variedad de contextos, incluidas plataformas comerciales de análisis de grandes volúmenes de datos, proyectos de investigación, intercambio de datos de ensayos clínicos y con fines de capacitación.
- *Herramienta de anonimización UTD* [19]: Es una herramienta de código abierto desarrollada en el *UT Dallas Data Security* y *Privacy Lab*, donde se implementan varios métodos de anonimización para uso público. Los algoritmos que ofrece se pueden usar tanto directamente contra un conjunto de datos como a través de librerías implementadas dentro de otras

aplicaciones. Utiliza métodos de anonimización diferentes, entre ellos la k-anonimidad.

Además de las herramientas mencionadas, se pueden encontrar otras herramientas en la sección "*Herramientas para anonimizar datos*" dentro del **Anexo I**.

II.5. Herramientas para la minería de datos

Para el estudio que se realiza en este trabajo es necesario poseer un conocimiento general sobre la minería de datos y las diferentes herramientas disponibles para su aplicación, debido a que mas allá de poder hacer público un conjunto de datos preservando la privacidad de las personas, lo que se espera es poder aprovechar los mismos para generar información de interés.

Las herramientas de minería de datos se utilizan para gestionar los datos e identificar las posibles tendencias y patrones más significativos. Los programas desarrollados para ello son cada vez más complejos y el abanico de herramientas cada vez mayor. Para tener una visión general se presentan algunas de las herramientas usadas para el "*data mining*"

- **RapidMiner (antes conocido como YALE)** [20]**:** RadipMiner permite realizar un análisis avanzado de los datos, a través de plantillas. Esta herramienta ofrece un servicio excelente, y está entre las mejores herramientas de data mining hoy en día. Además, dispone de la funcionalidad de pre-procesamiento y visualización de datos, análisis predictivo y modelos estadísticos, así como evaluación y despliegue de la información.
- **Weka** [21]**:** Herramienta basada en Java, que permite analizar datos y establecer modelos predictivos. Igual que RapidMiner, Weka realiza trabajos de data mining estándar, incluyendo pre-procesamiento de datos, clustering, clasificación, regresión, visualización y selección de características. Hoy en día Weka está disponible también como API permitiendo la integración con desarrollos en Java.
- **NLTK** [22]**:** Cuando se trata de tareas de procesamiento del lenguaje, la herramienta recomendada es NLTK, debido a que proporciona un conjunto de herramientas de procesamiento del lenguaje, incluyendo la minería de datos, aprendizaje automático, raspado de datos, análisis de los sentimientos y otras tareas de procesamiento del lenguaje. Debido a que está escrito en *Python*, se pueden construir aplicaciones sobre sí misma y personalizarlo para tareas pequeñas.

Además de las herramientas mencionadas anteriormente, es importante mencionar el lenguaje de programación **R** y la aplicación **RStudio**, que aunque no es una aplicación destinada únicamente a la minería de datos, sí que cuenta con una variedad de paquetes que le permiten aplicar técnicas de minería sobre diversos conjuntos de datos. Para fines de este trabajo se ha decidido utilizar Weka y R como herramientas de minería de datos y es en el **Capítulo III** donde se puede apreciar mayor información sobre estas herramientas.

II.6. Datos personales en el ámbito de salud

La Comisión Europea ha definido los datos personales que son considerados como sensibles y que están sujetos a condiciones de tratamiento específicas. Entre estos datos se mencionan aquellos que están relacionados con el ámbito de salud, los cuales deberán ser cuidadosamente tratados. De esta forma, la Comisión Europea define los siguientes datos sensibles [23]:

- Aquellos datos personales que revelen el origen racial o étnico, las opiniones políticas, las convicciones religiosas o filosóficas.
- Datos de la afiliación sindical.
- ***Datos genéticos, datos biométricos tratados únicamente para identificar un ser humano***.
- ***Datos relativos a la salud***.
- Datos relativos a la vida sexual u orientación sexual de una persona.

Como regla general se prohíbe el tratamiento de estos datos. Entre las excepciones estipuladas por la Unión Europea, se indica que es necesario que exista un consentimiento explícito por parte del interesado y con las finalidades especificadas o que se dé alguna de las siguientes circunstancias [24]:

- Cumplimiento de obligaciones y ejercicios de derechos en el ámbito del Derecho Laboral y de la seguridad y protección social.
- Protección de Intereses vitales del interesado.
- Tratamiento efectuado en el ámbito de fundaciones o asociaciones cuya finalidad sea política, filosófica, religiosa o sindical.
- Tratamiento de datos manifiestamente públicos.
- Tratamientos necesarios para la formulación, ejercicio o defensa de reclamaciones, o tratamientos efectuados por tribunales en el ejercicio de su función judicial.
- Por razón de interés público en el ámbito de la salud pública.
- Con fines de archivo e interés público, fines de investigación científica o histórica o fines estadísticos.

En España, según el artículo 7 y 8 de la Ley Orgánica de Protección de Datos (LOPD), deben ocurrir ambos casos para considerar una excepción a la ley, es decir, que exista consentimiento previo y debe darse alguna de las circunstancias antes mencionadas.

II.6.1. Datos personales relativos a la salud

Los datos relativos a la salud [24] son todos aquellos datos personales relacionados con la salud física o mental de una persona física, incluyendo los datos relacionados con la prestación de servicios de atención sanitaria y datos que puedan revelar información sobre el estado de salud de una persona.

En consecuencia, cualquier información relacionada con enfermedades, discapacidades, riesgo de padecer alguna enfermedad, historial médico, tratamientos clínicos, estado fisiológico, estado biomédico, entre otros, se considera como datos relativos a la salud. Debido a lo transcendental de los mismos, el Reglamento General

de Protección de Datos (RGPD) les otorga el adjetivo de "*Especialmente Protegidos*", añadiendo una serie de normas adicionales sobre estos para permitir su tratamiento.

II.6.2. Autoridades para la protección de datos en España y normativas existentes en la materia

Existen diversas autoridades reconocidas por el grupo de protección de datos de Europa [25] (*European Data Protection Board*), presentando en la Tabla 1 a la autoridad en España, dejando el resto de autoridades disponibles para posibles consultas dentro de la sección de "Autoridades para la protección de datos en Europa *Datos personales en el ámbito de salud"* en el **Anexo I**.

Autoridad en España reconocida por el grupo de protección de datos de Europa			
Autoridad	Email	Página Web	Miembro
España *Agencia Española de Protección de Datos (AEPD)*	internacional@agpd.es	https://www.agpd.es/	Ms María del Mar España Martí, Director of the Spanish Data Protection Agency

Tabla 1 - Autoridad española reconocida por el grupo de protección de datos de Europa, adaptado de [25]

La Agencia Española de Protección de Datos (AEPD) [26], o también conocida por sus siglas AEPD, es la autoridad pública independiente encargada de velar por la privacidad y la protección de datos de los ciudadanos de España. En el contexto actual exige que aquellos que tratan datos apuesten por la implantación de nuevas políticas proactivas de cumplimiento, una labor para la que contarán con todo el apoyo que la Agencia pueda ofrecerles.

Para ayudar en la protección de datos, la Agencia ha elaborado un Plan Estratégico 2015-2019 [27], cuya última modificación tiene fecha del 1 de Marzo de 2020, la finalidad de dicho plan es ofrecer una respuesta práctica a ciudadanos, empresas, profesionales y organismos públicos en relación a la protección de datos. Este documento supone un importante compromiso público por parte de la institución, marcando las que van a ser las líneas de trabajo prioritarias.

En la normativa existente en materia de protección de datos se estipula una variedad de artículos referente a los distintos datos personales que se pueden encontrar hoy en día, indicando las restricciones y condiciones necesarias para tratar dichos datos, en donde indica que *"La ley limitará el uso de la informática para garantizar el honor y la intimidad personal y familiar de los ciudadanos y el pleno ejercicio de sus derechos"* **[Disposición 16673 del BOE núm. 294 de 2018]** [28]. De los artículos encontrados en este *Boletín Oficial de Estado* se presentan extractos de los artículos 1, 5 y 6, los cuales son considerados relevantes en lo referente al tratamiento y privacidad de datos, y los cuales se deben conocer para cualquier tratamiento de datos en general.

*"**Artículo 1**: La presente ley orgánica tiene por objeto*

a) Adaptar el ordenamiento jurídico español al Reglamento (UE) 2016/679 del Parlamento Europeo y el Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de sus datos

personales y a la libre circulación de estos datos, y completar sus disposiciones. El derecho fundamental de las personas físicas a la protección de datos personales, amparado por el artículo 18.4 de la Constitución, se ejercerá con arreglo a lo establecido en el Reglamento (UE) 2016/679 y en esta ley orgánica."

"***Artículo 5**. Deber de confidencialidad.*

1. *Los responsables y encargados del tratamiento de datos, así como todas las personas que intervengan en cualquier fase de este estarán sujetas al deber de confidencialidad al que se refiere el artículo 5.1.f) del Reglamento (UE) 2016/679."*

*"**Artículo 6**. Tratamiento basado en el consentimiento del afectado.*

1. *De conformidad con lo dispuesto en el artículo 4.11 del Reglamento (UE) 2016/679, se entiende por consentimiento del afectado toda manifestación de voluntad libre, específica, informada e inequívoca por la que este acepta, ya sea mediante una declaración o una clara acción afirmativa, el tratamiento de datos personales que le conciernen.*
2. *Cuando se pretenda fundar el tratamiento de los datos en el consentimiento del afectado para una pluralidad de finalidades será preciso que conste de manera específica e inequívoca que dicho consentimiento se otorga para todas ellas.*
3. *No podrá supeditarse la ejecución del contrato a que el afectado consienta el tratamiento de los datos personales para finalidades que no guarden relación con el mantenimiento, desarrollo o control de la relación contractual."*

Llegados a este punto, se puede decir que se han abarcado todos los aspectos principales relacionados con el trabajo de investigación presente y de esta forma se le ha proporcionado al lector suficiente información sobre el estado del arte de la anonimización hoy en día, así como también se ha informado sobre las entidades, normativas y reglamentos existentes en el ámbito de la protección de datos, permitiendo así continuar con el **Capítulo III**, en donde se explican las herramientas y los recursos que sirvieron de apoyo para la realización de este trabajo.

CAPÍTULO III. Recursos y Herramientas.

En este capítulo se muestran los diferentes recursos y herramientas que resultaron de apoyo para la realización del trabajo de investigación, estructurado a través de 5 secciones que hacen referencia al área en el que las herramientas y recursos dieron soporte, como en la minería de datos, desarrollo, diseño, gestión del trabajo, entre otros.

III.1. Minería de datos

Para el análisis de minería realizado en la elaboración de este trabajo, fue necesario hacer uso de las siguientes herramientas:

- **Weka - *Waikato Environment for Knowledge Analysis, en español «entorno para análisis del conocimiento de la Universidad de Waikato»***. Versión 3.8.3 [29]. Se trabajó con WEKA, la cual resulta ser una de las aplicaciones para minería de datos más recomendadas hoy en día. Ofrece una gran variedad de posibilidades al momento de aplicar minería de datos, desde algoritmos y métodos de clasificación, como filtrado sobre el conjunto de datos, visualización de datos a nivel gráfico, entre muchas otras ventajas que ofrece para el tratamiento de datos. Además de contar con una librería para desarrollos en Java, la cual ha servido para el desarrollo de la herramienta de apoyo para la comparación de datos.

El uso de WEKA abarca desde el análisis previo de los conjuntos de datos, donde se identifican sus atributos, instancias, valores nulos, entre otros; hasta la etapa de clasificación y estudio de los conjuntos de datos, donde también se hace uso de la librería de Weka para desarrollos en Java.

- **RStudio**. Versión 1.2.1335 [30]. Es un entorno de desarrollo para trabajos con R, el cual es un lenguaje de programación principalmente orientado a los análisis estadísticos, por lo cual se decide hacer uso de este, dado que ofrece un buen soporte para los trabajos de análisis de datos, además de contar con una gran variedad de librerías y paquetes que permiten el estudio y tratamiento de datos.

Al igual que Weka, se hace uso de esta herramienta dado que, además de permitir el trabajo con diversos conjuntos de datos, en ella es posible aplicar algoritmos y técnicas que pueden ser personalizadas, creadas, y modificadas según lo crea conveniente el autor de este trabajo para lograr así un análisis óptimo de los datos a estudiar.
Estas herramientas son contempladas para realizar la minería de datos debido a su capacidad para trabajar con grandes volúmenes de datos, y también dada la facilidad que ofrecen junto a las diferentes colecciones de algoritmos asociados a la minería de datos, los cuales permiten hacer diferentes estudios y comparativas sobre diversos conjuntos.

Se plantea realizar un trabajo exploratorio de los conjuntos de datos partiendo con Weka, donde mediante su interfaz de “Explorador” se pueda visualizar el conjunto de

datos, sus atributos, sus instancias, generar gráficas ilustrativas sobre los diferentes atributos, y establecer mediante diferentes técnicas de clasificación un valor de precisión inicial.
RStudio, con su integración con el lenguaje R, permite hacer uso de diferentes paquetes y librerías que harán posible estudiar diferentes técnicas y algoritmos de anonimización, con lo cual se podrá complementar el estudio que se realiza con Weka. De esta forma, se trabaja de la mano de estas dos herramientas, aprovechando al máximo las capacidades que presentan, a fin de obtener un estudio claro y preciso sobre los conjuntos de datos seleccionados para este trabajo.

III.2. Diseño

Dado que en este trabajo de investigación se plantea como uno de los objetivos la realización de una herramienta que sirva de apoyo para la comparación de técnicas de anonimización, se decide que para el diseño de la misma se utilice de soporte una aplicación con la cual realizar un diseño inicial que permita apreciar las vistas y secuencias de lo que se quiere desarrollar, facilitando también la realización de una planificación óptima para el desarrollo a realizar.

- **Balsamiq Mockups 3.** Version: 3.5.17. Herramienta de maquetación la cual permite realizar diseños para aplicaciones móviles, web y de escritorio de una forma eficiente, rápida y con la obtención de resultados deseables y presentables a partir de los cuales se orientó el desarrollo de la aplicación en las fases siguientes. Fue creada por *Balsamiq Studios,* que es un proveedor de software independiente fundado en marzo de 2008 por Peldi Guilizzoni [31].

- **Hatchful Online Version:** es una herramienta gratuita creada por Shopify que te permite crear tu propio logo [32].

III.3. Desarrollo

Entre otras herramientas necesarias para la realización del presente trabajo, debemos mencionar aquellas seleccionadas para el desarrollo de la aplicación cuyo objetivo es funcionar como herramienta que permita la comparación entre las distintas técnicas y algoritmos más usados para anonimizar y minar datos.

- **Java 8**. Versión JDK 1.8.0_20. Lenguaje de programación usado para el desarrollo de la aplicación.
- **Java Swing.** Es un *framework* MVC para desarrollar interfaces gráficas para Java con independencia de la plataforma, el cual se obtiene mediante la versión de JDK de Java 8 utilizada [33]. Ofreciendo:
  - Independencia de plataforma.
  - Extensibilidad.
  - Personalizable.

Librerías. Se necesitó hacer uso de diferentes librerías para la realización de la aplicación, de las que principalmente debemos mencionar aquellas que hicieron posible la integración entre la aplicación y las herramientas de minería de datos (Weka y RStudio).

- Para integrar con Weka, se siguieron los pasos indicados en la sección de desarrollo de la wiki de Weka [21], donde se muestran diferentes formas de conectar con la API, de las cuales se decidió hacer uso del fichero .jar de Weka directamente.
- Para el trabajo con las gráficas dentro de Java, se usó la librería de JFreeChart-1.0.19.jar y jcommon-1.0.23.jar.
- Para la aplicación de técnicas de anonimización, se aprovecharon paquetes de R como *addNoise* y *localSuppression* las cuales están disponibles dentro del paquete *sdcMicro* para aplicar ruido y la técnica de k-anonimidad [34].

Patrón de diseño.

Para la realización de la aplicación se trabajó con el patrón de diseño de software Modelo Vista Controlador (MVC), basado en la separación de la lógica del producto y los datos de la interfaz del sistema, con la finalidad de diseñar y hacer la aplicación más sostenible y reusable.

Entorno de Desarrollo Integrado (IDE).

Para el desarrollo de la aplicación se hizo uso de dos IDEs, con el fin de aprovechar los diferentes servicios que estos ofrecen y agilizar el proceso de desarrollo.
Al principio del desarrollo se decidió trabajar con ***IntelliJ IDEA 2019***, pero durante el proceso del desarrollo visual de la aplicación se descubrió la posibilidad de agilizar el trabajo de la interfaz gráfica mediante el *plugin JFormDesigner* de la plataforma; lamentablemente el mismo requería tener una licencia del IDE que era de pago, por lo que se decidió migrar el trabajo a ***Eclipse*** con la ***versión 2020-03 (4.15.0)*** donde fue posible realizar el trabajo de la interfaz gráfica con la ayuda del *plugin Windows Builder* y siguiendo las recomendaciones de arquitectura del Patrón Modelo Vista Controlador (MVC).

III.4. Gestión y planificación

Para la planificación y gestión del desarrollo, se decidió trabajar con la herramienta **Asana**, la cual es una aplicación web y móvil diseñada para mejorar la comunicación y colaboración de equipos [35]. Se decide utilizar esta herramienta debido a que permite no sólo tener un control de las tareas y proyectos e información general de las actividades que se planifican, sino también permite gestionar los tiempos de planificación y estimación de tareas, logrando llevar a través de ella la metodología establecida para el trabajo, la metodología SCRUM, dividiendo así el desarrollo en distintas fases, resumidas en 8 *sprints*.

III.5. Control de versiones

Para el control de versiones se decidió trabajar con Git y Bitbucket, siendo el último un sistema de alojamiento web para proyectos que utilizan el sistema de control de versiones Mercurial y Git [36].

CAPÍTULO IV. Obtención y estudio de los conjuntos de datos

En este capítulo se presenta la metodología y el procedimiento seguido para la obtención de conjuntos de datos que resulten útiles para estudiar el impacto de las diversas técnicas de anonimización, permitiendo establecer en un futuro una comparativa entre los conjuntos sin anonimizar y los anonimizados.

IV.1. Metodología CRISP-DM

Para este trabajo se decidió hacer uso de la metodología CRISP-DM, la cual se utiliza en los trabajos con conjuntos de datos y proporciona de forma normalizada el ciclo de vida de un proyecto estándar de análisis de datos. El mismo se divide en seis fases mostradas en la Figura 3. La secuencia de estas fases no es rígida y se entiende que el trabajo en sí no se termina una vez que se despliega una solución, puesto que la información que se descubre en el proceso y la solución obtenida puede producir nuevas iteraciones del modelo [37].

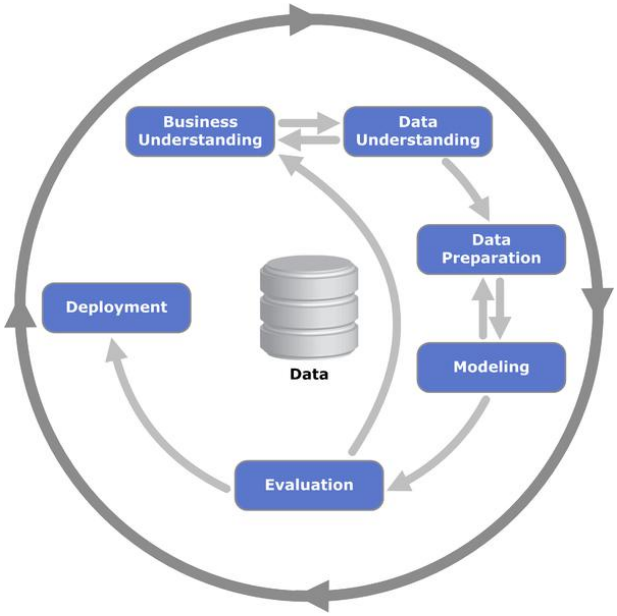


Figura 3 - Metodología CRISP-DM [37]

Siguiendo las diferentes fases mostradas en la Figura 3, se procede a la primera de ellas que busca definir las necesidades del negocio, que en este caso particular son las necesidades del estudio a realizar, donde la principal necesidad está en encontrar varios conjuntos de datos relacionados con la salud. Continuando así por una fase de pre-procesado y modelado, con lo cual se preparan los conjuntos de datos para el estudio, y finalmente por la etapa de evaluación la cual permitirá la obtención de conjuntos de datos con técnicas de anonimización aplicadas, permitiendo una comparación entre ellos.

IV.2. Selección y estudio del conjunto de datos de referencia

Para iniciar el estudio se decidió que sería idóneo usar un conjunto de datos base que sirviera de referencia para seleccionar los conjuntos de datos a estudiar más adelante, de forma que se buscaran conjuntos de datos con características similares a

este. Por lo tanto, se parte de una variedad de conjuntos de datos accesibles a través de las librerías de Weka y R, y se decide hacer uso del conjunto de datos *"diabetes.arff"* (Pima Indians Diabetes Database), debido a la diversidad de estudios relacionados con la minería de datos realizados con este conjunto de datos ([38], [39], [40]).

Características del conjunto de datos de referencia

A continuación se presenta en la Tabla 2 las características observadas en el conjunto de datos de Pima Indians diabetes

Conjunto de datos Pima Indians Diabetes	
Tipo de conjunto de datos	Multivariado
Número de Instancias	768
Área	Salud
Tipo de Atributos	Todos valores numéricos
Número de Atributos	8 más clase
Fecha de donación	9-May-90
Valores faltantes	No

Tabla 2 - Características del conjunto de datos de diabetes, basado en [41]

Contexto y Contenido

Este conjunto de datos es originalmente del "*National Institute of Diabetes and Digestive and Kidney Diseases*". El objetivo del conjunto de datos es predecir de forma diagnóstica si un paciente tiene diabetes o no, basándose en ciertas mediciones de diagnóstico incluidas en el conjunto de datos. Se añadieron varias restricciones en la selección de estas instancias de una base de datos más grande. En particular, todos los pacientes son mujeres de al menos 21 años de herencia indígena Pima. Las variables predictivas presentes incluyen:

1. *Número de embarazos.*
2. *Concentración de glucosa plasmática a 2 horas.*
3. *Presión arterial diastólica (mm Hg).*
4. *Grosor del pliegue de la piel del tríceps (mm).*
5. *Insulina sérica de 2 horas (mu U / ml).*
6. *Índice de masa corporal (kg / m) ^ 2).*
7. *Función de pedigrí de diabetes.*
8. *Edad (años).*

Pre-procesado de datos

El pre-procesamiento de datos es un paso significativo en el proceso de descubrimiento de conocimiento, ya que las decisiones de calidad deben basarse en datos de calidad.

Un total de 768 casos están disponibles en el conjunto de datos, algunos de los cuales contienen valores que carecen de sentido. La razón puede deberse a errores en la recopilación de los datos de los individuos o simplemente que los individuos no concedieron algunos de estos datos. Por lo tanto, se pueden observar casos como por ejemplo individuos con edad 0, con el IMC 0. Dichos registros se han eliminado del conjunto de datos. La Tabla 3 muestra el número de registros eliminados como parte del pre-procesado de datos.

Atributos pre-procesados (valores 0)	N. de instancias eliminadas
Presión Arterial Diastólica	35
Prueba de tolerancia a la glucosa	5
Índice de masa corporal	4
Grosor del pliegue de la piel del tríceps	192
Insulina sérica de 2 horas	140
Total	376

Tabla 3 - Pre-procesado del conjunto de datos de diabetes, adaptado [42]

Después de eliminar estas instancias, quedaron unos 392 casos sin valores faltantes. La Tabla 4 explica las propiedades estadísticas de los atributos asociados a estas instancias finales.

Atributo	Mínimo y Máximo	Desviación Media	Desviación Estándar
1	0 – 17	3.301	3.211
2	56-198	122.628	30.861
3	24 – 110	70.663	12.496
4	7-63	29.145	10.516
5	14 – 846	156.056	118.842
6	18.2 - 67.1	33.086	7.028
7	0.085 - 2.42	0.523	0.345
8	21 – 81	30.865	10.201

Tabla 4 - Propiedades de los atributos tras pre-procesado de datos, adaptado de [42]

IV.3. Selección y estudio de los conjuntos de datos para experimentación

Se procedió a seleccionar de diferentes fuentes dos conjuntos de datos que, además de estar disponibles, estuviesen relacionados con el ámbito de salud, para estudiarlos y seleccionar de entre los mismos aquellos en los cuales fuese posible aplicar estrategias de anonimización como las antes mencionadas, observando en ellos los datos aportados, sus limitaciones de uso, y características generales.
En la Tabla 5 se presenta una comparativa de 5 conjuntos de datos pre-seleccionados donde es posible observar de forma general las características principales observadas en los mismos, resaltando entre ellos aquellos seleccionados para el desarrollo de este trabajo.

Conjunto de datos	Nº Instancias	Nº Atributos	Atributo Objetivo	Anonimizada	Tipo de Anonimización observada
Heart Disease Dataset [43]	1025	14	**"target"**: Presencia/Ausencia de enfermedad cardíaca	Sí	Eliminación y generalización de cuasi-identificadores
Breast Cancer Wisconsin (Diagnostic) Data Set [44]	569	32	**"diagnosis"**: Cáncer benigno o maligno	Sí	Eliminación de datos que permitan re-identificación de los pacientes

Medical Appointment No Shows [45]	110527	14	**"No Show":** Asiste o no a la consulta	Sí	Eliminación de atributos pseudo-identificativos, como nombres, códigos postales, número de la seguridad social, entre otros
Heart Disease UCI Dataset [46]	303	76	**"target"**: Presencia/Ausencia de enfermedad cardíaca	Sí	Eliminación de cuasi-identificadores, y otros atributos médicos
Sample Insurance Claim Prediction [47]	1338	9	**"Insuranceclaim":** Hay/no hay reclamación de seguro	Sí	Eliminación y generalización de cuasi-identificadores

Tabla 5 - Tabla comparativa de los diferentes conjuntos de datos estudiados para el trabajo

El criterio para la selección de los mismos se basó en el interés propio del investigador sobre el objeto de estudio del conjunto de datos, así como el hecho de que el tipo de conjunto de datos tuviese cierta similitud al conjunto de referencia, como en el tipo de atributos, la clase objetivo, junto con la cantidad de datos disponibles para experimentación, el número de atributos, y la anonimización previamente realizada sobre los conjuntos de datos.

En el caso particular del conjunto de datos "*Heart Disease UCI Dataset"*, se decide trabajar con el mismo porque mediante la investigación se encontró que los datos observados en los dos conjuntos de datos de enfermedades cardíacas observados en la Tabla 5, coinciden entre ellos con la diferencia en que en uno de ellos se poseen instancias duplicadas, y en el otro se han eliminado las mismas, por lo que se ha decidido trabajar con el conjunto de datos "más limpio". Y finalmente, de la misma forma en que se analizó el conjunto de datos de referencia, se procede a analizar los conjuntos de datos seleccionados.

Características de los conjuntos de datos para el estudio

De igual manera en la que se estudiaron las características para el conjunto de datos de referencia, se muestra en la Tabla 6 y en la Tabla 7 las características asociadas a los 2 conjuntos de datos seleccionados.

1. Conjunto de datos Heart Disease UCI.

Conjunto de datos Heart Disease UCI	
Tipo de conjunto de datos	Multivariado
Número de Instancias	303
Área	Vida
Tipo de Atributos	Categórico, Entero, Real
Número de Atributos	75
Fecha de donación	1-Jul-88
Valores faltantes	Sí

Tabla 6 - Características del conjunto de datos Heart Disease UCI, basado en [41]

2. Conjunto de datos Breast Cancer Wisconsin (Diagnostic)

Conjunto de datos Heart Disease UCI	
Tipo de conjunto de datos	Multivariado
Número de Instancias	569
Área	Vida
Tipo de Atributos	Real
Número de Atributos	32
Fecha de donación	1-Nov-95
Valores faltantes	No

Tabla 7 - Características del conjunto de datos Breast Cancer Wisconsin, basado en [48]

Contexto y Contenido

1. Conjunto de datos Heart Disease UCI.

Esta base de datos contiene 76 atributos, pero todos los experimentos publicados se refieren al uso de un subconjunto de 14 de ellos. En particular, la base de datos de Cleveland es la única que ha sido utilizada por investigadores de *Machine Learning* para esta fecha. El campo "objetivo" se refiere a la presencia de enfermedad cardíaca en el paciente. Tiene un valor entero de 0 (sin presencia) a 4. Los experimentos con la base de datos de Cleveland se han concentrado en el simple intento de distinguir la presencia (valores 1, 2, 3,4) de la ausencia (valor 0) de enfermedad cardíaca [46]. Dentro del conjunto de datos se puede observar que los datos contienen atributos predictivos médicos y un atributo objetivo o clase, los cuales se listan a continuación:

1. *Edad.*
2. *Sexo (1 =masculino; 0 = femenino).*
3. *Tipo de dolor de pecho (1 = angina típica, 2 = angina atípica, 3 = dolor no anginal, 4 = asintótico).*
4. *Presión arterial en reposo (en mm Hg al ingreso al hospital).*
5. *Colesterol en mg / dl.*
6. *Azúcar en sangre en ayunas >120 mg / dl (1 = verdadero; 0 = falso).*
7. *Resultados electrocardiográficos en reposo (0 = normal, 1 = tener anormalidad en la onda ST-T, 2 = hipertrofia ventricular izquierda).*
8. *Frecuencia cardíaca máxima alcanzada.*
9. *Angina inducida por ejercicio (1 = sí; 0 = no).*
10. *Depresión inducida por el ejercicio ST relativo al descanso (1 = pendiente ascendente, 2 = plano, 3 = descenso).*
11. *La pendiente del segmento pico del ejercicio ST.*
12. *Número de vasos principales (0-3) coloreados por flourosopía.*
13. *El estado del corazón tras la prueba de talio (3 niveles: N-Normal, FD-Defecto fijo, RD-Defecto reversible).*
14. *Objetivo (1= Presencia de enfermedad cardíaca o 0= Ausencia de enfermedad).*

2. Conjunto de datos Breast Cancer Wisconsin (Diagnostic)

Esta base de datos contiene 32 atributos y 569 instancias. El campo objetivo o clase se refiere al diagnóstico del tumor (B = benigno, M = maligno). La intención es hacer uso de estos datos para intentar predecir el diagnóstico de un tumor de mama.

Las características se calculan a partir de una imagen digitalizada de un aspirado con aguja fina (FNA) de una masa mamaria, dichas características describen los núcleos celulares presentes en la imagen. Entre los atributos podemos encontrar:

1. *Número de Identificación*
2. *El diagnóstico de los tejidos mamarios (M = maligno, B = benigno)*
3. *Radio medio, media de distancias desde el centro a puntos en el perímetro*
4. *Textura media, desviación estándar de valores de escala de grises*
5. *Perímetro medio, tamaño medio del perímetro del tumor central*
6. *Área media, tamaño medio del área del tumor central*
7. *Suavidad media, media de variación local en longitudes de radio*
8. *Media de compacidad, media del perímetro ^ 2 / área - 1.0*
9. *Media de concavidad, media de las porciones cóncavas del contorno*
10. *Media de puntos cóncavos, media para el número de porciones cóncavas del contorno*
11. *Media de simetría*
12. *Media de la dimensión fractal*
13. *Error estándar de radio para la media de distancias desde el centro a puntos en el perímetro*
14. *Error estándar de textura para desviación estándar de valores de escala de grises*
15. *Error estándar de perímetro*
16. *Error estándar del área*
17. *Error estándar de suavidad para variación local en longitudes de radio*
18. *Error estándar de compacidad*
19. *Error estándar de concavidad, error por la gravedad de las partes cóncavas del contorno*
20. *error estándar de puntos cóncavos para el número de porciones cóncavas del contorno*
21. *Error estándar de simetría*
22. *Error estándar de dimensión fractal*
23. *"Peor" o mayor valor medio para la media de distancias desde el centro a puntos en el perímetro*
24. *"Peor" o mayor valor medio para la desviación estándar de los valores de escala de grises*
25. *“Peor” o mayor valor medio para la desviación estándar del perímetro*
26. *“Peor” o mayor valor medio para la desviación estándar del área*
27. *"Peor" o mayor valor medio para la variación local en longitudes de radio*
28. *"Peor" o mayor valor medio para el perímetro ^ 2 / área - 1.0*
29. *"Peor" o mayor valor medio para la gravedad de las partes cóncavas del contorno*
30. *"Peor" o valor medio más grande para el número de porciones cóncavas del contorno*
31. *“Peor” o mayor valor medio de simetría*
32. *"Peor" o mayor valor medio para "aproximación de la costa" – 1*

Pre-procesado de datos

Como se indicó anteriormente, el pre-procesado de datos es importante. Con este podemos identificar instancias con valores que carecen de lógica o que simplemente se encuentran vacíos dentro del conjunto de datos, como se observó para el caso del conjunto de datos de diabetes, de forma que se eliminan para cada conjunto de datos estas instancias con valores vacíos a fin de obtener un conjunto de datos más “limpio”.

1. Conjunto de datos Heart Disease UCI

Atributos pre-procesados (valores 0/nulo)	N. de instancias eliminadas
Número de vasos principales (0-3) coloreados por flourosopía	4
El estado del corazón tras la prueba de talio.	2
Total	6

Tabla 8 - Pre-procesado del conjunto de datos Heart Disease UCI, adaptado de [42]

Después de eliminar las instancias observadas en la Tabla 8, se obtiene un total de 297 instancias sin valores faltantes. La Tabla 9 explica las propiedades estadísticas de los atributos numéricos.

Atributo	Mínimo y Máximo	Desviación Media	Desviación Estándar
1	29 – 77	54.542	9.05
2	*Sexo: Atributo Nominal (2 valores posibles)*		
3	*CP: Atributo Nominal (4 valores posibles)*		
4	94-200	131.694	17.763
5	126-564	247.35	51.998
6	*FBS: Atributo Nominal (2 valores posibles)*		
7	*RESTECG: Atributo Nominal (3 valores posibles)*		
8	71-202	149.599	22.942
9	*EXANG: Atributo Nominal (2 valores posibles)*		
10	0-6.2	1.056	1.166
11	*SLOPE: Atributo Nominal (3 valores posibles)*		
12	*CA: Atributo Nominal (4 valores posibles)*		
13	*THAL: Atributo Nominal (3 valores posibles)*		

Tabla 9 - Propiedades de los atributos numéricos tras el pre-procesado de datos, adaptado de [42]

Para estudiar los atributos nominales, fue necesario aplicar el filtro de Weka "*NumericToNominal*" indicándole los atributos nominales resaltados en la Tabla 9, permitiendo así observar las propiedades de estos atributos como se muestra en la Tabla 10.

Atributo	Valores Posibles	Total
2	0	96
	1	201
3	1	23
	2	49
	3	83
	4	142
6	0	254
	1	43
9	0	200
	1	97
11	1	139
	2	137
	3	21
12	0	174
	1	65
	2	38
	3	20
13	3	164

7	0	147		6	18
	1	4		7	115
	2	146			

Tabla 10 - Propiedades de los atributos nominales tras el pre-procesado de datos

2. Conjunto de datos Breast Cancer Wisconsin (Diagnostic)

Atributos pre-procesados (valores 0/nulo/falsos)	N. de instancias eliminadas
Media de concavidad	15
Error estándar de concavidad	1
Erro estándar de dimensión fractal	3
Total	19

Tabla 11 - Pre-procesado del conjunto de datos Breast Cancer Wisconsin, adaptado de [42]

Después de eliminar las instancias observadas en la Tabla 11, se obtienen 550 instancias sin valores faltantes. La Tabla 12 explica las propiedades estadísticas de los atributos numéricos, en la misma se omiten los atributos 1 y 2, dado que el atributo 1 consiste en los identificadores asignados a las instancias del conjunto de datos, cuyo valor no resulta relevante para el estudio, y el atributo 2 corresponde a la clase de estudio, por lo que se deja para análisis posteriores.

Atributo	Mínimo y Máximo	Desviación Media	Desviación Estándar
3	6.981 – 28.11	14.127	3.524
4	9.71 – 39.28	19.29	4.301
5	43.79 – 188.5	91.969	24.299
6	143.5 – 2501	654.889	351.914
7	0.053 – 0.163	0.096	0.014
8	0.019 – 0.345	0.104	0.053
9	0 – 0.427	0.089	0.08
10	0 – 0.201	0.049	0.039
11	0.106 – 0.304	0.181	0.027
12	0.05 – 0.097	0.063	0.007
13	0.112 – 2.873	0.405	0.277
14	0.36 – 4.885	1.217	0.552
15	0.757 – 21.98	2.866	2.022
16	6.802 – 542.2	40.337	45.491
17	0.002 – 0.031	0.007	0.003
18	0.002 – 0.135	0.025	0.018
19	0 – 0.396	0.032	0.03
20	0 – 0.053	0.012	0.006
21	0.008 – 0.079	0.021	0.008
22	0.001 – 0.03	0.004	0.003
23	7.93 – 36.04	16.269	4.833
24	12.02 – 49.54	25.677	6.146
25	50.41 – 251.2	107.261	33.603
26	185.2 – 4254	880.583	569.357
27	0.071 – 0.223	0.132	0.023

28	0.027 – 1.058	0.254	0.157
29	0 – 1.252	0.272	0.209
30	0 – 0.291	0.115	0.066
31	0.157 – 0.664	0.29	0.062
32	0.055 – 0.207	0.084	0.018

Tabla 12 - Propiedades de los atributos numéricos tras pre-procesado de datos, adaptado de [42]

De esta forma es posible continuar con la siguiente fase de la investigación, en la cual se desarrolla la herramienta que sirve de apoyo para comparar los resultados a obtener tras aplicar técnicas de anonimización sobre los conjuntos de datos seleccionados anteriormente.

CAPÍTULO V. Desarrollo de una herramienta de apoyo a la evaluación de estrategias de anonimización sobre conjuntos de datos

A continuación se presentan las diferentes fases del desarrollo e investigación que llevaron a la realización de la aplicación ***CPDA - Comparator of precision in data anonymization***, la cual permite comparar diferentes técnicas de anonimización aplicadas sobre un mismo conjunto de datos, a fin de poder estudiar el impacto de estas técnicas sobre el conjunto de datos.

V.1. Alcance, requisitos funcionales y no funcionales y diagramas de la herramienta

Antes de iniciar el proceso para realizar el diseño de la aplicación fue necesario definir el alcance de la herramienta a desarrollar, estableciendo los requisitos a satisfacer, tomando en consideración que lo que se plantea es el desarrollo de una herramienta que sirva de apoyo para la comparación entre conjuntos de datos y diversas técnicas de anonimización aplicadas, se definen los requerimientos observados en la Tabla 13 y la Tabla 14.

Requerimiento	Descripción del requerimiento
RF1	La aplicación debe permitir la carga de conjuntos de datos mediante una interfaz gráfica.
RF2	La aplicación debe permitir la visualización del resumen asociado al conjunto de datos, donde se muestre el nombre del conjunto de datos, la cantidad de instancias y de atributos, similar a como ocurre en la aplicación Weka.
RF3	La aplicación debe permitir al usuario visualizar los atributos del conjunto de datos cargado, y las características asociadas a dichos atributos, de forma similar a como ocurre dentro de la aplicación Weka.
RF4	La aplicación debe permitir al usuario visualizar el conjunto de datos que ha cargado, permitiendo observar las instancias y sus valores para cada uno de los atributos.
RF5	La aplicación debe ofrecer al usuario la posibilidad de aplicar técnicas de anonimización sobre el conjunto de datos que se ha cargado.
RF6	La aplicación debe permitir al usuario aplicar más de una técnica de anonimización sobre el conjunto de datos cargado.
RF7	La aplicación debe permitir que el usuario seleccione las técnicas que desea comparar.
RF8	La aplicación debe permitir al usuario aplicar técnicas de anonimización mediante herramientas externas como por ejemplo R.
RF9	La aplicación debe permitir al usuario elegir entre un grupo de técnicas de clasificación de datos.
RF10	La aplicación deberá utilizar la técnica de clasificación seleccionada por el usuario para clasificar los conjuntos de datos obtenidos con las técnicas de anonimización y el conjunto de datos cargado al inicio.
RF11	La aplicación deberá mostrar al usuario los resultados obtenidos al clasificar los conjuntos de datos generados mediante la aplicación de las técnicas de anonimización

	seleccionadas previamente por el usuario para su comparación.
RF12	Los resultados deberán presentarse mediante la interfaz gráfica, permitiendo que el usuario visualice claramente los valores obtenidos para cada técnica de anonimización seleccionada.
RF13	La aplicación deberá permitir la descarga de los conjuntos de datos obtenidos tras aplicar las técnicas de anonimización seleccionadas por el usuario.
RF14	La descarga de los conjuntos de datos deberá realizarse sobre una carpeta identificada bajo el nombre de la aplicación, permitiendo que el usuario identifique cada conjunto como crea conveniente.
RF15	La aplicación deberá ofrecer la posibilidad de visualizar los resultados obtenidos mediante gráficas que permitan al usuario comparar visualmente las técnicas seleccionadas y los resultados obtenidos por cada una de ellas.

Tabla 13 - Requisitos funcionales de la herramienta CPDA

Requerimiento	Descripción del requerimiento
RNF1	El usuario tendrá la posibilidad de ejecutar la aplicación dentro de un ordenador con Sistema Operativo Windows.
RNF2	Se deberá ofrecer una interfaz amigable, intuitiva y conversacional donde el usuario podrá ejecutar las acciones que la aplicación ofrece de manera sencilla.
RNF3	La aplicación deberá conectar con la API de Weka para fines de minería de datos.
RNF4	La aplicación deberá ser un archivo ejecutable .jar que facilite a los usuarios el uso de la herramienta.
RNF5	La aplicación deberá ser fácil de utilizar.

Tabla 14 - Requisitos no funcionales de la herramienta CPDA

En consecuencia, se realizó el diagrama de arquitectura observado en la Figura 4, en el cual se puede ver que la aplicación está dividida en 3 secciones: la sección de la interfaz (*Vista*) con la cual interactúa el usuario tanto para insertar conjuntos de datos, como para obtener los resultados y los conjuntos de datos anonimizados; la sección del controlador (*Controller*), que se encarga de procesar la información solicitada por el usuario y proveer mediante el modelo de datos y de la librería de Weka la respuesta a la petición realizada por el usuario; y, finalmente, se encuentra la sección del modelo de datos (*Modelo*), donde se definen los datos con los que se trabaja.

También se realizó el diagrama de clases observado en la Figura 5, en el cual se representan aquellas clases de la aplicación involucradas en la lógica para aplicar técnicas de anonimización sobre un conjunto de datos en la aplicación CPDA. En este diagrama es posible observar las diferentes clases que interactúan en el proceso de anonimizar el conjunto de datos importado por el usuario, desde las clases que intervienen a nivel gráfico hasta aquellas clases que intervienen a nivel lógico para aplicar técnicas de anonimización y técnicas de clasificación.

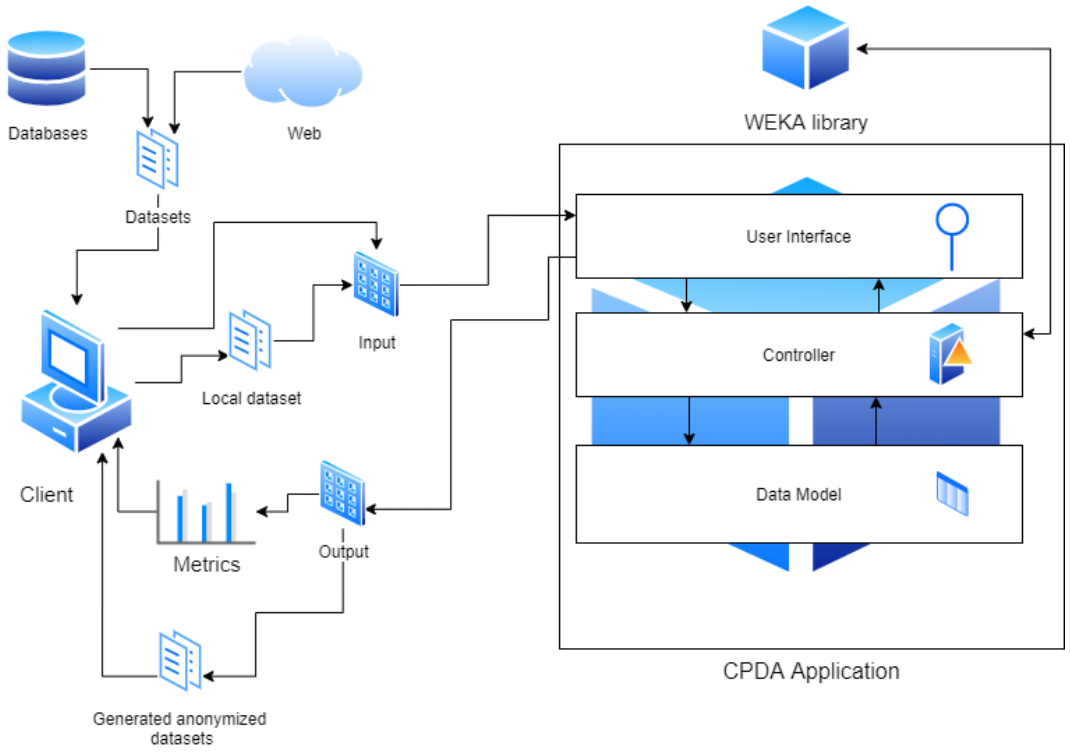


Figura 4 - Diagrama de arquitectura de la herramienta CPDA

Si se llegase a presentar la necesidad de modificar o añadir nuevas técnicas de anonimización, será necesario tener en cuenta el diagrama de clases que se muestra en la Figura 5. Lo primero a considerar es que las técnicas de anonimización son aplicadas a través de la clase "*Techniques*", por lo que para añadir o modificar la misma se podría o bien modificar directamente la clase, o lo más recomendable, crear una nueva clase que extienda de la clase original "*Techniques*" y añada o modifique en ella los métodos que se consideren. Luego será necesario añadir la lógica asociada en las clases que interactúan con dichas técnicas.

De forma similar ocurre para el caso en el que se desee añadir alguna técnica de clasificación diferente a las actuales de la herramienta, donde habrá que considerar que los algoritmos de clasificación vienen importados directamente de la clase "*classifiers*" de Weka. Por tanto, se deberá considerar si la nueva técnica de clasificación proviene de Weka o si se va a aplicar una técnica de clasificación propia; para el primer caso bastará con importar la nueva técnica de clasificación deseada en el controlador, mientras que para el segundo caso será necesario crear una clase correspondiente para la técnica de clasificación, donde se aplique toda la lógica asociada, y finalmente hacer uso de la clase creada dentro del controlador, añadiendo la posibilidad de usar la nueva técnica de clasificación en el método "*Classify*".

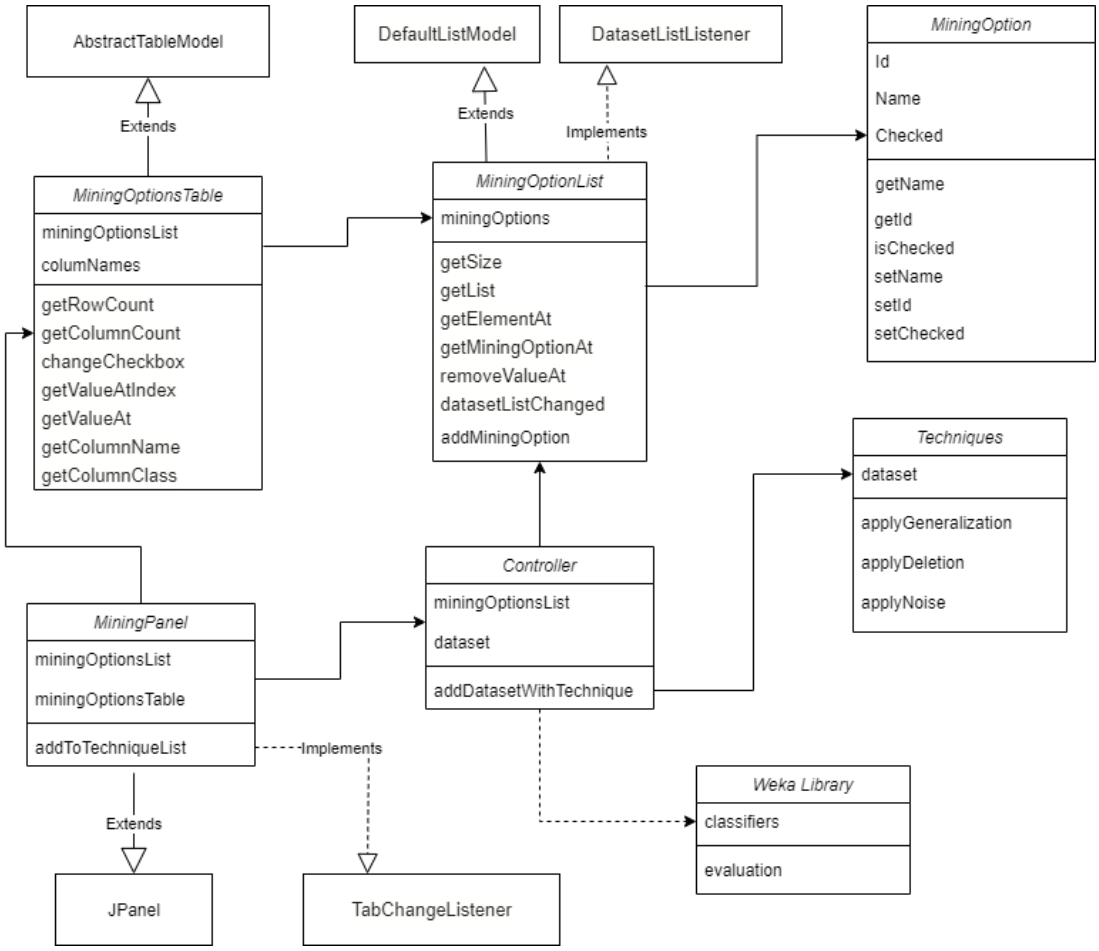


Figura 5 - Diagrama de clases para la aplicación de anonimización sobre un conjunto de datos en CPDA

V.2. Diseño Inicial de la herramienta gráfica

Una vez definido el alcance de la herramienta a desarrollar, se procedió con la realización del diseño de la herramienta, la cual se empezó a esbozar tras un proceso de lluvia de ideas cuyo resultado da base a la realización del diseño final.

Se realizó un boceto digital inicial de la aplicación con la ayuda de una herramienta para creación de *Mockups* (bocetos): *Balsamiq Mockups* [31]. En ella se planifican las diferentes vistas de la aplicación y se realizan los diseños de las mismas. Entre algunos de los diseños realizados se tienen los presentados en las Figuras 6 y 7, donde se visualizan algunas de las secciones esbozadas al principio del trabajo, así como la lógica planteada para la herramienta desde el inicio.

1. Sección de carga de datos

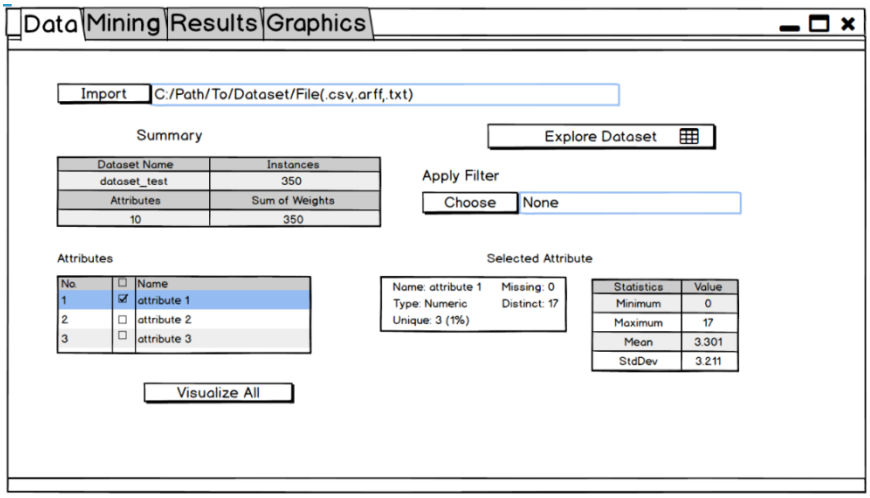


Figura 6 - Boceto inicial de la sección de carga de conjuntos de datos de la herramienta CPDA

2. Sección de minería de datos.

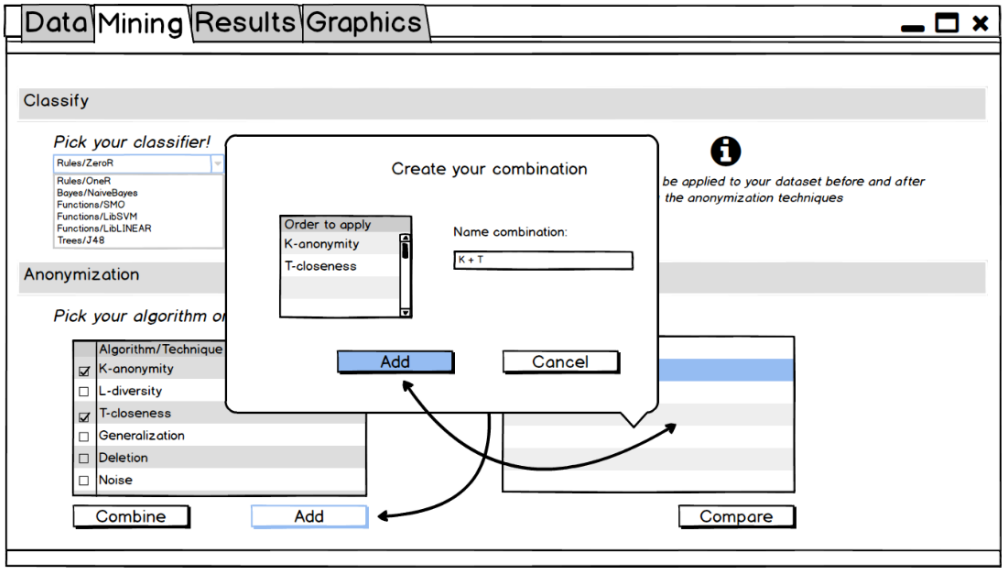


Figura 7 - Boceto inicial de la sección de minería de datos de la herramienta CPDA

Es de importancia destacar que el diseño que se muestra inicialmente es sólo un boceto inicial de lo que viene a ser la aplicación final. Este diseño tuvo algunas modificaciones a lo largo del desarrollo, como por ejemplo cosas que fueron descartadas en la implementación por diferentes motivos y sólo aquellas modificaciones que implicasen un cambio importante en la lógica básica de la aplicación fueron rediseñados, siempre buscando mantener o incluso mejorar la interfaz gráfica. Los bocetos iniciales y finales se pueden observar en el **Anexo II**.

V.3. Metodología para el desarrollo

Se decidió trabajar con una metodología basada en SCRUM [49], donde se definieron 5 fases (8 *sprints*) principales, con una durabilidad inicialmente definida para cada una de las etapas, y que fue evolucionando en base a los avances que se iban realizando en

cada iteración, y en donde se realizaron las reestimaciones necesarias según los obstáculos encontrados.
Se hizo uso de la herramienta Asana [35], mencionada anteriormente, con la que se da soporte y seguimiento a la metodología aplicada. En el **Anexo II** se puede observar en detalle cada uno de los *sprints*, con las tareas correspondientes, y los resultados obtenidos al final de cada s*print*.

V.4. Técnicas de minería y estrategias de anonimización utilizadas en la herramienta CPDA

Para este trabajo se habilitaron 6 técnicas de clasificación dentro de la aplicación ***CPDA***, las cuales serán usadas para la comparativa de los conjuntos de datos iniciales y los modificados con técnicas de anonimización. De igual forma se habilitaron 3 técnicas de anonimización por defecto las cuales se explican más adelante.

V.4.1. Técnicas de clasificación

Las técnicas de clasificación habilitadas en la herramienta (ver Tabla 15) son accesibles mediante la librería de Weka, la cual se ha integrado en el desarrollo de la aplicación, de forma que se hace uso directo de las técnicas disponibles en la librería. Para el trabajo realizado se aplican los clasificadores haciendo uso de la técnica de validación cruzada o *cross-validation* [50] con el valor por defecto de 10 *folds*.

Clasificador	Resumen
ZeroR	Aplica una predicción sobre la clase o categoría principal de un conjunto de datos. Resulta útil para determinar un valor base sobre el cual comparar otros métodos de clasificación [51].
***One Rule* - OneR**	Para cada predictor en los datos se genera una regla y se selecciona la misma con el error total más pequeño como "regla única" [52].
Naive Bayes	Tipo de clasificador en el que se asume que la presencia o ausencia de una característica particular no está relacionada con la presencia o ausencia de cualquier otra característica, dada la clase variable. [53]
***Sequential Minimal Optimization* - SMO**	Es un clasificador en el cual se utiliza el algoritmo secuencial de optimización mínima de John Platt, con el fin de entrenar un clasificador de vectores de soporte. Este tipo de implementación reemplaza globalmente todos los valores faltantes, transforma los atributos nominales en binarios y normaliza todos los atributos por defecto. [54]
***Support Vector Machine* - SVM (LibSVM / LibLINEAR)**	Es un algoritmo de aprendizaje supervisado que se puede utilizar tanto para técnicas de clasificación como de regresión. Se basa en la idea de encontrar un hiperplano que mejor divida el conjunto de datos en 2 clases [55].
J48	Se basa en implementar el algoritmo C4.5, el cual consiste en un algoritmo usado para generar un árbol de decisión para la clasificación de datos [56].

Tabla 15 - Técnicas de clasificación habilitadas dentro de la herramienta CPDA

Es importante mencionar que para el caso de la clasificación lineal se hace uso de dos librerías de Weka cuyos algoritmos en principio cumplen una misma función pero se diferencian en el proceso que siguen para cumplir su objetivo. Principalmente su diferencia está en que el algoritmo de SVM utiliza muchas transformaciones del kernel; para convertir un problema no lineal en uno lineal antes de aplicar la clasificación, mientras que LibLinear hace lo mismo sin hacer uso de las transformaciones del kernel, por lo tanto en aquellos casos en los que las transformaciones de kernel no sean necesarias LibLinear sería el clasificador lineal más recomendado.

V.4.2. Técnicas de anonimización

Dentro de la herramienta de CPDA se pueden comparar técnicas de anonimización aplicadas mediante la propia herramienta como técnicas aplicadas a través de herramientas externas como R.

V.4.2.1. Técnicas de anonimización mediante R

Entre el tipo de técnicas aplicadas usando aplicaciones externas se utilizaron 2: una para añadir ruido al conjunto de datos, y otra para aplicar la técnica de k-anonimización. Ambas técnicas provienen del paquete *sdcMicro* de R, para lo cual se utilizaron los scripts que se encuentran disponibles en el **Anexo IV**.

Para la técnica externa de ruido se hace uso del método "*addNoise*", donde se añade un ruido aleatorio a todo el conjunto de datos y como resultado de aplicar dicha técnica se genera un nuevo grupo de conjuntos de datos anonimizados, los cuales se encuentran disponibles en el repositorio en la ruta: */datasets/noisyR.*

Para la segunda técnica externa, donde se aplica el algoritmo de *localSuppression*, se aplica la técnica de k-anonimidad con k=2, y se almacenan los conjuntos de datos generados en el directorio local. Se aplicó dicha técnica a los 3 conjuntos de datos seleccionados obteniendo así 3 ficheros, los cuales también se encuentran disponibles en el repositorio bajo la ruta: */datasets/localSuppR.* Además de los conjuntos de datos anonimizados se generaron unas gráficas donde se muestra el número de supresiones que fueron necesarias para alcanzar el nivel de anonimidad con k=2, y las cuales quedan disponibles en el repositorio en la ruta: */datasets/localSuppR/plots/*.

V.4.2.2. Técnicas de anonimización disponibles en CPDA

En la herramienta CPDA se habilitaron 3 técnicas de anonimización, las cuales fueron implementadas dentro de la misma y cuyo funcionamiento se explica a continuación.

1. Generalización

Para la técnica de generalización se siguió un método en el cual se seleccionan cada una de las instancias del conjunto de datos en cuestión. Luego, por cada valor numérico en cada instancia se asigna un nuevo valor correspondiente a un rango asociado a la misma, de la forma siguiente: para cada atributo de la instancia, se escoge el valor mínimo y el valor máximo del atributo, se calcula la distancia entre el mínimo y el máximo y se divide el valor obtenido entre 4, de forma que se obtienen 4 rangos a partir de ese valor, y se usa el mínimo como valor inicial del rango 1; por ejemplo, si se tiene un atributo cuyo valor mínimo es 15 y el valor máximo es 80, se

dividirían los valores posibles en 4 rangos, distanciados por el resultado de dividir el valor máximo menos el valor mínimo entre 4.

Por tanto, en el ejemplo se tendría que la distancia sería el resultado de dividir (80 – 15) / 4, es decir 65 / 4 = 16.25. Luego, partiendo del valor mínimo del atributo dado, en este caso 15, tendríamos los siguientes rangos: rango 1: [15 – 31.25), rango 2: [31.25 – 47.5), rango 3: [47.5 – 63.75) y rango 4: [63.75 – 80]. Y una vez que se tienen los rangos correspondientes al atributo, lo que se realiza es una sustitución del valor inicial en la instancia para dicho atributo, por el número del rango al que pertenece, es decir 1, 2, 3 ó 4.

Con este método no sólo se realiza una generalización sobre los atributos, sino que también se anonimizan los rangos de tal forma que no es posible identificar mediante el valor del atributo el posible valor original del dato. De esta forma, al aplicar generalización sobre un atributo como la edad, el usuario que tenga acceso al conjunto de datos anonimizado no podrá saber las diferentes edades recopiladas en el conjunto de datos original: sólo podrá ver valores entre 1 y 4, correspondientes a los rangos a los que pertenece la edad en cada una de las instancias.

2. Ruido

Para la técnica de ruido se procede de forma similar a la técnica usada en la generalización, seleccionando cada una de las instancias y recorriendo cada una de sus columnas (atributos). Si el atributo es numérico se procede a aplicar el ruido, el cual consiste en generar un primer valor aleatorio *"r"*, luego obtener un valor mínimo partiendo de la desviación estándar del atributo divido por 4, obtener el valor máximo que es igual a la desviación estándar de ese atributo, y finalmente generar el valor del ruido a aplicar, el cual consistirá en el valor mínimo más la diferencia entre el valor máximo y mínimo obtenidos multiplicado por el valor aleatorio *"r"* generado inicialmente.

Una vez que se obtiene el valor del ruido (*noise*), simplemente se aplica el mismo a todos los valores de la columna en cuestión, sumando el ruido al valor inicial y generando así un nuevo valor para cada atributo de cada instancia.
La lógica usada para generar el ruido en este caso parte de analizar diferentes técnicas usadas en otros algoritmos de ruido, en donde parten del valor de la desviación estándar para generar un valor aleatorio.

3. Eliminación

Para la técnica de eliminación se decidió aplicar la eliminación de registros, donde se eliminarían aquellos registros cuyos atributos fuesen únicos dentro del conjunto de datos. Para lograr esto se realiza una iteración por columnas (atributos) y se van recorriendo las filas (instancias) del conjunto de datos validando si hay valores que no se repitan del atributo en cuestión, si existe algún atributo que no se repita, se selecciona esa fila y se elimina, hasta haber recorrido todas las instancias del conjunto de datos, una vez llegado este punto se repite la misma lógica para el siguiente atributo en la lista hasta haber validado todos los atributos. El código utilizado para aplicar las técnicas explicadas anteriormente queda disponible dentro del **Anexo IV** de este trabajo.

V.5. Resultado y entregables

Una vez finalizada la etapa de desarrollo de la aplicación se obtiene como resultado la herramienta CPDA, a partir de la cual se extrae el código fuente, el fichero ejecutable y los conjuntos de datos obtenidos mediante R, los cuales se presentan como resultados y entregables generados del desarrollo.

El código fuente (ver Figura 8) de la misma queda publicado en el repositorio privado generado para el control de versiones del desarrollo realizado en la siguiente dirección: https://bitbucket.org/vlinayo/tfm/src/master/

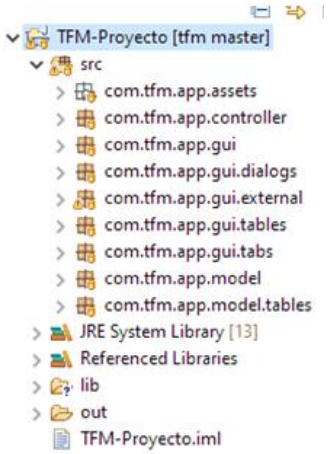


Figura 8 - Estructura del proyecto de la herramienta CPDA

En el repositorio también se puede encontrar el archivo ejecutable *CPDA-App.jar* generado bajo el fichero "*exec*" (ver Figura 9).

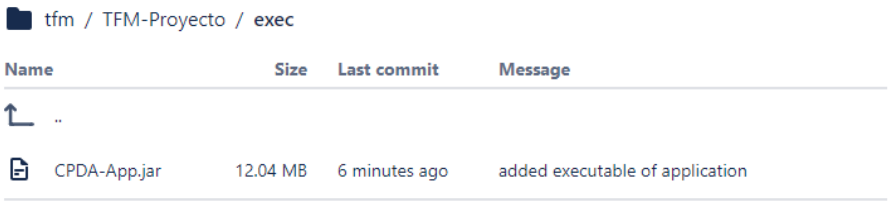


Figura 9 - ejecutable de la herramienta CPDA dentro del repositorio de Bitbucket

En donde se podrá visualizar el resultado final obtenido del desarrollo de la herramienta, y con la cual se hará la comparación de diversas técnicas aplicadas a los conjuntos de datos seleccionados en el **Capítulo IV**. De igual forma queda reflejada la interfaz y el uso de la herramienta en el **Anexo IV**.

CAPÍTULO VI. Resultados y análisis de resultados

Tras el trabajo previamente realizado, se da inicio a la etapa final del presente trabajo, donde a partir de diferentes técnicas de clasificación se analizan los resultados obtenidos de aplicar diversas técnicas de anonimización sobre los conjuntos de datos seleccionados en el **Capítulo IV**.

Con la intención de tener una aproximación inicial de los valores óptimos, principalmente de precisión, que se pueden obtener tras aplicar distintas técnicas de anonimización sobre los conjuntos de datos seleccionados, se ha decidido verificar que la aplicación desarrollada genera los mismos resultados que se generan en Weka, dado que en la aplicación de CPDA no se re-implementa ninguna técnica de clasificación, sino que se accede a la librería de Weka y se utilizan las técnicas disponibles en ella. Por lo tanto se espera que los resultados obtenidos mediante la aplicación de CPDA sean iguales a los obtenidos mediante Weka.

Para verificar lo anterior se realiza un análisis inicial con el conjunto de datos de referencia (Pima diabetes) comparando los resultados obtenidos por ambas aplicaciones (ver Tabla 16).

Pima Diabetes (diabetes.arff)	Weka ("Accuracy" - Exactitud)	CPDA ("Accuracy" - Exactitud)
ZeroR	66.837	66.837
OneR	76.786	76.786
Naive Bayes	76.276	76.276
LibSVM	66.837	66.837
SMO	77.806	77.806
LibLINEAR	62.5	62.5
J48	79.592	79.592

Tabla 16 - Comparativa de clasificadores Weka vs CPDA para diabetes.arff

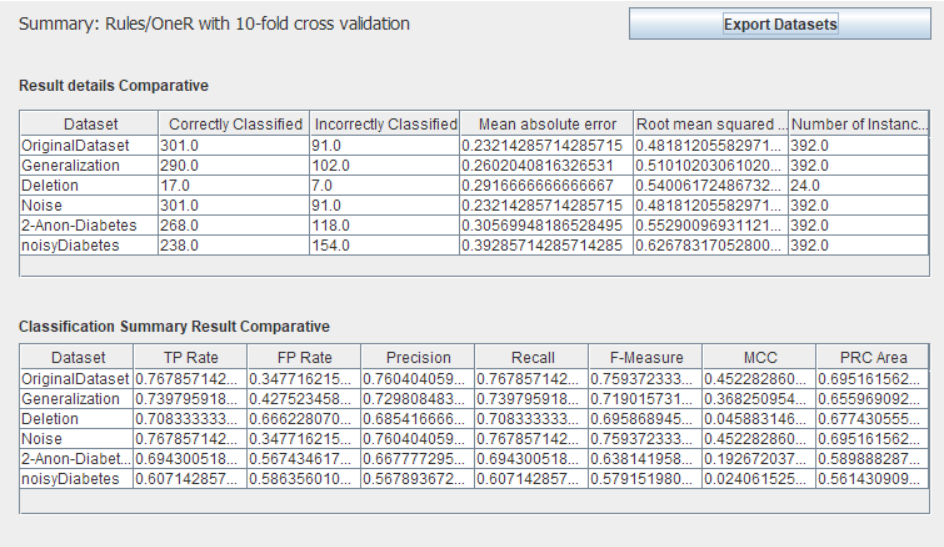
Al observar la Tabla 16 se puede concluir que la herramienta CPDA implementa correctamente la librería de Weka. Lo interesante de estos datos es observar los peores y mejores valores de exactitud alcanzados para cada caso, dado que estos valores serán la base para comparar las técnicas de anonimización a usar a continuación. En consecuencia se presenta a modo resumen la Tabla 17 donde se visualizan los mejores y peores valores de exactitud obtenidos para cada conjunto de datos.

Conjunto de datos	Mejor valor de exactitud obtenido	Peor valor de exactitud obtenido
Diabetes	79.592 - Clasificador: J48	62.5 - Clasificador: LibLINEAR
Heart Disease	83.502 - Clasificador: Naive Bayes	53.872 - Clasificador: ZeroR
Breast Cancer	97.455- Clasificador: SMO	60.182 - Clasificador: LibLINEAR

Tabla 17 - Comparativa de mejores y peores resultados de exactitud obtenidos para cada conjunto de datos sin anonimizar

Partiendo de las técnicas explicadas anteriormente y haciendo uso de los conjuntos de datos originales seleccionados, se realiza una serie de experimentos mediante la aplicación de CPDA, en donde se generan nuevos conjuntos de datos a partir de los conjuntos de datos originales. Los nuevos conjuntos de datos serán el resultado de aplicar diversas técnicas de anonimización sobre el conjunto de datos original.

De esta forma se comparan los resultados obtenidos en el conjunto de datos original y los resultados que se obtienen con cada conjunto de datos generado al aplicar diversas técnicas de clasificación sobre ellos. Con la aplicación de CPDA, esta comparación resulta sencilla de hacer dado que simplemente se indica dentro de la aplicación los conjuntos de datos a generar y comparar, una vez dentro de la sección de resultados de la misma se pueden observar 2 tablas donde se comparan los resultados obtenidos, como en el ejemplo que se observa en la Figura 10.


Summary: Rules/OneR with 10-fold cross validation

Export Datasets

Result details Comparative

Dataset	Correctly Classified	Incorrectly Classified	Mean absolute error	Root mean squared ...	Number of Instanc...
OriginalDataset	301.0	91.0	0.23214285714285715	0.48181205582971...	392.0
Generalization	290.0	102.0	0.2602040816326531	0.51010203061020...	392.0
Deletion	17.0	7.0	0.2916666666666667	0.54006172486732...	24.0
Noise	301.0	91.0	0.23214285714285715	0.48181205582971...	392.0
2-Anon-Diabetes	268.0	118.0	0.30569948186528495	0.55290096931121...	392.0
noisyDiabetes	238.0	154.0	0.39285714285714285	0.62678317052800...	392.0

Classification Summary Result Comparative

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	PRC Area
OriginalDataset	0.767857142...	0.347716215...	0.760404059...	0.767857142...	0.759372333...	0.452282860...	0.695161562...
Generalization	0.739795918...	0.427523458...	0.729808483...	0.739795918...	0.719015731...	0.368250954...	0.655969092...
Deletion	0.708333333...	0.666228070...	0.685416666...	0.708333333...	0.695868945...	0.045883146...	0.677430555...
Noise	0.767857142...	0.347716215...	0.760404059...	0.767857142...	0.759372333...	0.452282860...	0.695161562...
2-Anon-Diabet...	0.694300518...	0.567434617...	0.667777295...	0.694300518...	0.638141958...	0.192672037...	0.589888287...
noisyDiabetes	0.607142857...	0.586356010...	0.567893672...	0.607142857...	0.579151980...	0.024061525...	0.561430909...

Figura 10 - Captura de los resultados obtenidos mediante CPDA

En la herramienta se logra generar una variedad de métricas que pueden ser de interés para comparar los resultados obtenidos por cada una de las técnicas de clasificación aplicadas. Esto permitirá a los usuarios tener mayor información respecto al impacto que una técnica de anonimización puede generar sobre un conjunto de datos. Algunas de las métricas consideradas al estudiar los resultados obtenidos y generar las gráficas

comparativas en la herramienta fueron las métricas de exactitud (*accuracy*), precisión (*precision*), exhausitividad (*recall*) y medida F (*F-measure*). Estos valores sirven para comparar el nivel de confianza que se puede tener sobre el clasificador usado y la técnica dada, debido a que con ellos se puede establecer una comparación directa con los resultados obtenidos de los conjuntos de datos originales y así poder analizar el impacto que han tenido dichas técnicas sobre cada conjunto de datos original.

Teniendo en cuenta el hecho de que la herramienta CPDA desarrollada se centra en técnicas de clasificación, dentro de las diversas técnicas de minería existentes, se decide comparar principalmente el valor de la ***exactitud***, debido a que con esta técnica se calcula la exactitud del clasificador al predecir instancias, y con ella es posible comparar si se pierde o no el nivel de exactitud obtenido de los conjuntos de datos originales con respecto a lo obtenido al aplicar las diversas técnicas de anonimización.

De esta forma se generan 3 tablas (Tabla 18, 19 y 20), una para cada conjunto de datos, donde se comparan las diferentes técnicas aplicadas y el valor de exactitud obtenido para cada una de ellas, así como 2 gráficos de barras para cada conjunto de datos, donde se comparan visualmente los valores obtenidos en correspondencia con el clasificador usado en cada caso.

El análisis detallado realizado en este capítulo, se centra en los resultados obtenidos del conjunto de datos “Pima Diabetes”. De igual forma, en el **Anexo IV**, se pueden observar los análisis correspondientes a los resultados obtenidos de los conjuntos de datos restantes, así como también, se pueden visualizar en detalle los resultados obtenidos para cada conjunto de datos y clasificador dado, donde se aprecian los 4 valores principales mencionados anteriormente, además de otros valores que pueden resultar de interés para el usuario final.

Resultados obtenidos para el conjunto de datos Pima Diabetes

Diabetes	ZeroR	OneR	Naive Bayes	SMO	LibSVM	LibLINEAR	J48
Inicial	66.837	76.786	76.276	77.806	66.837	62.5	79.592
Generalización	66.837	73.979	77.551	80.102	77.551	79.847	75.765
Ruido	66.837	76.786	77.041	77.806	66.837	61.480	79.592
Eliminación	79.167	70.833	75	83.333	79.167	69.912	62.5
Ruido R script	66.837	60.714	73.214	67.092	66.837	59.184	65.051
K -Anonimización	67.617	69.430	75.389	73.834	67.617	54.145	72.798

Tabla 18 - Comparativa del nivel de exactitud (accuracy) para el conjunto de datos de “Pima Diabetes” y las diversas técnicas de anonimización aplicadas

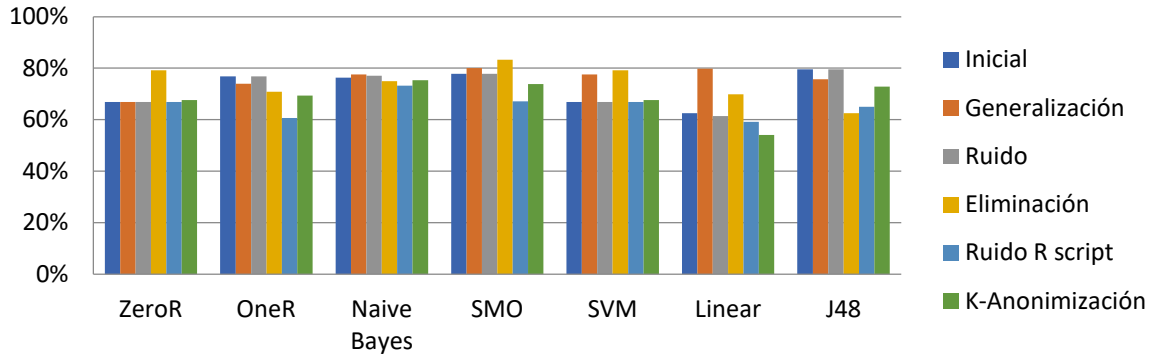


Figura 11 – Valores de exactitud obtenidos para cada clasificador y técnicas aplicadas al conjunto de datos diabetes

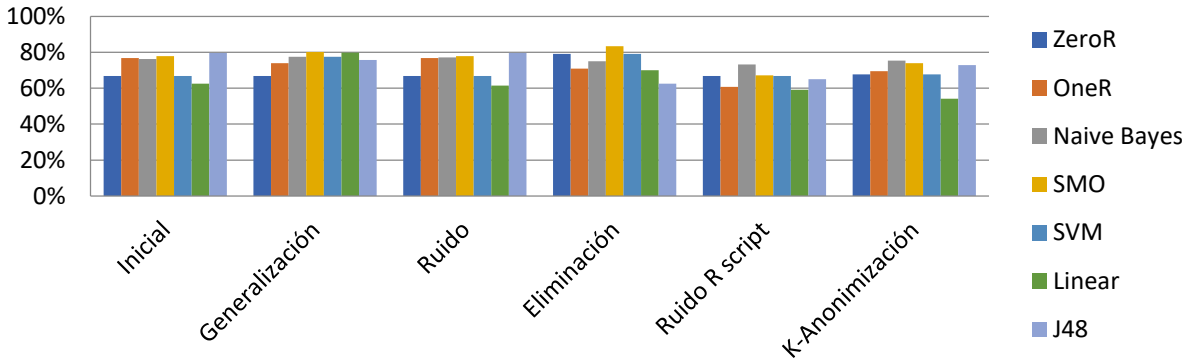


Figura 12 - Valores de exactitud obtenidos para cada técnica aplicada al conjunto de datos diabetes de los clasificadores disponibles

Tras observar los resultados obtenidos para el conjunto de datos de diabetes surgen 2 preguntas principales, que se comentan a continuación.

1. ***Para cada una de las técnicas de anonimización aplicadas, ¿se obtiene un resultado igual, mejor o peor al obtenido con el conjunto de datos original?***

Si se observan los resultados obtenidos anteriormente para el conjunto de datos de *Pima Diabetes* se puede concluir lo siguiente:

- Para el clasificador ZeroR, el nivel de exactitud alcanzado para casi todas las técnicas es igual o mejor al obtenido con el conjunto de datos original, con lo que en un principio las técnicas de anonimización aplicadas no parecen afectar negativamente.
- Para el clasificador OneR, al contrario que ZeroR, obtiene resultados iguales o inferiores al obtenido con el conjunto de datos inicial, por lo que parece que para este caso las técnicas de anonimización sí afectan negativamente el resultado de exactitud obtenido.
- Para el clasificador de Naive Bayes, se obtiene un resultado mixto, donde algunas de las técnicas de anonimización obtienen un resultado de exactitud levemente inferior al obtenido en el conjunto de datos original, mientras que otras obtienen un valor levemente superior, por lo que en general se puede observar que las técnicas obtienen valores que oscilan entre el 70% y 77% de

exactitud, indicando que el nivel de compromiso entre la privacidad de los datos y la exactitud del clasificador es pequeño en comparación a otros casos.

- Para el clasificador de SMO, se puede observar nuevamente cierta variación en los resultados: parte de las técnicas parecen no tener impacto negativo sobre la exactitud obtenida, mientras que otras técnicas muestran un claro impacto negativo sobre el conjunto de datos.
- Para el clasificador SVM ocurre algo similar al caso de ZeroR: las técnicas obtienen resultados iguales o superiores a los obtenidos con el conjunto de datos original.
- Para el clasificador LibLinear, se obtiene en su mayoría un impacto negativo debido a que la mayoría de los conjuntos de datos anonimizados muestran un resultado de exactitud inferior al obtenido inicialmente.
- Para el clasificador J48, se obtienen resultados igual o inferiores para los conjuntos de datos anonimizados, por lo que el impacto de anonimización para este clasificador es claramente negativo.

2. ***¿Se podría establecer alguna relación entre los mejores y peores valores obtenidos para cada conjunto de datos?***

A simple vista no sería posible establecer una relación entre los mejores y peores valores de exactitud obtenidos para cada conjunto de datos porque, si se observa en detalle la Figura 12, es posible apreciar que para los peores valores de exactitud obtenidos ocurre que 3 de 5 técnicas de anonimización obtienen el peor resultado de exactitud del clasificador LibLinear al igual que en el conjunto de datos original.
Por otro lado, no es posible aplicar una relación similar para los mejores valores obtenidos, dado que mientras que el mejor valor del conjunto de datos original proviene del clasificador J48, para el resto de técnicas de anonimización no hay una relación clara, dado que se obtienen los mejores valores a partir de 4 clasificadores diferentes, como son J48, SMO, Naive Bayes y LibLinear.

Resultados obtenidos para el conjunto de datos Heart Disease UCI

Heart Disease	ZeroR	OneR	Naive Bayes	SMO	LibSVM	LibLINEAR	J48
Inicial	53.872	74.074	83.502	82.828	54.882	73.737	77.778
Generalización	53.872	74.074	83.838	83.838	82.828	84.512	81.145
Ruido	53.872	74.074	83.502	82.828	54.882	69.360	77.778
Eliminación	59.880	66.467	78.443	80.838	59.880	61.677	73.054
Ruido R script	53.872	55.892	72.054	71.380	53.872	61.953	61.279
K -Anonimización	54.671	74.048	86.159	86.851	65.052	74.394	85.121

Tabla 19 - Comparativa del nivel de exactitud (accuracy) para el conjunto de datos de "Heart Disease UCI" y las diversas técnicas de anonimización aplicadas

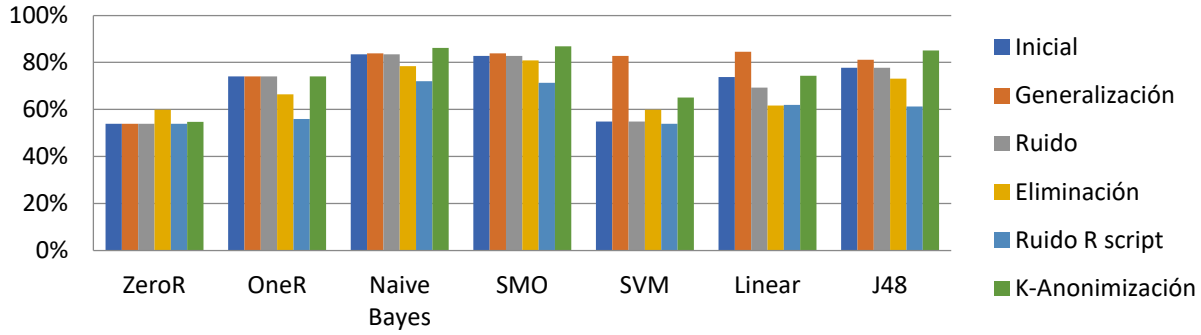


Figura 13 - Valores de exactitud obtenidos para cada clasificador y técnicas aplicadas al conjunto de datos heart disease

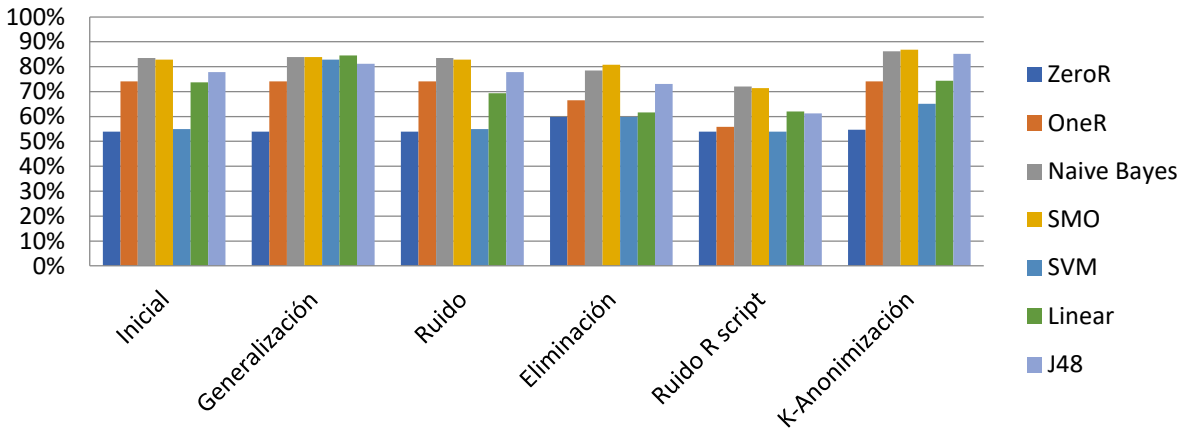


Figura 14 - Valores de exactitud obtenidos para cada técnica aplicada al conjunto de datos heart disease de los clasificadores disponibles

Resultados obtenidos para el conjunto de datos Breast Cancer Wisconsin

Breast Cancer	ZeroR	OneR	Naive Bayes	SMO	LibSVM	LibLINEAR	J48
Inicial	61.455	90	92.727	97.455	61.455	60.182	92.182
Generalización	61.455	89.273	92.727	96.182	97.091	96	94
Ruido	61.455	90	92.727	97.455	61.455	57.091	92.182
Eliminación	-	-	-	-	-	-	-
Ruido R script	61.455	65.091	88	89.273	61.455	48.545	76.182
K -Anonimización	61.566	79.053	91.621	93.078	61.566	51.730	90.710

Tabla 20 - Comparativa del nivel de exactitud (accuracy) para el conjunto de datos de "Breast Cancer Wisconsin" y las diversas técnicas de anonimización aplicadas

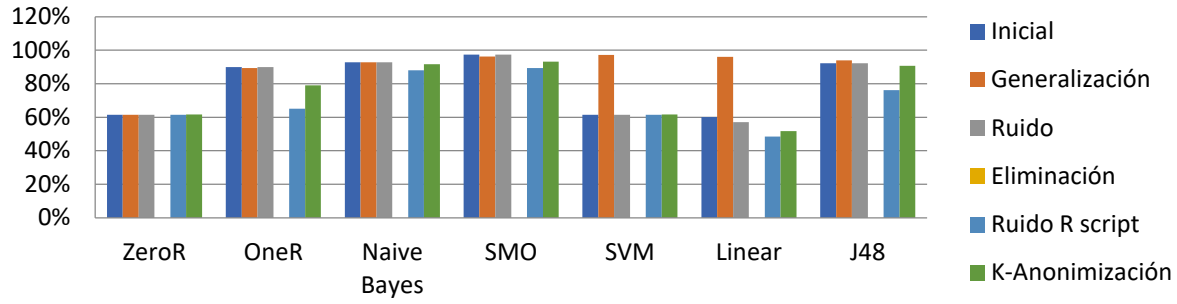


Figura 15 - Valores de exactitud obtenidos para cada clasificador y técnicas aplicadas al conjunto de datos breast cancer

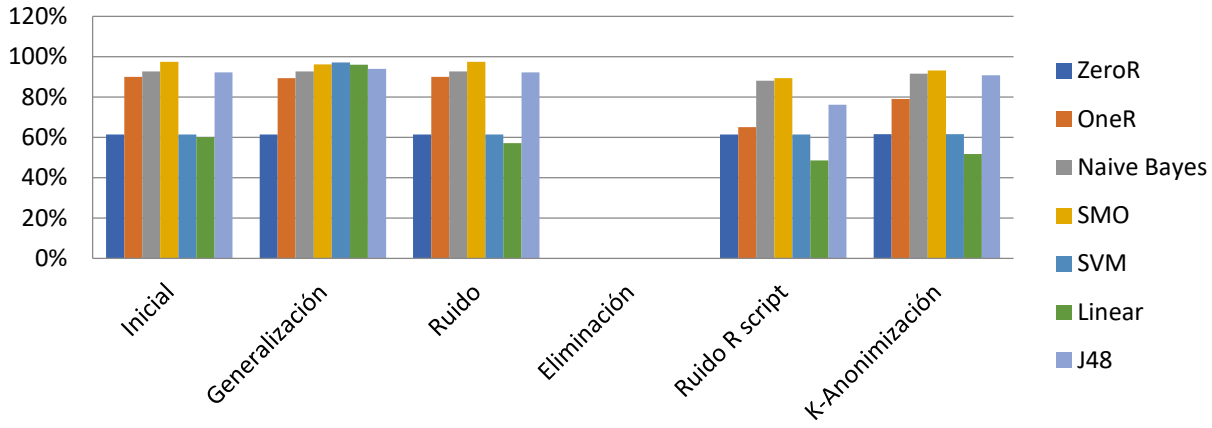


Figura 16 - Valores de exactitud obtenidos para cada técnica aplicada al conjunto de datos breast cancer de los clasificadores disponibles

Las diferencias obtenidas entre las técnicas y el clasificador usado se deben principalmente a que cada técnica de anonimización aplica una modificación de mayor o menor impacto sobre el conjunto de datos inicial, por lo que el conjunto de datos original es modificado. Para explicar uno de los motivos por los que se obtienen los resultados anteriores se puede hacer referencia a lo que ocurre con el caso de generalización, donde agrupar los valores de las instancias en grupos más genéricos puede facilitar la predicción e incluso ocasionar que se obtengan mejores resultados que con el conjunto de datos original. De forma que es posible que para ciertos tipos de técnicas de clasificación el uso de técnicas de anonimización tenga un impacto positivo al modificar el conjunto de datos original en un conjunto de datos más sencillo de clasificar, como se puede observar en los casos de los clasificadores lineales vistos anteriormente.

Por otro lado, también se obtienen resultados que muestran un impacto negativo tras la anonimización de los conjuntos de datos. Como ejemplo se puede hacer referencia a técnicas como la de eliminación, principalmente haciendo hincapié sobre la técnica de eliminación que se ha implementado dentro de la herramienta de CPDA, debido a que se eliminan todas aquellas instancias únicas dentro del conjunto de datos y en consecuencia pueden ocurrir casos como el visto con el conjunto de datos de *Breast*

Cancer Wisconsin, donde el algoritmo de eliminación devuelve un conjunto de datos vacío debido a que no encuentra instancias que no sean únicas dentro del conjunto dado. Lo anterior termina ocasionando una pérdida de información que, al establecer clasificaciones con fines predictivos o al aplicar estudios estadísticos de interés, se limitan parcial o totalmente los posibles estudios a realizar sobre el conjunto de datos, causando que se dejen de lado ciertos datos que podrían influir significativamente sobre los posibles estudios a realizar.

De forma que es posible observar que los resultados que se obtienen son consecuencia de la evaluación usada junto con la técnica de anonimización aplicada, es decir, tanto el tipo de clasificador seleccionado como la técnica de anonimización aplicada juegan un papel importante al momento de generar los resultados observados anteriormente. Por lo tanto según se combinen estas técnicas y según el tipo de conjunto de datos ante el cual nos podamos encontrar, se podrá observar un mayor o menor impacto sobre el nivel de exactitud alcanzado.

A partir de estas observaciones se puede concluir que el nivel de impacto que tendrá una técnica de anonimización para un clasificador dado dependerá del tipo de anonimización aplicada, la técnica de clasificación seleccionada y la cantidad de instancias de datos que resulten afectadas por la técnica de anonimización. Cuanto menor sea la modificación aplicada a los datos o instancias mayor será la probabilidad de obtener un resultado más fiel con respecto a los resultados obtenidos con el conjunto de datos original.

CAPÍTULO VII. Conclusiones y trabajo futuro

En este trabajo se ha realizado un estudio en el que se busca entender el estado del arte respecto a la anonimización de datos y en donde se intenta comprender el posible impacto que tienen dichas técnicas de anonimización sobre la minería de datos, específicamente sobre la clasificación de datos y sobre aquellos conjuntos de datos relacionados con la salud. No sólo se observa el impacto generado por las técnicas de anonimización, sino que se busca permitir la comparación de las mismas mediante una herramienta de apoyo desarrollada como parte del trabajo, con lo cual es posible llegar a ciertas conclusiones referentes al estudio realizado.

Aunque para cada conjunto de datos se observan situaciones diferentes con respecto a las técnicas aplicadas y los resultados obtenidos tras aplicar los métodos de clasificación, es posible decir que, aunque con algunas excepciones, la mayoría de los resultados obtenidos tras aplicar técnicas de clasificación sobre los conjuntos de datos anonimizados sufren un impacto negativo sobre el valor de exactitud obtenido para cada caso. Dichos resultados son coherentes debido a que no existe una relación directa entre los conjuntos de datos estudiados, por lo que cada conjunto de datos es independiente del otro, lo que explica que para cada técnica podamos observar resultados diferentes.

Lamentablemente, en la mayoría de los casos prácticos, al trabajar con datos de salud, nos encontraremos ante la imposibilidad de trabajar con conjuntos de datos sin anonimizar. Por tanto, ante la necesidad de aprovechar los datos que se tienen será necesario valorar hasta qué punto aplicar una o más técnicas de anonimización, sin perder la capacidad de predicción a la vez que se protegen los individuos representados en las instancias de nuestro conjunto de datos.

Ante el estudio realizado en este trabajo de investigación se puede recomendar que siempre que se tenga que aplicar alguna técnica de anonimización sobre un conjunto de datos se tenga un vasto conocimiento sobre los datos en cuestión, de forma que se puedan modificar los mismos mediante un criterio lógico asociado al conjunto de datos, y no mediante un sistema automatizado, indicando así a las diversas técnicas de anonimización mediante parámetros, los diferentes atributos sensibles, los cuasi identificadores, entre otros. Esto permitirá realizar un descarte con sentido sobre los datos.

VII.1.Conclusiones personales y dificultades encontradas

Finalmente se puede concluir a modo personal que este trabajo ha representado un gran reto, habiendo estudiado un tema que resulta interesante no sólo a nivel académico sino también fuera del mismo. Especialmente ante un mundo que está cada vez más digitalizado y donde cada vez existen más datos y más información al alcance de todos, en el que cada día resulta más importante preservar la información de los individuos y mantener el anonimato de los mismos, pero en el que a la vez se busca generar conocimiento partiendo de la disponibilidad de dicha información.

En este trabajo se han encontrado una variedad de retos, desde entender y aprender a trabajar con conjuntos de datos y diversas técnicas de clasificación para la minería de datos, como la realización de la herramienta CPDA, donde se trabajó por primera vez

con Java Swing y librerías como la de Weka y JFreeChart. También se encontraron algunas dificultades, entre las cuales se pueden mencionar:

- La obtención de conjuntos de datos con un mínimo nivel de anonimización. Hoy en día es complicado encontrar datos públicos que no tengan cierto nivel de anonimización y aun lo es más cuando se trata de datos asociados al ámbito de salud, debido a la sensibilidad de estos datos y a la necesidad que existe de proteger a los individuos asociados a ellos.
- Integración de la herramienta CPDA con herramientas externas. Se intentó integrar la herramienta CPDA desarrollada para hacer uso de las técnicas de anonimización disponibles en R. Parte de la integración fue lograda con éxito, donde fue posible ejecutar desde la herramienta comandos de R, pero se encontró gran dificultad para recuperar en Java los resultados obtenidos mediante R, debido a la forma de manejar los conjuntos de datos en R y la forma en la que se manejan dentro de CPDA y Weka.

Llegados a este punto, es importante resaltar también los esfuerzos que se han tenido que realizar durante el desarrollo de este trabajo, tomando en consideración que se contó con una disponibilidad de unas 2 horas de dedicación al día aproximadamente, en donde se invirtió un total de ***680 horas*** distribuidas de la siguiente forma:

- Cursos realizados para facilitar el desarrollo e investigación cuyo contenido se encuentra en el **Anexo I**: se invirtió un total de ***60 horas*** aproximadamente distribuidas entre teoría y práctica.
- Planificación del trabajo, establecimiento de objetivos y alcance, investigación inicial para dar base a los conceptos a tratar en el estado del arte: se invirtió un total de ***140 horas aproximadamente***.
- Búsqueda, selección y estudio de conjuntos de datos relativos a la salud: se invirtió un total de ***60 horas aproximadamente***.
- Desarrollo de la herramienta CPDA: se invirtió un total de ***304 horas aproximadamente***.
- Generación de los resultados finales y realización del informe: se invirtió un total de ***116 horas aproximadamente***.

VII.2. Trabajo Futuro

A partir de la realización de este trabajo nacen múltiples posibilidades de trabajos futuros. Se espera que el trabajo desarrollado en este TFM sirva como base para futuros trabajos e investigaciones llevadas a cabo por grupos de investigación de la universidad de Zaragoza, en particular al grupo COSMOS (Computer Science for Complex System Modelling). Además, podría servir de aplicación en contextos docentes, como base de estudio en asignaturas como Manipulación y Análisis de Grandes Volúmenes de Datos del Máster Universitario en Ingeniería Informática. También podría dar base a futuros trabajos de fin de grado o máster en el cual se pueda continuar con la investigación realizada. Algunos de los posibles pasos a seguir son los siguientes:

1. **Continuación del desarrollo de la aplicación CPDA.**

Partiendo de la primera versión de la aplicación CPDA es posible añadir nuevas funcionalidades a la misma, como por ejemplo:

- Añadir nuevas técnicas de anonimización internas a CPDA.
- Modificar la técnica de eliminación actual de CPDA, donde en vez de aplicar una eliminación completa de la instancia se elimine aquel atributo que pueda contribuir a la revelación de la identidad de un individuo.
- Añadir la lógica para permitir al usuario insertar los valores de los parámetros que poseen las técnicas de anonimización y de clasificación usadas en CPDA.
- Añadir la posibilidad de aplicar filtros sobre los conjuntos de datos en CPDA.
- Añadir la posibilidad de utilizar otras técnicas de minería de datos distintas a las de clasificación, por ejemplo, predicción de valores numericos.
- Integrar con herramientas externas, como por ejemplo la usada para este trabajo (R) y aplicar directamente las técnicas de anonimización desde la aplicación.
- Integrar con herramientas externas diferentes a las usadas en este trabajo, como por ejemplo la herramienta ARX, que ofrece diversas técnicas de anonimización para aplicar en conjuntos de datos, y además cuenta con una API que facilita su integración con entornos de desarrollo.

2. **Añadir una sección de re-identificación a CPDA.**

Sería un estudio interesante, partiendo de lo visto en este trabajo, investigar el riesgo de re-identificación [57] presente en las diferentes técnicas de anonimización aplicadas, investigando sobre los diversos algoritmos que permiten calcular el riesgo de identificación, y de esta forma añadir una sección nueva a la aplicación de CPDA, donde se pueda comparar el riesgo de cada técnica.

3. **Extender la evaluación experimental.**

Sería relevante extender la evaluación experimental realizada en este trabajo y evaluar el impacto práctico de aplicar técnicas de anonimización en casos reales. También sería interesante utilizar conjuntos de datos textuales y considerar técnicas de anonimización de textos (por ejemplo, aplicadas a historias clínicas de pacientes), lo que requeriría aplicar técnicas de minería de textos.

Bibliografía

[1] Real Academia Española, «Anonimizar,» Real Academia Española, Diciembre 2019. [En línea]. Available: https://dle.rae.es/?id=2jjMiRi. [Último acceso: Septiembre 2020].

[2] Agencia Española de Protección de Datos, «Orientaciones y garantías en los procedimientos de ANONIMIZACIÓN de datos personales,» 2016. [En línea]. Available: https://www.aepd.es/media/guias/guia-orientaciones-procedimientos-anonimizacion.pdf. [Último acceso: Septiembre 2020].

[3] Real Academia Española, «Dato,» Real Academia Española, Diciembre 2019. [En línea]. Available: https://dle.rae.es/?id=Bskzsq5|BsnXzV1. [Último acceso: Septiembre 2020].

[4] Fundación Wikimedia, Inc., «Minería de Datos - Wikipedia, la enciclopedia libre,» Wikipedia, 26 Agosto 2020. [En línea]. Available: https://es.wikipedia.org/wiki/Miner%C3%ADa_de_datos. [Último acceso: Septiembre 2020].

[5] Real Academia Española, «Protección - Diccionario de la lengua española,» Real Academia Española, Diciembre 2019. [En línea]. Available: https://dle.rae.es/?id=URUdTVs. [Último acceso: Septiembre 2020].

[6] Real Academia Española, «Procesamiento - Diccionario de la lengua española,» Real Academia Española, Diciembre 2019. [En línea]. Available: https://dle.rae.es/?id=UFLxCoW. [Último acceso: Septiembre 2020].

[7] K. El Emam, «The de-identification of personaly identifiable information.,» *The Biometric Consortium,* 2014.

[8] K. El Emam, S. Jabbouri, S. Sams, Y. Drouet y M. Power, «Evaluating common de-identifcation heuristics for personal health information.,» *Journal of Medical Internet Research,* 2006.

[9] Parlamento Europeo, «Reglamento general de protección de datos,» *Agencia Estatal Boletin Oficial del Estado,* p. 88, 2016.

[10] Grupo de trabajo sobre protección de datos del artículo 19, «Dictamen 05/2014 sobre técnicas de anonimización,» *AEPD,* 2014.

[11] Real Academia Española, «Generalizar - Diccionario de la lengua española,» Real Academia Española, Diciembre 2019. [En línea]. Available: https://dle.rae.es/?id=J3yUWHw. [Último acceso: Septiembre 2020].

[12] L. Adkinson y P. Dago, «Anonimización, ciberseguridad y ciberprivacidad,» Gradiant, 19 Octubre 2017. [En línea]. Available: https://www.gradiant.org/blog/anonimizacion-ciberseguridad-ciberprivacidad-2/. [Último acceso: Septiembre 2020].

[13] Privitar LTD, «K- Anonymity: An Introduction,» PRIVITAR, 07 Abril 2017. [En línea]. Available: https://www.privitar.com/listing/k-anonymity-an-introduction. [Último acceso: Septiembre 2020].

[14] A. Machanavajjhala, J. Gehrke, D. Kifer y M. Venkitasubramaniam, «ℓ-Diversity: Privacy Beyond k-Anonymity,» Department of Computer Science, Cornell University, 24 Abril 2006. [En línea]. Available: https://www.utdallas.edu/~muratk/courses/privacy08f_files/ldiversity.pdf. [Último acceso: Septiembre 2020].

[15] N. Li, T. Li y S. Venkatasubramanian, «t-Closeness: Privacy Beyond k-Anonymity and l-Diversity,» Department of Computer Science, Purdue University - AT&T Labs – Research, 04 Junio 2007. [En línea]. Available: https://www.cs.purdue.edu/homes/ninghui/papers/t_closeness_icde07.pdf. [Último acceso: Septiembre 2020].

[16] Real Academia Española, «Aleatorización - Diccionario de la lengua española,» Real Academia Española, Diciembre 2019. [En línea]. Available: https://dle.rae.es/?id=1g7dltd. [Último acceso: Septiembre 2020].

[17] K. Lubowicka, «The ultimate guide to data anonymization in analytics,» Piwik Pro, 03 Agosto 2020. [En línea]. Available: https://piwik.pro/blog/the-ultimate-guide-to-data-anonymization-in-analytics/. [Último acceso: Septiembre 2020].

[18] ARX, «ARX - Data Anonymization Tool,» ARX, 05 Septiembre 2020. [En línea]. Available: https://arx.deidentifier.org/. [Último acceso: Septiembre 2020].

[19] AEPD, «La K-Anonimidad como medida de la privacidad,» Agencia Española Protección Datos, Septiembre 2019. [En línea]. Available: aepd.es/sites/default/files/2019-09/nota-tecnica-kanonimidad.pdf. [Último acceso: Septiembre 2020].

[20] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz y T. Euler, «Yale: Rapid prototyping for complex data mining tasks.,» ACM, 17 Mayo 2014. [En línea]. Available: https://www.researchgate.net/profile/Ralf_Klinkenberg/publication/220017671_YALE_Rapid_Prototyping_for_Complex_Data_Mining_Tasks/links/0fcfd51498434090d1000000/YALE-Rapid-Prototyping-for-Complex-Data-Mining-Tasks.pdf. [Último acceso: Septiembre 2020].

[21] E. Frank, M. A. Hall y I. H. Witten, «Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques",» The WEKA Workbench, Diciembre 2016. [En línea]. Available: https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf. [Último acceso: Septiembre 2020].

[22] E. Loper y S. Bird, «NLTK: the Natural Language Toolkit,» ACM, Julio 2002. [En línea]. Available: https://dl.acm.org/doi/10.3115/1118108.1118117. [Último acceso: Septiembre 2020].

[23] Comisión Europea, «¿Qué datos personales se consideran sensibles?,» Comisión Europea, 18 Diciembre 2019. [En línea]. Available: https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/legal-grounds-processing-data/sensitive-data/what-personal-data-considered-sensitive_es. [Último acceso: Septiembre 2020].

[24] Grupo Ático 34, «¿Cuáles son los datos sensibles en la RGPD y LOPD?,» Grupo Ático34 SL · Empresa líder en Protección de Datos, 6 Febrero 2018. [En línea]. Available: https://protecciondatos-lopd.com/empresas/datos-especialmente-protegidos-sensibles/. [Último acceso: Septiembre 2020].

[25] European Data Protection Board, «Members | European Data Protection Board,» EDPB - European Data Protection Board, 10 Enero 2018. [En línea]. Available: https://edpb.europa.eu/about-edpb/board/members_en. [Último acceso: Septiembre 2020].

[26] Agencia Española de Protección de Datos, «Agencia Española de Protección de Datos,» Agencia Española de Protección de Datos, Septiembre 2020. [En línea]. Available: https://www.aepd.es/. [Último acceso: Septiembre 2020].

[27] Agencia Española de Protección de Datos, «Plan Estratégico AEPD 2015-2019,» AEPD, 01 Marzo 2020. [En línea]. Available: https://www.aepd.es/es/la-agencia/plan-estrategico-aepd-2015-2019. [Último acceso: Septiembre 2020].

[28] BOE, «Ley Orgánica de protección de datos y garantía a los derechos digitales,» Agencia Estatal Boletín Oficial de Estado, 06 Diciembre 2018. [En línea]. Available: https://boe.es/boe/dias/2018/12/06/pdfs/BOE-A-2018-16673.pdf. [Último acceso: Septiembre 2020].

[29] I. H. Witten, E. Frank, M. H. Len Trigg, G. Holmes y S. J. Cunningham, «Weka 3 - Data Mining with Open Source Machine Learning Software in Java,» Ian H. Witten, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo Cunningham, Junio 2015. [En línea]. Available: https://www.cs.waikato.ac.nz/ml/weka/. [Último acceso: Septiembre 2020].

[30] RStudio, «RStudio,» RStudio, 27 Mayo 2020. [En línea]. Available: https://www.rstudio.com/products/rstudio/. [Último acceso: Septiembre 2020].

[31] Wikimedia Foundation, Inc., «Balsamiq,» Wikipedia, 17 Agosto 2020. [En línea]. Available: https://en.wikipedia.org/wiki/Balsamiq. [Último acceso: Septiembre 2020].

[32] Shopify, «Hatchful by Shopify,» Shopify, 27 Noviembre 2018. [En línea]. Available: https://hatchful.shopify.com/. [Último acceso: Septiembre 2020].

[33] Fundación Wikimedia, Inc., «Swing (Java),» Wikipedia, 22 Octubre 2019. [En línea]. Available: https://es.wikipedia.org/wiki/Swing_(biblioteca_gr%C3%A1fica). [Último acceso: Septiembre 2020].

[34] M. Templ, A. Kowarik y Bernhard, «Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro,» *Journal of Statistical Software,* pp. 1--36, 2015.

[35] Fundación Wikimedia, Inc, «Asana (software),» Wikipedia, 17 Julio 2020. [En línea]. Available: https://es.wikipedia.org/wiki/Asana_(software). [Último acceso: Septiembre 2020].

[36] Fundación Wikimedia, Inc., «Bitbucket,» Wikipedia, 26 Agosto 2020. [En línea]. Available: https://es.wikipedia.org/wiki/Bitbucket. [Último acceso: Septiembre 2020].

[37] Fundación Wikimedia, Inc., «Cross Industry Standard Process for Data Mining,» Wikipedia, 28 Enero 2020. [En línea]. Available: https://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining. [Último acceso: Septiembre 2020].

[38] P. Mittal y N. S. Gill, «A comparative analysis of classification techniques on medical datasets,» *International Journal of Research in Engineering and Technology* , 2014.

[39] K. Ahmed y H. Rauf, «A PPDM Framework to Analyze Privacy and Utility Trade-Off for MNQIA Anonymization Algorithm,» *International Journal of Advanced Research Trends in Engineering and Technology,* Abril 2017.

[40] M. T. Ogedengbe y C. O. Egbunu, «CSE-DT Features Selection Technique for Diabetes Classification,» *ARQII Publication,* 2020.

[41] A. Janosi, W. Steinbrunn, M. Pfisterer y R. Detrano, «Heart Disease Data Set,» The UCI Machine Learning Repository, 01 Julio 1988. [En línea]. Available: https://archive.ics.uci.edu/ml/datasets/heart+Disease. [Último acceso: Septiembre 2020].

[42] R. S. Kamath, «WEKA Approach for Exploration Mining in Diabetic Patients Database,» 18 Diciembre 2013. [En línea]. Available: https://www.researchgate.net/publication/270397035_WEKA_Approach_for_Exploration_Mining_in_Diabetic_Patients_Database. [Último acceso: Septiembre 2020].

[43] D. Lapp, «Heart Disease Dataset,» Kaggle, 06 Junio 2019. [En línea]. Available: https://www.kaggle.com/johnsmith88/heart-disease-dataset. [Último acceso: Septiembre 2020].

[44] Kaggle, «Breast Cancer Wisconsin (Diagnostic) Data Set,» Kaggle, 25 Septiembre 2016. [En línea]. Available: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data. [Último acceso: Septiembre 2020].

[45] Kaggle, «Medical Appointments No Shows,» Kaggle, 20 Agosto 2017. [En línea]. Available: https://www.kaggle.com/joniarroba/noshowappointments. [Último acceso: Septiembre 2020].

[46] R. Feldman, «Heart Disease UCI,» Kaggle, 25 Junio 2018. [En línea]. Available: https://www.kaggle.com/ronitf/heart-disease-uci. [Último acceso: Septiembre 2020].

[47] Kaggle, «Sample Insurance Claim Prediction Dataset,» Kaggle, 04 Junio 2018. [En línea]. Available: https://www.kaggle.com/easonlai/sample-insurance-claim-prediction-dataset#insurance2.csv. [Último acceso: Septiembre 2020].

[48] D. W. H. Wolberg, W. N. Street y O. L. Mangasarian, «Breast Cancer Wisconsin (Diagnostic) Data Set,» The UCI Machine Learning Repository, 01 Noviembre 1995. [En línea]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic). [Último acceso: Septiembre 2020].

[49] K. Schwaber y J. Sutherland, «La Guía de Scrum,» Julio 2016. [En línea]. Available: https://www.scrumguides.org/docs/scrumguide/v2016/2016-Scrum-Guide-Spanish.pdf#zoom=100. [Último acceso: Septiembre 2020].

[50] Fundación Wikimedia, Inc., «Validación cruzada,» Wikipedia, 05 Marzo 2020. [En línea]. Available: https://es.wikipedia.org/wiki/Validaci%C3%B3n_cruzada. [Último acceso: Septiembre 2020].

[51] Data Science, «ZeroR,» Data Science, 04 Abril 2018. [En línea]. Available:

http://datascience.esy.es/wiki/zeror/. [Último acceso: Septiembre 2020].

[52] D. S. Sayad, «OneR,» Dr. Saed Sayad, 2020. [En línea]. Available: https://www.saedsayad.com/oner.htm. [Último acceso: Septiembre 2020].

[53] Fundación Wikimedia, Inc., «Clasificador bayesiano ingenuo,» Wikipedia, 09 Septiembre 2020. [En línea]. Available: https://es.wikipedia.org/wiki/Clasificador_bayesiano_ingenuo. [Último acceso: Septiembre 2020].

[54] E. Frank, S. Legg y S. Inglis, «Class SMO,» The WEKA Workbench, 20 Diciembre 2019. [En línea]. Available: https://weka.sourceforge.io/doc.dev/weka/classifiers/functions/SMO.html. [Último acceso: Septiembre 2020].

[55] Fundación Wikimedia, Inc., «Máquinas de vectores de soporte,» Wikipedia, 27 Julio 2020. [En línea]. Available: https://es.wikipedia.org/wiki/M%C3%A1quinas_de_vectores_de_soporte#SVM_Multiclase. [Último acceso: Septiembre 2020].

[56] Fundación Wikimedia, Inc., «C4.5,» Wikipedia, 08 Febrero 2020. [En línea]. Available: https://es.wikipedia.org/wiki/C4.5. [Último acceso: Septiembre 2020].

[57] F. K. Dankar, K. E. Emam, A. Neisa y T. Roffey, «Estimating the re-identification risk of clinical,» *BMC Medical Informatics and Decision Making,* 2012.

[58] Eurolopd, «Tipos de datos personales en la LOPD y RGPD,» Eurolopd, 2020. [En línea]. Available: https://eurolopd.com/rgpd-lopd-tipos-de-datos-personales/. [Último acceso: Septiembre 2020].

[59] Reduce, «Cómo se clasifican los datos según la LOPD,» Reduce, 01 Octubre 2019. [En línea]. Available: https://reduce.es/se-clasifican-los-datos-personales-segun-la-lopd/. [Último acceso: Septiembre 2020].

[60] P. Samarati y L. Sweeney, «Protecting Privacy when Disclosing Information: k-Anonymity,» *Electronic Privacy Information Center,* 1998.

[61] V. Ciriani, S. D. C. d. Vimercati, S. Foresti y P. Samarati, «K-Anonymity,» Universit`a degli Studi di Milano, 26013 Crema, Italia, Enero 2007. [En línea]. Available: http://spdp.di.unimi.it/papers/k-Anonymity.pdf. [Último acceso: Septiembre 2020].

[62] Wikimedia Foundation, Inc., «Differential Privacy,» Wikipedia, 02 Septiembre 2020. [En línea]. Available: https://en.wikipedia.org/wiki/Differential_privacy. [Último acceso: Septiembre 2020].

[63] T. Zhu, «Explainer: what is differential privacy and how can it protect your data?,» ASOCIACION THE CONVERSATION ESPAÑA, 18 Marzo 2018. [En línea]. Available: https://theconversation.com/explainer-what-is-differential-privacy-and-how-can-it-protect-your-data-90686. [Último acceso: Septiembre 2020].

[64] Oracle, «Oracle Data Masking and Subsetting Pack,» Oracle, 19 Abril 2018. [En línea]. Available: https://www.oracle.com/database/technologies/security/data-masking-subsetting.html. [Último acceso: Septiembre 2020].

[65] Oracle, «Oracle Database Security Assessment Tool,» Oracle, 19 Marzo 2020. [En línea]. Available: https://www.oracle.com/database/technologies/security/dbsat.html. [Último acceso: Septiembre 2020].

[66] Microsoft, «Enmascaramiento estático de datos,» Microsoft, 02 Mayo 2019. [En línea]. Available: https://docs.microsoft.com/es-es/sql/relational-databases/security/static-data-masking?view=sql-server-2017. [Último acceso: Septiembre 2020].

[67] Microsoft, «Enmascaramiento de datos dinámico - SQL Server | Microsoft Docs,» Microsoft, 02 Mayo 2019. [En línea]. Available: https://docs.microsoft.com/es-es/sql/relational-databases/security/dynamic-data-masking?view=sql-server-2017. [Último acceso: Septiembre 2020].

[68] 2000–2020 CYBERTEC PostgreSQL International GmbH, «Data Masking for POSTGRESQL,» Cybertec, 21 Febrero 2020. [En línea]. Available: https://www.cybertec-postgresql.com/en/products/data-masking-for-postgresql/. [Último acceso: Septiembre 2020].

[69] MongoDB, Inc, «IRI Voracity,» MongoDB, 02 Febrero 2016. [En línea]. Available: https://www.mongodb.com/partners/iri-voracity. [Último acceso: Septiembre 2020].

[70] Innovative Routines International (IRI), Inc., «Static Data Masking,» Innovative Routines International (IRI), Inc., 2020. [En línea]. Available: https://www.iri.com/solutions/data-masking/static-data-masking/overview. [Último acceso: Septiembre 2020].

[71] OpenAIRE, «Amnesia - Data Anonymization made easy,» Amnesia, 2019. [En línea]. Available: https://amnesia.openaire.eu/. [Último acceso: Septiembre 2020].

[72] Eyedea Recognition s. r. o., «Image data anonymization,» Eyedea Recognition, 12 Mayo 2017. [En línea]. Available: http://www.eyedea.cz/image-data-anonymization/. [Último acceso: Septiembre 2020].

[73] Net 2000 Ltd. Authors of the Data Masker Software, «Datamasker,» Net2000Ltd, 2016. [En línea]. Available: http://www.datamasker.com/. [Último acceso: Septiembre 2020].

[74] A.-e.-e. A. Hussien, N. Hamza y H. A. Hefny, «Attacks on Anonymization-Based Privacy-Preserving:,» Scientific Research, Abril 2013. [En línea]. Available: http://file.scirp.org/pdf/JIS_2013042311170360.pdf. [Último acceso: Septiembre 2020].

[75] B. C. M. F. P. C. K. H. C.-K. L. Noman Mohammed, «ACM Digital Library,» University of Ontario Institute of Technology y Hong Kong Red Cross Blood Transfusion Service, Octubre 2010. [En línea]. Available: https://dl.acm.org/citation.cfm?doid=1857947.1857950. [Último acceso: Julio 2019].

[76] H. C. Koh y G. Tan, «Data Mining Applications in Healthcare,» J Healthc Inf Manag, 19 2005. [En línea]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.3184&rep=rep1&type=pdf. [Último acceso: Septiembre 2020].

[77] K. E. Emam, S. Rodgers y B. Malin, «Anonymising and sharing individual patient data,» British Medical Journal, 20 Marzo 2015. [En línea]. Available:

https://www.bmj.com/content/bmj/350/bmj.h1139.full.pdf. [Último acceso: Septiembre 2020].

[78] The University of Waikato, «Data Mining with Weka,» Department of Computer Science at the University of Waikato, New Zealand, 2013. [En línea]. Available: https://www.cs.waikato.ac.nz/ml/weka/mooc/dataminingwithweka/. [Último acceso: Agosto 2020].

[79] The University of Waikato, «More Data Mining with Weka,» Department of Computer Science at the University of Waikato, New Zealand, 2014. [En línea]. Available: https://www.cs.waikato.ac.nz/ml/weka/mooc/moredataminingwithweka/. [Último acceso: Agosto 2020].

[80] J. Purcell, «Java Swing (GUI) Programming: From Beginner to Expert,» Udemy, Agosto 2015. [En línea]. Available: https://www.udemy.com/course/java-swing-complete/. [Último acceso: Septiembre 2020].

[81] «A survey of state-of-the-art in anonymity metrics,» ACM Digital Library, Octubre 2008. [En línea]. Available: https://dl.acm.org/citation.cfm?doid=1456441.1456453. [Último acceso: Julio 2019].

[82] «Precisión y exactitud,» Wikipedia, 2020. [En línea]. Available: https://es.wikipedia.org/wiki/Precisi%C3%B3n_y_exactitud. [Último acceso: 2020].

[83] «Precisión y exhaustividad,» Wikipedia, 2019. [En línea]. Available: https://es.wikipedia.org/wiki/Precisi%C3%B3n_y_exhaustividad#Exhaustividad. [Último acceso: 2020].

[84] Wikimedia Foundation, Inc., «Dimensión (Almacén de datos),» Wikipedia, 15 Julio 2020. [En línea]. Available: https://es.qwe.wiki/wiki/Dimension_(data_warehouse). [Último acceso: Septiembre 2020].

[85] Fundación Wikimedia, Inc., «Valor-F,» Wikipedia, 2019. [En línea]. Available: https://es.wikipedia.org/wiki/Valor-F. [Último acceso: 2020].

[86] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel y B. Wiswedel, «KNIME - the Konstanz Information Miner: Version 2.0 and Beyond,» ACM, Noviembre 2009. [En línea]. Available: http://doi.acm.org/10.1145/1656274.1656280. [Último acceso: Septiembre 2020].

[87] J. Demsar, T. Curk, A. Erjavec, C. Gorup, T. Hocevar, M. Milutinovic, M. Mozina, M. Polajnar, M. Toplak, A. Stari, M. Stajdohar, L. Umek, L. Zagar, J. Zbontar, M. Zitnik y B. Zupan, «Orange: Data Mining Toolbox in Python,» University of Ljubljana, Agosto 2013. [En línea]. Available: https://jmlr.org/papers/volume14/demsar13a/demsar13a.pdf. [Último acceso: Septiembre 2020].

ANEXO I. Información complementaria sobre el estado del arte

A continuación se presenta información que se ha considerado interesante para complementar el **Capítulo II** de este trabajo, donde se presenta el estado del arte respecto a la clasificación de datos, la anonimización de datos, las herramientas usadas para la minería de datos, entre otros.

I.1. Clasificación y tipos de datos

A continuación se muestra información complementaria sobre la clasificación y tipos de datos mencionados en el **Capítulo II**.

Tipos de datos personales en el RGPD y LOPD.

De acuerdo con el RGPD UE 679/2016 y la LOPDGDD 3/2018, los datos personales se definen, como "cualquier tipo de información numérica, alfabética, gráfica, fotográfica, acústica o de cualquier otro tipo que concierne a personas físicas identificadas o identificables" [58]. Para la ley, un dato personal es todo aquel cuya información refiera a una persona física identificada (se posee una relación directa entre el dato y la persona, como el DNI) ó identificable (no hay una relación directa entre el dato y la persona, pero se podría investigar y asociar el dato con una única persona). Y dentro de todos los posibles datos que existen, se definen diversos tipos con intención de clasificar los mismos.

- **Datos especialmente protegidos**: Son aquellos datos que contienen información de un individuo referente a su ideología, creencias, origen racial o étnico, salud, vida sexual, datos genéticos y biométricos (Art.9 del RGPD) así como datos relativos a condenas, infracciones penales (Art.10 del RGPD) [58].

- **Datos identificativos**: Son aquellos datos donde se posee información que permite identificar directamente a un individuo dentro de la sociedad, como el DNI, dirección postal o electrónica, imagen, voz, número de Seguridad Social, nombres y apellidos, entre otros.

- **Datos relativos a las características personales**: Son aquellos datos que poseen información característica de los individuos, que hacen referencia a información como estado civil, datos de familia, fechas de nacimiento, lugar de nacimiento, edad, sexo, nacionalidad, lengua materna, características físicas o antropométricas.

- **Datos relativos a las circunstancias sociales**: Son aquellos datos que hacen referencia a las circunstancias sociales de un individuo como las características de alojamiento, vivienda, situación familiar, propiedades, posesiones, aficiones y

estilos de vida, pertenencia a clubes y asociaciones, licencias, permisos y autorizaciones.

- **Datos académicos y profesionales**: Son aquellos datos que hacen referencia a la vida académica de una persona, como la formación, titulaciones, historial del estudiante, experiencia profesional, pertenencia a colegios o asociaciones profesionales.

- **Detalles de empleo**: Aquellos datos relacionados a la vida laboral de un individuo como su profesión, puesto de trabajo, datos no económicos de nómina, historial del trabajador, entre otros.

- **Datos que aportan información comercial**: Son aquellos datos relacionados a las actividades comerciales que puede tener una persona, como actividades y negocios, licencias comerciales, subscripciones a publicaciones o medios de comunicación, creaciones artísticas, literarias, científicas o técnicas.

- **Datos económicos, financieros y de seguros**: Aquellos datos que hacen referencia a la economía de un individuo como los ingresos, rentas, inversiones, bienes patrimoniales, créditos, préstamos, planes de pensiones, jubilación, entre otros.

Clasificación de datos según la LOPD [59]

Los datos personales también se pueden clasificar en función de las medidas de seguridad que se deben adoptar para su protección en:

- **Datos de nivel Básico:** Son aquellos datos identificativos, características personales, circunstancias sociales o familiares, así como de empleo o puestos de trabajo anteriores. Un ejemplo de datos básicos serían los contenidos en un currículo.

- **Datos de nivel medio:** Son aquellos que tienen que ver con las infracciones administrativas o penales, información financiera, los datos de los que sean responsables las administraciones tributarias, de las entidades gestoras y servicios de la Seguridad Social o mutuas de accidentes y enfermedades profesionales. También se incluyen en esta categoría aquellos que permitan evaluar determinados aspectos de la personalidad o comportamiento de las personas.

- **Datos de nivel alto:** Son los que se refieren a ideología, afiliación sindical, religión o vida sexual, los que contengan datos derivados de actos de violencia de género así como aquellos que hayan sido recabados para fines policiales.

I.2. Seudonimización

El siguiente material ha sido extraído del Dictamen 05/2014 sobre técnicas de anonimización adoptado el 10 de abril del 2014, por el grupo de trabajo sobre protección de datos del artículo 29 [10] para aportar mayor información al lector sobre la seudonimización

La seudonimización consiste en la sustitución de un atributo (normalmente un atributo único) por otro en un registro. Por consiguiente, sigue existiendo una alta probabilidad de identificar a la persona física de manera indirecta; en otras palabras, el uso exclusivo de la seudonimización no garantiza un conjunto de datos anónimo.

La seudonimización reduce la vinculabilidad de un conjunto de datos con la identidad del interesado; se trata, por tanto, de una medida de seguridad útil, pero no es un método de anonimización. El resultado de la seudonimización puede ser independiente del valor inicial o bien derivarse de los valores originales de un atributo o conjunto de atributos, como por ejemplo en el caso de funciones hash o sistemas de cifrado.

Las técnicas de seudonimización más utilizadas son las siguientes:

- Cifrado con clave secreta: En esta técnica, el poseedor de la clave puede re-identificar al interesado con suma facilidad. Para ello, le basta con descifrar el conjunto de datos, ya que este contiene los datos personales, aunque sea en forma cifrada. Si se aplican los sistemas de cifrado más avanzados, tan solo es posible descifrar los datos si se conoce la clave.

- Función hash: Se trata de una función que devuelve un resultado de tamaño fijo a partir de un valor de entrada de cualquier tamaño (esta entrada puede estar formada por un solo atributo o por un conjunto de atributos). Esta función no es reversible, es decir, no existe el riesgo de revertir el resultado, como en el caso del cifrado. Sin embargo, si se conoce el rango de los valores de entrada de la función hash, se pueden pasar estos valores por la función a fin de obtener el valor real de un registro determinado. Por ejemplo, si se aplica la función hash al número de identificación nacional para seudonimizar un conjunto de datos, dicho atributo se puede obtener simplemente ejecutando la función con todos los posibles valores de entrada y comparando los resultados con los valores del conjunto de datos.

- Función con clave almacenada: Se trata de un tipo de función hash que hace uso de una clave secreta a modo de valor de entrada suplementario. El responsable del tratamiento puede reproducir la ejecución de la función con el atributo y la clave secreta, sin embargo para los atacantes que no conocen la clave, lo tendrían mucho más difícil debido a que el número de combinaciones

que habría que probar sería tan grande que convertiría este procedimiento en impracticable.

- Cifrado determinista o función hash con clave con borrado de clave: Esta técnica equivale a generar un número aleatorio a modo de seudónimo para cada atributo de la base de datos y, posteriormente, borrar la tabla de correspondencia. Esta solución reduce el riesgo de vinculabilidad entre los datos personales contenidos en el conjunto de datos y los datos personales relativos a la misma persona contenidos en otro conjunto de datos en el que se usa un seudónimo diferente.

- Descomposición en tokens: Esta técnica se usa típicamente en el sector financiero (aunque no exclusivamente en él) para reemplazar los números de identificación de tarjetas por valores que son de poca utilidad para los atacantes.

Garantías de la seudonimización

- *Singularización*: Aún es posible singularizar registros de las personas, ya que la persona queda identificada por un atributo único, que es el resultado de la función de seudonimización (es decir, el atributo seudonimizado).

- *Vinculabilidad*: La vinculabilidad entre registros que usan el mismo atributo seudonimizado para referirse a la misma persona sigue resultando sencillo. Aunque se utilizaran atributos seudonimizados diferentes para el mismo interesado, la vinculabilidad todavía sería posible a través de otros atributos. La única forma de que no haya ninguna referencia cruzada obvia entre dos conjuntos de datos que usan atributos seudonimizados diferentes es que no pueda usarse ningún otro atributo del conjunto de datos para identificar al interesado y que se haya eliminado cualquier vínculo entre el atributo original y el atributo seudonimizado (también por borrado de los datos originales).

- *Inferencia*: Se pueden llevar a cabo ataques por inferencia a la identidad real del interesado en el conjunto de datos o en diversas bases de datos que usen el mismo atributo seudonimizado para una persona, o bien en el caso de que los seudónimos sean auto descriptivos y no enmascaren adecuadamente la identidad del interesado.

I.3. Técnicas de anonimización más usadas hoy

A continuación se muestra información complementaria sobre las técnicas de anonimización más usadas hoy en día, las cuales fueron mencionadas en el **Capítulo II**, y las cuales se explican en mayor detalle a continuación.

I.3.1.Generalización

Consiste en el reemplazo de datos por otros menos específicos, pero semánticamente consistentes. La generalización es la acción y efecto de generalizar, que, según la RAE, es "*Abstraer lo que es común y esencial a muchas cosas, para formar un concepto general que las comprenda todas*" [11].

La generalización es conocida como uno de los tipos de anonimización debido a que a través de algoritmos y técnicas de generalización es posible reducir la precisión de los datos sin perder su utilidad, se puede observar un ejemplo sencillo en la Figura 1**¡Error! No se encuentra el origen de la referencia.**, en donde se generalizan dos atributos, el sexo y la edad. En un estudio en el que se desea analizar el año de nacimiento de una persona, se puede modificar el atributo que contiene la fecha de nacimiento (día, mes y año) a simplemente contener el año; de esta forma se reduce la probabilidad de identificar a los individuos del estudio, mientras que aún es posible analizar los datos para el objetivo pensado.

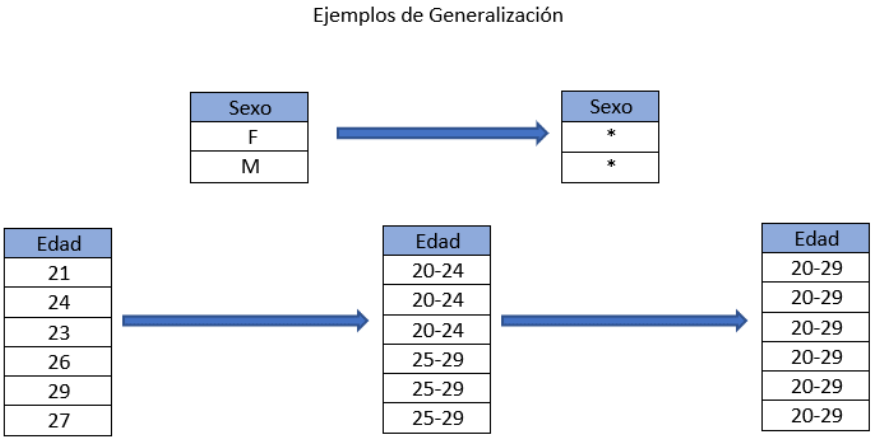


Figura 17 - Ejemplo de Generalización, adaptado de [12]

Existen distintas formas de generalizar datos, de las cuales se pueden mencionar entre las más usadas y comunes hoy en día se tiene la agregación, la K-anonimidad, la diversidad L, y proximidad T.

I.3.1.1. Agregación

Consiste simplemente en agrupar los datos de las personas, buscando impedir que un individuo pueda ser identificado dentro del grupo, como en el ejemplo observado en la Figura 17, donde se pueden observar formas de generalización, en el caso del atributo de Edad, simplemente se genera una agrupación por edades separadas en un rango de 4 o 9 años según el nivel de agrupación que se realice, en el primer caso podría ser más sencillo para un atacante dar con una persona específica, debido a que tiene menos registros de datos en común con su objetivo, por ejemplo, si se busca a una persona de 22 años, ya se pueden descartar las últimas 3 filas de datos, y luego mediante otros atributos se podría intentar de identificar a la persona, claramente el segundo nivel de agrupación abarca muchas más filas por lo que dificulta un poco más la posibilidad de identificar a una persona, más no es del todo imposible si se tiene cierto conocimiento sobre la persona a identificar.

I.3.1.2. K-Anonimidad

La k-anonimidad es un concepto clave que se introdujo para abordar el riesgo de la re-identificación de datos anónimos a través del enlace a otros conjuntos de datos. El modelo de privacidad "*k-anonymity*" fue propuesto por primera vez en 1998 por Latanya Sweeney en su documento "*Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression*" [60]. En él establece que para lograr la k-anonimidad, debe haber al menos k individuos en el conjunto de datos donde compartan un conjunto de atributos susceptibles a ser identificados para cada individuo. La *k-anonymity* o K-anonimidad se puede describir como una garantía de "esconderse entre la multitud": si cada individuo es parte de un grupo más grande, entonces cualquiera de los registros de este grupo podría corresponder a una sola persona [13].

*"**Requisito para la k anonimidad:** Cada publicación de datos debe ser tal que cada combinación de valores de cuasi identificadores pueda coincidir indistintamente con al menos k encuestados"* [61].

"***k-anonymity**. Sea T (A1, ..., Am) una tabla, y Q sea un cuasi-identificador asociado a ella, se dice que T satisface el anonimato k con respecto a QI si cada secuencia de valores en T [QI] aparece al menos con k apariciones en T [QI]*" [61].

Veamos un ejemplo, en la Figura 18, algunos de los atributos que podrían usarse para identificar a un individuo son el nombre, código postal, edad, y género; estos son considerados como cuasi-identificadores, ya que podrían encontrarse en otras fuentes de datos o juntarse con otros atributos y se podría identificar a un individuo. Una enfermedad, por ejemplo, sería un atributo sensible que suele ser objetivo de estudio, y es de aquellos datos cuya privacidad tiene mayor prioridad para los individuos que la padecen.

Name	Código Postal	Edad	Género	Enfermedad
Alejandro	28001	22	Hombre	Cardiovascular
Jorge	28005	23	Hombre	Respiratoria
Rodrigo	28009	18	Hombre	Saludable
Irene	18001	47	Mujer	Cáncer
Andrea	18003	42	Mujer	Saludable
Patricia	18005	56	Mujer	Cardiovascular
Manuel	50017	23	Hombre	Respiratoria
Andres	50015	29	Hombre	Hígado
Sara	50016	18	Mujer	Cáncer

Figura 18 - Datos de pacientes para k-anonimizar, adaptado de [61]

En la Figura 19, se muestran los datos anonimizados para lograr el anonimato k con k = 3, como se puede observar, esto se logró mediante la generalización de algunos atributos cuasi identificadores y con la omisión de otros. En este pequeño ejemplo, los datos se han distorsionado significativamente, pero cuanto mayor sea el conjunto de datos, menor será la distorsión necesaria para alcanzar el nivel deseado de k.

Si se observa cada uno de los atributos se puede ver que algunos se han omitido totalmente sustituyéndolos por un *, mientras que otros se han agrupado como los códigos postales que se agruparon por los primero 3 dígitos, de esa forma se busca garantizar que no sea posible identificar a un individuo dentro de los datos, en el caso

del género ocurre algo interesante, donde sólo las últimas 3 filas fueron omitidas, esto es debido a que dentro de ese grupo sólo existe un individuo “Mujer” lo que haría vulnerable a identificación a dicha persona.

Name	Código Postal	Edad	Género	Enfermedad
*	280*	22	Hombre	Cardiovascular
*	280*	23	Hombre	Respiratoria
*	280*	18	Hombre	Saludable
*	180*	47	Mujer	Cáncer
*	180*	42	Mujer	Saludable
*	180*	56	Mujer	Cardiovascular
*	500*	23	*	Respiratoria
*	500*	29	*	Hígado
*	500*	18	*	Cáncer

Figura 19 - Datos k-anonimizados, k=3. Adaptado de [61]

Si bien el anonimato k puede proporcionar algunas garantías útiles, la técnica viene con las siguientes condiciones de uso:

1. ***Las columnas confidenciales de interés no deben revelar información que se haya omitido en las columnas generalizadas.*** Por ejemplo, ciertas enfermedades son exclusivas de hombres o mujeres que luego pueden revelar un atributo de género eliminado anteriormente.
2. ***Los valores en las columnas sensibles no deben ser todos iguales para un grupo particular de k.*** Si los valores sensibles son todos iguales para un conjunto de registros k que comparten atributos de cuasi identificadores, entonces este conjunto de datos todavía es vulnerable a un ataque de homogeneidad, el cual consiste en la aparición de un único valor para un atributo sensible dentro del conjunto de datos k-anonimizados, lo que permitiría romper la privacidad; en el ejemplo visto en la Figura 19, un caso de posible ataque de homogeneidad se daría en el individuo cuya enfermedad es de “Hígado”, lo que podría permitir que un atacante identificara al individuo en cuestión.
3. ***La dimensión[1] de los datos debe ser suficientemente baja. Si los datos son de alta dimensión, como los datos de series de tiempo, es bastante difícil dar la misma garantía de privacidad que con los datos de baja dimensión***. Para tipos de dato, como los de transacciones o de ubicación puede ser posible identificar a una persona de manera única mediante el encadenamiento de múltiples puntos de datos. Además, como la dimensión de los datos aumenta a menudo, los puntos de datos están muy distribuidos. Esto hace que sea difícil agrupar registros sin distorsionar mucho los datos para lograr el anonimato k. Combinando este enfoque con la minimización de datos y solo haciendo públicas

[1] es una estructura que clasifica los hechos y medidas con el fin de permitir a los usuarios responder a las preguntas de negocio. Dimensiones utilizadas comúnmente son las personas, los productos, el lugar y el tiempo. [84]

las columnas que las personas realmente necesitan, la dimensión se puede reducir a niveles manejables.

De igual forma la técnica sigue siendo una herramienta poderosa cuando se aplica de manera adecuada y con las garantías adecuadas. Forma una parte importante del arsenal de tecnologías para mejorar la privacidad, junto con técnicas alternativas como *Differential Privacy Algorithm* [13].

I.3.1.3. Diversidad ℓ

En la sección anterior, mostramos que el anonimato k es susceptible de ataques de homogeneidad y de conocimiento de fondo, donde el último consiste en ataques donde se posee información o conocimiento sobre uno o varios individuos, y el atacante es capaz de aislar a estos dentro del conjunto de datos debido a ese conocimiento que posee sobre ellos; por lo tanto, se necesita una definición más fuerte de privacidad. La *ℓ-diversity* o ℓ-diversidad proporciona privacidad incluso cuando el proveedor de los datos no sabe qué tipo de conocimiento posee el atacante. La idea principal detrás de la ℓ-diversidad es el requisito de que los valores de los atributos sensibles estén bien representados en cada grupo. De esa forma dados los conocimientos previos de un atacante, una tabla publicada T podría revelar información de dos maneras importantes: divulgación positiva y revelación negativa.

*"**Divulgación positiva:** La publicación de la tabla T⋆ que se derivó de T resulta en una revelación positiva si el atacante puede identificar correctamente el valor de un atributo sensible con alta probabilidad"* [14]

***"Revelación negativa:** La publicación de la tabla T⋆ que se derivó de T da como resultado una revelación negativa si el atacante puede descartar correctamente algunos valores posibles del atributo sensible (con alta probabilidad)"* [14]

De esta forma, es posible dar con la definición ideal de privacidad partiendo del siguiente principio:

***"Principio no informativo**: La tabla publicada debe proporcionar al atacante poca información adicional más allá del "background knowledge". En otras palabras, no debe haber una gran diferencia entre los conocimientos anteriores de los datos en la tabla y los conocimientos posteriores"* [14]

Esto es posible observarlo mediante el siguiente ejemplo (caso), en donde se tienen datos de pacientes mostrados en la Figura 20, que son posteriormente k-anonimizados (k=4) en la Figura 21.

A pesar del conocimiento de fondo previo que pueda tener un individuo, si hay valores sensibles ℓ "bien representados" en un bloque q⋆, entonces dicho individuo necesita ℓ - 1 piezas de conocimiento de fondo suficientemente relevantes como para eliminar ℓ-1 valores sensibles posibles e inferir una revelación positiva [14]. Por lo tanto, al establecer el parámetro ℓ, el que publica los datos puede determinar cuánta protección se proporciona contra el conocimiento de fondo, incluso si este conocimiento de fondo es desconocido para sí mismo. Al juntar estos dos argumentos, llegamos al siguiente principio.

"Principio de diversidad ℓ. *Un bloque $q^\star$ es diverso sí contiene al menos ℓ valores "bien representados" para el atributo sensible S. Una tabla es diversa sí cada bloque $q^\star$ es diverso"* [14]**.**

	Datos No Sensibles			Datos Sensibles
	Código Postal	Edad	Nacionalidad	Condición
1	13053	28	Ruso	Enfermedad del corazón
2	13068	29	Americano	Enfermedad del corazón
3	13068	21	Japonés	Infección Viral
4	13053	23	Americano	Infección Viral
5	14853	50	Indio	Cáncer
6	14853	55	Ruso	Enfermedad del corazón
7	14850	47	Americano	Infección Viral
8	14850	49	Americano	Infección Viral
9	13053	31	Americano	Cáncer
10	13053	37	Indio	Cáncer
11	13068	36	Japonés	Cáncer
12	13068	35	Americano	Cáncer

Figura 20 - Inpatient Microdata", adaptado de [14]

	Datos No Sensibles			Datos Sensibles
	Código Postal	Edad	Nacionalidad	Condición
1	130**	< 30	*	Enfermedad del corazón
2	130**	< 30	*	Enfermedad del corazón
3	130**	< 30	*	Infección Viral
4	130**	< 30	*	Infección Viral
5	1485*	≥ 40	*	Cáncer
6	1485*	≥ 40	*	Enfermedad del corazón
7	1485*	≥ 40	*	Infección Viral
8	1485*	≥ 40	*	Infección Viral
9	130**	3*	*	Cáncer
10	130**	3*	*	Cáncer
11	130**	3*	*	Cáncer
12	130**	3*	*	Cáncer

Figura 21 - Anonimidad 4 aplicada a la Figura 20, adaptado [14]

Considerando el ejemplo de registro de pacientes dado en la Figura 20, se presenta una versión 3-diversa de la tabla en la Figura 22, la cual al compararla con la tabla 4-anónima en la Figura 21, vemos que los ataques contra la tabla 4-anónima se evitan con la tabla 3-diversa. De esta forma no es posible inferir de la tabla 3-diversa mediante conocimiento de fondo ciertas afirmaciones, digamos por ejemplo una persona que conoce a un individuo de nombre Bob, que es un estadounidense de 31 años del código postal 13053, esta persona no podría inferir que Bob tiene cáncer, o si conoce también a un japonés en el código postal 13068 no podría inferir que este tenga una infección viral o cáncer.

	Datos No Sensibles			Datos Sensibles
	Código Postal	Edad	Nacionalidad	Condición
1	1305*	≤ 40	*	Enfermedad del corazón
4	1305*	≤ 40	*	Infección Viral
9	1305*	≤ 40	*	Cáncer
10	1305*	≤ 40	*	Cáncer
5	1485*	≥ 40	*	Cáncer
6	1485*	≥ 40	*	Enfermedad del corazón
7	1485*	≥ 40	*	Infección Viral
8	1485*	≥ 40	*	Infección Viral
2	1306*	≤ 40	*	Enfermedad del corazón
3	1306*	≤ 40	*	Infección Viral
11	1306*	≤ 40	*	Cáncer
12	1306*	≤ 40	*	Cáncer

Figura 22 - Diversidad 3 aplicada a Figura 21, adaptado de [14]

I.3.1.4. Proximidad T

La proximidad T [15] busca generalizar los datos de forma que la distribución de cuasi identificadores en cada clase de equivalencia sea similar a la distribución de datos original. La idea de la proximidad es que la distribución de datos confidenciales en todos los grupos no está demasiado lejos de la distribución en toda la población. La "t" viene de exigir que las distribuciones no sean más que una distancia t aparte en el sentido que definiremos a continuación. Si dentro de un grupo de datos, aquellos datos considerados sensibles/privados no destacan entre el resto de datos sensibles, se frustra el ataque de homogeneidad y el ataque de conocimiento de fondo.

*"**Principio de la proximidad t**: Se dice que una clase de equivalencia tiene cercanía t si la distancia entre la distribución de un atributo sensible en esta clase y la distribución del atributo en toda la tabla no es mayor que un umbral t. Se dice que una tabla tiene t-cercanía si todas las clases de equivalencia tienen t-cercanía".*

Por supuesto, requerir que los atributos no se diferencien demasiado entre ellos limitaría la cantidad de información útil que se libera, ya que limita la información sobre la correlación de los atributos cuasi identificadores y los atributos confidenciales. Sin embargo, esto es precisamente lo que hay que limitar. Si un observador obtiene una imagen demasiado clara de esta correlación, entonces se produce la revelación del atributo. El parámetro t en *t-closeness* permite compensar entre utilidad y privacidad.

La proximidad T surge a partir de los diversos ataques que han sufrido tablas anonimizadas mediante diversidad ℓ; se explican estos ataques partiendo del siguiente ejemplo:

	Código Postal	Edad	Salario	Enfermedad
1	47677	29	3K	Úlcera gástrica
2	47602	22	4K	Gastritis
3	47678	27	5K	Cáncer de estómago
4	47905	43	6K	Gastritis
5	47909	52	11K	Gripe
6	47906	47	8K	Bronquitis
7	47605	30	7K	Bronquitis
8	47673	36	9K	Neumonía
9	47607	32	10K	Cáncer de estómago

Figura 23 - Tabla original de salarios/enfermedad, adaptado [15]

Dada la Figura 23, donde se presentan datos de enfermedades y los salarios de los pacientes, se procede a realizar un primer método de anonimización de tipo diversidad ℓ, con lo cual se obtienen los datos observados en la Figura 24.

	Código Postal	Edad	Salario	Enfermedad
1	476**	2*	3K	Úlcera gástrica
2	476**	2*	4K	Gastritis
3	476**	2*	5K	Cáncer de estómago
4	4790*	≥ 40	6K	Gastritis
5	4790*	≥ 40	11K	Gripe
6	4790*	≥ 40	8K	Bronquitis
7	476**	3*	7K	Bronquitis
8	476**	3*	9K	Neumonía
9	476**	3*	10K	Cáncer de estómago

Figura 24 - Versión 3-diversa de la tabla en la figura 7, adaptado de [15]

Entre los ataques a tablas ℓ diversas se pueden mencionar:

- *Skewness Attack (Ataque de sesgo)*: Cuando la distribución general es sesgada, el valor de ℓ diversidad se cumple y aún así no impide la obtención y divulgación de atributos.
- *Similarity Attack (Ataque por similaridad):* Cuando los valores de atributos sensibles en una clase de equivalencia son distintos, pero semánticamente similares, un atacante puede aprender información importante.

En el ejemplo de la Figura 24 hay dos atributos sensibles: Salario y Enfermedad. Suponiendo que se sabe que el registro de una persona corresponde a uno de los tres primeros registros; entonces se sabe que el salario de esa persona está en el rango [3K-5K] y es posible inferir que su salario es relativamente bajo. Este ataque se aplica no solo a atributos numéricos como "Salario", sino también a atributos categóricos como "Enfermedad". Saber que el registro de una persona pertenece a la primera clase de equivalencia permite concluir que la misma tiene algunos problemas relacionados con el estómago.

Esta fuga de información confidencial se produce porque, si bien el requisito de diversidad asegura la "diversidad" de los valores confidenciales en cada grupo, no tiene en cuenta la proximidad semántica de estos valores. En resumen, las distribuciones que tienen el mismo nivel de diversidad pueden proporcionar niveles de privacidad muy diferentes, porque existen relaciones semánticas entre los valores de los atributos, porque los valores diferentes tienen niveles de sensibilidad muy diferentes y porque la privacidad también se ve afectada por la relación con la distribución general.

De esta forma retomamos el ejemplo de la Figura 23 y la Figura 24, y vemos lo que ocurre cuando se aplica la técnica de *t-closeness* en la Figura 25, definiendo unas distancias base para los atributos en cuestión, razonamiento que se ve claramente reflejado en "*T-Closeness: Privacy Beyond k-Anonymity and ℓ-Diversity*" [15], obteniendo así una tabla que muestra los mismos datos con un *0,167-closeness* asociados al atributo del salario, y un *0,278-closeness* referente al atributo de enfermedad. Con la técnica aplicada se logra prevenir el ataque por similitud evitando que se pueda asociar datos a individuos mediante inferencia.

Con lo anterior se puede afirmar que mediante la proximidad t o *t-closeness* se logra proteger los datos de la revelación de atributos, pero no lidia con la revelación de identidad. De esta forma, puede ser recomendable usar más de una técnica de anonimización para lograr proteger los datos de los posibles ataques.

	Código Postal	Edad	Salario	Enfermedad
1	4767*	≤ 40	3K	Úlcera gástrica
3	4767*	≤ 40	5K	Cáncer de estómago
8	4767*	≤ 40	9K	Neumonía
4	4790*	≥ 40	6K	Gastritis
5	4790*	≥ 40	11K	Gripe
6	4790*	≥ 40	8K	Bronquitis
2	4760*	≤ 40	4K	Gastritis
7	4760*	≤ 40	7K	Bronquitis
9	4760*	≤ 40	10K	Cáncer de estómago

Figura 25 - Tabla con 0,167-closenees en salario y 0,278-closeness para enfermedad, adaptado de [15]

I.3.2. Aleatorización

Consiste en la acción y efecto de aleatorizar, lo cual consiste según la RAE en, “Someter algo o a alguien a un proceso aleatorio” [16]. La aleatorización es una técnica usada en la anonimización de datos que consiste en aplicar una modificación aleatoria a los datos. Se conocen tres métodos principales de aleatorización, la inyección de ruido, las permutaciones y la privacidad diferencial.

I.3.2.1. Adición de ruido

Consiste en modificar los atributos del conjunto de datos para que sean menos exactos, conservando no obstante su distribución general. Al tratar un conjunto de datos, se podría suponer que los valores encontrados en este son exactos, pero esto es cierto hasta un punto. Por ejemplo, si la altura de una persona se mide originalmente hasta el centímetro más próximo, el conjunto de datos anonimizados podría contener valores con una exactitud de ± 10 cm, con lo cual los valores reflejados no son lo valores originales, pero sirven como representación de la realidad [10].

Si se utiliza esta técnica de manera competente, un tercero no debe poder identificar a una persona ni tampoco debe ser capaz de restaurar los datos o de averiguar cómo se han modificado. Normalmente, esta técnica se usa de forma combinada con otras técnicas de anonimización, como la eliminación de datos identificativos y de cuasi identificadores. El nivel de ruido que se aplique dependerá de la cantidad y el tipo de información que se requiera, así como del impacto que tenga la revelación de los datos protegidos en la privacidad de las personas [10].

I.3.2.2. Permutación

La técnica de permutación, es una forma de aleatorización, en donde se busca mezclar los valores de los atributos de una tabla, con el fin de crear un vínculo ficticio entre los datos. Se considera una estrategia útil para aquellos casos en los que sea importante conservar la distribución exacta de los atributos en el conjunto de datos.

La permutación puede considerarse como una forma de adición de ruido, la diferencia radica en que en la forma clásica de adición de ruido, los atributos se sustituyen por valores aleatorios, en cambio en la permutación, no se añade ruido sobre los datos, sino que se modifica la asociación que existe entre ellos y el individuo al que pertenecen.

La tarea de generar un ruido que sea consistente no es fácil, puesto que si la modificación de los valores de los atributos es muy pequeña, puede ocurrir que no se obtenga el grado de privacidad deseado; con la técnica de permutación, se intercambian los valores contenidos en el conjunto de datos, trasladándolos de un registro a otro. Al permutar los datos de esta manera, se garantiza que el rango y la distribución de los valores se mantengan idénticos al conjunto de datos original, pero se perderían las correlaciones entre los valores y las personas. Si dos o más atributos tienen una relación lógica o una correlación estadística y se permutan independientemente del resto, dicha relación quedará destruida. Por consiguiente, sería importante permutar un conjunto de atributos que estén relacionados entre sí a fin de no romper la relación lógica entre los datos [10].

Al igual que ocurre con la adición de ruido, la permutación por sí sola no permite obtener la anonimización, por lo que siempre debe combinarse con el procedimiento de eliminación de atributos obvios o cuasi identificadores [10].

I.3.2.3. Privacidad diferencial

Consiste en introducir ruido en las respuestas, protegiendo la privacidad de los sujetos presentes en los datos; aunque se parece a la inserción de ruido, la privacidad diferencial adopta un enfoque diferente. Mientras que, en la práctica, la inserción de ruido ocurre antes de difundir el conjunto de datos, la privacidad diferencial, por el contrario, puede usarse cuando el responsable del tratamiento de datos genera vistas anonimizadas de un conjunto de datos, al mismo tiempo que conserva una copia de los datos originales. Estas vistas anonimizadas normalmente son generadas mediante un grupo de consultas provenientes de un tercero. Dicho grupo de consultas contiene algo de ruido aleatorio que se añade de manera deliberada con posterioridad. La privacidad diferencial indica al responsable del tratamiento cuánto ruido debe añadir, y en qué forma, para obtener las garantías de privacidad necesarias. En este contexto, es especialmente importante una supervisión continua (como mínimo de cada nueva consulta) para evaluar cualquier posibilidad de identificación de una persona en el conjunto de resultados obtenidos de las consultas. Sin embargo, conviene aclarar que las técnicas de privacidad diferencial no modifican los datos originales. Por lo tanto, mientras se conserven los datos originales, el responsable del tratamiento es capaz de identificar a las personas a partir de los resultados de las consultas de privacidad diferencial mediante el conjunto de los medios que pueden ser razonablemente utilizados. Estos resultados también deben considerarse como datos personales [10].

Una de las ventajas de este enfoque consiste en que los datos son entregados en forma de respuesta a una consulta concreta, y no simplemente como consecuencia de la publicación de un único conjunto de datos. Es importante saber que a una consulta también se le pueden aplicar técnicas de anonimización, incluidas la adición de ruido y la sustitución, con el fin de aumentar la protección de la privacidad [10].

De esa forma, la privacidad diferencial resulta ser una técnica estadística que tiene como objetivo proporcionar medios para maximizar la precisión de las consultas de las bases de datos estadísticas mientras se mide el impacto en la privacidad de las personas cuya información se encuentra en la base de datos [62].

El principal mecanismo para lograrlo es agregar ruido aleatorio a los datos agregados. Como ejemplo, digamos que se quiere mostrar las rutas más populares de un parque. Partiendo de seguir las rutas de 100 personas que caminan regularmente por el parque y si caminan por el sendero o por el pasto, el estudio puede decir que el número de personas que prefieren cruzar el césped está entre 59 y 61 personas, en lugar del número exacto de 60. El número inexacto puede preservar la privacidad de un individuo, pero tendrá un impacto muy pequeño en el patrón que indica que alrededor del 60% de las personas prefieren tomar un atajo [63].

I.4. Técnicas aplicadas por gestores de bases de datos

Ante la evidente vulnerabilidad a los abundantes y diversos ataques que puede enfrentar un gestor de bases de datos se han desarrollado diversas técnicas para proteger y enmascarar datos sensibles que en estos se puedan almacenar.
Entre los gestores que poseen técnicas propias de seguridad sobre los datos podemos mencionar:

a) **Oracle**: Esta entre los gestores de bases de datos más usados y conocidos en el mercado, y ofrece entre sus productos:
   - "*Oracle Data Masking and Subsetting Pack*" [64], el cual es un paquete para el gestor de datos en donde la lógica de enmascaramiento y subconjunto de datos desarrollado por Oracle permite extraer copias completas o subconjuntos de datos de la aplicación de la base de datos, ofuscarlos y compartirlos con socios dentro y fuera de la empresa. Mediante este paquete se preserva la integridad de la base de datos asegurando la continuidad de las aplicaciones.
   - "*Oracle Database Security Assessment Tool*" [65], La herramienta de evaluación de seguridad de la base de datos de Oracle es una herramienta de línea de comandos independiente que acelera el proceso de evaluación y cumplimiento normativo mediante la recopilación de los tipos relevantes de información de configuración de la base de datos y la evaluación del estado de seguridad actual para proporcionar recomendaciones sobre cómo mitigar los riesgos identificados. De esta forma, ayuda a identificar las áreas donde la configuración, operación o implementación de una base de datos introduce riesgos y recomienda cambios y controles para mitigar los mismos.

b) **SQL Server:** También uno de los gestores más populares hoy en día, en el que se han generado soluciones para la protección de datos, entre algunas de ellas se pueden mencionar:
   - "*Static data masking*" [66]: Es una funcionalidad de SQL Server Management Studio la cual permite a los usuarios crear una copia enmascarada de una base de datos. La funcionalidad se desarrolló para las organizaciones que necesitan compartir datos, algunos de ellos confidenciales, entre equipos o con otras organizaciones.
   - "*Dynamic data masking*" [67] : El "*dynamic data masking*" (DDM) o enmascaramiento dinámico de datos logra limitar la exposición de información confidencial al ocultarla de los usuarios sin privilegios. Se puede usar para simplificar considerablemente el diseño y la codificación de la seguridad en una aplicación. Esta técnica consiste en limitar la

exposición de la información confidencial, con lo que se impide que los usuarios vean datos a los que no deberían poder acceder.

Un ejemplo del enmascaramiento estático se puede ver en la Figura 26 donde antes de enmascarar los datos, se tiene una columna que contiene los D.N.I. de un grupo de personas. Después del enmascaramiento, los cinco primeros dígitos de cada número se han reemplazado por números generados aleatoriamente, quedando así unos valores de D.N.I. que no corresponden al valor original, protegiendo así los datos reales.

Documento Nacional de Identidad (D.N.I) – Antes de enmascarar	Documento Nacional de Identidad (D.N.I) – Enmascarado
95070981K	11122281K
45311316M	43214216M
45311316M	56424616M
43468001W	89574201W
15088433L	17634533L

Figura 26 - Enmascaramiento estático de datos, adaptado de [67].

El enmascaramiento dinámico de datos no pretende evitar que los usuarios de la base de datos se conecten directamente a ella y ejecuten consultas exhaustivas que expongan información confidencial. Para ello se recomienda complementar el uso del enmascaramiento de datos con otras características de seguridad de SQL Server (auditoría, cifrado, seguridad de nivel de fila...), permitiendo así proteger mejor la información confidencial en la base de datos.
Se puede observar en la Figura 27 el ejemplo mostrado en la Figura 26, pero en este caso se usa "*dynamic data masking*" el cual posee un interesante parecido a la técnica de agregación vista en secciones anteriores.

Documento Nacional de Identidad (D.N.I) – Antes de enmascarar	Documento Nacional de Identidad (D.N.I) – Enmascarado
95070981K	******81K
45311316M	******16M
45311316M	******16M
43468001W	******01W
15088433C	******33C

Figura 27 - Enmascaramiento dinámico de datos, adaptado de [67]

c) **PostgreSQL**: Otro de los gestores más populares de hoy en día, ofrece una solución para la privacidad de datos conocida como "*Data Masking for PostgreSQL*" [68], la cual consiste en una extensión para PostgreSQL proporcionada por Cybertec, permitiendo limitar la exposición de datos sensibles mediante el enmascaramiento de los mismos.

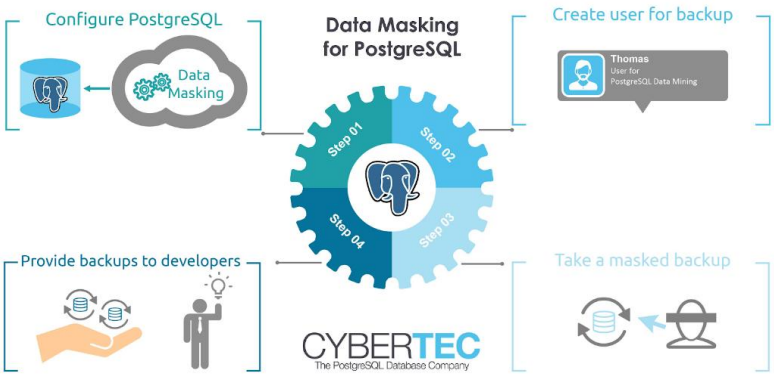


Figura 28 - Enmascaramiento de datos en PostgreSQL, extraído de [68]

d) **MongoDB**: Uno de los gestores más recientes, y uno de los más usados en el mundo de las bases de datos no relacionales, hace uso de la solución conocida como "*IRI Voracity*" [69], la cual es una plataforma de gestión de datos de alto rendimiento para MongoDB y otras fuentes. Esta ofrece el enmascaramiento de datos persistentes, o enmascaramiento de datos estáticos (SDM) [70], como método principal para proteger elementos de datos específicos. Estos "elementos" suelen ser valores de la base de datos que se consideran confidenciales. Estos campos pueden contener información de identificación personal (*Personal identifiable information o PII*), información de salud protegida (*Protected health information o PHI*), números de cuenta principales (*Principal account number o PAN*), secretos comerciales u otros valores privados.

I.5. Herramientas para anonimizar datos

Además de las técnicas mencionadas en el **Capítulo II**, se pueden mencionar otras herramientas como:

- *Amnesia* [71]: Amnesia es una herramienta de anonimización de datos, que permite eliminar la información de identificación de los datos. Amnesia no solo elimina identificadores directos como nombres, números de la seguridad social, etc., sino que también transforma identificadores secundarios como fechas de nacimiento y códigos postales para que no sea posible identificar un individuo en los datos. Amnesia soporta el anonimato k y el anonimato k^m.
- *Anonymizer* [72]: El software detecta caras y matrículas de automóviles en diversas escalas y orientaciones, y aplica filtros borrosos para que las caras no puedan identificarse y las matrículas sean ilegibles. Las funciones principales de la biblioteca anonimizan archivos de imagen, buffers JPEG y buffers de imagen RGB. Anonymizer está diseñado para ser usado para anonimizar fotos de Google Street View, fotos de cámaras web, fotos

enviadas por el usuario o cualquier otra foto donde se necesite proteger la privacidad.

- *Data Masker* [73]: En un paquete fácil de usar y proporciona todas las herramientas necesarias para sustituir, codificar u ofuscar los datos de prueba. Esto lo realiza mediante la definición de reglas simples y comprensibles para operar con los datos. La recopilación de estas reglas realiza una serie de acciones conocidas, probadas y repetibles con solo presionar un botón.

I.6. Trabajos relacionados con la anonimización y protección de datos en diferentes ámbitos

La anonimización de datos es un campo muy estudiado hoy en día, tanto por la necesidad de proteger a las personas físicas, como por la normativa legal existente en materia de protección de datos.

En consecuencia, se han realizado numerosos estudios relacionados con la búsqueda de las mejores técnicas, algoritmos, y metodologías que permitan lograr la mejor protección posible sobre los datos.

Algunos estudios relacionados que podemos destacar:

- "***Attacks on Anonymization-Based Privacy-Preserving: A Survey for Data Mining and Data***" [74]: En este artículo se presenta un estudio de las técnicas de ataques más comunes para la anonimización basada en PPDM (*Private Preserving Data Publishing – Publicación de datos preservando la privacidad*) y PPDP (*Privacy Preserving Data Mining - Privacidad preservando la minería de datos*), explicando sus efectos sobre la privacidad de datos.
- "***Centralized and Distributed Anonymization for High-Dimensional Healthcare Data***" [75]: En este artículo se estudian las preocupaciones en el ámbito de la privacidad cuando se piensa compartir información de pacientes entre un servicio de transfusión de sangre y los hospitales públicos. Y propone un nuevo modelo de privacidad llamado *LKC-privacy* para superar los desafíos y presentar dos algoritmos de anonimización para lograr la privacidad de LKC en escenarios centralizados y distribuidos.
- "***Data Mining Applications in Healthcare***" [76]: Este artículo explora las aplicaciones de minería de datos en salud. En particular, analiza la minería de datos y sus aplicaciones dentro de la atención médica en áreas importantes como la evaluación de la efectividad del tratamiento, la gestión de la atención médica, la gestión de las relaciones con los clientes y la detección de fraude y abuso.
- "***Anonymising and sharing individual patient data***" [77]: Este artículo describe los conceptos y principios clave para el anonimato de los datos de salud a la vez que garantiza que siga siendo adecuado para un análisis significativo.

I.7. Autoridades para la protección de datos en Europa

Entre las autoridades reconocidas por el grupo de protección de datos de Europa [25] (European Data Protection Board) se encuentran:

Autoridades para la protección de datos en Europa			
	Contacto	Página Web	Miembro
European Data Protection Supervisor	edps@edps.europa.eu	http://www.edps.europa.eu/EDPSWEB/	Mr Giovanni BUTTARELLI, European Data Protection Supervisor
Francia *Commission Nationale de l'Informatique et des Libertés – CNIL*	https://www.cnil.fr/en/contact-cnil	http://www.cnil.fr/	Ms Marie-Laure DENIS, President of CNIL
Alemania *Die Bundesbeauftragte für den Datenschutz und die Informationsfreiheit*	poststelle@bfdi.bund.de	http://www.bfdi.bund.de/	Mr Ulrich KELBER Federal Commissioner for Data Protection and Freedom of Information
España *Agencia Española de Protección de Datos (AEPD)*	internacional@agpd.es	https://www.agpd.es/	Ms María del Mar España Martí, Director of the Spanish Data Protection Agency
Bélgica Autorité de la protection des données - Gegevensbescher mingsautoriteit (APD-GBA)	contact@apd-gba.be	https://www.autoriteprotectiondonnees.be/	Mr. David Stevens, President
Italia Garante per la protezione dei dati personali	protocollo@gpdp.it	http://www.garanteprivacy.it/	Mr Antonello SORO, President of Garante per la protezione dei dati personali
Portugal Comissão Nacional de Protecção de Dados – CNPD	geral@cnpd.pt	http://www.cnpd.pt/	Ms Filipa CALVÃO, President, Comissão Nacional de Protecção de Dados

Tabla 21 - Autoridades para la protección de datos en Europa, extraído de [32]

I.8. Estudios previos realizados para facilitar el desarrollo e investigación del trabajo

Para la realización de este trabajo fue necesario realizar un par de cursos de los cuales se obtuvo el conocimiento necesario para trabajar con la herramienta de Weka y para el desarrollo con Java Swing.

Para el trabajo con la herramienta de Weka se realizaron 2 cursos, el primero "*Data Mining with Weka"* [78] *y* el segundo "*More Data Mining with Weka*" [79]. Los cuales están disponibles en la página web oficial de la herramienta https://www.cs.waikato.ac.nz/~ml/weka/courses.html y cuyo contenido se muestra a continuación.

- *Data Mining with Weka*
    1. *Getting started with Weka.*
        - ✓ *Introduction.*
        - ✓ *Exploring the Explorer.*
        - ✓ *Exploring datasets.*
        - ✓ *Building a classifier.*
        - ✓ *Using a filter.*
        - ✓ *Visualizing your data.*
        - ✓ *Questions answered.*
    2. *Evaluation.*
        - ✓ *Be a classifier!.*
        - ✓ *Training and testing.*
        - ✓ *Repeated training and testing.*
        - ✓ *Baseline accuracy.*
        - ✓ *Cross-validation.*
        - ✓ *Cross-validation results.*
        - ✓ *Questions answered.*
    3. *Simple classifiers.*
        - ✓ *Simplicity first.*
        - ✓ *Overfitting.*
        - ✓ *Using probabilities.*
        - ✓ *Decision tres.*
        - ✓ *Pruning decision tres.*
        - ✓ *Nearest neighbor.*
        - ✓ *Questions answered.*
    4. *More classifiers.*
        - ✓ *Classification boundaries.*
        - ✓ *Linear regression.*
        - ✓ *Classification by regression.*
        - ✓ *Logistic regression.*

- ✓ *Support vector machines.*
- ✓ *Ensemble learning.*
- ✓ *Questions answered.*

5. *Putting it all together.*
   - ✓ *The data mining process.*
   - ✓ *Pitfalls and pratfalls.*
   - ✓ *Data mining and ethics.*
   - ✓ *Summary.*
   - ✓ *Questions answered.*

- *More Data Mining with Weka*
  1. *Exploring Weka's interfaces, and working with big data.*
  2. *Discretization and text classification.*
  3. *Classification rules, association rules, and clustering .*
  4. *Selecting attributes and counting the cost.*
  5. *Neural networks, learning curves, and performance optimization.*

Para el desarrollo de la aplicación con Java Swing se realizó mediante la plataforma de Udemy el curso "*Java Swing (GUI) Programming: From Beginner to Expert*" [80], el cual está disponible en la página web oficial de Udemy y cuyo contenido se divide en 3 secciones.

1. Sección 1: *Desktop applications*.
   - ✓ *Introduction: About the Course, Plus Some Useful Resources.*
   - ✓ *Creating a Basic Swing Application.*
   - ✓ *Adding Components: Layouts, Buttons and Text Areas.*
   - ✓ *Responding to Button Clicks.*
   - ✓ *Custom Components.*
   - ✓ *Simple Toolbars.*
   - ✓ *Communication Between Components.*
   - ✓ *Listeners and Events: Using Interfaces to Cleanly Separate Components.*
   - ✓ *Setting Component Sizes.*
   - ✓ *Setting Borders.*
   - ✓ *Text Fields and Labels.*
   - ✓ *Laying Out Controls with GridBagLayout.*
   - ✓ *Custom Events and Form Submission.*
   - ✓ *List Boxes.*
   - ✓ *Working With List Box Data.*
   - ✓ *Combo Boxes.*
   - ✓ *Checkboxes.*
   - ✓ *Radio Buttons.*
   - ✓ *Menus.*
   - ✓ *Using Checkboxes in Menus.*
   - ✓ *Mnemonics and Accelerators.*

- ✓ *Message Boxes.*
- ✓ *Open/Save File Dialogs.*
- ✓ *Filtering File Choosers.*
- ✓ *Model-View-Controller: Creating a Data Model.*
- ✓ *Model-View-Controller: Creating a Controller.*
- ✓ *Creating Tables.*
- ✓ *Serialization: Saving Objects to Files.*
- ✓ *Popup Menus.*
- ✓ *Selecting Rows in Tables.*
- ✓ *Deleting Rows in Tables.*
- ✓ *Dialogs.*
- ✓ *Spinners: Specialised Controls for Entering Numbers.*
- ✓ *Password Fields.*
- ✓ *Saving Program Data: Preferences.*
- ✓ *Arranging and Designing Dialogs and Forms.*
- ✓ *JDBC: Retrieving Values from Databases.*
- ✓ *JDBC: Saving to a Database.*
- ✓ *JDBC: Updating Databases.*
- ✓ *JDBC: Loading Data from Databases.*
- ✓ *Wiring in the Database Code: Bringing It All Together.*
- ✓ *Intercepting the Window Closing Event.*
- ✓ *Using Images and Icons.*
- ✓ *Draggable Toolbars: Using the JToolbar class.*
- ✓ *Split Panes: Creating Resizeable Separate Areas.*
- ✓ *Tabs: Using Tabbed Panes.*
- ✓ *Tree Views Using JTree.*
- ✓ *Tree Selection Events.*
- ✓ *Associating Data With Tree Nodes.*
- ✓ *Tree Node Icons.*
- ✓ *Custom Tree Cell Renderers: Using Checkboxes in Trees.*
- ✓ *Custom Tree Cell Editors: Editing Tree Nodes Using Checkboxes.*
- ✓ *Detecting Tree Node Editor Changes.*
- ✓ *A Simulated Message Server.*
- ✓ *Multithreading in Swing: The SwingWorker class.*
- ✓ *Modal Dialogs.*
- ✓ *Progress Bars.*
- ✓ *Distributing Your Application: Runnable Jars.*
- ✓ *Adding Text to Progress Bars.*
- ✓ *Cancelling SwingWorker Threads.*
- ✓ *Setting the Cursor.*
- ✓ *Multiple Nested Split Panes.*
- ✓ *Responding to Tab Selections.*

- ✓ *Custom List Renderers.*
- ✓ *Responding to List Selections.*
- ✓ *Changing the Font Using Logical Fonts.*
- ✓ *Loading Font Files.*
- ✓ *Configuring the Database Connection.*
- ✓ *Editable Tables.*
- ✓ *Using Checkboxes in Table Cells.*
- ✓ *Custom Table Cell Renderers.*
- ✓ *Using Custom Editors in Table Cells.*

2. Sección 2: *Applets and animations*.
   - ✓ *About Applets and the Following Section.*
   - ✓ *Drawing Custom Components.*
   - ✓ *Drawing Shapes: Exploring the Graphics API.*
   - ✓ *Deploying Applets.*
   - ✓ *Timers: Using the Swing Timer Class.*
   - ✓ *Basic Animation.*
   - ✓ *Smoothing Your Animations With Double Buffering.*
   - ✓ *Mouse Listeners.*
   - ✓ *Hiding the Cursor.*
   - ✓ *Key Listeners.*
   - ✓ *Detecting Component Resizing.*
   - ✓ *Using Visual Designers: The Free Window Builder Pro Plugin.*
   - ✓ *CardLayout: Switching Between Completely Different Views.*
   - ✓ *Detecting Collisions Between Shapes.*
   - ✓ *Turning an Applet into a Desktop App.*
3. Sección 3: *Appendix*.
   - ✓ *Setting the Look and Feel.*
   - ✓ *Source Code – Complete Projects.*
   - ✓ *Projtects – The Source Code Projects You See in the Tutorials.*
   - ✓ *"Swing Test" Database.*

Este curso enseña cómo crear aplicaciones de escritorio y basadas en la web utilizando Java Swing, el kit de herramientas de interfaz de usuario integrado de Java. Entre otras cosas, se estudian casi todos los *widgets* de Swing y se estudian los conceptos de JDBC para el acceso a bases de datos, la API de gráficos, la arquitectura modelo-vista-controlador (MVC), el modelo dirigido por eventos, entre otros.

ANEXO II. Desarrollo de la Herramienta CPDA

En este anexo se muestran los diferentes diseños ideados y se explica la metodología utilizada para el desarrollo de la herramienta.

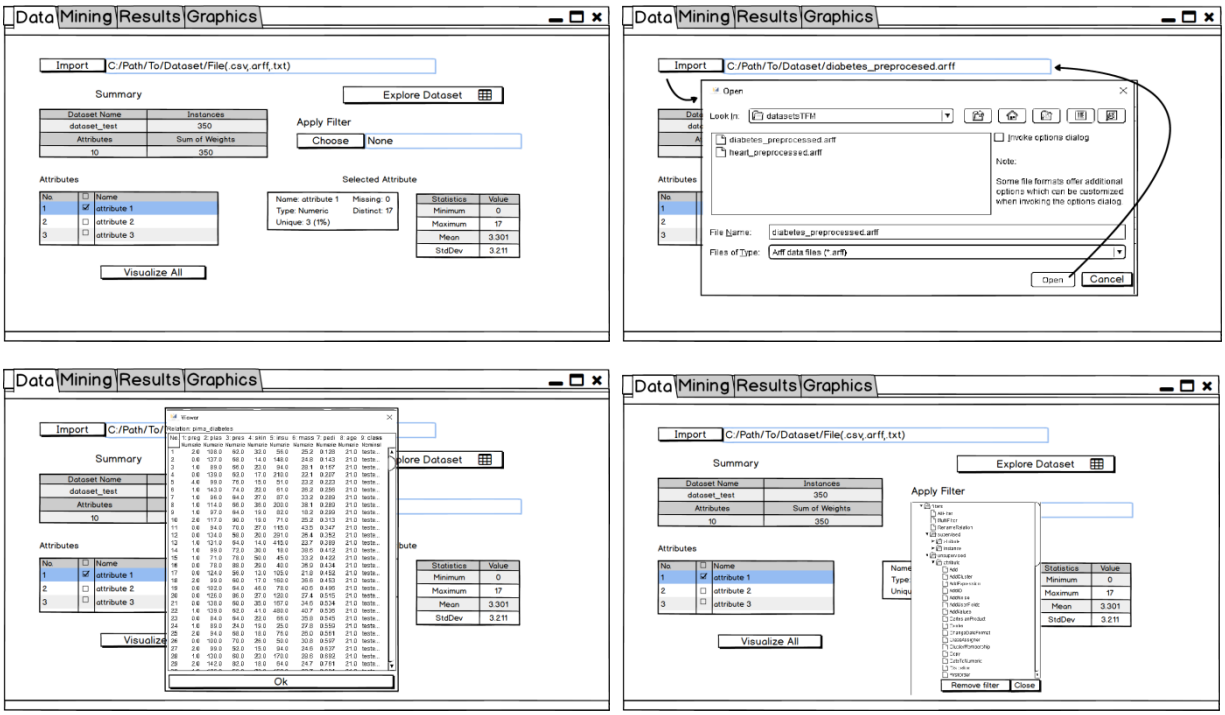
II.1. Bocetos digitales realizados con la herramienta Balsamiq Mockups.

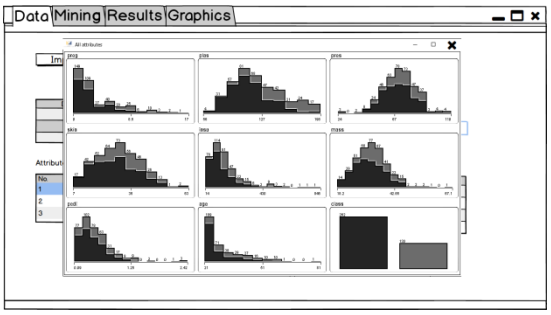
A continuación se muestran los bocetos iniciales y finales diseñados para dar de la herramienta, dividiendo cada caso en 4 secciones asociadas a las pantallas correspondientes.

II.1.1. Diseño Inicial

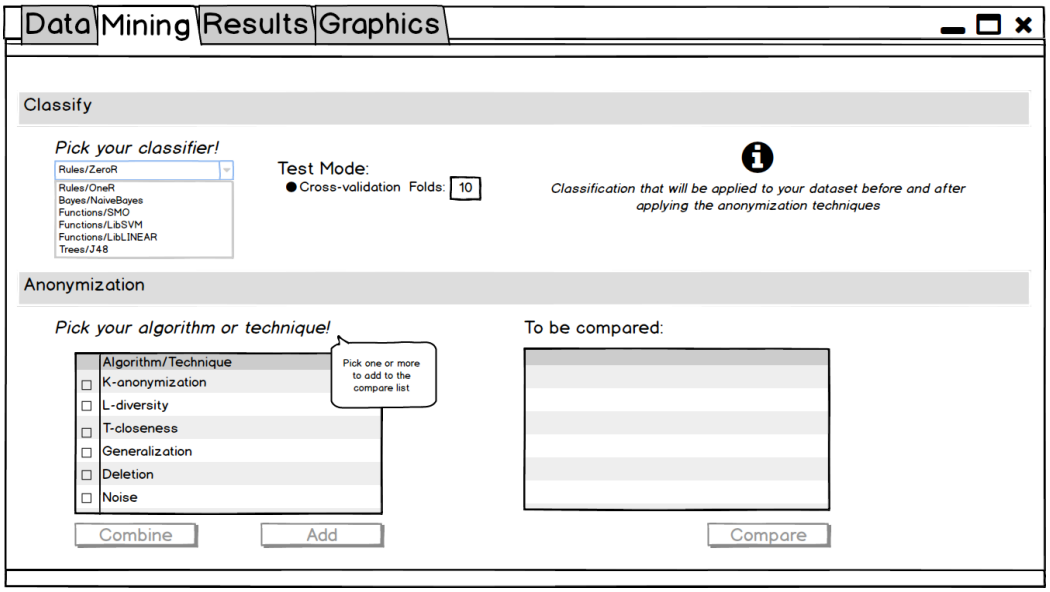
II.1.1.1. Pantalla de bienvenida

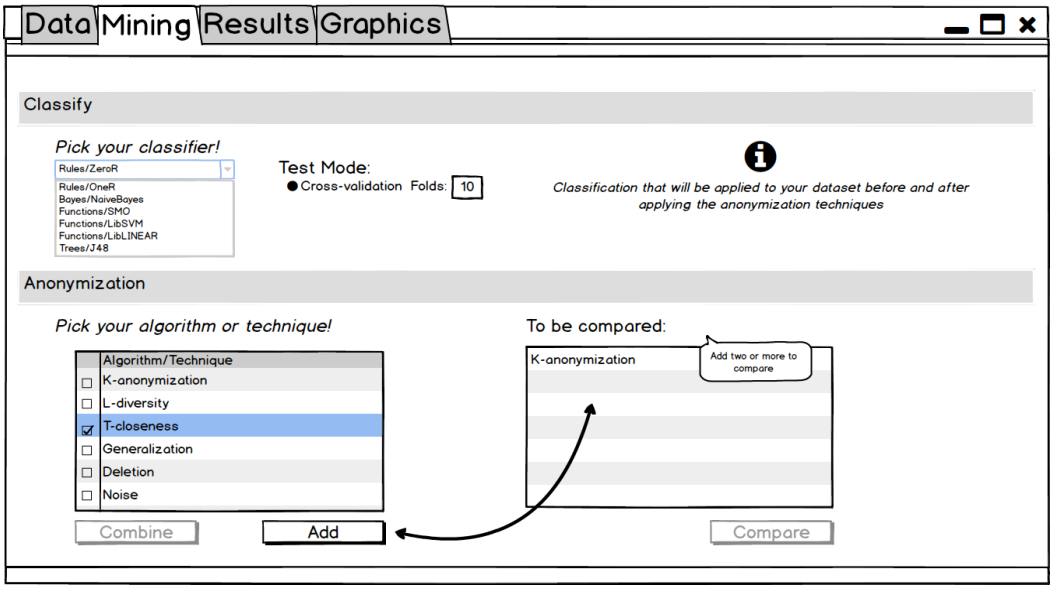
II.1.1.2. Pantalla de carga y visualización de los conjuntos de datos

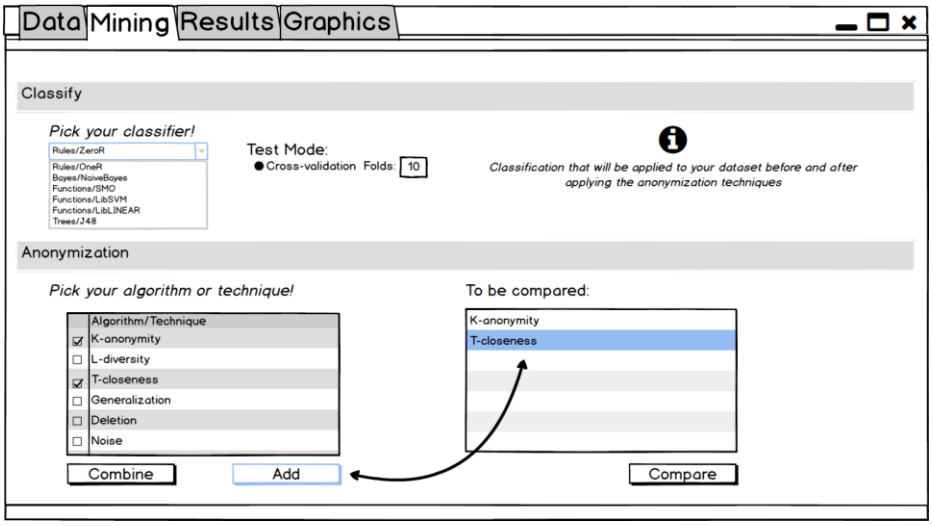


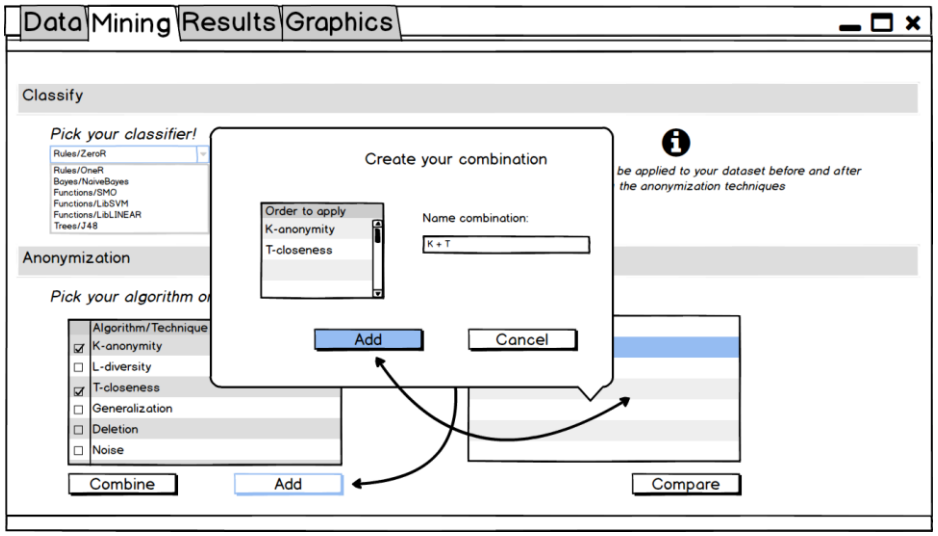


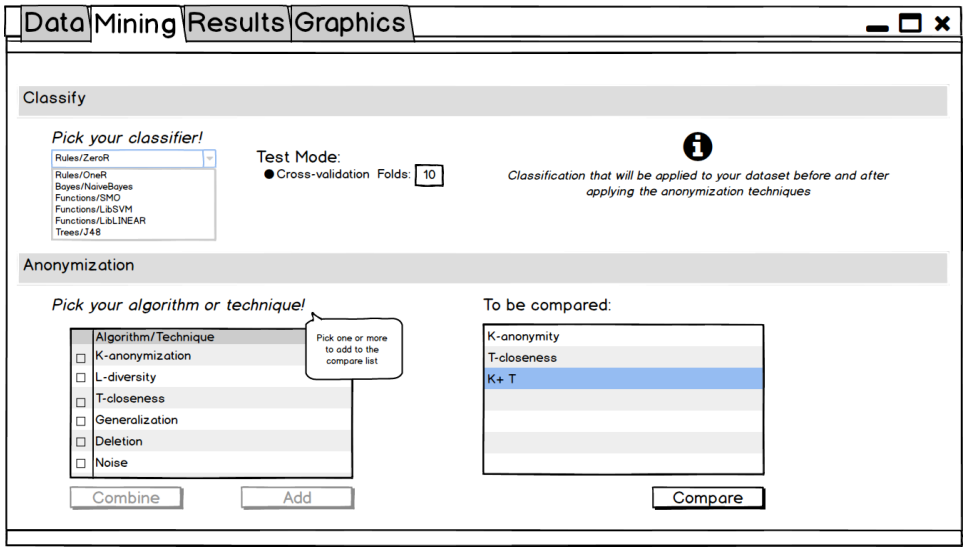
II.1.1.3. Pantalla para la minería de datos



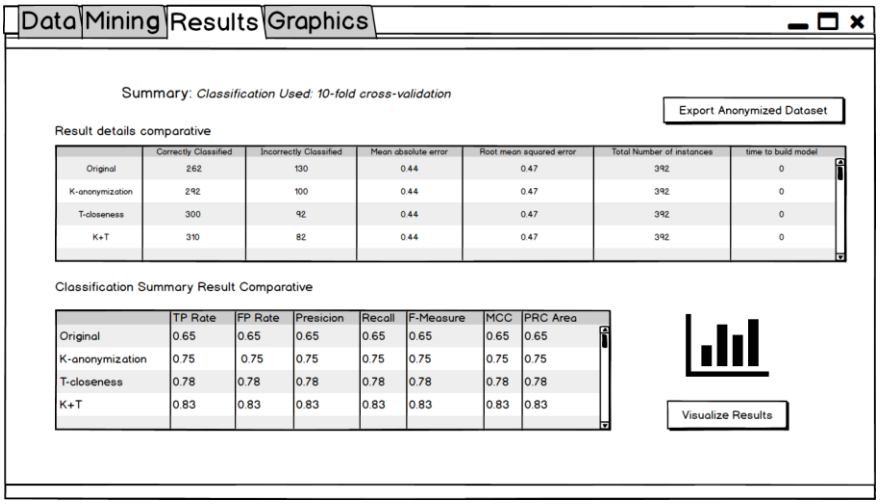


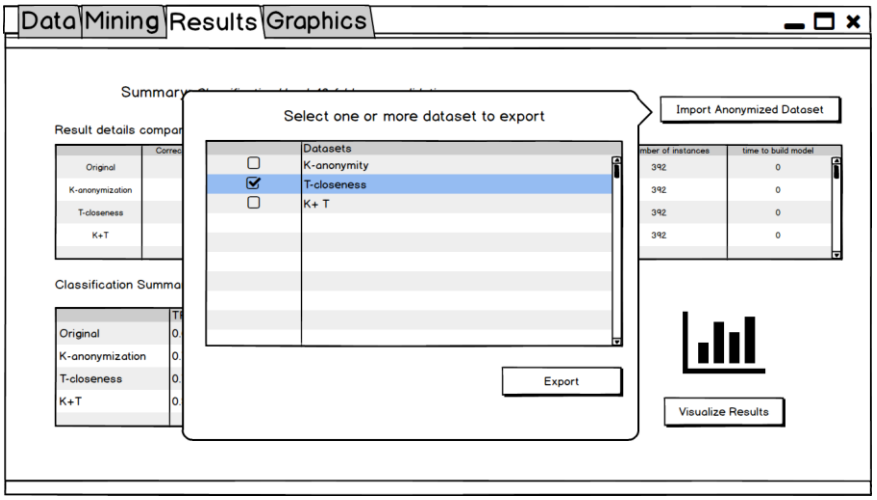


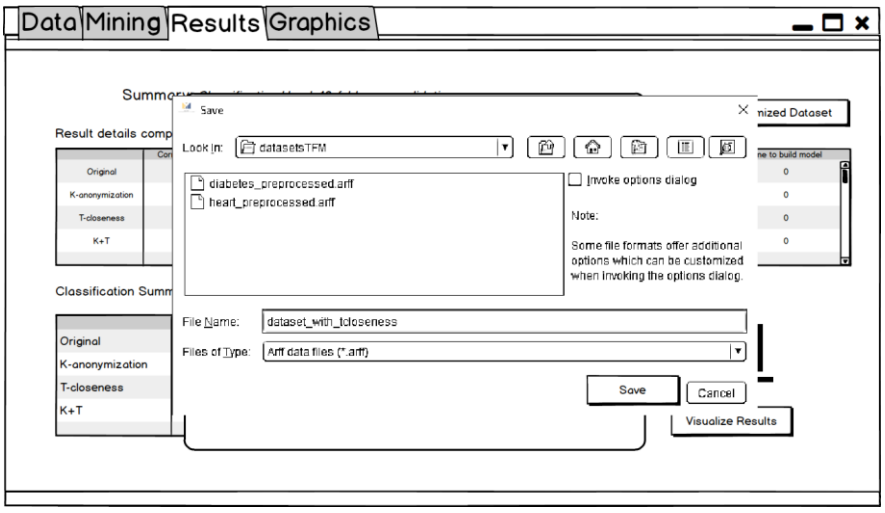




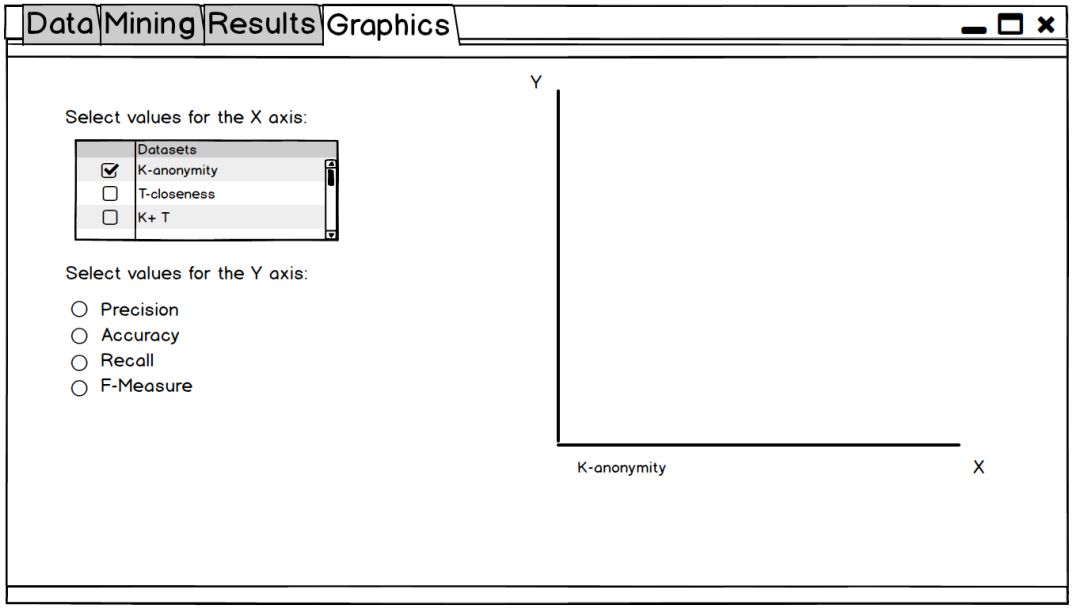
II.1.1.4. Pantalla para la visualización de resultados

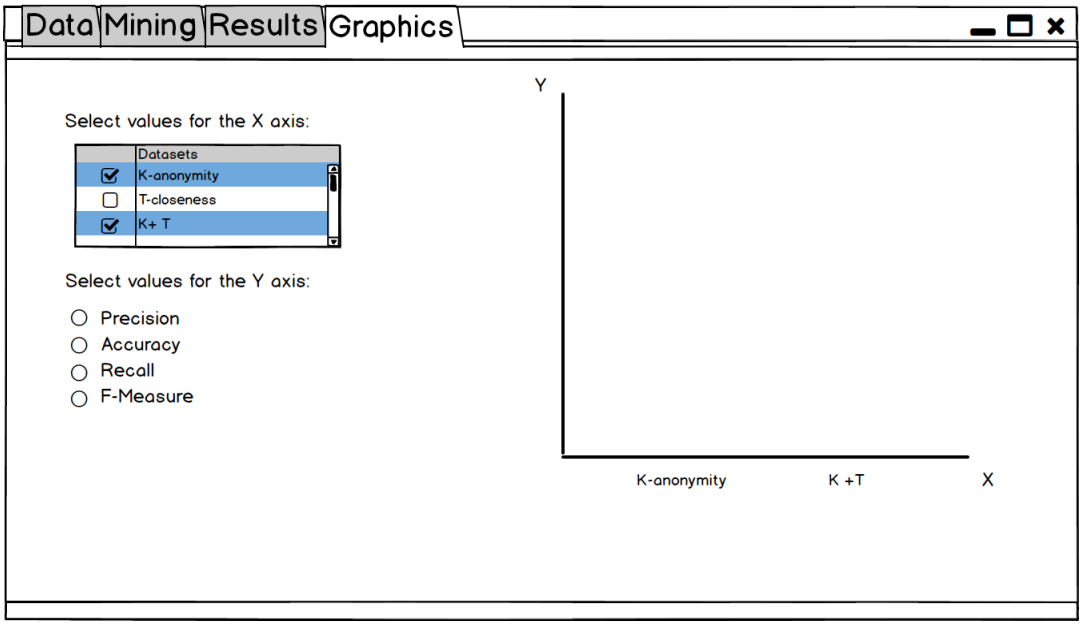


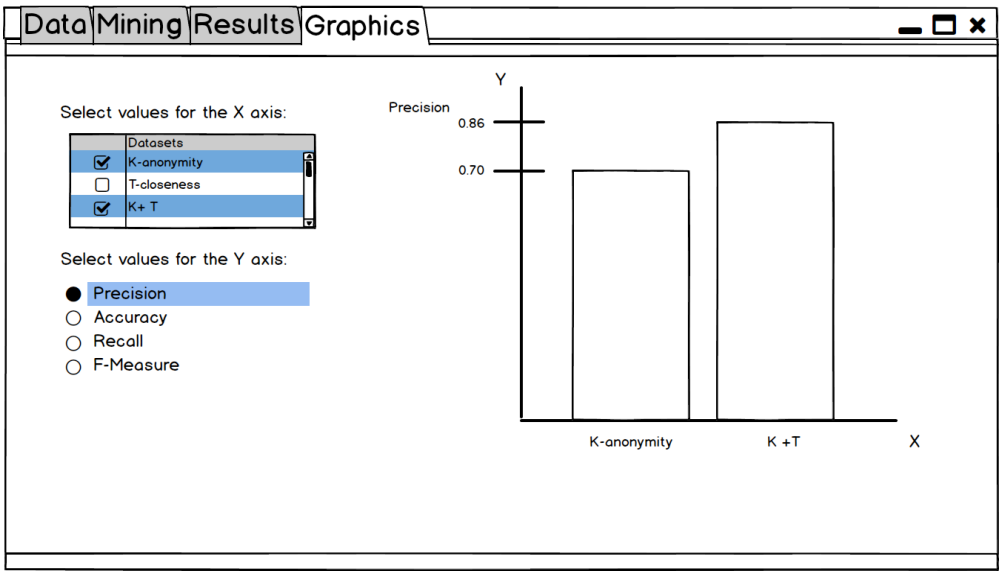




II.1.1.5. Pantalla para la visualización gráfica de los resultados



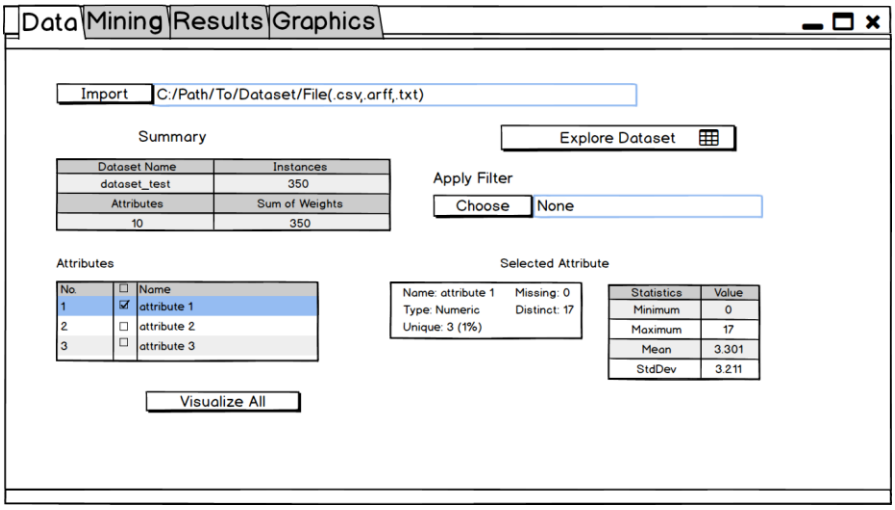


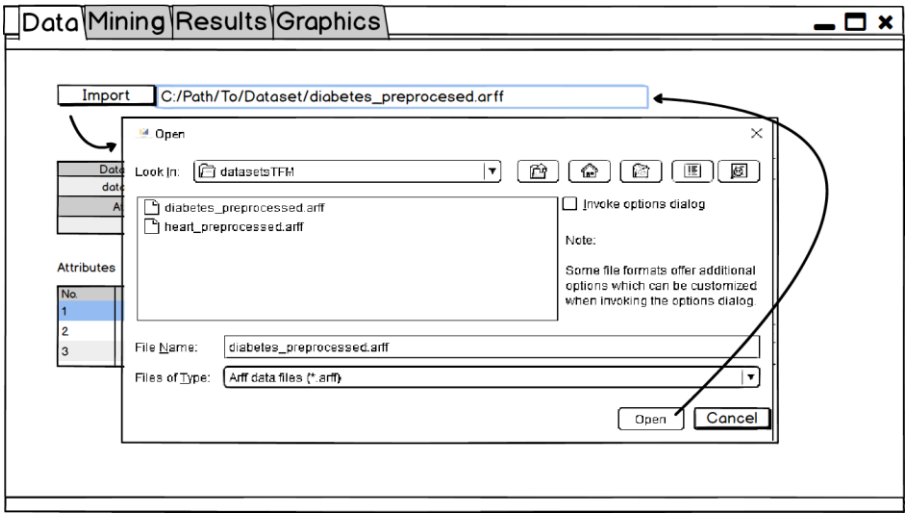


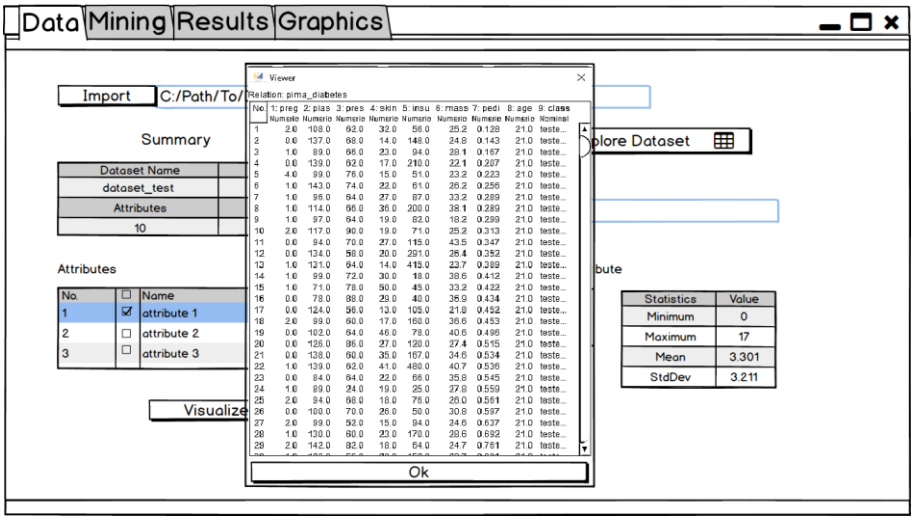
II.1.2. Diseño Final

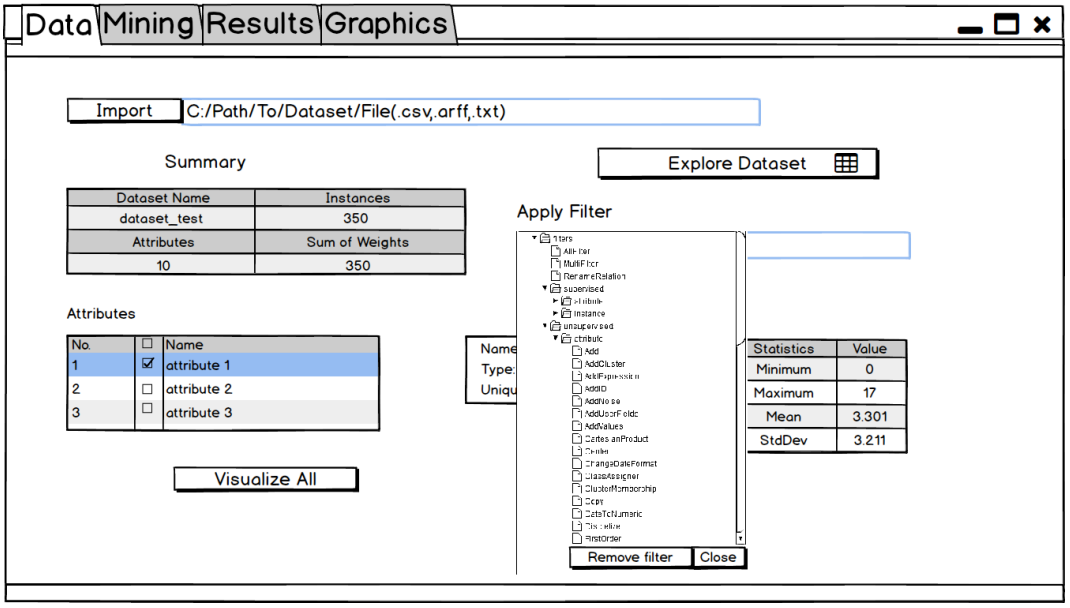
II.1.2.1. Pantalla de bienvenida

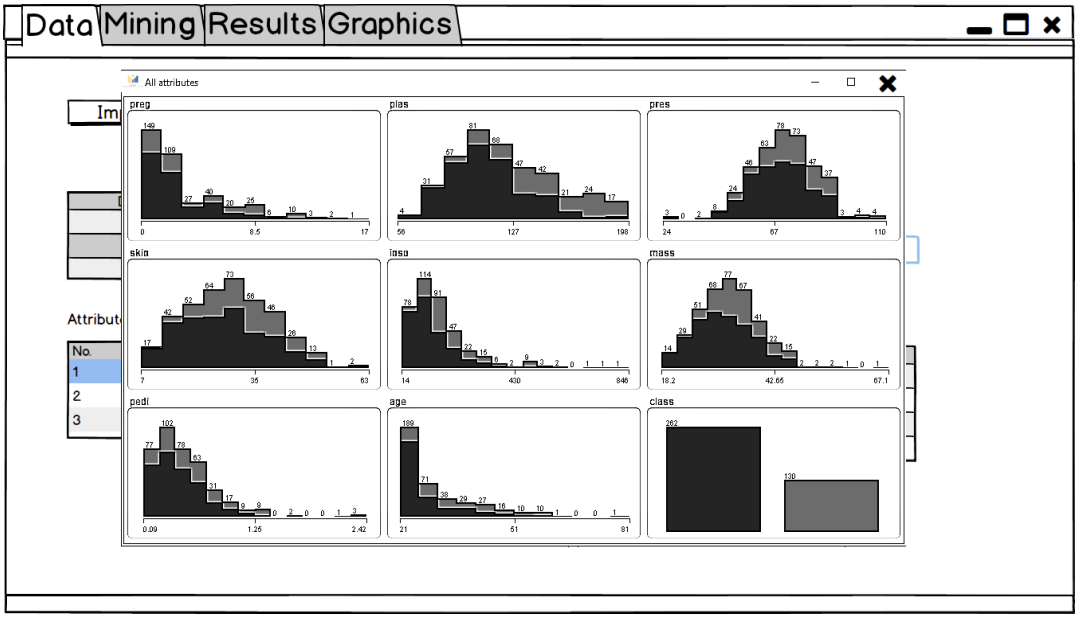
II.1.2.2. Pantalla de carga y visualización del conjunto de los conjuntos de datos



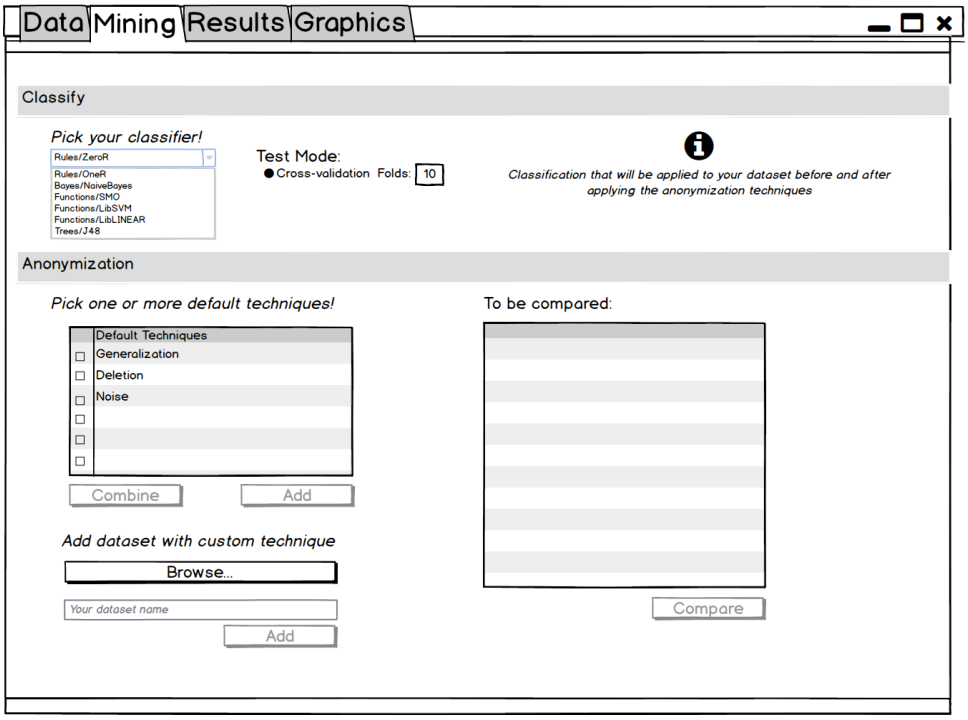


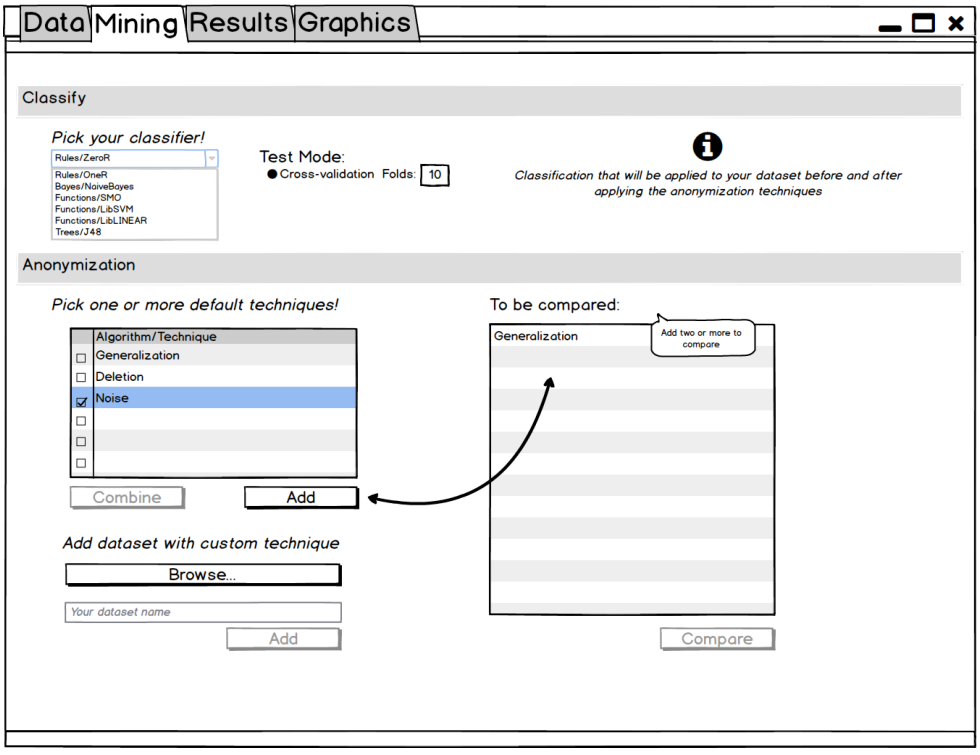


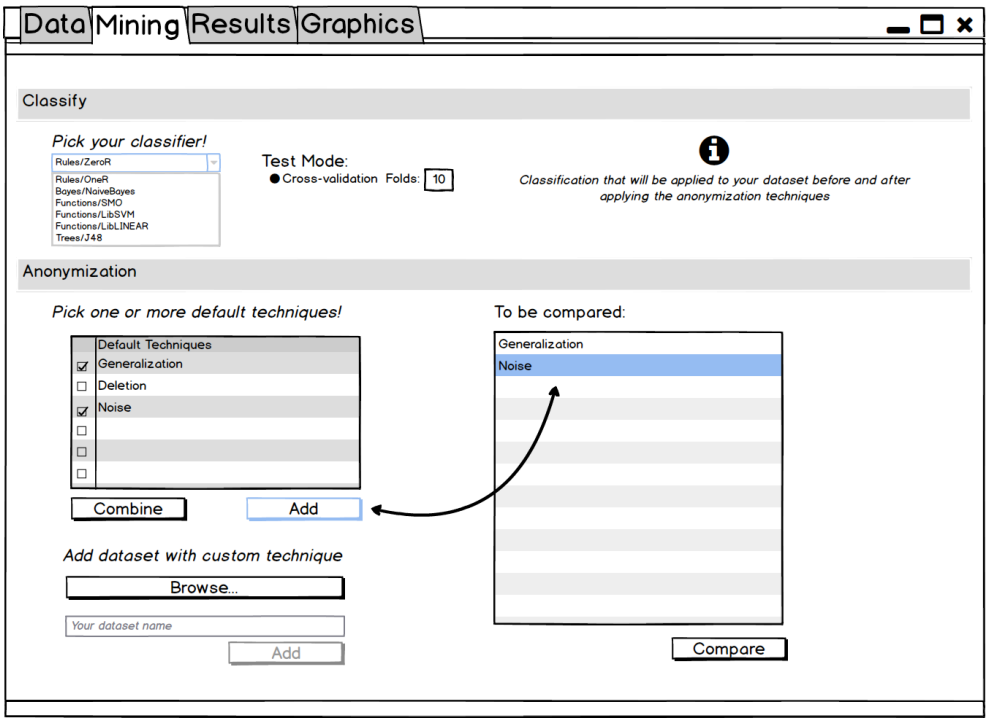


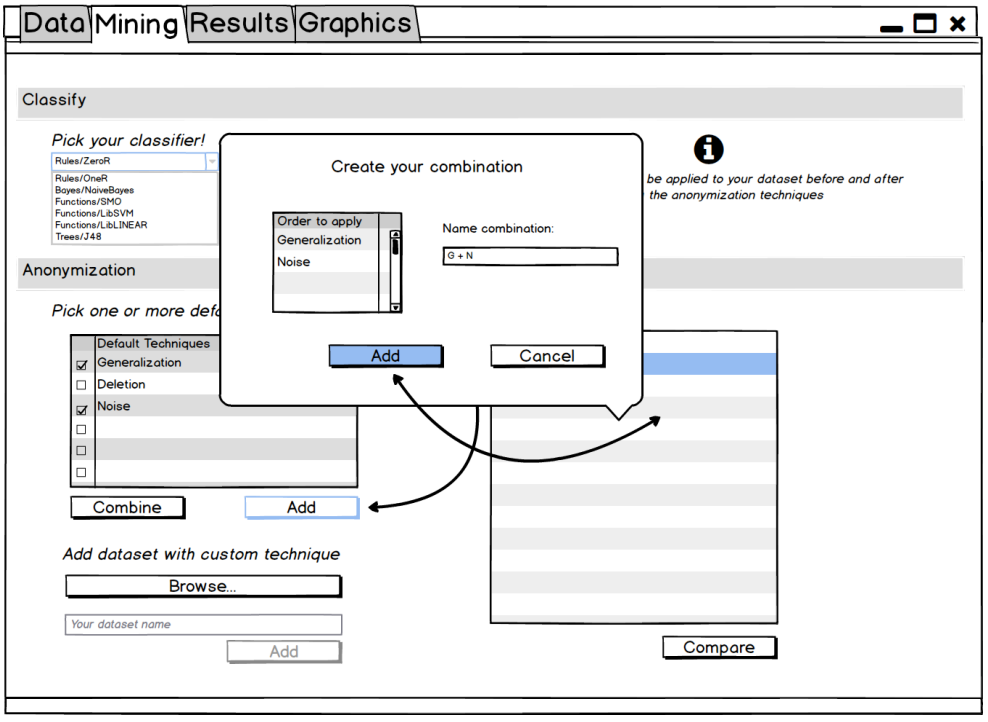


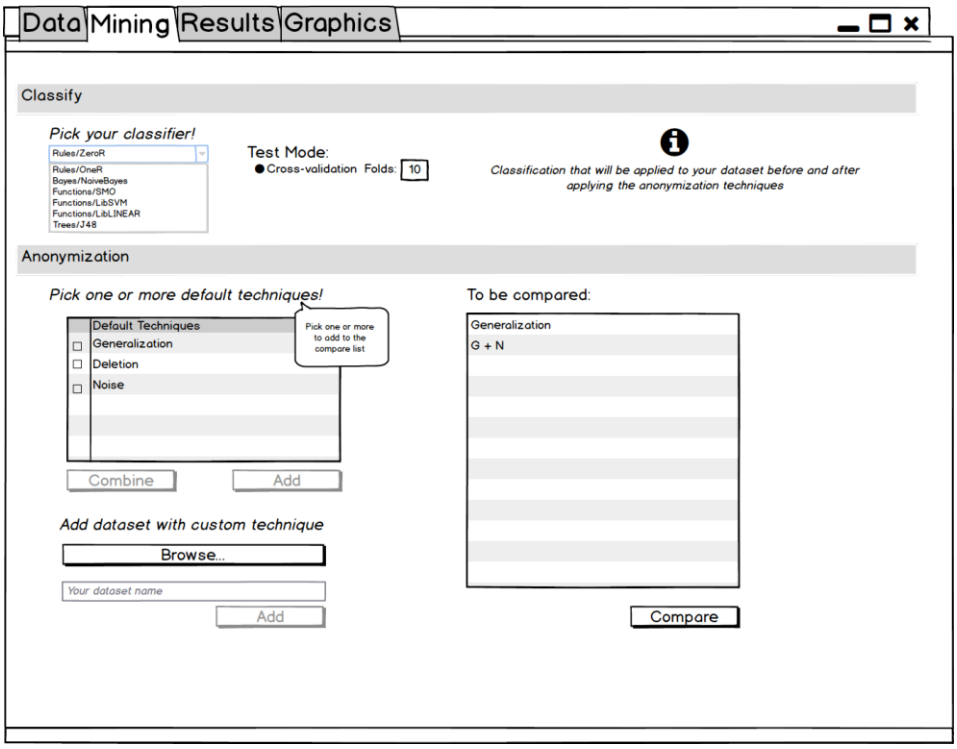
II.1.2.3. Pantalla para la minería de datos

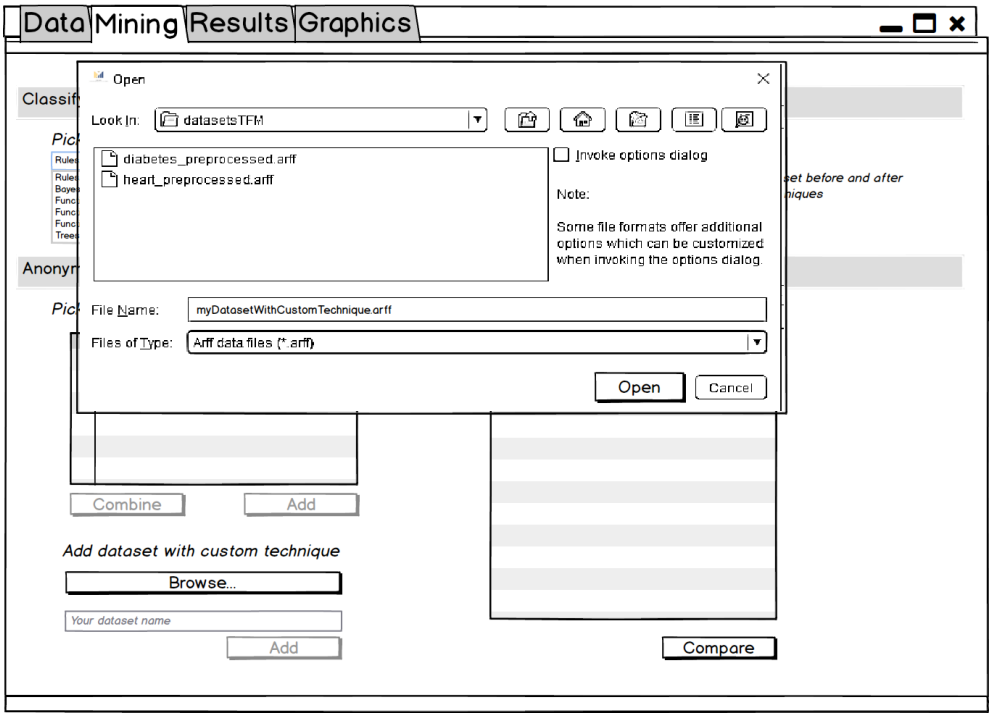


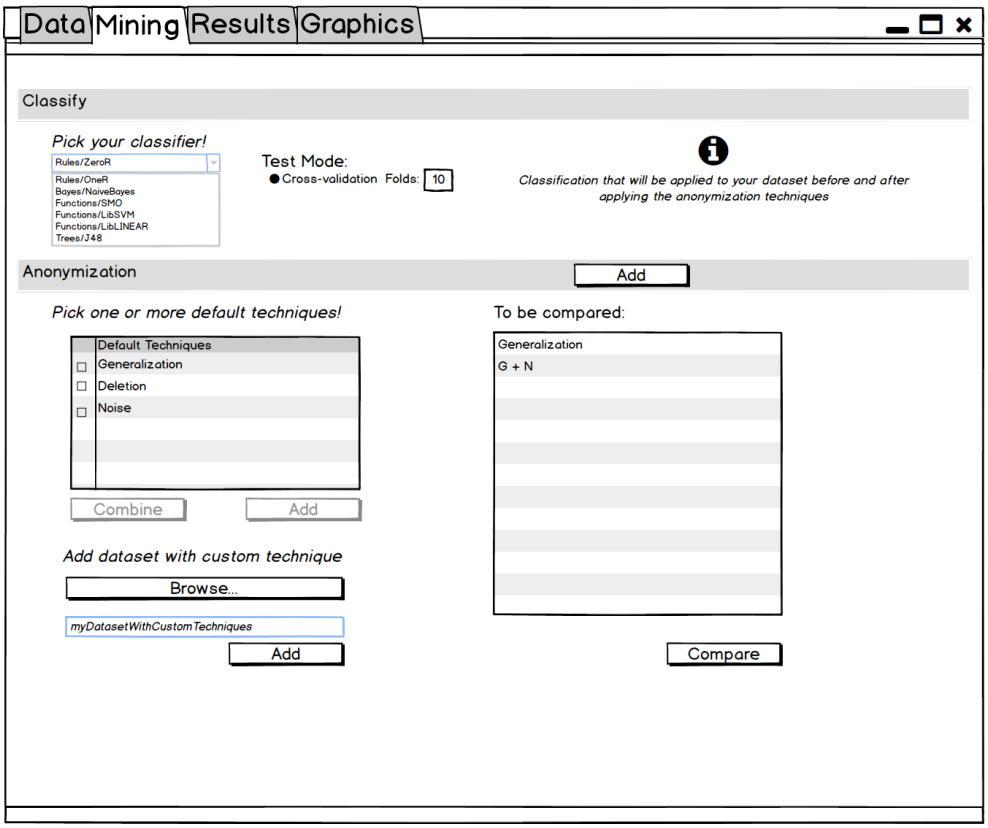


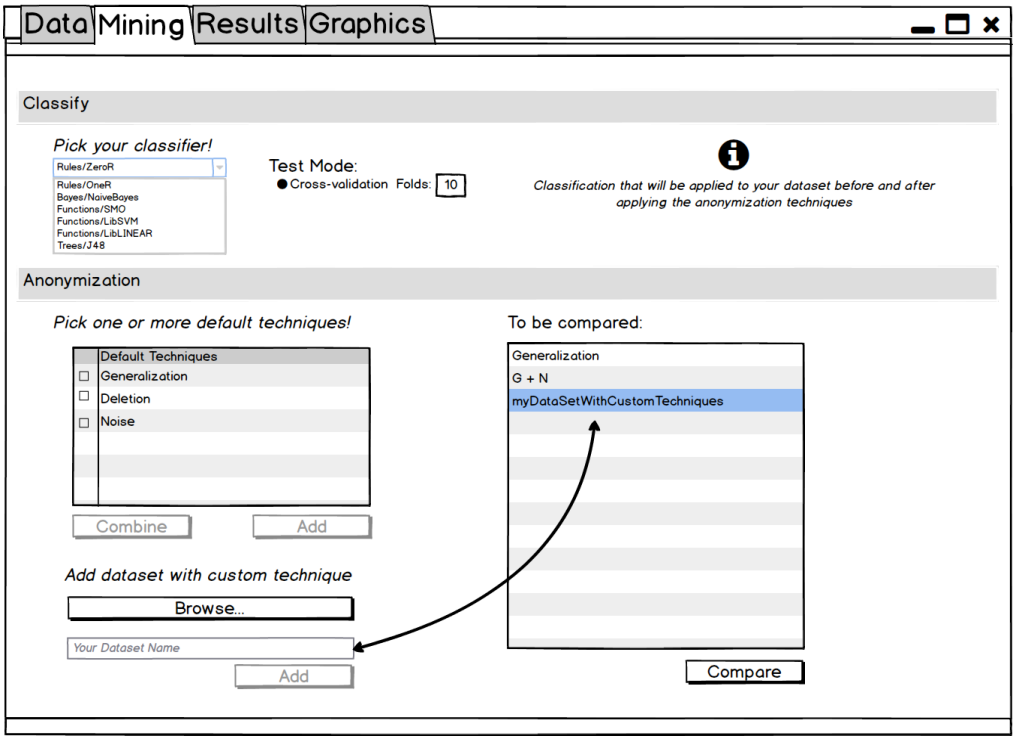




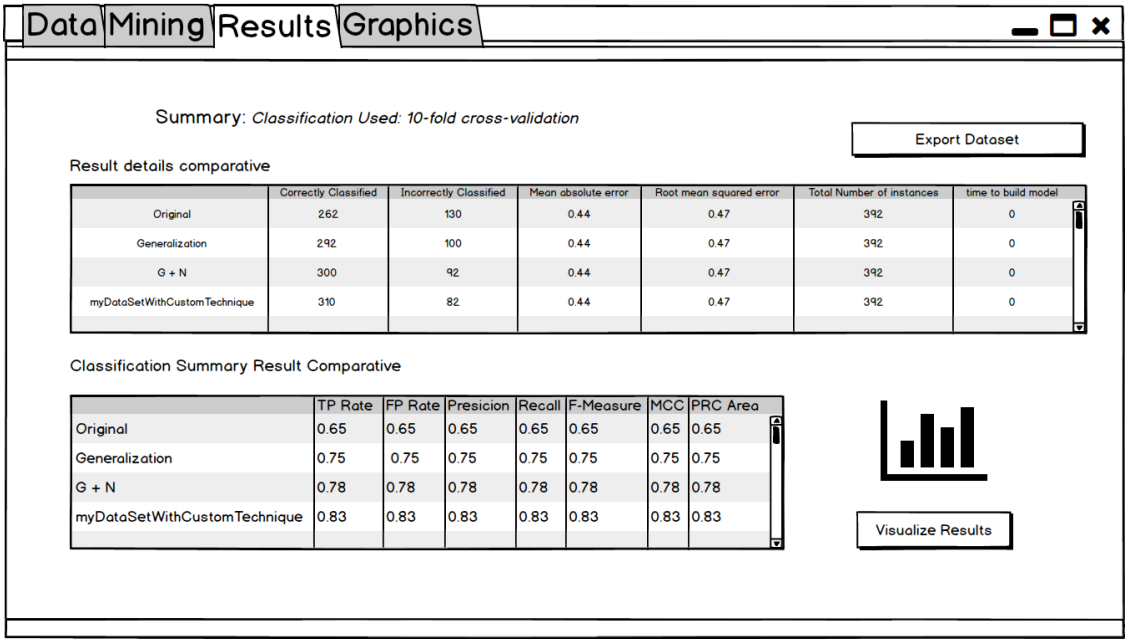


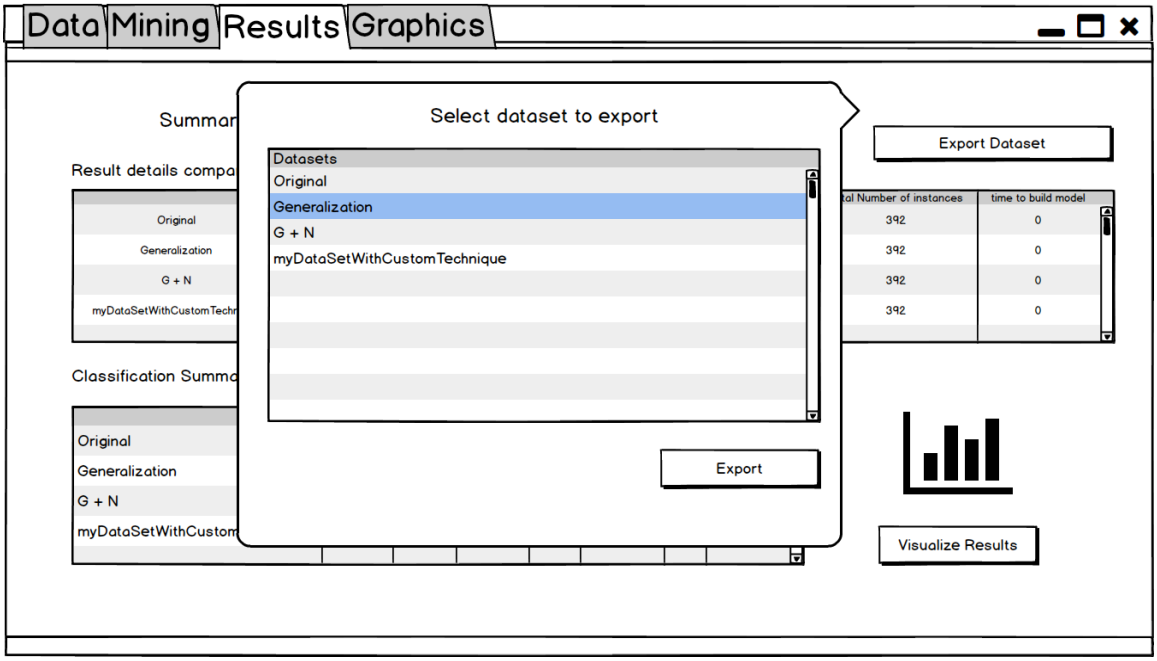


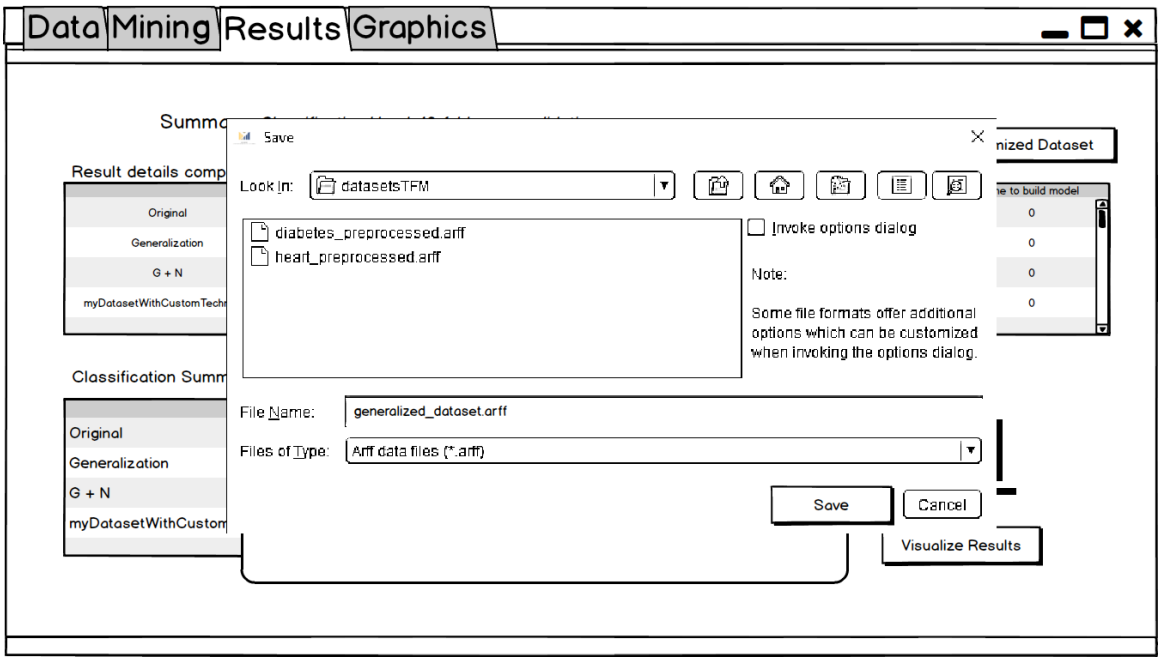




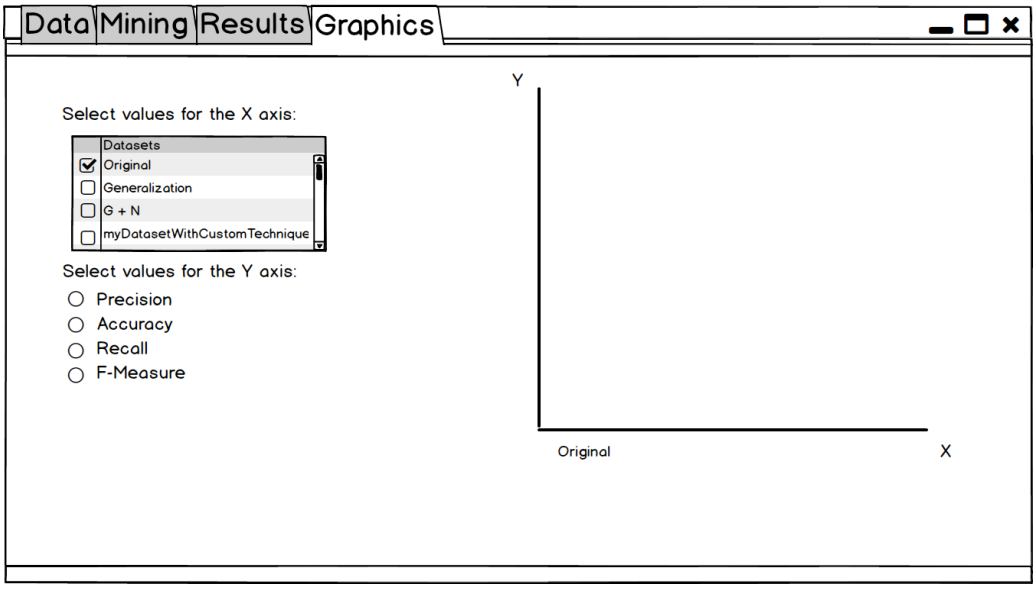
II.1.2.4. Pantalla para la visualización de los resultados

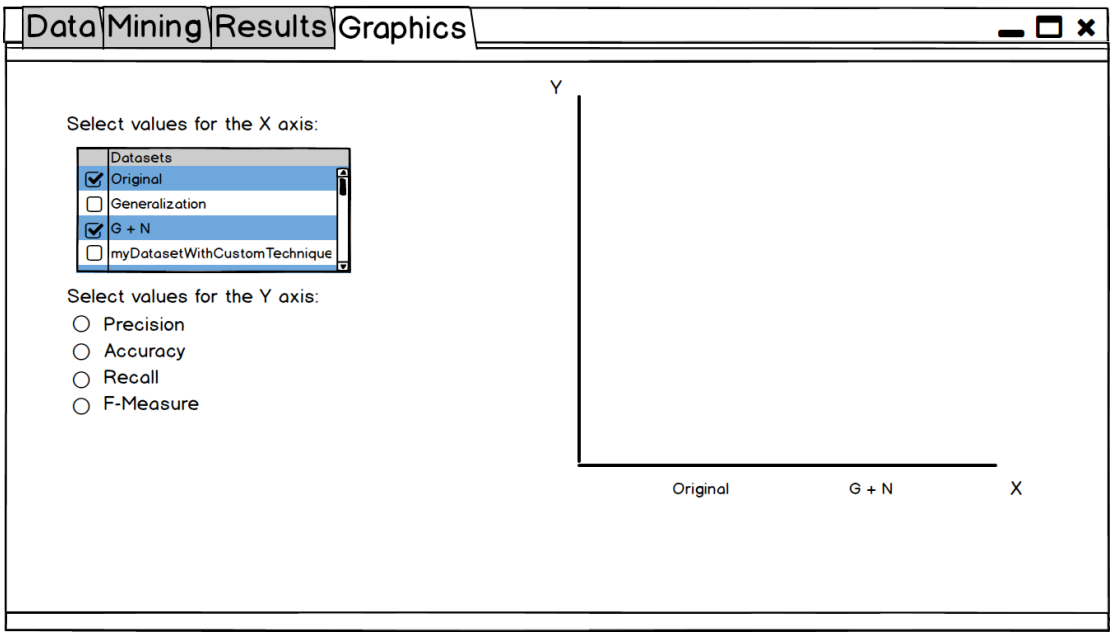


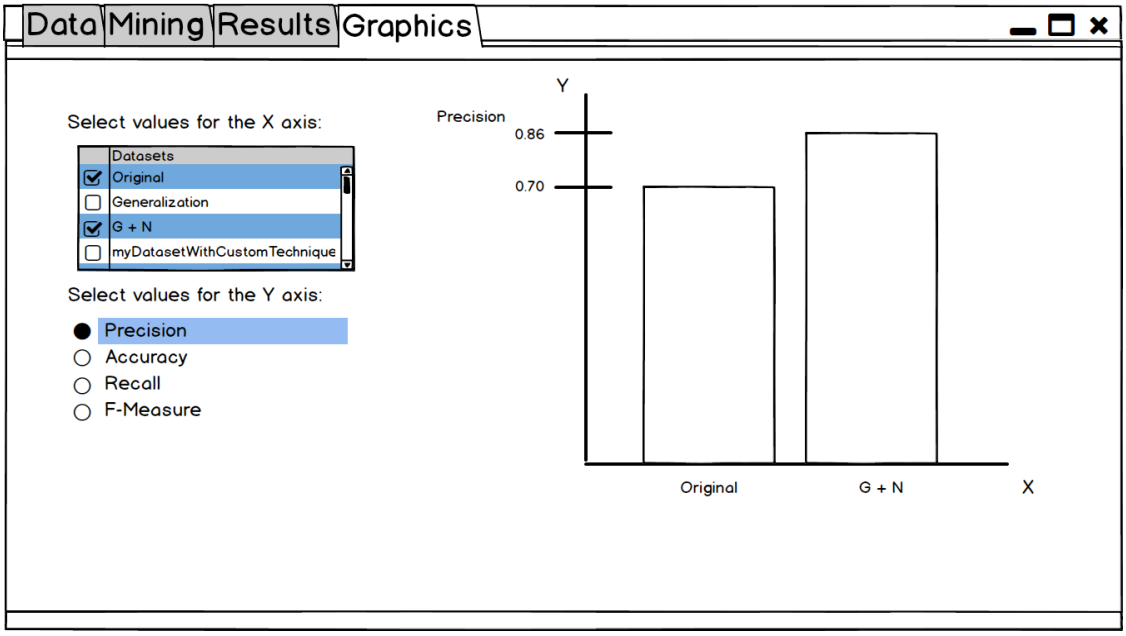




II.1.2.5. Pantalla para la visualización gráfica de los resultados







II.2. Metodología para el desarrollo

Para el desarrollo de la aplicación CPDA, se decidió seguir a modo orientativo una metodología *agile* debido a las ventajas que ofrece como el desarrollo auto-organizado, la división del trabajo en tareas o historias de usuario, adaptabilidad y flexibilidad ante los cambios, entre otros.

II.2.1. Descripción de la metodología

Para el desarrollo de la herramienta CPDA, se decidió trabajar con la metodología ágil de desarrollo SCRUM [49], el cual es un modelo de referencia que define un conjunto de prácticas y roles, que puede tomarse como punto de partida para definir el proceso de desarrollo que se ejecutará durante un proyecto.

Para el caso particular de este trabajo, no se definieron roles, dado que es un trabajo unipersonal, por lo cual no tendría sentido la definición de roles como *Scrum Master*, *Product Owner* o *Team/Developer*.

Se decide usar la metodología a modo orientativo, como un modelo de apoyo, permitiendo tener una disciplina en el trabajo a realizar, con unos objetivos y tiempos definidos, permitiendo observar así el progreso a lo largo del tiempo, y los resultados obtenidos al finalizar cada una de las fases.

II.2.2. Implementación de la metodología SCRUM

En este trabajo se decidió realizar una división del mismo en dos etapas, la de investigación y la del desarrollo. Es en la segunda etapa en donde entra en juego el uso de la metodología SCRUM, dado que todo lo investigado en la primera etapa serviría de base para poder definir la herramienta a desarrollar y el alcance deseado de la misma. Dentro de la segunda etapa del trabajo, se procedió a definir las 5 fases del trabajo, las cuales fueron:

- Diseño de *mockups* (bocetos).
- Vista de la sección de datos.
- Vista de la sección de minería de datos.
- Vista de la sección de resultados.
- Vista de la sección gráfica.

Partiendo de estas 5 fases se definieron inicialmente 5 *sprints*, uno por cada fase, contemplando una duración de 15 días para cada *sprint*.

Dicha planificación fue modificada a medida que se progresaba en el desarrollo, debido a que se observó una sub-estimación en los tiempos contemplados para las fases 2 y 3. De esa forma la planificación sufrió un reajuste y las fases fueron reestimadas y valoradas en un total de 2 *sprints* por cada fase, es decir, un total de 30 días para las fases 2 y 3 del trabajo.

De esta forma la planificación quedó en un total de 7 *sprints*, más un *sprint* final de 20 días, para la realización de cambios recomendados, desarrollos adicionales deseables, pruebas, entre otros, quedando así un total de 8 *sprints*, cuya planificación y estimación puede verse reflejada a continuación:

- ***Sprint #1 – Diseño Mockups: del 02/02/2020 al 16/02/2020***

El objetivo principal de este *sprint* fue la realización del diseño inicial de la herramienta, los entregables obtenidos en este *sprint* se pueden observar en el Apéndice A, donde se visualizan los diseños realizados para dar base al inicio del desarrollo.

Entre las tareas a realizar dentro del *sprint* se contemplaron las siguientes:

- Diseño Logo app
- Diseño Splash Screen
- Diseño Vista Inicial (Data Section)
- Diseño Vista Mining Section
- Diseño Vista Results Section
- Diseño Vista Gráficos Comparativos

Figura 29 - Captura de pantalla del sprint 1 en Asana

- ***Sprint #2 – Data Section: del 17/02/2020 al 03/03/2020***

El objetivo principal del *sprint* 2 fue además de lograr establecer la conexión con la librería de Weka para el trabajo con los conjuntos de datos desde Java, inicializar la sección de la carga de conjuntos de datos, lo cual implica la lógica para cargar un conjunto de datos dentro de la aplicación y mostrar la información asociada al mismo a través de la interfaz gráfica diseñada.

Las tareas y subtareas asociadas a este *sprint* fueron las siguientes:

- Import Logic and Design (Lógica de importación y diseño)
  - Diseñar botón para importar conjuntos de datos.
  - Abrir una nueva ventana mostrando el sistema de archivos local al hacer *clic* en el botón de importar.

        - Abrir el sistema de archivos en una ruta por defecto asociada a la aplicación.
        - Al seleccionar un fichero, se mostrará la ruta del mismo en la casilla *Path* mostrada al lado del botón de importar.
    - Dataset Information load Logic and Design
        - Al cargar el conjunto de datos, mostrar información resumen: nombre, instancias, atributos, etc.
        - Se mostrará una tabla con los atributos asociados al conjunto de datos.
        - Se habilitará el botón de Explore/Visualize data.

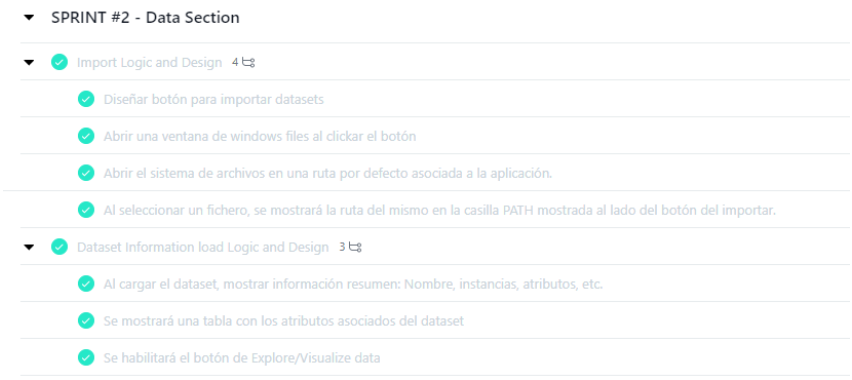


Figura 30 - Captura de pantalla del sprint 2 en Asana

- ***Sprint #3 – Data Section: del 05/03/2020 a 20/03/2020***

El *sprint* 3 consistió en la continuación del *sprint* anterior. En este se tenía como objetivo generar un primer entregable de la sección de datos, es decir, se esperaba que al finalizar el *sprint* toda la lógica asociada a la carga y lectura de conjuntos de datos estuviera finalizada, permitiendo al usuario observar todos los datos asociados al conjunto seleccionado: las instancias, atributos, la información asociada a cada uno de los mismos, e incluso una visualización general del conjunto de datos cargado.

Las tareas y sub-tareas asociadas a este *sprint* fueron las siguientes:

- Dataset information load logic and design
    - El botón de visualizar datos mostrará las instancias como la herramienta de weka, que permite visualizar un conjunto de datos.
    - El botón inferior “Visualize All” se habilitará al cargar un conjunto de datos.
- Dataset attributes display and selection logic and design
    - Mostrar atributos en la tabla correspondiente.
    - Hacer atributos seleccionables.

- Al seleccionar un atributo tipo numeric, mostrar la información asociada en la tabla asociada “attributes details”.
- Al seleccionar un atributo tipo nominal, mostrar la información asociada en la tabla asociada “attributes details”.
- Al seleccionar otro atributo, actualizar tabla con información respectiva.
- Al seleccionar un atributo, mostrar la información resumen del mismo.

Figura 31 - Captura de pantalla del sprint 3 en Asana

- ***Sprint #4 – Mining Section: del 22/03/2020 a 05/04/2020***

En el *sprint* 4 se da inicio al desarrollo de la sección de minería de datos de la herramienta, en ella se espera poder aplicar técnicas tanto de clasificación como de anonimización al conjunto de datos que se ha cargado en la sección de datos anterior.

Las tareas y subtareas definidas para este *sprint* fueron las siguientes:

- Subdividir en 2 secciones la interfaz: Técnicas de anonimización y clasificadores
- Sección del clasificador:
  - Título y subtítulo de la sección.
  - Botón tipo “dropdown” que liste los clasificadores disponibles.
  - Se deberá indicar visualmente que el *Test Mode* es el de *cross-validation*, y se deberá dejar un *input* numérico que permita al usuario indicar el número de *folds*.
  - La lista de clasificadores será limitada a unos 5 clasificadores aproximadamente, llamando a la librería de Weka para aplicar el clasificador sobre el conjunto de datos original y el/los conjuntos de datos que se originen de las técnicas de anonimización.
- Sección de técnicas de anonimización:

- Listar técnicas para aplicar en el conjunto de datos.
- Será posible seleccionar 1 o más técnicas. Mostrar mensaje informativo que lo indique.
- Al seleccionar 1 o más técnicas se habilitará el botón de añadir.
- Botón de añadir, que agregará la/s técnica/s seleccionada/s a la lista de "*datasets to generate*"
- Al clicar con botón derecho un elemento de la lista con las técnicas seleccionadas, se elimina el elemento de la lista.
- Añadir *tooltip* sobre lista de técnicas seleccionadas para saber cómo eliminar de la lista.
- Si selecciona 2 o más técnicas, se habilitará el botón de "*combine*".
- Si se deseleccionan las técnicas, hasta quedar 1 o ninguna, se desactiva el botón de "*combine*".
- El botón de combinar abrirá una ventana mostrando las técnicas seleccionadas.
- El botón añadir agregará la combinación a la lista de "*datasets to generate*".
- El botón cancelar volverá a la vista anterior.

Figura 32 - Captura de pantalla del sprint 4 en Asana

- ***Sprint #5 – Mining: 07/04/2020 a 22/04/2020***

En el *sprint* 5 se continúa con el desarrollo del *sprint* anterior. En este se tiene como entregable la sección de minería de datos finalizada; es en esta sección donde el usuario puede, además de estudiar mediante una técnica de clasificación el conjunto de datos dado, aplicar una serie de técnicas usadas para la anonimización de datos, e incluso subir el conjunto de datos original con técnicas aplicadas por el usuario con otras herramientas y aplicaciones externas, como Weka y R.

Las tareas y subtareas asociadas a este *sprint* fueron las siguientes:

- Sección de Técnicas de anonimización:
  - En la ventana de combinar, se podrá indicar el orden en que aplicar las técnicas, y se dará un nombre a la combinación.
  - Bloquear las secciones hasta que exista un conjunto de datos cargado.
- Botón Comparar:
  - Se tendrá que validar que el numero de *folds* es correcto; de lo contrario se abrirá un mensaje de error
  - El botón se habilitara cuando la lista de conjuntos de datos a comparar no esté vacía.
- Modificación Vista:
  - Añadir al panel la opción de subir conjuntos de datos propios.
  - Los conjuntos de datos que se carguen a la aplicación podrán ser añadidos mediante un botón al listado a comparar.

Figura 33 - Captura de pantalla del sprint 5 en Asana

- ***Sprint #6 – Results Section: 24/04/2020 a 08/05/2020***

En el *sprint* 6 se da inicio al desarrollo de la sección de resultados. De este *sprint* se espera como resultado obtener la vista de los resultados, donde se mostrarán unas tablas que permitan comparar el conjunto de datos original con los conjuntos de datos obtenidos de la aplicación de las técnicas de anonimización, de forma que el usuario pueda comparar los mismos y observar cuál de las técnicas aplicadas genera mejores resultados.

Las tareas y sub-tareas asociadas a este *sprint* fueron las siguientes:

- Re-estructuración:
  - Guardar un listado de conjuntos de datos (original y los que añada el usuario con sus técnicas).
  - Anadir lógica para listar los conjuntos de datos subidos por el usuario.

- Al entrar en la sección de minería de datos (*mining tab*), debe salir de forma automática el conjunto de datos subido por el usuario dentro del listado de conjuntos de datos a comparar.
- Anadir lógica para que el conjunto de datos original no se pueda eliminar de la lista.
- Anadir lógica para mostrar la técnica a aplicar en el listado de técnicas seleccionadas.
- Añadir lógica para aplicar generalización a un conjunto de datos.
- Añadir lógica para aplicar ruido a un conjunto de datos.
- Aplicar eliminación a un conjunto de datos.

- Diseño Vista Resultados:
  - Diseñar paneles de la vista de los resultados.
  - Mostrar los resultados obtenidos en 2 tablas comparativas, donde se aprecien los valores asociados a cada conjunto de datos.
- Resumen General
  - Mostrar en las tablas correspondientes los resultados de cada conjunto de datos.
  - La lista de estos resultados podrá tener un *scroll* vertical, y se mostrará según la cantidad de técnicas comparadas.
  - Añadir opción de exportar conjuntos de datos.
  - Mostrar a modo resumen el clasificador usado y las técnicas aplicadas.
  - Añadir vista *popup* con lista de conjuntos de datos para seleccionar y exportar.
  - Añadir lógica para exportar conjuntos de datos.
  - Anadir lógica para exportar como arff o como csv.
  - Mostrar mensaje de confirmación que indique el directorio en el que se ha descargado el fichero, o de error en caso de haber encontrado problemas.

Figura 34 - Captura de pantalla del sprint 6 en Asana

- ***Sprint 7 – Graphical Section: 10/05/2020 a 25/05/2020***

Considerado como el *sprint* final del desarrollo, es en este *sprint* donde se tiene como objetivo el desarrollo de la vista gráfica de la aplicación, donde mediante gráficas el usuario pueda visualizar los conjuntos de datos obtenidos, y en base a alguna característica obtenida de la minería de datos se pueda comparar los conjuntos de datos.

Las tareas y sub-tareas asociadas a éste *sprint* fueron las siguientes:

- Resumen General:
  - Añadir vista popup con lista de conjuntos de datos para seleccionar y exportar.
  - Añadir lógica para exportar conjuntos de datos.
  - Anadir lógica para exportar como arff o como csv.
  - Mostrar mensaje de confirmación que indique el directorio en el que se ha descargado el fichero, o de error en caso de haber encontrado problemas.
- Diseño vista gráficas:
  - Configurar un *loader* específico para ficheros con extensiones tipo csv.
  - Diseñar vista gráficos.
  - Listar conjuntos de datos seleccionables para añadir a la gráfica comparativa (multi selección)
  - Listar los valores del eje Y seleccionables (único valor por vez, usando *radio buttons*).
  - Pintar gráfico de barras usando la librería JFreeChart.
  - Añadir elementos seleccionados al eje X de la gráfica.
  - Añadir elemento seleccionado al eje Y, y cambiar según selección.
  - Pintar barras asociadas al valor X, Y correspondiente.

Figura 35 - Captura de pantalla del sprint 7 en Asana

- ***Sprint #8 - Desirables: 27/05/2020 al 17/06/2020***

El *sprint* 8 es el *sprint* creado a partir de varios elementos que fueron considerados como deseables durante el desarrollo de los *sprints* anteriores. En este *sprint* se incluyen cosas como la realización de un ejecutable de la aplicación y la preparación de la entrega final de la app, donde se incluye la limpieza y documentación del código, entre otros.

Figura 36 - Captura de pantalla del sprint 8 en Asana

ANEXO III. Manual de Usuario de CPDA

En este trabajo se ha desarrollado una aplicación en Java titulada como **CPDA – *Comparator of precisión in data anonymization,*** dicha aplicación puede ser ejecutada directamente sobre un ordenador con sistema operativo Windows, para lo cual será necesario descargar el fichero "CPDA-App.jar" que se encuentra dentro del repositorio de código en *Bitbucket* mencionado en el **Capítulo V** bajo la carpeta "*exec*".

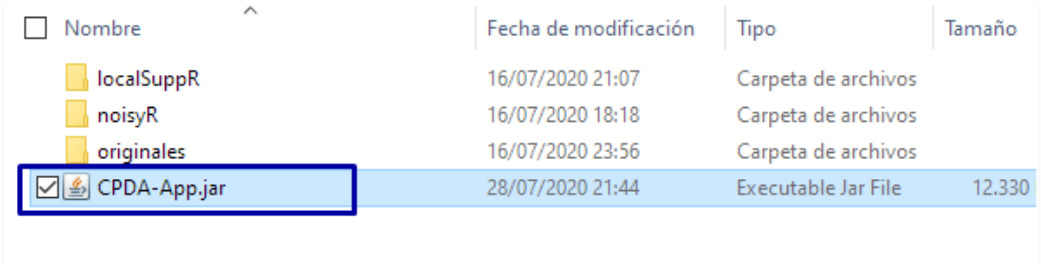


Figura 37 – Manual de usuario: Ejecutable de la herramienta CPDA

Una vez que se descargue el fichero en el ordenador tal y como se visualiza en la Figura 37, se podrá ejecutar el mismo y así empezar a hacer uso de la herramienta. Al iniciar la aplicación CPDA, lo primero que se observa es la pantalla inicial de carga que se muestra en la Figura 38, pasando directamente a la ventana de carga de datos (*Data tab*) que se observa en la Figura 39, donde el usuario deberá cargar un conjunto de datos para poder visualizar información del mismo en la ventana actual, y además para poder habilitar la sección siguiente de minería de datos (*Mining tab*).

Figura 38 – Manual de usuario: Pantalla de bienvenida de CPDA

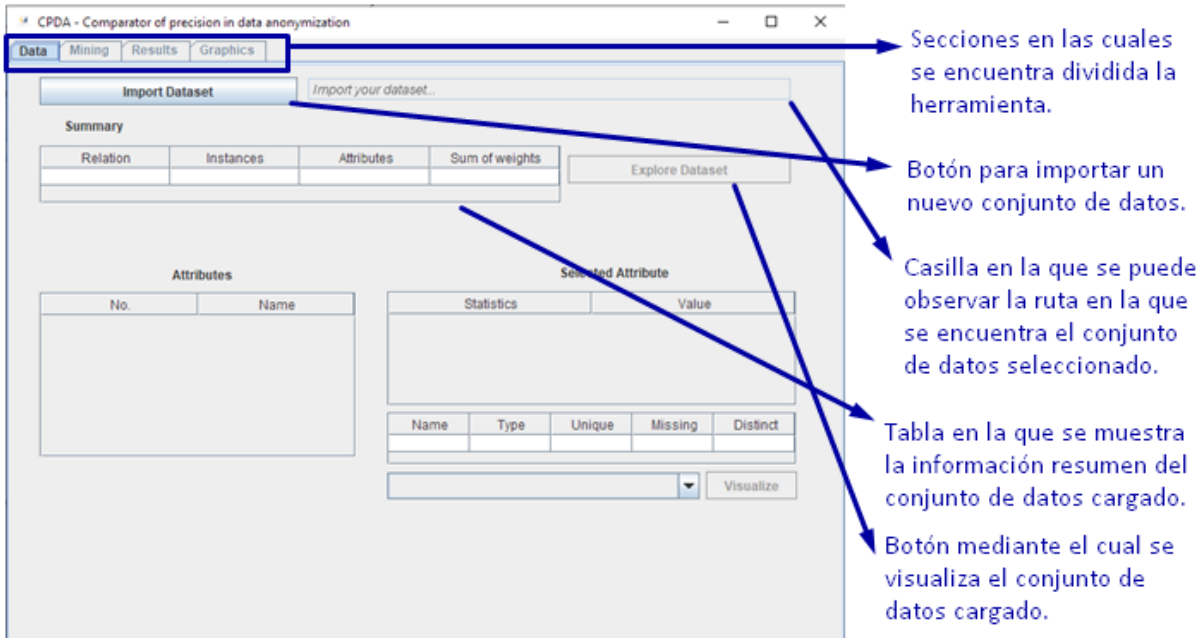


Figura 39 – Manual de usuario: Pantalla de carga de datos antes de importar un conjunto de datos a la herramienta CPDA

III.1. Sección de datos (*Data tab*)

A continuación se explica paso a paso las acciones que puede realizar un usuario dentro de la sección de datos (*Data tab*).

1. El usuario deberá hacer clic sobre el botón “Import Dataset”.

2. La herramienta abrirá una ventana nueva en la cual el usuario podrá ver su sistema de archivos tal y como se observa en la Figura 40, la ventana usará como ruta por defecto la de “*Documents*” del usuario. A continuación el usuario deberá seleccionar el conjunto de datos que desea cargar a la herramienta, con la restricción de que sólo podrá cargar conjuntos de datos con la extensión *.arff* y *.csv*.

3. Una vez que se ha seleccionado un conjunto de datos, la ventana se actualizará automáticamente con la información asociada al mismo, de forma que se poblarán los datos de las tablas “*Summary*” y “*Attributes*”, así como también se actualizará el campo de la ruta del fichero a la del conjunto de datos seleccionado, y se asignará por defecto el último atributo como el atributo “clase”/objetivo para efectos de minería de datos (propósitos de predicción) tal y como se observa en la Figura 41.

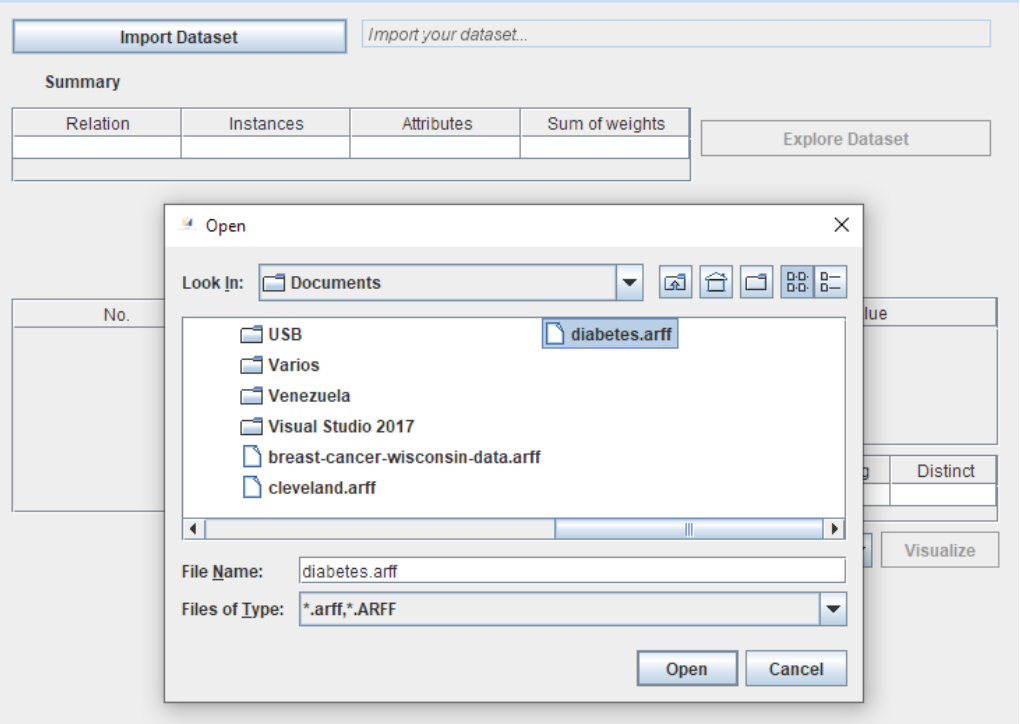


Figura 40 - Manual de usuario: Carga de un conjunto de datos

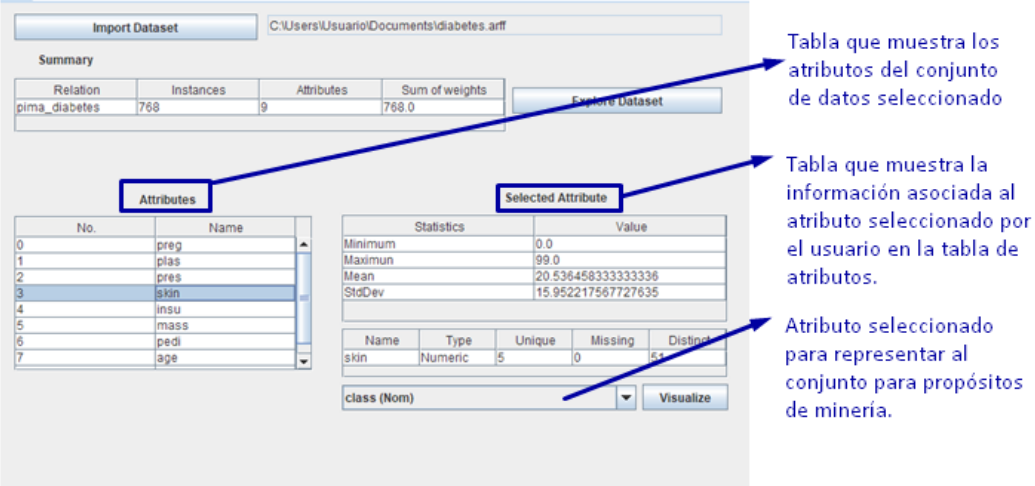


Figura 41 - Manual de usuario: Ventana de datos actualizada tras la carga de un conjunto de datos

4. Una vez que se ha cargado un conjunto de datos y se visualiza la información del mismo, el usuario bien puede explorar en detalle los valores asociados a cada uno de los atributos del conjunto de datos, para ello tendrá que seleccionar uno de los atributos encontrados en la tabla "*Attributes*" y de forma automática se mostrarán los datos asociados al atributo seleccionado en la tabla "*Selected Attribute*".

5. El usuario también podrá hacer uso de una funcionalidad de la herramienta que se habilita una vez cargado un conjunto de datos, la cual se accede mediante el botón de “Explore Dataset”, una vez que el usuario hace *clic* en el mismo se abrirá una nueva ventana en donde podrá visualizar el conjunto de datos que ha cargado, tal y como se muestra en el ejemplo de la Figura 42.

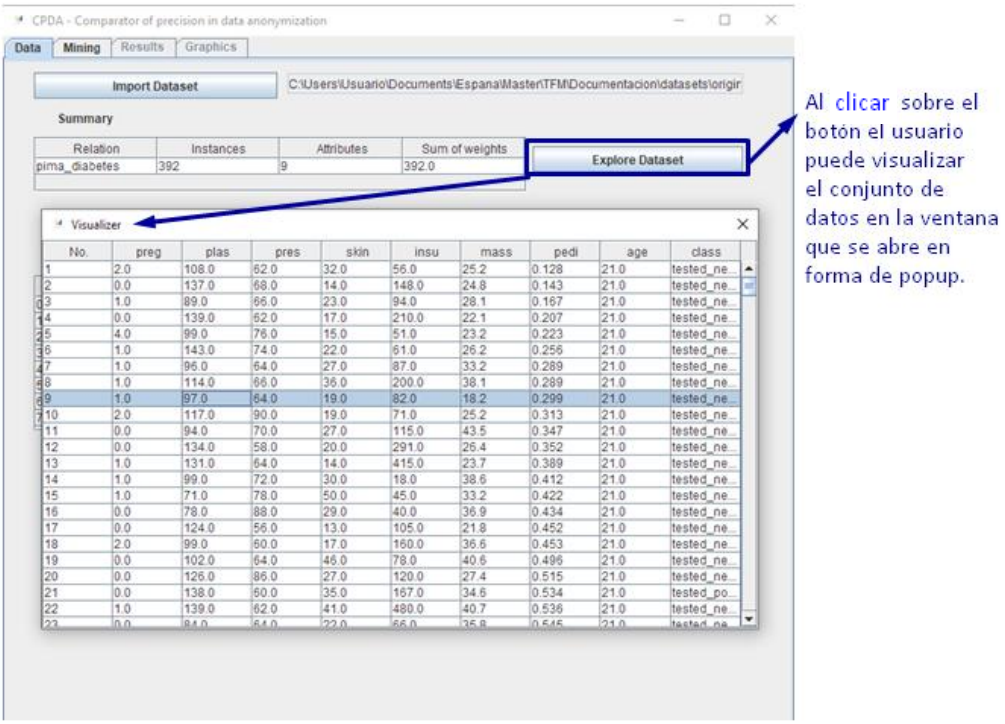


Figura 42 - Manual de usuario: Ventana para la visualización del conjunto de datos cargado

6. Por último, dentro de la pestaña de “*Data*”, el usuario podrá visualizar de forma gráfica la distribución de los valores de todos los atributos contenidos en el conjunto de datos en relación al atributo seleccionado como “clase”.

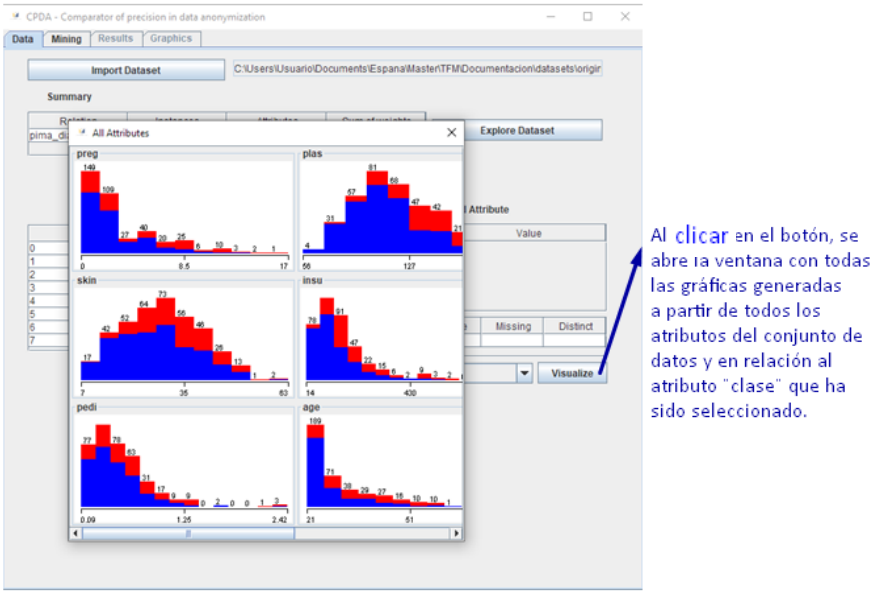


Figura 43 - Manual de usuario: Visualización gráfica de los atributos y su distribución

III.2. Sección de minería (*Mining tab*)

Una vez que el usuario ha finalizado la exploración del conjunto de datos dentro de la pestaña "*Data*" podrá cambiar a la pestaña de "*Mining*" la cual estará ahora habilitada y dentro de la cual podrá realizar una serie de acciones las cuales se explican a continuación paso por paso.

1. Como se puede observar en la Figura 44 el usuario podrá visualizar 2 secciones principales dentro de la ventana, donde la primera sección hace referencia a la clasificación de datos y a la configuración de la misma para los propósitos de minería de datos, y luego se observa una sección asociada a la anonimización.

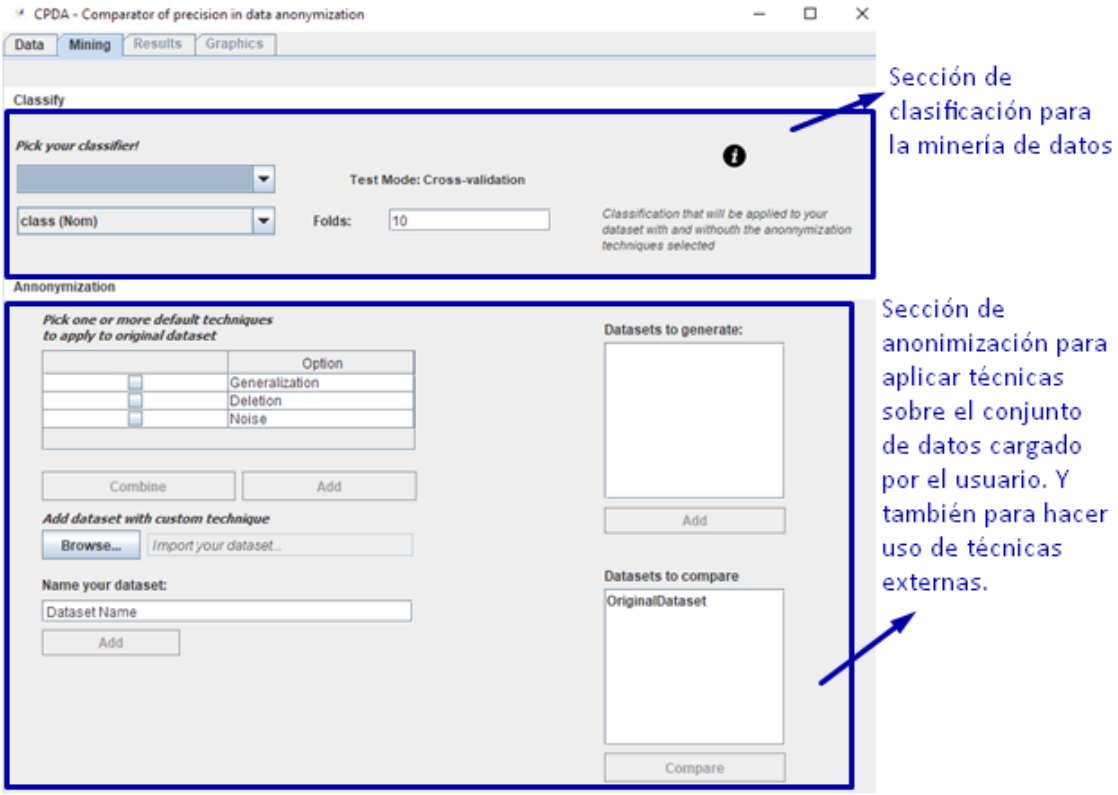


Figura 44 - Manual de usuario: Pantalla inicial de minería de datos de la herramienta

2. Dentro de la sección de clasificación el usuario podrá seleccionar un tipo de clasificador a utilizar dentro de las 7 opciones habilitadas las cuales se observan en la Figura 45, también podrá indicar el atributo nominal que desea establecer como atributo "clase"/objetivo y finalmente podrá indicar el número de "*folds*" a utilizar en la evaluación cruzada o también conocida como "*cross-validation*" que será el tipo de evaluación usada por defecto dentro de la herramienta.

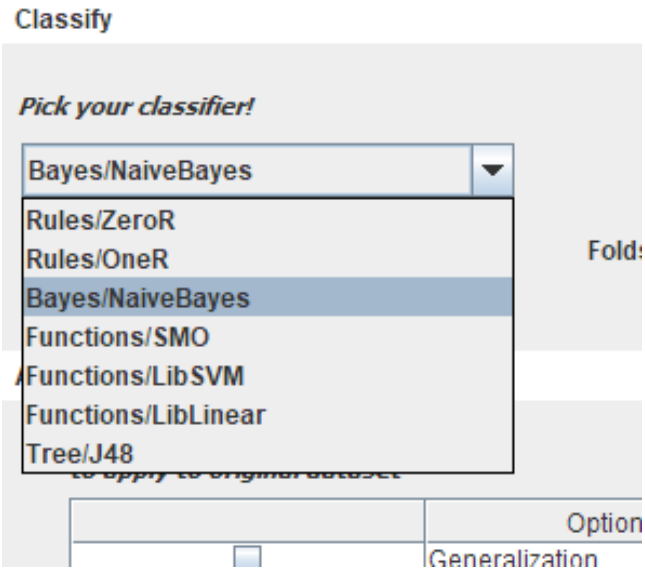


Figura 45 - Manual de usuario: Clasificadores disponibles en la herramienta

3. Dentro de la sección de anonimización el usuario podrá observar una tabla con 3 técnicas de anonimización a elegir, dicha tabla permite la selección múltiple de las mismas con lo cual el usuario puede añadir simultáneamente 2 o más técnicas. Una vez que se selecciona una de las técnicas, el botón de "*Add*" se habilita, permitiendo al usuario añadir la técnica seleccionada a la lista de "*Datasets to generate*". Si el usuario selecciona simultáneamente dos o más técnicas, el botón de "*Combine*" se habilitará.

4. Para combinar técnicas el usuario deberá hacer *clic* en el botón "*Combine*", el cual hará que la herramienta abra una ventana en la que se muestran las técnicas seleccionadas y en donde el usuario podrá indicar el orden en el que desea que se apliquen las técnicas sobre el conjunto de datos y también nombrar la combinación creada para añadirla a la lista de "*Datasets to generate*", dicha ventana que permite combinar técnicas puede visualizarse en la Figura 46.

5. Una vez que el usuario ha añadido las técnicas de anonimización que desea aplicar, las podrá visualizar en la lista "*Datasets to generate*". Debajo de dicha lista se encuentra un botón "*Add*" el cual se habilitará siempre que exista al menos una técnica de anonimización añadida a la lista. Si el botón se encuentra habilitado y el usuario hace *clic* en el mismo (ver Figura 47), la herramienta procederá a aplicar las técnicas de anonimización seleccionadas sobre el conjunto de datos que ha cargado el usuario y generará un nuevo conjunto de datos para cada técnica de anonimización seleccionada.

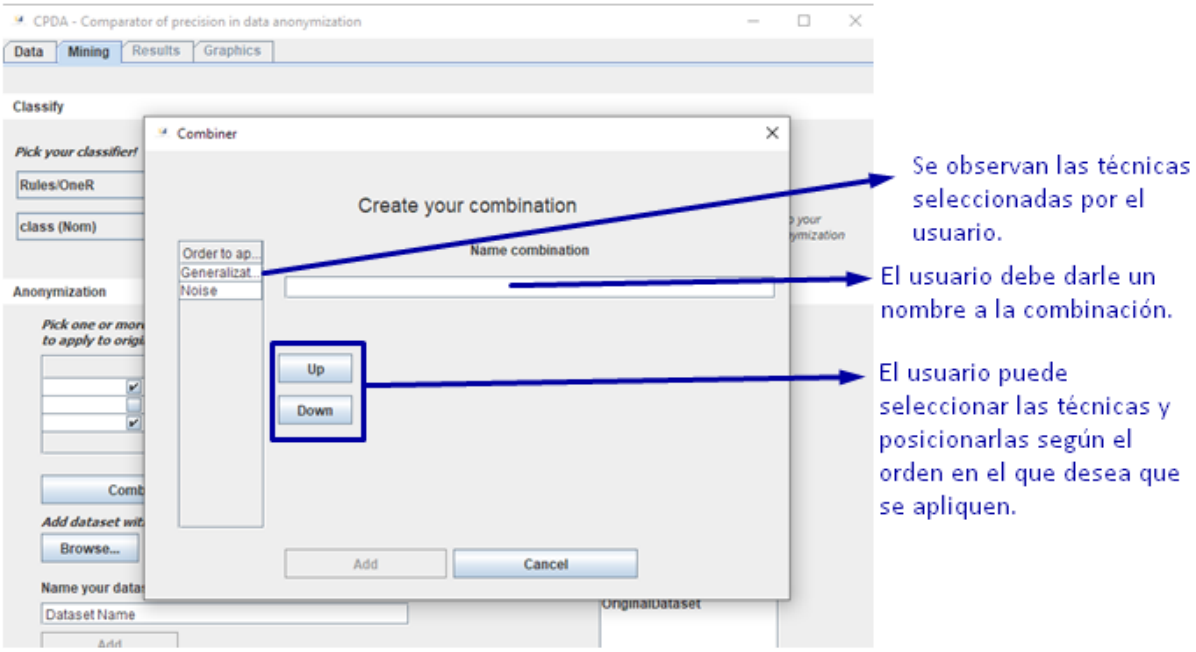


Figura 46 - Manual de usuario: Ventana para combinar técnicas de anonimización

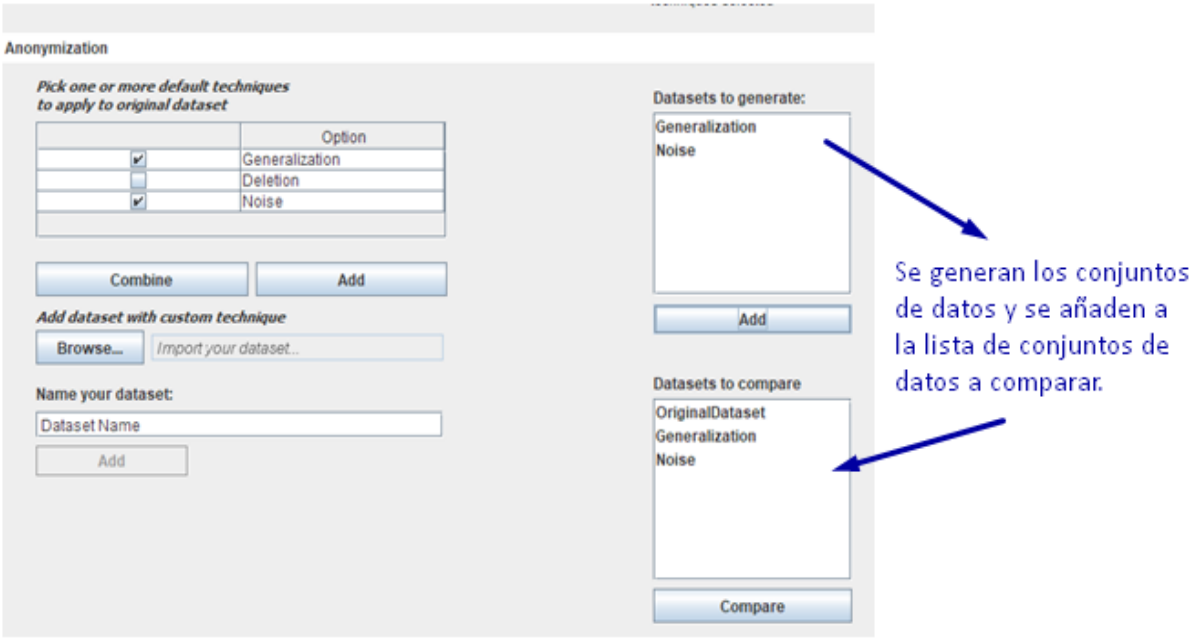


Figura 47 - Manual de usuario: Generación de conjuntos de datos para comparar a partir de las técnicas de anonimización seleccionadas

6. Además de las técnicas de anonimización que ofrece la herramienta, el usuario puede añadir sus propias técnicas a la lista de "*Datasets to compare*", para esto el usuario podrá hacer uso de alguna herramienta externa, como por ejemplo R, en la cual aplique una técnica de anonimización sobre el conjunto de datos con

el que está trabajando en la herramienta. Una vez que se ha aplicado la técnica de anonimización en R, el usuario deberá guardar el conjunto de datos resultante con la extensión .arff o .csv y proceder así a subir a la herramienta de CPDA dicho conjunto de datos anonimizados tal y como se muestra en la Figura 48.

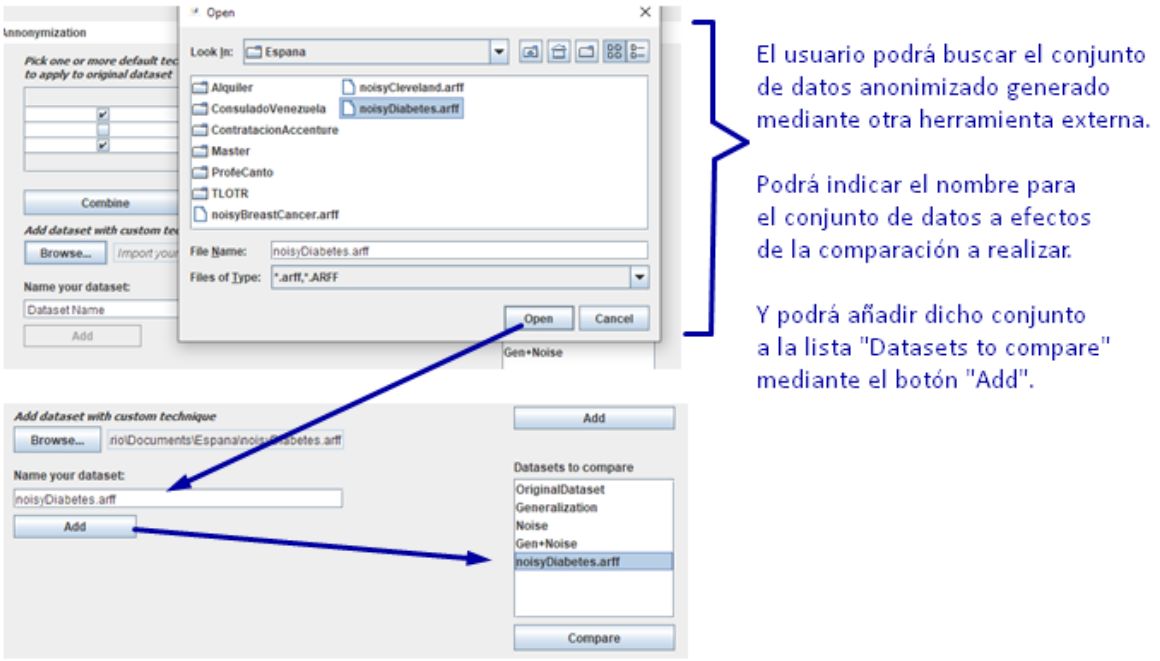


Figura 48 - Manual de usuario: Adición de técnicas de anonimización generadas mediante herramientas externas

III.3. Sección de visualización de resultados (*Results tab*)

Una vez que el usuario ha añadido todas las técnicas de anonimización deseadas a la lista de "*Datasets to compare*" podrá hacer *clic* sobre el botón de "*Compare*" el cual hará que la herramienta aplique la técnica de clasificación seleccionada sobre cada conjunto de datos y en consecuencia se habilitarán las pestañas de "*Results*" y "*Graphics*" donde el usuario podrá visualizar los resultados obtenidos para cada técnica de anonimización seleccionada.

Dentro de la pestaña de "*Results*" el usuario podrá:

1. Visualizar la técnica de clasificación aplicada y el número de *folds* utilizados en la evaluación.

2. Observar en una primera tabla el detalle de los resultados obtenidos para cada conjunto de datos, donde el usuario podrá comparar valores como el número de instancias correctamente clasificadas, las incorrectamente clasificadas, la media de error absoluta, entre otros.

3. Observar en una segunda tabla información más detallada en relación a la clasificación aplicada y las métricas que se obtienen de cada conjunto de datos como por ejemplo el *TP Rate*, *FP Rate*, *Precision*, *Recall*, *F-Measure*, entre otros.

En la Figura 49 se puede observar un ejemplo de los puntos mencionados anteriormente, donde se comparan 5 técnicas de anonimización mediante la técnica de clasificación de *Naive Bayes* con *10-fold cross validation*.

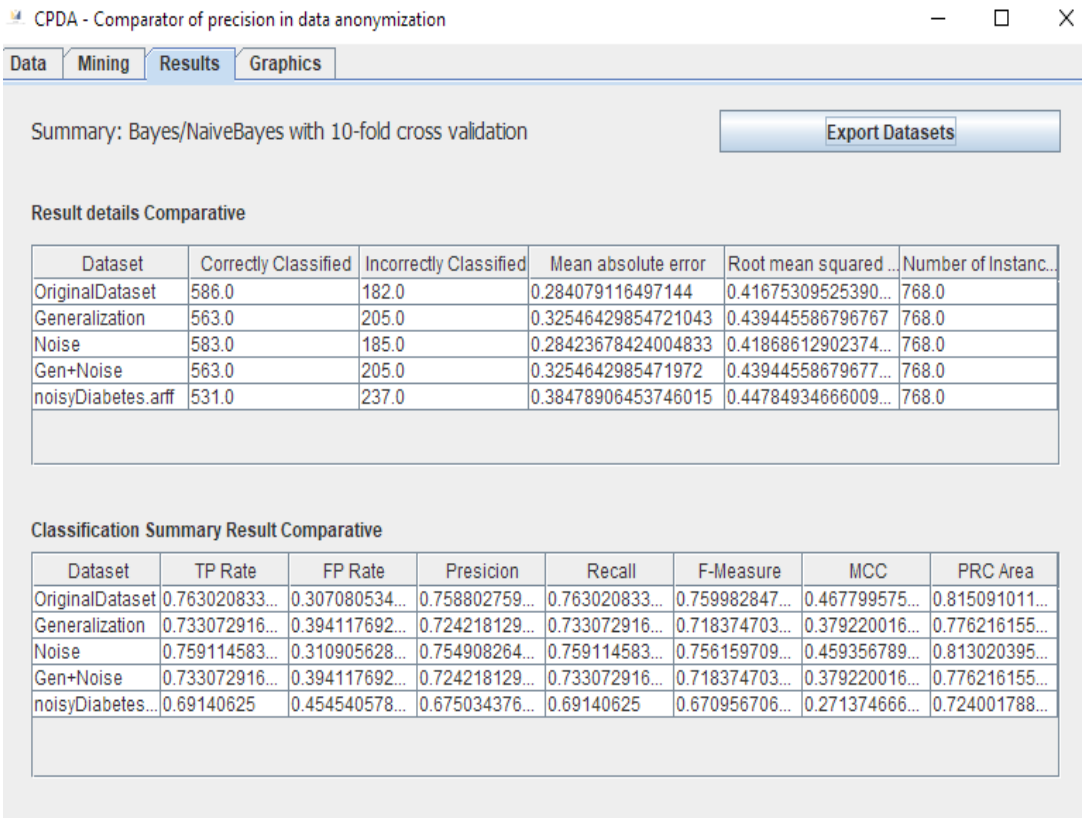


Figura 49 - Manual de usuario: Presentación de los resultados obtenidos mediante tablas comparativas en la herramienta.

4. Además de la posibilidad de observar las tablas comparativas, el usuario tendrá la posibilidad de exportar los conjuntos de datos que se han comparado, para ello deberá hacer clic en el botón "*Export Datasets*", con el cual se abrirá una nueva ventana mostrando en una lista los conjuntos de datos generados mediante la aplicación de las técnicas de anonimización seleccionadas.

5. Una vez que el usuario selecciona el conjunto de datos a exportar y presiona en el botón de "*Export*" se abrirá el sistema de archivos en una ruta creada por defecto por la aplicación (*~Documents/CPDA*), en donde se crea una carpeta con el nombre de la misma y en donde el usuario puede o bien guardar el

conjunto de datos directamente en la ruta por defecto o guardar el conjunto de datos en una ruta alterna que considere apropiada. Dicho proceso de exportar se puede visualizar en el ejemplo mostrado por la Figura 50.

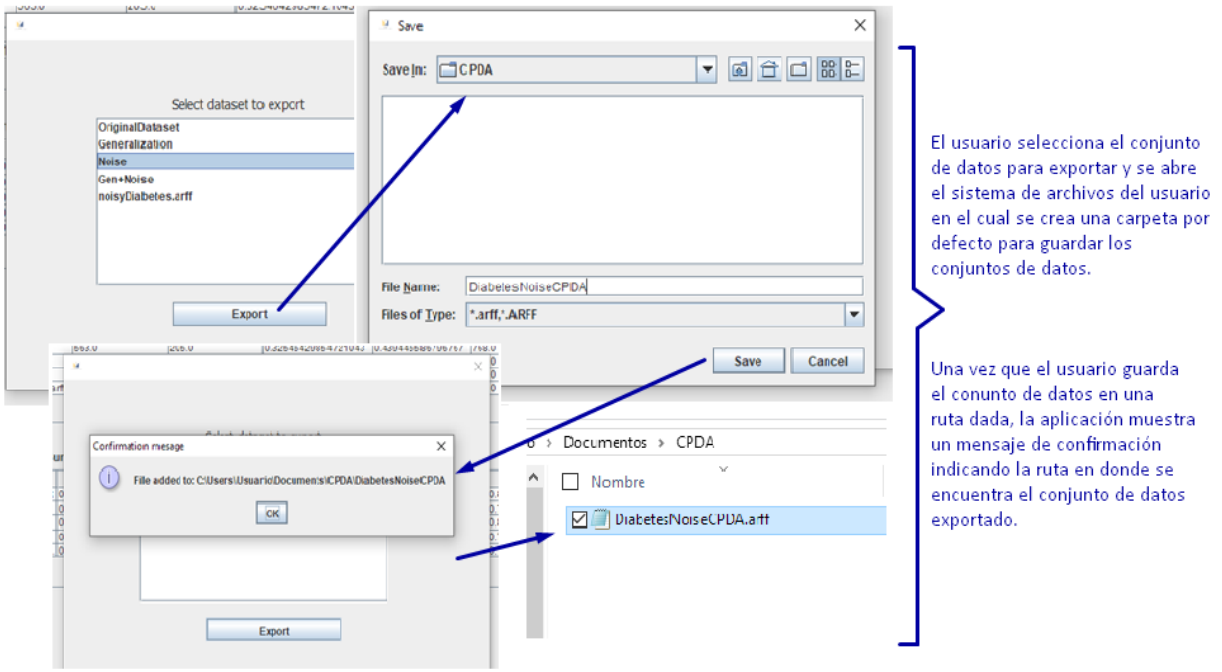


Figura 50 - Manual de usuario: Proceso para exportar los conjuntos de datos anonimizados

III.4. Sección de visualización gráfica de resultados (*Graphics tab*)

En la pestaña de “Graphics” se puede visualizar mediante un gráfico de barras algunas de las métricas observadas en la pestaña de “*Results*”, se eligieron específicamente 4 de las métricas más relevantes a tomar en cuenta al momento de aplicar minería de datos.

En esta pestaña el usuario deberá elegir las técnicas que desea comparar, las cuales conformarán el eje X de la gráfica y también deberá seleccionar la métrica sobre la cual desea hacer la comparación (eje Y de la gráfica), como en la Figura 51 donde se selecciona la métrica “*Precision*” y se seleccionan todas las técnica de anonimización, finalmente para visualizar la gráfica el usuario deberá *clicar* en el botón “*Show charts*”.

Con intención de complementar el manual de usuario presentado anteriormente, se añade a continuación el flujo de trabajo de la herramienta, el cual puede observarse en la Figura 52, donde se puede observar el proceso por el que atraviesa un conjunto de datos al ser anonimizado mediante las técnicas que ofrece la herramienta o mediante una herramienta externa, y la forma en la que se realiza la minería de datos sobre dicho conjunto.

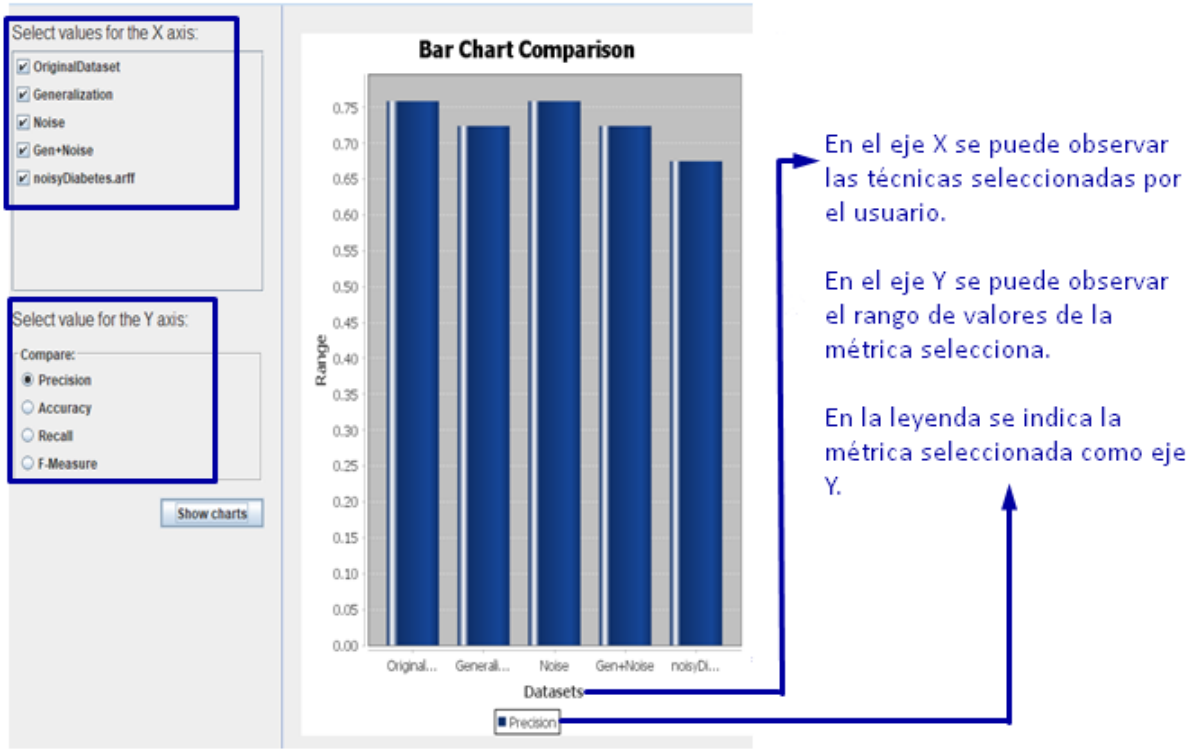


Figura 51 - Manual de usuario: Comparativa de resultados mediante gráfico de barras

Como es posible observar en la Figura 52 y en el manual de usuario, existen dos formas de comparar técnicas dentro de la herramienta, una de ellas es mediante las técnicas disponibles en la herramienta de CPDA, y luego está la posibilidad de usar técnicas de anonimización externas. Para lo cual el usuario debe de aplicar la técnica de anonimización deseada a través de la herramienta externa, generar el fichero con el conjunto de datos anonimizado mediante dicha herramienta externa e importar el mismo dentro de la herramienta.

El usuario podrá comparar tantas técnicas como considere, ya sea usando las técnicas de anonimización de la herramienta o usando técnicas de anonimización disponibles en herramientas externas. Y a continuación se muestran los resultados obtenidos para el caso de estudio visto en el trabajo.

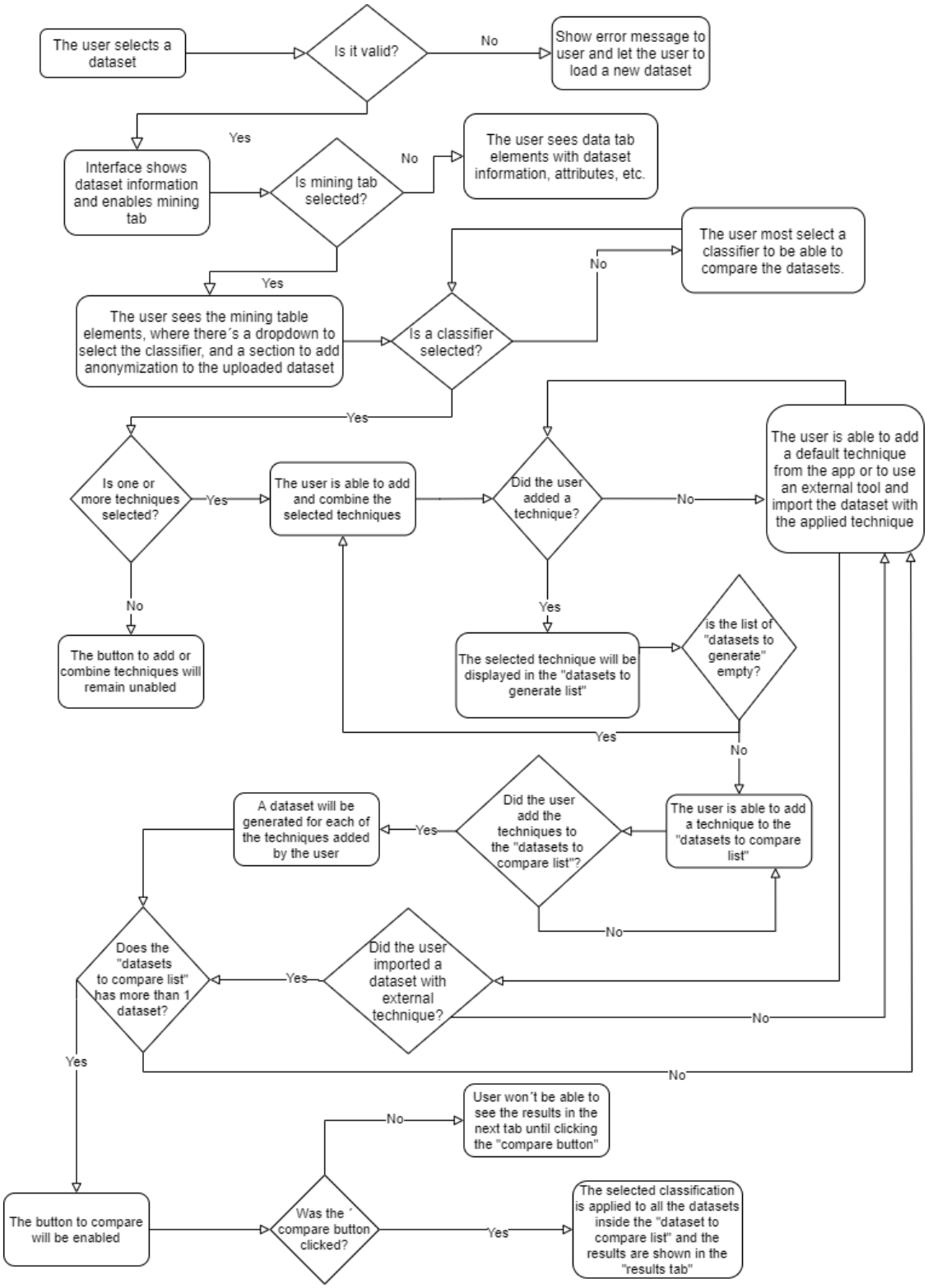


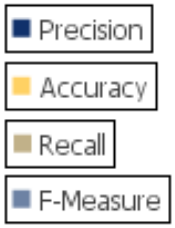
Figura 52 - Flujo de Trabajo de la herramienta CPDA en la sesión de minería de datos

ANEXO IV. Resultados y extractos de código

En este anexo se presenta mediante el uso de capturas de pantalla los resultados obtenidos en la herramienta CPDA. Y también se muestran extractos del código utilizado para aplicar técnicas de anonimización, tanto dentro de la herramienta de CPDA como mediante de la herramienta externa R.

IV.1. Resultados obtenidos en la herramienta CPDA para el caso de ejemplo estudiado en el actual trabajo

A continuación se presenta de forma gráfica mediante capturas realizadas sobre la aplicación CPDA, los resultados obtenidos para cada conjunto de datos, con los 7 clasificadores escogidos en el trabajo y las 5 técnicas aplicadas a cada uno de los conjuntos de datos.
De esta forma es posible apreciar la obtención de los resultados que se muestran resumidos en el **Capítulo V**. Es importante destacar que para las gráficas que se muestran como parte de los resultados obtenidos con cada tipo de clasificador se tiene como leyenda la siguiente:



Leyenda de las gráficas de barras de CPDA.

IV.1.1. Conjunto de datos "*Pima Diabetes*"

IV.1.1.1. Clasificador: ZeroR

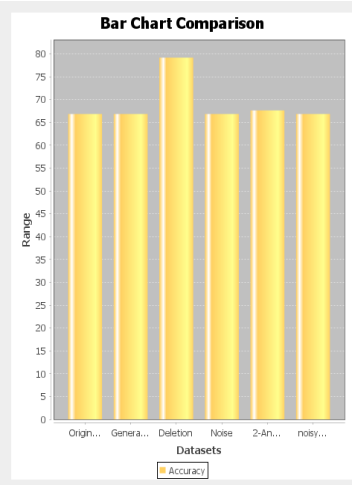
Summary: Rules/ZeroR with 10-fold cross validation — Export Datasets

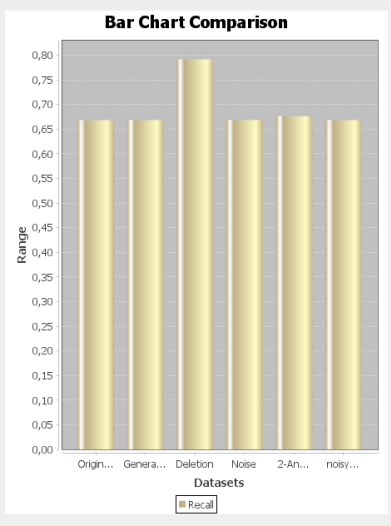
Result details Comparative

Dataset	Correctly Classified	Incorrectly Classified	Mean absolute error	Root mean squared ...	Number of Instanc...
OriginalDataset	262.0	130.0	0.4436284372904081	0.47080360080080...	392.0
Generalization	262.0	130.0	0.4436284372904081	0.47080360080080...	392.0
Deletion	19.0	5.0	0.348731884057971	0.41448342370246...	24.0
Noise	262.0	130.0	0.4436284372904081	0.47080360080080...	392.0
2-Anon-Diabetes	261.0	125.0	0.4383045563193152	0.46796302204148...	392.0
noisyDiabetes	262.0	130.0	0.4436284372904081	0.47080360080080...	392.0

Classification Summary Result Comparative

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	PRC Area
OriginalDataset	0.668367346...	0.668367346...	NaN	0.668367346...	NaN	NaN	0.555352804...
Generalization	0.668367346...	0.668367346...	NaN	0.668367346...	NaN	NaN	0.555352804...
Deletion	0.791666666...	0.791666666...	NaN	0.791666666...	NaN	NaN	0.603963744...
Noise	0.668367346...	0.668367346...	NaN	0.668367346...	NaN	NaN	0.555352804...
2-Anon-Diabet...	0.676165803...	0.676165803...	NaN	0.676165803...	NaN	NaN	0.548945419...
noisyDiabetes	0.668367346...	0.668367346...	NaN	0.668367346...	NaN	NaN	0.555352804...





IV.1.1.2. Clasificador: OneR

Summary: Rules/OneR with 10-fold cross validation

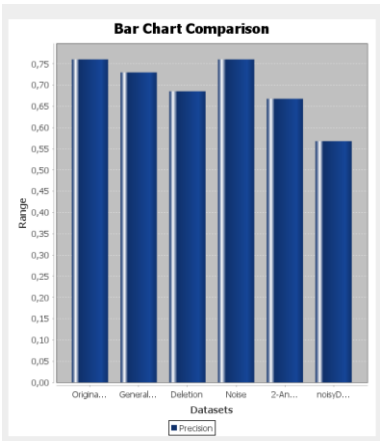
Export Datasets

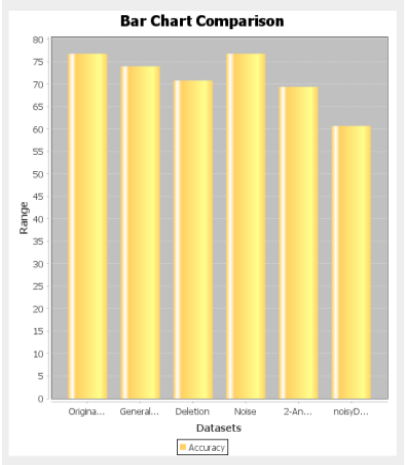
Result details Comparative

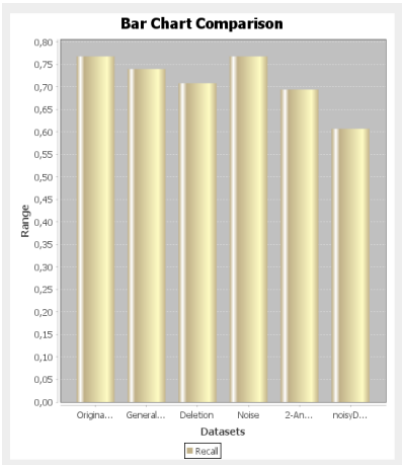
Dataset	Correctly Classified	Incorrectly Classified	Mean absolute error	Root mean squared ...	Number of Instanc...
OriginalDataset	301.0	91.0	0.23214285714285715	0.48181205582971...	392.0
Generalization	290.0	102.0	0.2602040816326531	0.51010203061020...	392.0
Deletion	17.0	7.0	0.2916666666666667	0.54006172486732...	24.0
Noise	301.0	91.0	0.23214285714285715	0.48181205582971...	392.0
2-Anon-Diabetes	268.0	118.0	0.30569948186528495	0.55290096931121...	392.0
noisyDiabetes	238.0	154.0	0.39285714285714285	0.62678317052800...	392.0

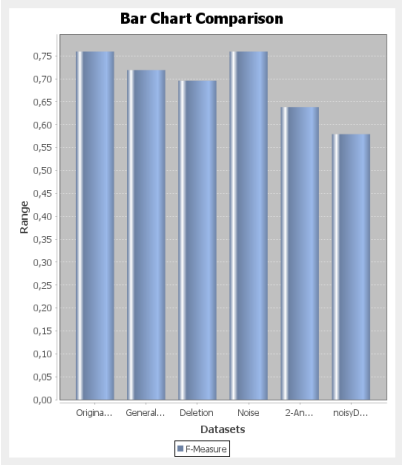
Classification Summary Result Comparative

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	PRC Area
OriginalDataset	0.767857142...	0.347716215...	0.760404059...	0.767857142...	0.759372333...	0.452282860...	0.695161562...
Generalization	0.739795918...	0.427523458...	0.729808483...	0.739795918...	0.719015731...	0.368250954...	0.655969092...
Deletion	0.708333333...	0.666228070...	0.685416666...	0.708333333...	0.695868945...	0.045883146...	0.677430555...
Noise	0.767857142...	0.347716215...	0.760404059...	0.767857142...	0.759372333...	0.452282860...	0.695161562...
2-Anon-Diabet...	0.694300518...	0.567434617...	0.667777295...	0.694300518...	0.638141958...	0.192672037...	0.589888287...
noisyDiabetes	0.607142857...	0.586356010...	0.567893672...	0.607142857...	0.579151980...	0.024061525...	0.561430909...









IV.1.1.3. Clasificador: Naive Bayes

Summary: Bayes/NaiveBayes with 10-fold cross validation

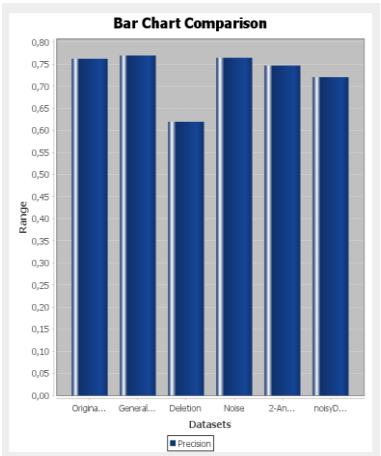
Export Datasets

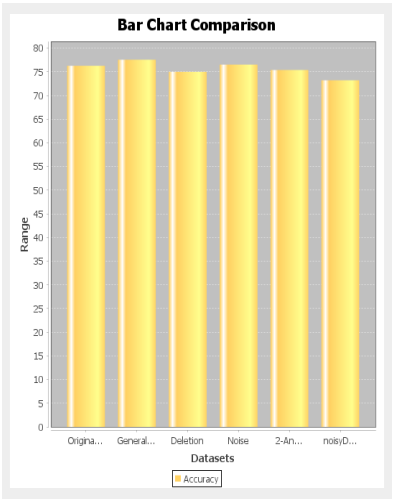
Result details Comparative

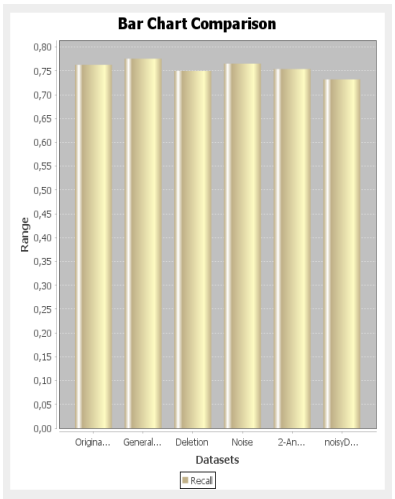
Dataset	Correctly Classified	Incorrectly Classified	Mean absolute error	Root mean squared er...	Number of Instances
OriginalDataset	299.0	93.0	0.25775426145865143	0.4316392196638404	392.0
Generalization	304.0	88.0	0.2584646209212247	0.42156851871911366	392.0
Deletion	18.0	6.0	0.24118526661945741 4	0.46846328844508767	24.0
Noise	300.0	92.0	0.2580621256533314	0.43336398731623893	392.0
2-Anon-Diabetes	291.0	95.0	0.2896278442344698	0.42951907878670154	392.0
noisyDiabetes	287.0	105.0	0.35984941066194404	0.4348967789072139	392.0

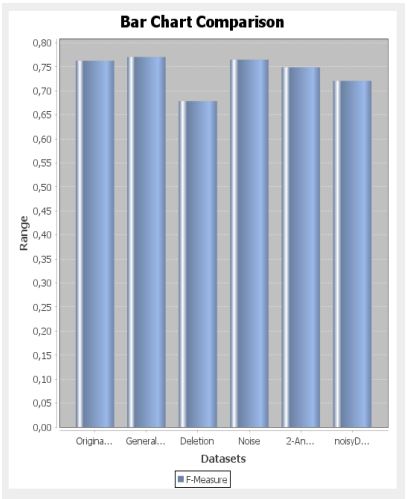
Classification Summary Result Comparative

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	PRC Area
OriginalDataset	0.7627551020...	0.29986608266...	0.7623013216...	0.76275510204...	0.7625230478...	0.46379543641...	0.8292546466...
Generalization	0.7755102040...	0.32066581183...	0.7695101251...	0.77551020408...	0.7702831132...	0.47661122857...	0.8249798441...
Deletion	0.75	0.80263157894...	0.6195652173...	0.75	0.6785714285...	-0.1069656374...	0.6615725260...
Noise	0.7653061224...	0.29860030917...	0.7644207540...	0.76530612244...	0.7648426423...	0.46855285019...	0.8262925172...
2-Anon-Diabetes	0.7538860103...	0.35131129771...	0.7468538122...	0.75388601036...	0.7487888069...	0.41965024024...	0.8034872532...
noisyDiabetes	0.7321428571...	0.40031666806...	0.7205898504...	0.73214285714...	0.7206077916...	0.36202295254...	0.7459827137...









IV.1.1.4. Clasificador: SMO

Summary: Functions/SMO with 10-fold cross validation

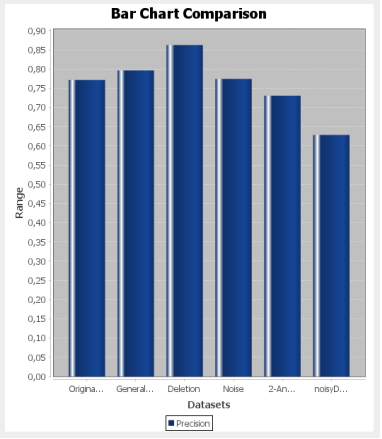
Export Datasets

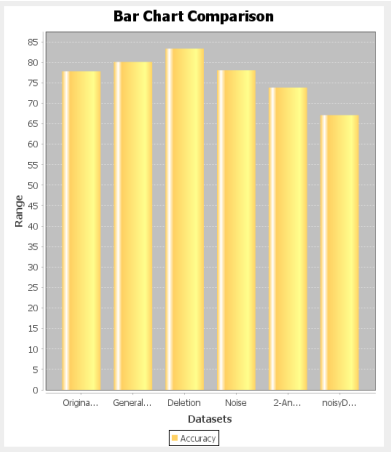
Result details Comparative

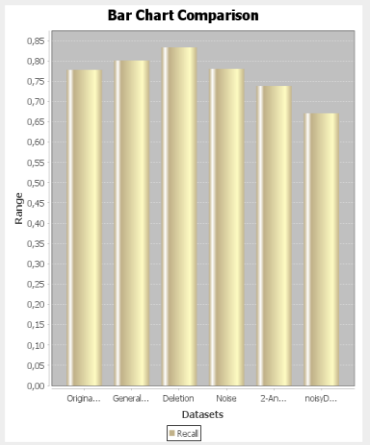
Dataset	Correctly Classified	Incorrectly Classified	Mean absolute error	Root mean squared er...	Number of Instances
OriginalDataset	305.0	87.0	0.22193877551020408	0.4711037842240328	392.0
Generalization	314.0	78.0	0.1989795918367347	0.44607128559988557	392.0
Deletion	20.0	4.0	0.16666666666666666	0.408248290463863	24.0
Noise	306.0	86.0	0.2193877551020408	0.46838846602157147	392.0
2-Anon-Diabetes	285.0	101.0	0.261658031088082 9	0.5115252008338229	392.0
noisyDiabetes	263.0	129.0	0.32908163265306123	0.5736563715788933	392.0

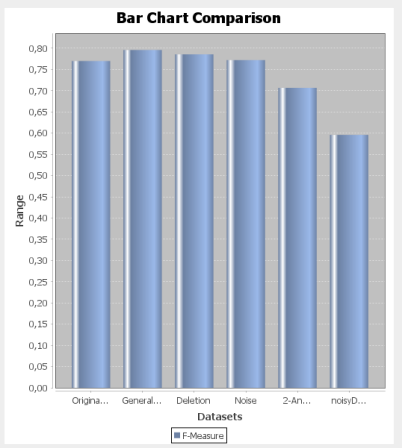
Classification Summary Result Comparative

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	PRC Area
OriginalDataset	0.7780612244...	0.33490209354...	0.7715646327...	0.77806122448...	0.7699493737...	0.47706201967...	0.7053679006...
Generalization	0.8010204081...	0.29250602178...	0.7963508564...	0.80102040816...	0.7958326187...	0.53586392166...	0.7330561084...
Deletion	0.8333333333...	0.63333333333...	0.8623188405...	0.83333333333...	0.7857142857...	0.40646942234...	0.7303743961...
Noise	0.7806122448...	0.33363632005...	0.7743849594...	0.78061224489...	0.7722444494...	0.48274193592...	0.7074222641...
2-Anon-Diabetes	0.7383419689...	0.47130748615...	0.7306970978...	0.73834196891...	0.7068620464...	0.34487219512...	0.6368899149...
noisyDiabetes	0.6709183673...	0.60509335266...	0.6284692145...	0.67091836734...	0.5958880306...	0.11483620676...	0.5742084111...









IV.1.1.5. Clasificador: LibSVM

Summary: Functions/LibSVM with 10-fold cross validation

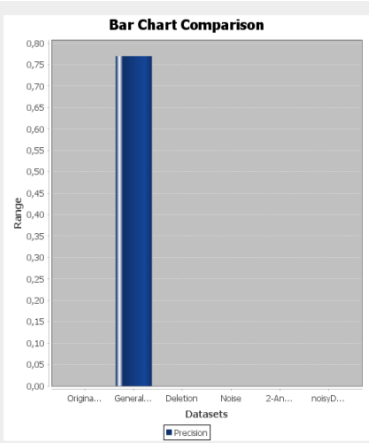
Export Datasets

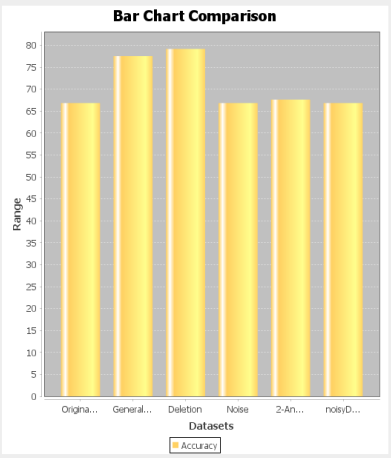
Result details Comparative

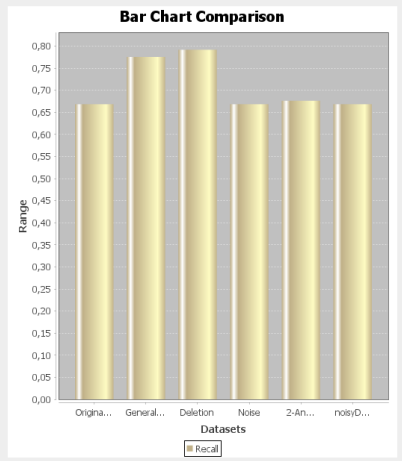
Dataset	Correctly Classified	Incorrectly Classified	Mean absolute error	Root mean squared er...	Number of Instances
OriginalDataset	262.0	130.0	0.33163265306122447	0.5758755534498964	392.0
Generalization	304.0	88.0	0.22448979591836735	0.47380354147934284	392.0
Deletion	19.0	5.0	0.20833333333333334	0.45643546458763845	24.0
Noise	262.0	130.0	0.33163265306122447	0.5758755534498964	392.0
2-Anon-Diabetes	261.0	125.0	0.3238341968911917	0.5690643170074817	392.0
noisyDiabetes	262.0	130.0	0.33163265306122447	0.5758755534498964	392.0

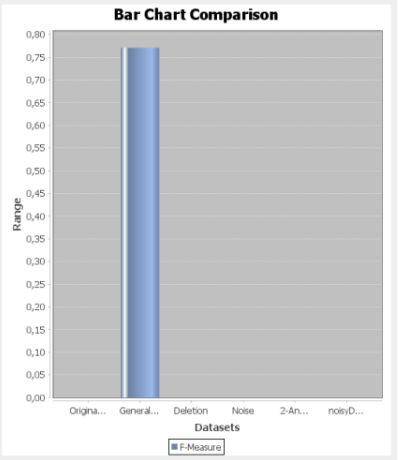
Classification Summary Result Comparative

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	PRC Area
OriginalDataset	0.6683673469...	0.66836734693...	NaN	0.66836734693...	NaN	NaN	0.5566951270...
Generalization	0.7755102040...	0.31679029803...	0.7698524052...	0.77551020408...	0.7708891003...	0.47805942134...	0.7090281393...
Deletion	0.7916666666...	0.79166666666...	NaN	0.79166666666...	NaN	NaN	0.6701388888...
Noise	0.6683673469...	0.66836734693...	NaN	0.66836734693...	NaN	NaN	0.5566951270...
2-Anon-Diabetes	0.6761658031...	0.67616580310...	NaN	0.67616580310...	NaN	NaN	0.5534656867...
noisyDiabetes	0.6683673469...	0.66836734693...	NaN	0.66836734693...	NaN	NaN	0.5566951270...









IV.1.1.6. Clasificador: LibLinear

Summary: Functions/LibLinear with 10-fold cross validation

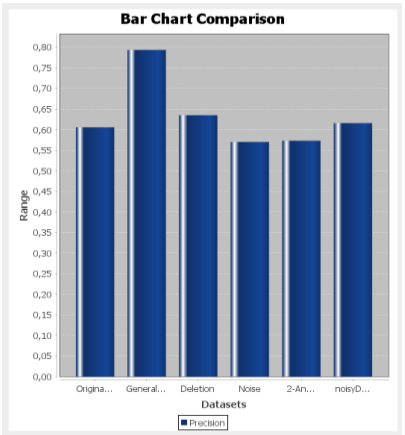
Export Datasets

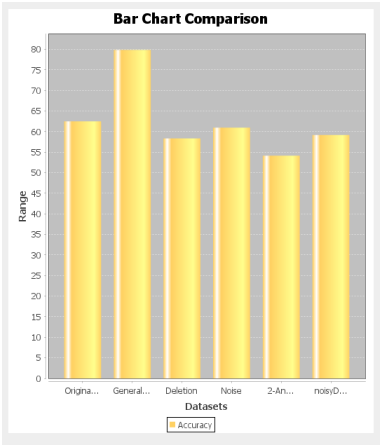
Result details Comparative

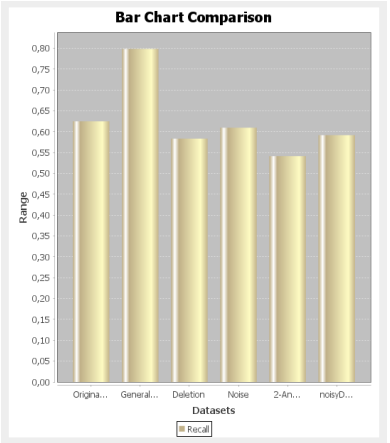
Dataset	Correctly Classified	Incorrectly Classified	Mean absolute error	Root mean squared er...	Number of Instances
OriginalDataset	245.0	147.0	0.375	0.6123724356957945	392.0
Generalization	313.0	79.0	0.20153061224489796	0.4489216103563048	392.0
Deletion	14.0	10.0	0.4166666666666667	0.6454972243679028	24.0
Noise	239.0	153.0	0.3903061224489796	0.6247448458762822	392.0
2-Anon-Diabetes	209.0	177.0	0.4585492227979275	0.6771626265513533	392.0
noisyDiabetes	232.0	160.0	0.40816326530612246	0.638876564999399	392.0

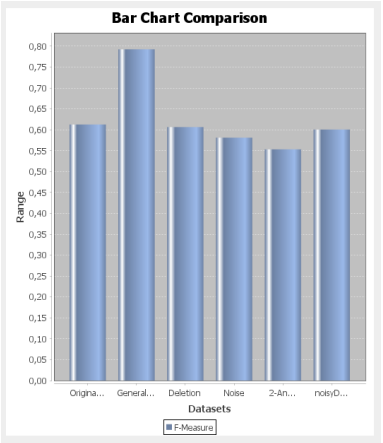
Classification Summary Result Comparative

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	PRC Area
OriginalDataset	0.625	0.52323840281...	0.6056119192...	0.625	0.6124603724...	0.10885257392...	0.5818771667...
Generalization	0.7984693877...	0.30152282286...	0.7936251155...	0.79846938775...	0.7923382584...	0.52831922781...	0.7287124394...
Deletion	0.5833333333...	0.69912280701...	0.6351540616...	0.58333333333...	0.6064814814...	-0.1034564094...	0.6528069561...
Noise	0.6096938775...	0.58509023691...	0.5700418192...	0.60969387755...	0.5810624540...	0.02860875726...	0.5623299977...
2-Anon-Diabetes	0.5414507772...	0.51558104540...	0.5730360745...	0.54145077720...	0.5531296669...	0.02457270548...	0.5587392706...
noisyDiabetes	0.5918367346...	0.45443215454...	0.6158350193...	0.59183673469...	0.6005550063...	0.13161929113...	0.5909910135...









IV.1.1.7. Clasificador: J48

Summary: Tree/J48 with 10-fold cross validation

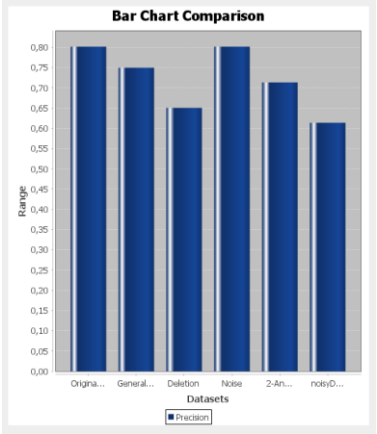
Export Datasets

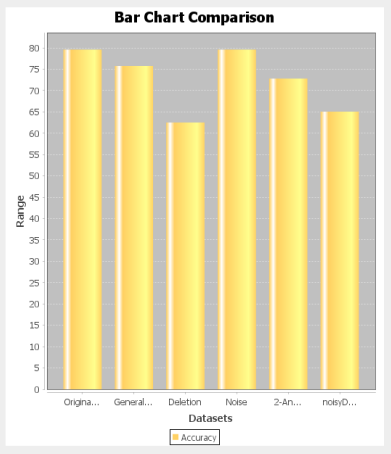
Result details Comparative

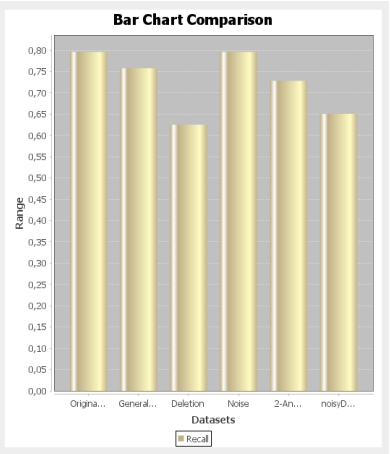
Dataset	Correctly Classified	Incorrectly Classified	Mean absolute error	Root mean squared er...	Number of Instances
OriginalDataset	312.0	80.0	0.2676862071578874	0.41001402766612377	392.0
Generalization	297.0	95.0	0.31770284034475355	0.42958020729611107	392.0
Deletion	15.0	9.0	0.37010971055088704	0.5764548702826701	24.0
Noise	312.0	80.0	0.2676862071578874	0.41001402766612377	392.0
2-Anon-Diabetes	281.0	105.0	0.38485595926533916	0.43919281393746584	392.0
noisyDiabetes	255.0	137.0	0.41742913525382935	0.5007319102933802	392.0

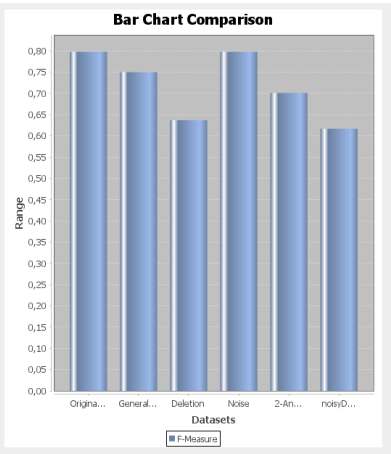
Classification Summary Result Comparative

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	PRC Area
OriginalDataset	0.7959183673...	0.23302934796...	0.8016717447...	0.79591836734...	0.7980285864...	0.55135449805...	0.7506302859...
Generalization	0.7576530612...	0.35277930902...	0.7497274170...	0.75765306122...	0.7502801841...	0.43043377948...	0.7269509422...
Deletion	0.625	0.68815789473...	0.6504629629...	0.625	0.6369778869...	-0.0592348877...	0.6912377450...
Noise	0.7959183673...	0.23302934796...	0.8016717447...	0.79591836734...	0.7980285864...	0.55135449805...	0.7506302859...
2-Anon-Diabetes	0.7279792746...	0.46793329759...	0.7134746654...	0.72797927461...	0.7012000322...	0.31941652701...	0.7089986513...
noisyDiabetes	0.6505102040...	0.55321131975...	0.6137857407...	0.65051020408...	0.6169012291...	0.11894563021...	0.6055059951...









IV.1.2. Conjunto de datos "*Heart Disease UCI*"

IV.1.2.1. Clasificador: ZeroR

Summary: Rules/ZeroR with 10-fold cross validation

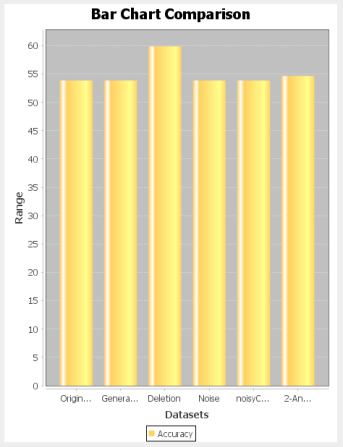
Export Datasets

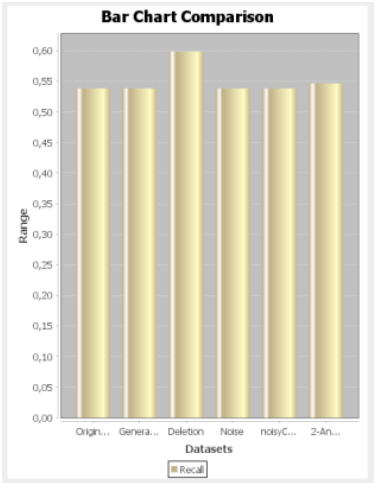
Result details Comparative

Dataset	Correctly Classified	Incorrectly Classified	Mean absolute error	Root mean squared ...	Number of Instanc...
OriginalDataset	160.0	137.0	0.4970377045965646	0.49851462225151...	297.0
Generalization	160.0	137.0	0.4970377045965646	0.49851462225151...	297.0
Deletion	100.0	67.0	0.48078069866459056	0.49020576853777...	167.0
Noise	160.0	137.0	0.4970377045965646	0.49851462225151...	297.0
noisyCleveland	160.0	137.0	0.4970377045965646	0.49851462225151...	297.0
2-Anon-Cleveland	158.0	131.0	0.4956958062420326	0.49784108705968...	297.0

Classification Summary Result Comparative

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	PRC Area
OriginalDataset	0.538720538...	0.538720538...	NaN	0.538720538...	NaN	NaN	0.499259478...
Generalization	0.538720538...	0.538720538...	NaN	0.538720538...	NaN	NaN	0.499259478...
Deletion	0.598802395...	0.598802395...	NaN	0.598802395...	NaN	NaN	0.512270808...
Noise	0.538720538...	0.538720538...	NaN	0.538720538...	NaN	NaN	0.499259478...
noisyCleveland	0.538720538...	0.538720538...	NaN	0.538720538...	NaN	NaN	0.499259478...
2-Anon-Clevel...	0.546712802...	0.546712802...	NaN	0.546712802...	NaN	NaN	0.486910743...





IV.1.2.2. Clasificador: OneR

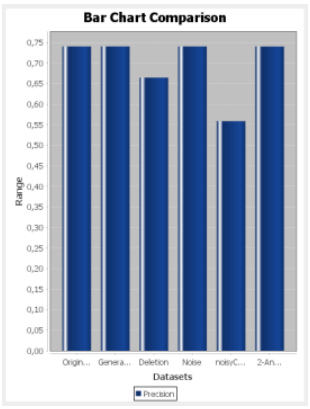
Summary: Rules/OneR with 10-fold cross validation — Export Datasets

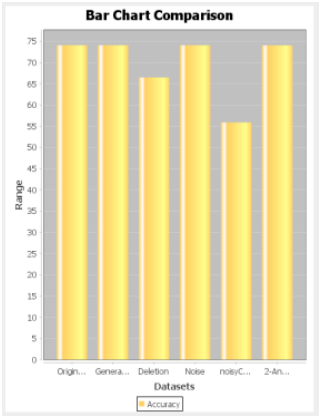
Result details Comparative

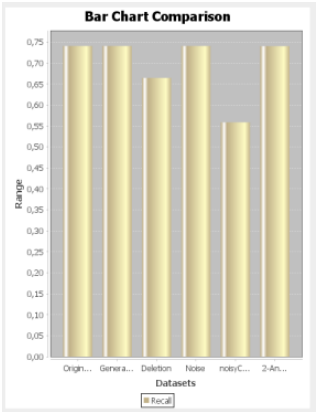
Dataset	Correctly Classified	Incorrectly Classified	Mean absolute error	Root mean squared ...	Number of Instanc...
OriginalDataset	220.0	77.0	0.25925925925925924	0.50917507721731...	297.0
Generalization	220.0	77.0	0.25925925925925924	0.50917507721731...	297.0
Deletion	111.0	56.0	0.33532934131736525	0.57907628281372...	167.0
Noise	220.0	77.0	0.25925925925925924	0.50917507721731...	297.0
noisyCleveland	166.0	131.0	0.44107744107744107	0.664136613263748	297.0
2-Anon-Cleveland	214.0	75.0	0.25951557093425603	0.50942670810849...	297.0

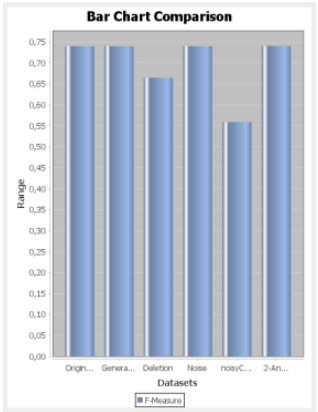
Classification Summary Result Comparative

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	PRC Area
OriginalDataset	0.740740740...	0.267109353...	0.740401736...	0.740740740...	0.739884324...	0.476773833...	0.677183816...
Generalization	0.740740740...	0.267109353...	0.740401736...	0.740740740...	0.739884324...	0.476773833...	0.677183816...
Deletion	0.664670658...	0.362581106...	0.664670658...	0.664670658...	0.664670658...	0.302089552...	0.614020913...
Noise	0.740740740...	0.267109353...	0.740401736...	0.740740740...	0.739884324...	0.476773833...	0.677183816...
noisyCleveland	0.558922558...	0.446924383...	0.558693591...	0.558922558...	0.558801761...	0.112059538...	0.533950695...
2-Anon-Clevel...	0.740484429...	0.264737980...	0.740326568...	0.740484429...	0.740396603...	0.476068567...	0.660030830...









IV.1.2.3. Clasificador: Naive Bayes

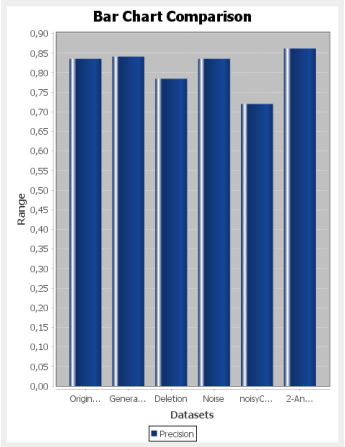
Summary: Bayes/NaiveBayes with 10-fold cross validation — Export Datasets

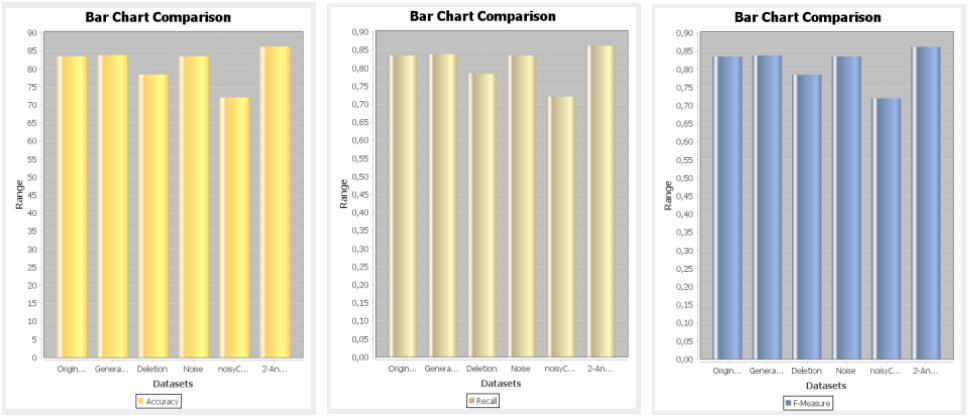
Result details Comparative

Dataset	Correctly Classified	Incorrectly Classified	Mean absolute error	Root mean squared ...	Number of Instanc...
OriginalDataset	248.0	49.0	0.18855084567214167	0.36434932052573...	297.0
Generalization	249.0	48.0	0.19606285946629687	0.36482448430091...	297.0
Deletion	131.0	36.0	0.249504431279293	0.40823242910639...	167.0
Noise	248.0	49.0	0.18856420615025352	0.36469588598515...	297.0
noisyCleveland	214.0	83.0	0.3427384753703102	0.447322892843135	297.0
2-Anon-Cleveland	249.0	40.0	0.1583486566758974	0.33265428068937...	297.0

Classification Summary Result Comparative

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	PRC Area
OriginalDataset	0.835016835...	0.171695667...	0.835481914...	0.835016835...	0.834471842...	0.667722167...	0.894324314...
Generalization	0.838383838...	0.171960480...	0.841055179...	0.838383838...	0.837220488...	0.676116393...	0.893738725...
Deletion	0.784431137...	0.233087854...	0.784431137...	0.784431137...	0.784431137...	0.551343283...	0.835496108...
Noise	0.835016835...	0.171695667...	0.835481914...	0.835016835...	0.834471842...	0.667722167...	0.893097028...
noisyCleveland	0.720538720...	0.288604413...	0.720136480...	0.720538720...	0.719338089...	0.435666981...	0.762750096...
2-Anon-Clevel...	0.861591695...	0.144759151...	0.861679804...	0.861591695...	0.861259241...	0.720286002...	0.889443712...





IV.1.2.4. Clasificador: SMO

Summary: Functions/SMO with 10-fold cross validation

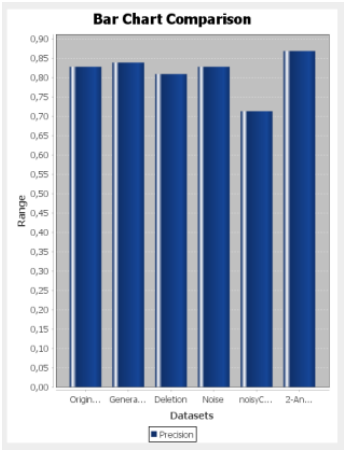
Export Datasets

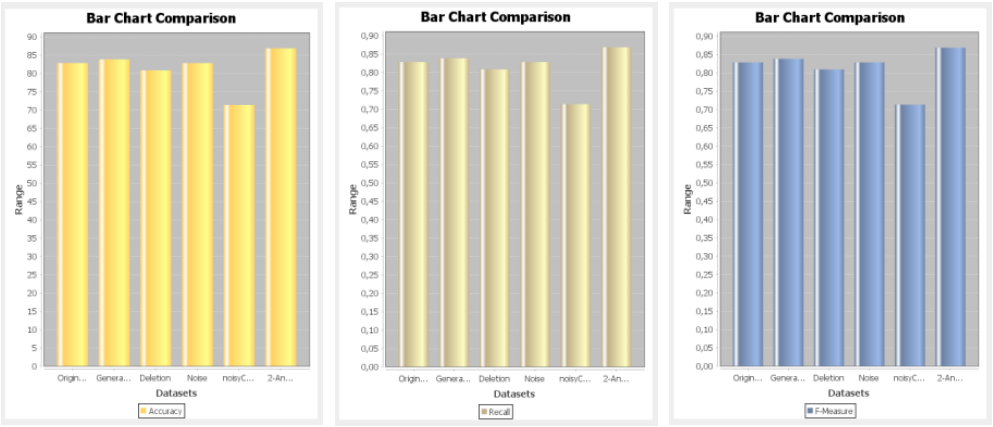
Result details Comparative

Dataset	Correctly Classified	Incorrectly Classified	Mean absolute error	Root mean squared ...	Number of Instanc...
OriginalDataset	246.0	51.0	0.1717171717171717	0.41438770700537...	297.0
Generalization	249.0	48.0	0.16161616161616163	0.40201512610368...	297.0
Deletion	135.0	32.0	0.19161676646706588	0.43774052413166...	167.0
Noise	246.0	51.0	0.1717171717171717	0.41438770700537...	297.0
noisyCleveland	212.0	85.0	0.28619528619528617	0.53497222936829...	297.0
2-Anon-Cleveland	251.0	38.0	0.1314878892733564	0.36261258840993...	297.0

Classification Summary Result Comparative

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	PRC Area
OriginalDataset	0.828282828...	0.177461660...	0.828387048...	0.828282828...	0.827869805...	0.653961276...	0.771003236...
Generalization	0.838383838...	0.168812670...	0.839090300...	0.838383838...	0.837771655...	0.674668910...	0.782499572...
Deletion	0.808383233...	0.202263830...	0.809500785...	0.808383233...	0.808824418...	0.603334284...	0.752586400...
Noise	0.828282828...	0.177461660...	0.828387048...	0.828282828...	0.827869805...	0.653961276...	0.771003236...
noisyCleveland	0.713804713...	0.294370407...	0.713235971...	0.713804713...	0.712859319...	0.422217166...	0.651527848...
2-Anon-Clevel...	0.868512110...	0.139021338...	0.868971709...	0.868512110...	0.868064742...	0.734475636...	0.796819387...





IV.1.2.5. Clasificador: LibSVM

Summary: Functions/LibSVM with 10-fold cross validation

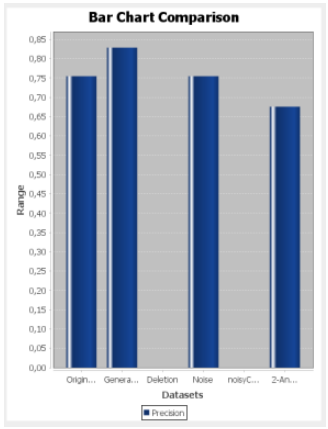
Export Datasets

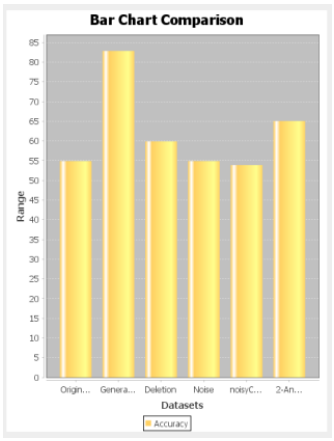
Result details Comparative

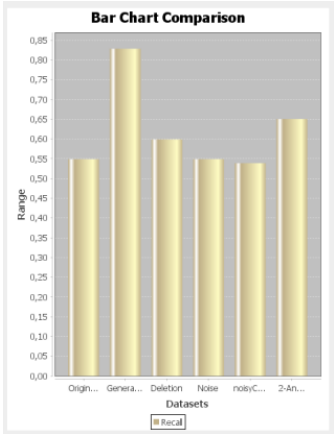
Dataset	Correctly Classified	Incorrectly Classified	Mean absolute error	Root mean squared ...	Number of Instanc...
OriginalDataset	163.0	134.0	0.4511784511784512	0.67169818458772...	297.0
Generalization	246.0	51.0	0.1717171717171717	0.41438770700537...	297.0
Deletion	100.0	67.0	0.40119760479041916	0.63340161413625...	167.0
Noise	163.0	134.0	0.4511784511784512	0.67169818458772...	297.0
noisyCleveland	160.0	137.0	0.4612794612794613	0.679175574707646	297.0
2-Anon-Cleveland	188.0	101.0	0.3494809688581315	0.59116915418358...	297.0

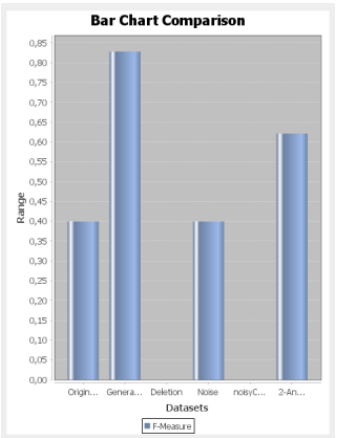
Classification Summary Result Comparative

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	PRC Area
OriginalDataset	0.548821548...	0.526923738...	0.754460706...	0.548821548...	0.399484036...	0.109165817...	0.511401608...
Generalization	0.828282828...	0.177461660...	0.828387048...	0.828282828...	0.827869805...	0.653961276...	0.771003236...
Deletion	0.598802395...	0.598802395...	NaN	0.598802395...	NaN	NaN	0.519523826...
Noise	0.548821548...	0.526923738...	0.754460706...	0.548821548...	0.399484036...	0.109165817...	0.511401608...
noisyCleveland	0.538720538...	0.538720538...	NaN	0.538720538...	NaN	NaN	0.502998560...
2-Anon-Clevel...	0.650519031...	0.399335341...	0.675489475...	0.650519031...	0.621048171...	0.301144114...	0.573110408...









IV.1.2.6. Clasificador: LibLinear

Summary: Functions/LibLinear with 10-fold cross validation

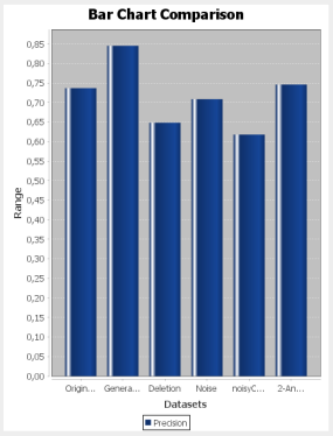
Export Datasets

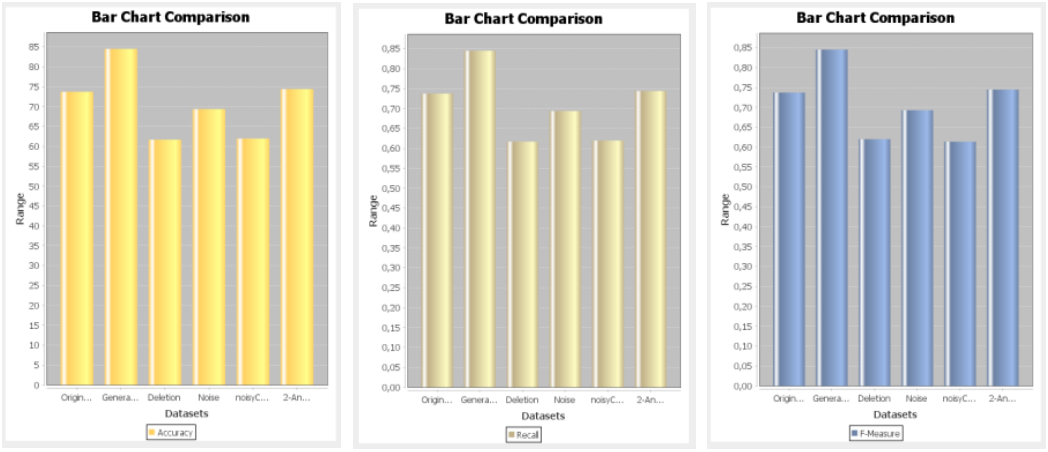
Result details Comparative

Dataset	Correctly Classified	Incorrectly Classified	Mean absolute error	Root mean squared ...	Number of Instanc...
OriginalDataset	219.0	78.0	0.26262626262626265	0.51247074319053...	297.0
Generalization	251.0	46.0	0.15488215488215487	0.39355070179349...	297.0
Deletion	103.0	64.0	0.38323353293413176	0.61905858602730...	167.0
Noise	206.0	91.0	0.3063973063973064	0.55353166702304...	297.0
noisyCleveland	184.0	113.0	0.38047138047138046	0.61682362184937...	297.0
2-Anon-Cleveland	215.0	74.0	0.2560553633217993	0.50601913335544...	297.0

Classification Summary Result Comparative

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	PRC Area
OriginalDataset	0.737373737...	0.268943080...	0.736933391...	0.737373737...	0.736850697...	0.470300990...	0.674368055...
Generalization	0.845117845...	0.159898867...	0.845197739...	0.845117845...	0.844809385...	0.687954896...	0.790886872...
Deletion	0.616766467...	0.355273929...	0.648434767...	0.616766467...	0.619743252...	0.257376688...	0.598258327...
Noise	0.693602693...	0.292781525...	0.708753740...	0.693602693...	0.692455125...	0.403872797...	0.643136665...
noisyCleveland	0.619528619...	0.399227524...	0.618021570...	0.619528619...	0.613535825...	0.227415354...	0.570595437...
2-Anon-Clevel...	0.743944636...	0.255346704...	0.746019765...	0.743944636...	0.744412655...	0.486813693...	0.665804335...





IV.1.2.7. Clasificador: J48

Summary: Tree/J48 with 10-fold cross validation

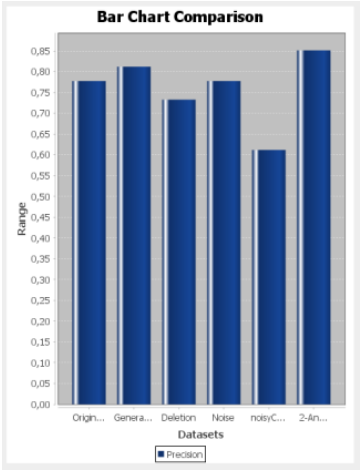
Export Datasets

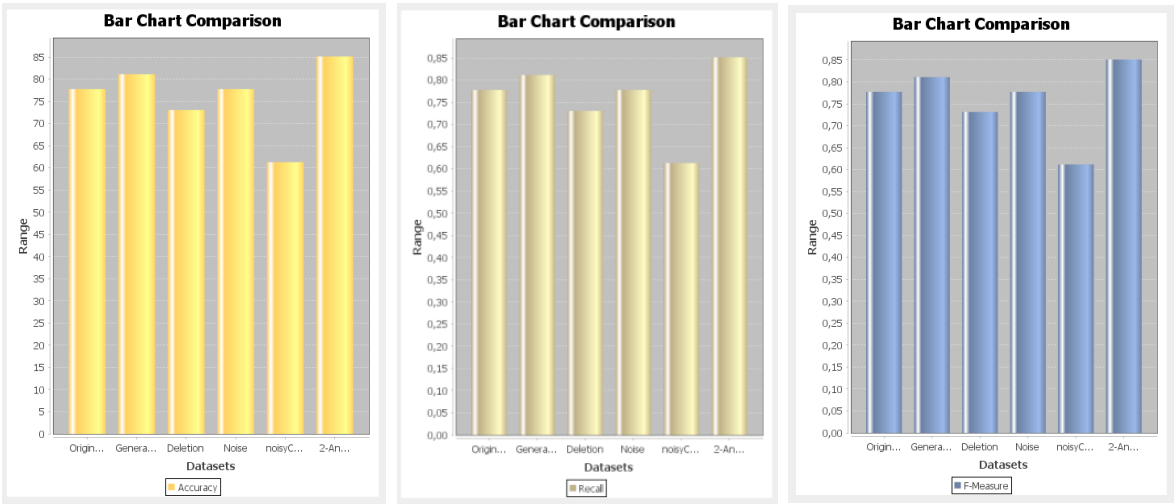
Result details Comparative

Dataset	Correctly Classified	Incorrectly Classified	Mean absolute error	Root mean squared ...	Number of Instanc...
OriginalDataset	231.0	66.0	0.25762954776841257	0.444917935333655	297.0
Generalization	241.0	56.0	0.23086353513890837	0.38633946959153...	297.0
Deletion	122.0	45.0	0.3073092430482052	0.49448321207555...	167.0
Noise	231.0	66.0	0.25762954776841257	0.444917935333655	297.0
noisyCleveland	182.0	115.0	0.4034100885947855	0.56709213746156...	297.0
2-Anon-Cleveland	246.0	43.0	0.22364054941217884	0.33481445758066...	297.0

Classification Summary Result Comparative

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	PRC Area
OriginalDataset	0.777777777...	0.228051500...	0.777532521...	0.777777777...	0.777335205...	0.551921205...	0.709397427...
Generalization	0.811447811...	0.196073723...	0.811886036...	0.811447811...	0.810733598...	0.620059204...	0.842652368...
Deletion	0.730538922...	0.283971757...	0.732808903...	0.730538922...	0.731437491...	0.443597715...	0.667446931...
Noise	0.777777777...	0.228051500...	0.777532521...	0.777777777...	0.777335205...	0.551921205...	0.709397427...
noisyCleveland	0.612794612...	0.395550087...	0.611613447...	0.612794612...	0.611863286...	0.218292695...	0.590578699...
2-Anon-Clevel...	0.851211072...	0.154670344...	0.851149578...	0.851211072...	0.850922501...	0.699248643...	0.861581038...





IV.1.3. Conjunto de datos "*Breast Cancer Wisconsin*"

IV.1.3.1. Clasificador: ZeroR

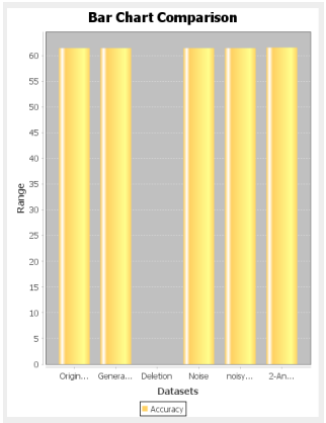
Summary: Rules/ZeroR with 10-fold cross validation — Export Datasets

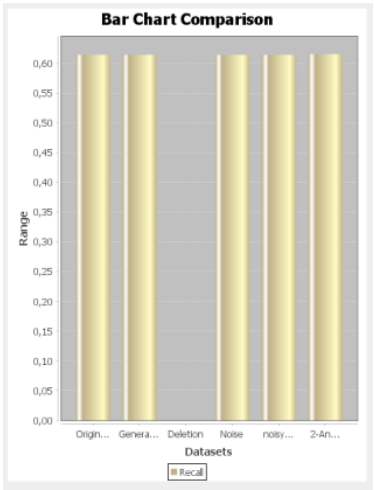
Result details Comparative

Dataset	Correctly Classified	Incorrectly Classified	Mean absolute error	Root mean squared ...	Number of Instanc...
OriginalDataset	338.0	212.0	0.4738759831717569	0.48671542575565...	550.0
Generalization	338.0	212.0	0.4738759831717569	0.48671542575565...	550.0
Deletion	0.0	0.0	0.0	0.0	0.0
Noise	338.0	212.0	0.4738759831717569	0.48671542575565...	550.0
noisyBreastCancer	338.0	212.0	0.4738759831717569	0.48671542575565...	550.0
2-Anon-BreastCa...	338.0	211.0	0.4733584094614091	0.48644554977870...	550.0

Classification Summary Result Comparative

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	PRC Area
OriginalDataset	0.614545454...	0.614545454...	NaN	0.614545454...	NaN	NaN	0.523385123...
Generalization	0.614545454...	0.614545454...	NaN	0.614545454...	NaN	NaN	0.523385123...
Deletion	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Noise	0.614545454...	0.614545454...	NaN	0.614545454...	NaN	NaN	0.523385123...
noisyBreastC...	0.614545454...	0.614545454...	NaN	0.614545454...	NaN	NaN	0.523385123...
2-Anon-Breast...	0.615664845...	0.615664845...	NaN	0.615664845...	NaN	NaN	0.523424408...





IV.1.3.2. Clasificador: OneR

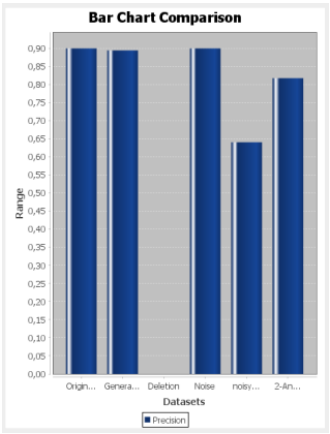
Summary: Rules/OneR with 10-fold cross validation — Export Datasets

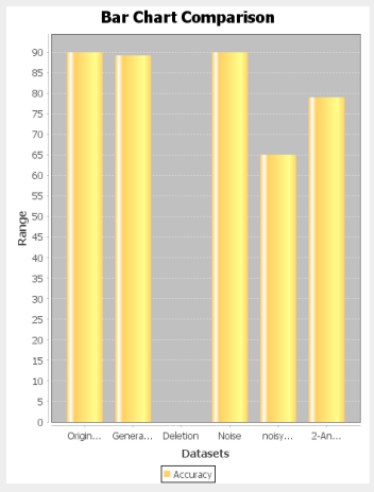
Result details Comparative

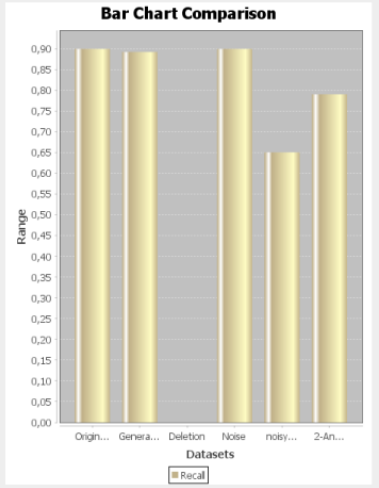
Dataset	Correctly Classified	Incorrectly Classified	Mean absolute error	Root mean squared ...	Number of Instanc...
OriginalDataset	495.0	55.0	0.1	0.31622776601683...	550.0
Generalization	491.0	59.0	0.10727272727272727	0.32752515517548...	550.0
Deletion	0.0	0.0	0.0	0.0	0.0
Noise	495.0	55.0	0.1	0.31622776601683...	550.0
noisyBreastCancer	358.0	192.0	0.3490909090909091	0.59083915670079...	550.0
2-Anon-BreastCa...	434.0	115.0	0.20947176684881602	0.45768085698313...	550.0

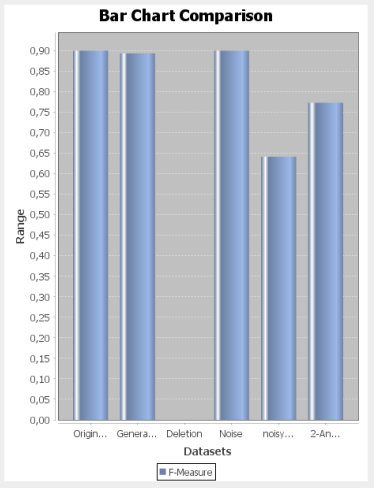
Classification Summary Result Comparative

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	PRC Area
OriginalDataset	0.9	0.113715529...	0.899785320...	0.9	0.899864618...	0.788416393...	0.859740161...
Generalization	0.892727272...	0.109485115...	0.894398396...	0.892727272...	0.893196376...	0.776418772...	0.854572257...
Deletion	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Noise	0.9	0.113715529...	0.899785320...	0.9	0.899864618...	0.788416393...	0.859740161...
noisyBreastC...	0.650909090...	0.426447775...	0.640383037...	0.650909090...	0.641365231...	0.237181855...	0.592737272...
2-Anon-Breast...	0.790528233...	0.317744363...	0.817572488...	0.790528233...	0.773005576...	0.563364252...	0.712064398...









IV.1.3.3. Clasificador: Naïve Bayes

Summary: Bayes/NaiveBayes with 10-fold cross validation

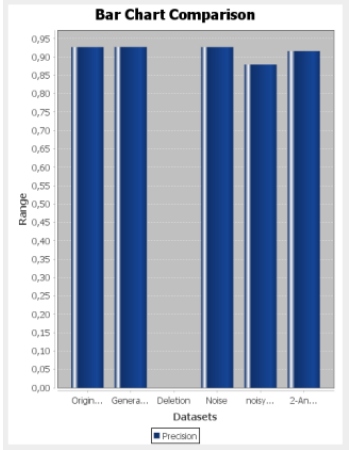
Export Datasets

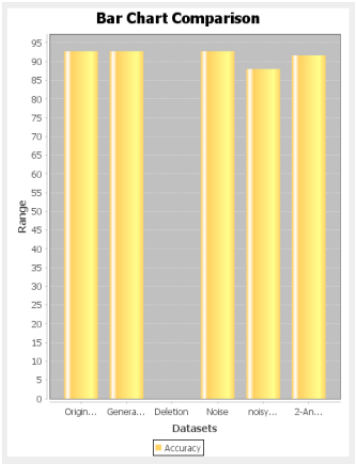
Result details Comparative

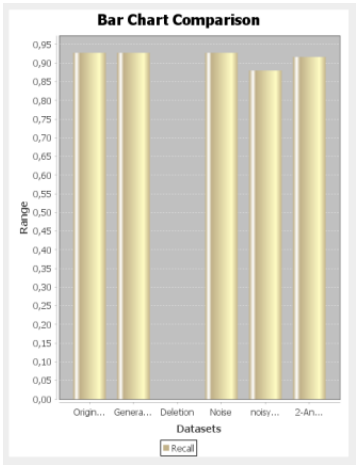
Dataset	Correctly Classified	Incorrectly Classified	Mean absolute error	Root mean squared ...	Number of Instanc...
OriginalDataset	510.0	40.0	0.07407823185271367	0.26576560656740...	550.0
Generalization	510.0	40.0	0.06947584111043616	0.24793887993067...	550.0
Deletion	0.0	0.0	0.0	0.0	0.0
Noise	510.0	40.0	0.07414068138059407	0.26583278432693...	550.0
noisyBreastCancer	484.0	66.0	0.12836376139317454	0.30849756576551...	550.0
2-Anon-BreastCa...	503.0	46.0	0.08501455221533535	0.27725172381496...	550.0

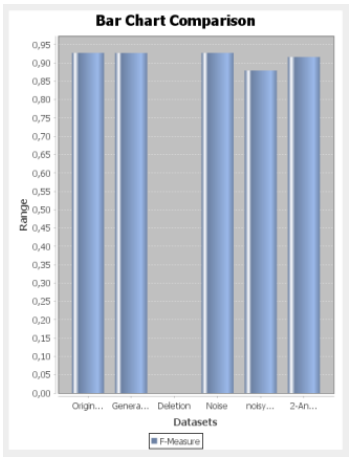
Classification Summary Result Comparative

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	PRC Area
OriginalDataset	0.927272727...	0.084300750...	0.927111192...	0.927272727...	0.927140310...	0.846047405...	0.969572687...
Generalization	0.927272727...	0.087817552...	0.927140840...	0.927272727...	0.926998732...	0.845805798...	0.977074928...
Deletion	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Noise	0.927272727...	0.084300750...	0.927111192...	0.927272727...	0.927140310...	0.846047405...	0.969601733...
noisyBreastC...	0.88	0.145602322...	0.879550045...	0.88	0.879168610...	0.744652494...	0.917233820...
2-Anon-Breast...	0.916211293...	0.093263369...	0.916211293...	0.916211293...	0.916211293...	0.822947923...	0.958125983...









IV.1.3.4. Clasificador: SMO

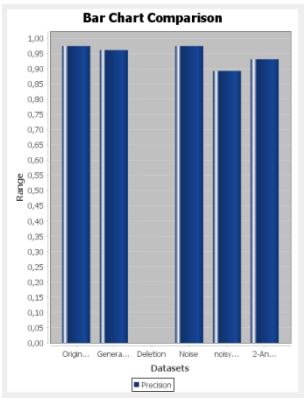
Summary: Functions/SMO with 10-fold cross validation — Export Datasets

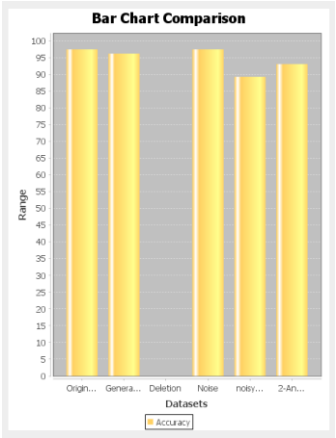
Result details Comparative

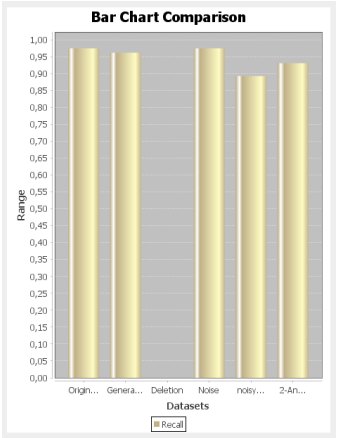
Dataset	Correctly Classified	Incorrectly Classified	Mean absolute error	Root mean squared ...	Number of Instanc...
OriginalDataset	536.0	14.0	0.025454545454545455	0.15954480704349...	550.0
Generalization	529.0	21.0	0.038181818181818185	0.19540168418367...	550.0
Deletion	0.0	0.0	0.0	0.0	0.0
Noise	536.0	14.0	0.025454545454545455	0.15954480704349...	550.0
noisyBreastCancer	491.0	59.0	0.10727272727272727	0.32752515517548...	550.0
2-Anon-BreastCa...	511.0	38.0	0.0692167577413479	0.26309077851826...	550.0

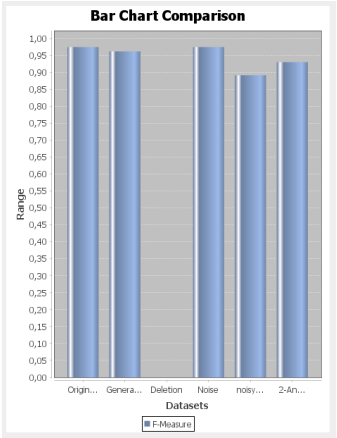
Classification Summary Result Comparative

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	PRC Area
OriginalDataset	0.974545454...	0.037066387...	0.974992395...	0.974545454...	0.974423561...	0.946504216...	0.960523891...
Generalization	0.961818181...	0.046807575...	0.961811395...	0.961818181...	0.961730534...	0.919221504...	0.943143932...
Deletion	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Noise	0.974545454...	0.037066387...	0.974992395...	0.974545454...	0.974423561...	0.946504216...	0.960523891...
noisyBreastC...	0.892727272...	0.139377937...	0.893494022...	0.892727272...	0.891415731...	0.772037129...	0.845884090...
2-Anon-Breast...	0.930783242...	0.093070462...	0.931663183...	0.930783242...	0.930138908...	0.853475512...	0.895652515...









IV.1.3.5. Clasificador: LibSVM

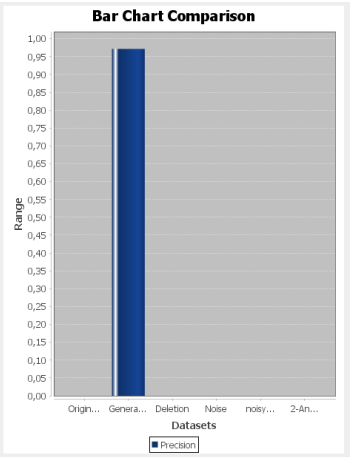
Summary: Functions/LibSVM with 10-fold cross validation — Export Datasets

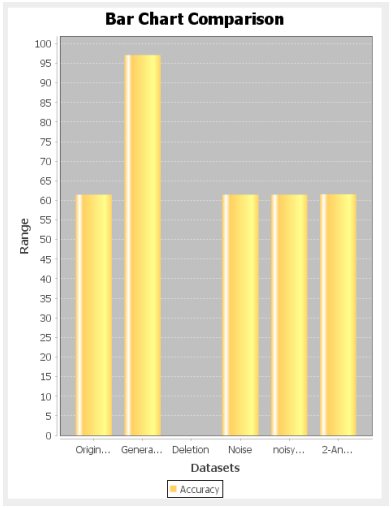
Result details Comparative

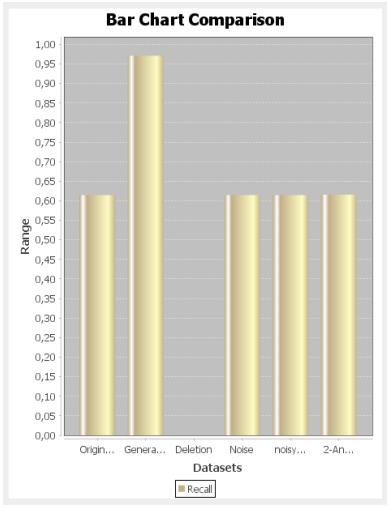
Dataset	Correctly Classified	Incorrectly Classified	Mean absolute error	Root mean squared ...	Number of Instanc...
OriginalDataset	338.0	212.0	0.38545454545454544	0.620849857416868	550.0
Generalization	534.0	16.0	0.02909090909090909	0.17056057308448...	550.0
Deletion	0.0	0.0	0.0	0.0	0.0
Noise	338.0	212.0	0.38545454545454544	0.620849857416868	550.0
noisyBreastCancer	338.0	212.0	0.38545454545454544	0.620849857416868	550.0
2-Anon-BreastCa...	338.0	211.0	0.3843351548269581	0.61994770330001...	550.0

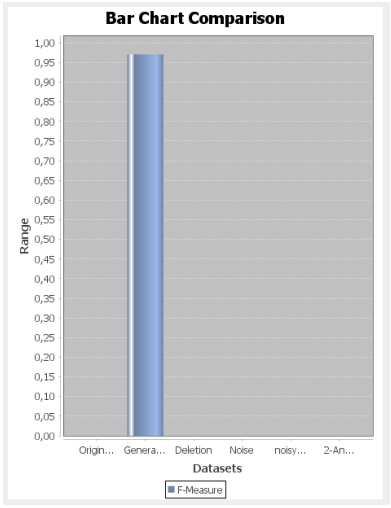
Classification Summary Result Comparative

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	PRC Area
OriginalDataset	0.614545454...	0.614545454...	NaN	0.614545454...	NaN	NaN	0.526241322...
Generalization	0.970909090...	0.042863986...	0.971563636...	0.970909090...	0.970739136...	0.938958385...	0.954919008...
Deletion	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Noise	0.614545454...	0.614545454...	NaN	0.614545454...	NaN	NaN	0.526241322...
noisyBreastC...	0.614545454...	0.614545454...	NaN	0.614545454...	NaN	NaN	0.526241322...
2-Anon-Breast...	0.615664845...	0.615664845...	NaN	0.615664845...	NaN	NaN	0.525798973...









IV.1.3.6. Clasificador: LibLinear

Summary: Functions/LibLinear with 10-fold cross validation

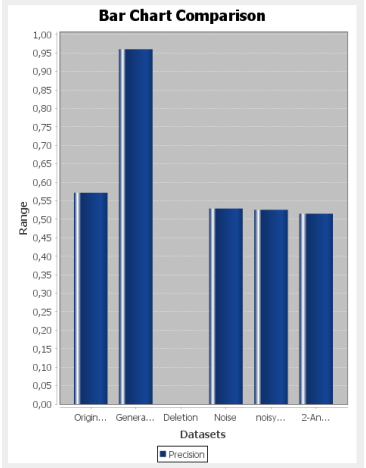
Export Datasets

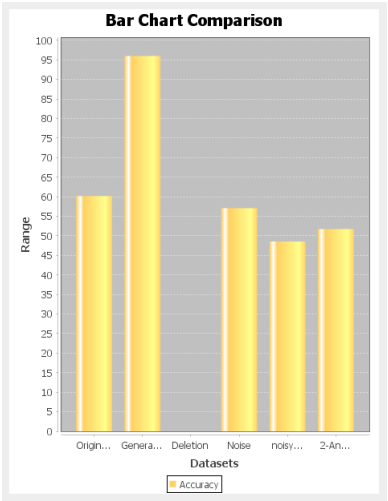
Result details Comparative

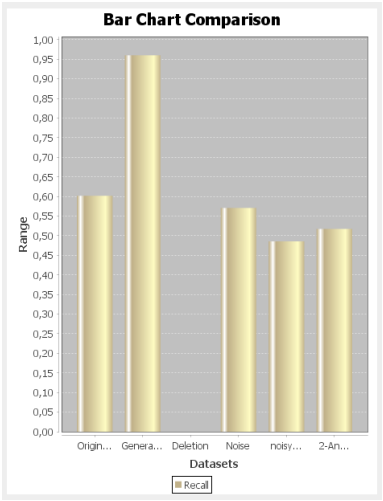
Dataset	Correctly Classified	Incorrectly Classified	Mean absolute error	Root mean squared ...	Number of Instanc...
OriginalDataset	331.0	219.0	0.3981818181818182	0.63101649596648...	550.0
Generalization	528.0	22.0	0.04	0.2	550.0
Deletion	0.0	0.0	0.0	0.0	0.0
Noise	314.0	236.0	0.4290909090909091	0.65505031035097...	550.0
noisyBreastCancer	267.0	283.0	0.5145454545454545	0.717318237984686	550.0
2-Anon-BreastCa...	284.0	265.0	0.48269581056466304	0.69476313270399...	550.0

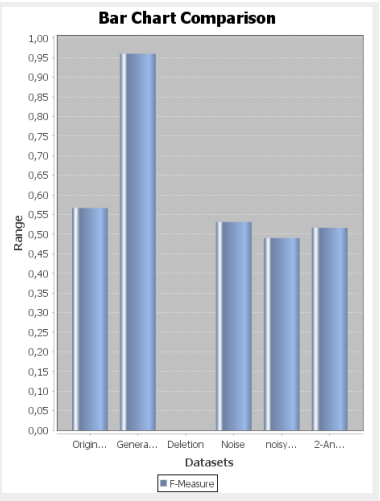
Classification Summary Result Comparative

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	PRC Area
OriginalDataset	0.601818181...	0.529332974...	0.571734951...	0.601818181...	0.566604286...	0.087316380...	0.545217573...
Generalization	0.96	0.047947973...	0.959961763...	0.96	0.959927170...	0.915379480...	0.940773442...
Deletion	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Noise	0.570909090...	0.566303754...	0.529181818...	0.570909090...	0.531075554...	0.005603572...	0.527339669...
noisyBreastC...	0.485454545...	0.486263968...	0.525840868...	0.485454545...	0.490252732...	-7.939880241...	0.526049736...
2-Anon-Breast...	0.517304189...	0.541730000...	0.515141330...	0.517304189...	0.516182340...	-0.024539623...	0.520014744...









IV.1.3.7. Clasificador: J48

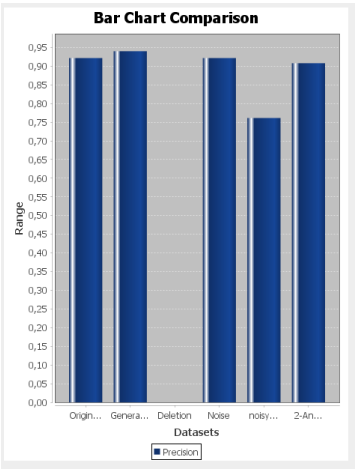
Summary: Tree/J48 with 10-fold cross validation — Export Datasets

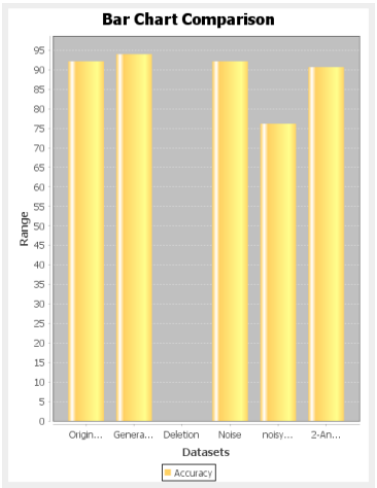
Result details Comparative

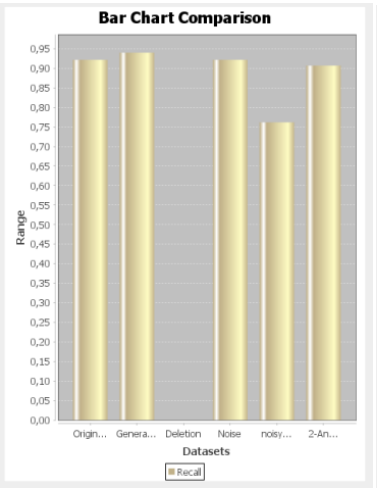
Dataset	Correctly Classified	Incorrectly Classified	Mean absolute error	Root mean squared ...	Number of Instanc...
OriginalDataset	507.0	43.0	0.08318081733159834	0.26728986197493...	550.0
Generalization	517.0	33.0	0.07933366378866814	0.23396755357870...	550.0
Deletion	0.0	0.0	0.0	0.0	0.0
Noise	507.0	43.0	0.08318081733159834	0.26728986197493...	550.0
noisyBreastCancer	419.0	131.0	0.240861981887047	0.47129637936944...	550.0
2-Anon-BreastCa...	498.0	51.0	0.2073998609832138	0.27844119022910...	550.0

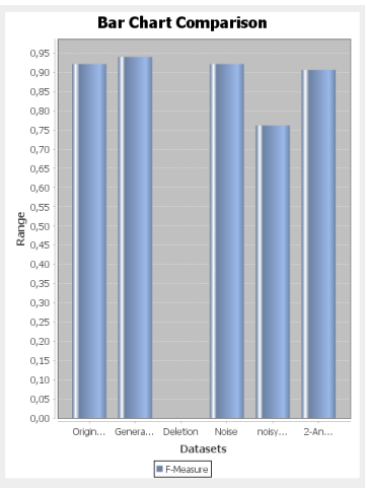
Classification Summary Result Comparative

Dataset	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	PRC Area
OriginalDataset	0.921818181...	0.092997147...	0.921622931...	0.921818181...	0.921562618...	0.834258358...	0.901196924...
Generalization	0.94	0.069284358...	0.939896448...	0.94	0.939918771...	0.873076490...	0.918410539...
Deletion	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Noise	0.921818181...	0.092997147...	0.921622931...	0.921818181...	0.921562618...	0.834258358...	0.901196924...
noisyBreastC...	0.761818181...	0.267205030...	0.761219604...	0.761818181...	0.761495727...	0.495954240...	0.723931465...
2-Anon-Breast...	0.907103825...	0.122098637...	0.907919772...	0.907103825...	0.906070821...	0.802690420...	0.957498016...









IV.2. Scripts para aplicar técnicas de anonimización mediante la herramienta externa R

En el trabajo realizado se decidió comparar conjuntos de datos obtenidos mediante la herramienta externa R. En ella se decide hacer uso de dos librerías que permiten aplicar técnicas de anonimización como la generación de ruido y como la k-anonimización. A continuación se deja a modo de ejemplo los scripts aplicados sobre el conjunto de datos de "*Pima Diabetes*", la misma lógica fue usada para aplicar estas técnicas al resto de conjuntos de datos.

IV.2.1. Script para aplicar k-anonimización

Como es posible visualizar en el código a continuación, se hace uso del paquete de *sdcMicro* para aplicar el método de "*localSuppression*" donde se utiliza la técnica de k-

anonimización con el nivel de k indicado en los parámetros de entrada, para este caso específico se utilizó el valor de k=2 y se aplica la anonimización para todos los atributos del conjunto de datos.

```
library(foreign)
library(dplyr)
library(sdcMicro)

# path to file/name of file with dataset.
pathToData <-
 "C:\\Users\\Usuario\\Documents\\Espana\\Master\\TFM\\Documentacion\\D
atasets\\originales\\diabetes.arff"

data <- read.arff(pathToData)

## Local Suppression:
## k - value of k anonimity to achieve,
## keyVar - attributes that should be used for local suppresion.
localS <- localSuppression(data, k=2, keyVar=c(1,2,3,4,5,6,7,8,9))
localS
plot(localS)

# save dataset into specified folder
newDataset<- as.data.frame(localS$xAnon)
path<-
 "C:\\Users\\Usuario\\Documents\\Espana\\Master\\TFM\\Documentacion\\d
atasets\\localSuppR"
nameNewDataset<- "2-Anon-Diabetes.arff"
write.arff(newDataset, file = file.path(path, nameNewDataset))
```

IV.2.2. Script para añadir ruido

Al igual que en el caso de la k-anonimización se decide hacer uso del paquete de *sdcMicro*, la diferencia es que en este caso se hace uso del método "*addNoise*", mediante el cual se añade un ruido aleatorio a todos los atributos numéricos del conjunto de datos.

```
library(foreign)
library(sdcMicro)
library(dplyr)

# path to file/name of file with dataset.
pathToData <-
 "C:\\Users\\Usuario\\Documents\\Espana\\Master\\TFM\\Documentacion\\D
atasets\\Originales\\diabetes.arff"

data <- read.arff(pathToData)
numericData<- select_if(data, is.numeric)
nonNumeric<- select_if(data, is.factor)

# apply noise to dataset
noise <- addNoise(numericData)
noisyData <- as.data.frame(noise$xm)
merge<- cbind(noisyData,nonNumeric)

# save dataset into specified folder
path<-
 "C:\\Users\\Usuario\\Documents\\Espana\\Master\\TFM\\Documentacion\\D
atasets\\noisyR"
nameNewDataset<- "noisyDiabetes.arff"
write.arff(merge, file = file.path(path, nameNewDataset))
```

IV.3. Extracto del código utilizado para aplicar técnicas de anonimización en la herramienta.

IV.3.1. Método de generalización

```
/**
 * Method to apply generalization to dataset
 * @param dataset
 * @return dataset with generalization.
 */
public Instances applyGeneralization(Instances dataset){

    Instances copyInstance = new Instances(dataset);
    //for each instance in the dataset
    for(int i =0 ; i< copyInstance.size(); i++) {

        for(int j=0 ;j< copyInstance.numAttributes(); j++ ) {
            Attribute att = copyInstance.get(i).attribute(j);
            if(att.isNumeric()) {

                //if value is NAN.. leave it as is
                Double value = copyInstance.get(i).value(j);

                if(!value.isNaN()) {
                    double min = this.controller.getDataset().attributeStats(j).n
umericStats.min;
                    double max = this.controller.getDataset().attributeStats(j).n
umericStats.max;

                    double distance = (max -
 min) / 4; //4 subsets to be created for each attribute

                    Range range1 = new Range(min, min + distance);
                    Range range2 = new Range(min + distance, min + distance + dis
tance);
                    Range range3 = new Range(min + distance + distance , min + di
stance + distance + distance );
                    Range range4 = new Range(min + distance + distance + distance
, max );

                    ArrayList<Range> ranges = new ArrayList<Range>();
                    ranges.add(range1);
                    ranges.add(range2);
                    ranges.add(range3);
                    ranges.add(range4);

                    copyInstance.get(i).setValue(j, getRange(copyInstance.get(i).
value(j), ranges) );
                }
```

IV.3.2. Método de ruido

```
/**
 * Method to apply noise to dataset
 * @param dataset
 * @return dataset with noise
 */
public Instances applyNoise(Instances dataset) {

    Instances copyInstance = new Instances(dataset);
        for(int j=0 ;j< copyInstance.numAttributes(); j++ ) {
            if(copyInstance.attribute(j).isNumeric()) {
                Random r = new Random();
                double low = this.controller.getDataset().attributeStats(j).numer
icStats.stdDev / 4 ;
                double high = this.controller.getDataset().attributeStats(j).nume
ricStats.stdDev;
                double noise = low + (high - low) * r.nextDouble();
                //apply noise to the attribute column j for each instance i
                for(int i=0; i<copyInstance.size(); i++ ) {
                    copyInstance.get(i).setValue(j, copyInstance.get(i).value(j)
+ noise);
                }
            }
        }

    return copyInstance;
}
```

IV.3.3. Extracto del método de eliminación

```
/**
 * Method to apply deletion technique to dataset
 * @param dataset
 * @return dataset with deletion
 */
public Instances applyDeletion(Instances dataset) {

    Instances copyInstance = new Instances(dataset);
    //look into each column of the dataset (attributes)
    for(int j=0; j < copyInstance.numAttributes(); j++ ) {
        ArrayList<Object> checked = new ArrayList <>();
        //For each row of the column
        for(int i=0 ; i <= copyInstance.size() - 1; i++) {
            Attribute att = copyInstance.get(i).attribute(j);
                //if attribute is numeric continue..
                if(att.isNumeric() && !att.name().equals("id")) {
                    Double value = copyInstance.get(i).value(j);
                    //if value is not NAN and is not checked already
                    if(!value.isNaN() && !checked.contains(value) ) {
                        // check if the value repeats itself in the rest of the c
olumn
                        int repeat = 0;
                        for(int k=i+1;  k < copyInstance.size(); k++ ) {
                            //if value repeats itself add it to checked list and
end loop
                            if(value == copyInstance.get(k).value(j)) {
                                repeat = repeat + 1;
                                //add value to verified list
                                checked.add(value);
                                break;
                            }
                        }
                        // if the row contains an attribute that is unique for th
e complete set remove it
                        if(repeat == 0) {
                            copyInstance.remove(i);
                            i--;
                        }
                    }
                }
            }
        }

    return copyInstance;
}
```

IV.4. Análisis complementario de los resultados

A continuación se presenta a modo complementario los análisis detallados de los resultados obtenidos para los conjuntos de datos *Heart Disease UCI* y *Breast Cancer Wisconsin* de la misma forma en la que se analizaron los resultados obtenidos del conjunto de datos de diabetes.

IV.4.1. Análisis para el conjunto de datos "*Heart Disease UCI*"

Tras observar los resultados obtenidos para el conjunto de datos de *Heart Disease UCI* surgen 2 preguntas principales, que se comentan a continuación.

1. ***Para cada una de las técnicas de anonimización aplicadas, ¿se obtiene un resultado igual, mejor o peor al obtenido con el conjunto de datos original?***

Si se observan los resultados obtenidos anteriormente para el conjunto de datos de *Heart Disease UCI* se puede concluir lo siguiente:

- Para el clasificador ZeroR, de la misma forma que ocurrió para el conjunto de datos de diabetes, podemos ver que el nivel de exactitud alcanzado para casi todas las técnicas es igual o superior al nivel obtenido con el conjunto de datos original, con lo que en un principio las técnicas de anonimización aplicadas no parecen afectar negativamente.

- Para el clasificador OneR, ocurre similar al caso visto en el conjunto de datos de diabetes: se obtienen resultados iguales o inferiores al obtenido con el conjunto de datos original, por lo que parece que para este caso las técnicas de anonimización sí afectan negativamente el resultado de exactitud obtenido.

- Para el clasificador de Naive Bayes, también se obtiene un resultado mixto, como con el conjunto de datos de diabetes, con la diferencia que para los valores de exactitud obtenidos en los conjuntos de datos anonimizados se observa un mayor grado de separación, tanto superior como inferior, con respecto al valor del conjunto de datos original.

- Para el clasificador de SMO, se puede observar nuevamente cierta variación en los resultados, aunque menor a la observada en el conjunto de datos de diabetes, donde la mayoría de las técnicas no muestran un impacto negativo sobre la exactitud obtenida.

- Para el clasificador SVM ocurre de forma similar a lo observado con el conjunto de datos de diabetes. Podemos ver que las técnicas obtienen resultados aproximadamente iguales o superiores al obtenido con el conjunto de datos original.

- Para el clasificador LibLinear, de forma similar a lo ocurrido en el conjunto de datos de diabetes, se obtiene en su mayoría un impacto negativo debido a que la mayoría de los conjuntos de datos anonimizados muestran un resultado de

exactitud inferior al obtenido inicialmente. Sólo una de las técnicas genera un valor de exactitud aproximado al original; el resto se diferencia notablemente, tanto por encima como por debajo del valor inicial.

- Para el clasificador J48 ocurre lo contrario a lo visto para el conjunto de datos de diabetes. En este caso se obtienen resultados de exactitud igual o superiores para los conjuntos de datos anonimizados, por lo que el impacto de anonimización para este clasificador no podría considerarse negativo.

2. ***¿Se podría establecer alguna relación entre los mejores y peores valores obtenidos para cada conjunto de datos?***

Si se observa en detalle la Figura 14, es posible apreciar que todos los conjuntos de datos tienen en común que el peor resultado obtenido de exactitud ha sido obtenido mediante el clasificador ZeroR.

Por otro lado para los mejores valores de exactitud ocurre algo similar a lo observado en el conjunto de datos de diabetes, y es que no hay un clasificador en común para los mejores valores, sino que estos se obtienen a partir de 3 clasificadores diferentes como SMO, Naive Bayes y LibLinear, en donde destacan principalmente los clasificadores SMO y Naive Bayes, debido a que estos generan los valores más altos de exactitud para 5 de los 6 conjuntos de datos comparados.

Por lo tanto, no es posible establecer una relación directa entre los mejores y peores valores de exactitud obtenidos para cada conjunto de datos y clasificador usado.

IV.4.2. Análisis para el conjunto de datos "*Breast Cancer Wisconsin*"

Al igual que con los conjuntos de datos anteriores, al observar los resultados obtenidos para el conjunto de datos de *Breast Cancer Wisconsin* se busca responder a 2 preguntas principales, pero antes es importante destacar que para este conjunto de datos ocurre una situación inesperada, y es que la técnica de anonimización de eliminación usada en la aplicación CPDA elimina el 100% de las instancias. Esto ocurre porque la técnica de eliminación implementada en la herramienta es una técnica bastante violenta que elimina de forma automática todas aquellas instancias únicas dentro del conjunto de datos; en consecuencia para este caso particular ocurre que se eliminan todas las instancias. Ante estos resultados obtenidos se decide descartar los resultados generados por esta técnica de anonimización en el análisis final y las conclusiones.

1. ***Para cada una de las técnicas de anonimización aplicadas, ¿se obtiene un resultado igual, mejor o peor al obtenido con el conjunto de datos original?***

Si se observan los resultados obtenidos anteriormente para el conjunto de datos de *Breast Cancer Wisconsin* se puede concluir lo siguiente:

- Para el clasificador ZeroR, de la misma forma que ocurrió para los conjuntos de datos anteriores, podemos ver que el nivel de exactitud alcanzado para casi todas las técnicas es igual o levemente superior al nivel obtenido con el conjunto

de datos inicial, con lo que en un principio las técnicas de anonimización aplicadas no parecen afectar negativamente a este conjunto de datos.

- Para el clasificador OneR ocurre similar a lo observado en los conjuntos de datos anteriores, donde se obtienen resultados iguales o inferiores al obtenido con el conjunto de datos inicial, por lo que parece que para este caso las técnicas de anonimización sí afectan negativamente el resultado de exactitud obtenido y claramente se puede observar que con algunas técnicas el impacto es mayor que con otras.

- Para el clasificador de Naive Bayes, ocurre distinto a lo que se observó en los conjuntos de datos anteriores, debido a que en este caso se obtienen valores iguales o por debajo al valor de exactitud inicial, por lo que en este caso todas las técnicas aplicadas generan un impacto negativo, de mayor o menor magnitud según cada técnica.

- Para el clasificador de SMO, ocurre nuevamente un resultado diferente a lo visto en los conjuntos de datos anteriores. Se pude observar que las técnicas generan valores de exactitud de impacto principalmente negativo, dado que los valores de exactitud obtenidos son en su mayoría inferiores al valor obtenido con el conjunto de datos original.

- Para el clasificador SVM ocurre de forma similar a lo observado con los dos conjuntos de datos anteriores, donde se puede ver que las técnicas obtienen resultados aproximadamente iguales o superiores al obtenido con el conjunto de datos original.

- Para el clasificador LibLinear, similar a lo ocurrido en los conjuntos de datos anteriores, se obtiene en su mayoría un impacto negativo, debido a que la mayoría de los conjuntos de datos anonimizados muestran un resultado de exactitud inferior al obtenido inicialmente.

- Para el clasificador J48, se observa una situación diferente a lo visto en los conjuntos de datos anteriores, dado que se obtienen resultados mixtos, es decir, para algunas técnicas se obtienen resultados inferiores al original, mientras que para otras se obtienen valores aproximadamente iguales o superiores, por lo cual no podría decirse que en este clasificador las técnicas de anonimización han tenido un impacto 100% negativo.

2. ***¿Se podría establecer alguna relación entre los mejores y peores valores obtenidos para cada conjunto de datos?***

Si se observa en detalle la Figura 16, es posible apreciar que todos los conjuntos de datos tienen en común que el peor resultado obtenido de exactitud se obtiene mediante el clasificador LibLinear.

Por otro lado, para los mejores valores de exactitud ocurre algo similar a lo observado en los conjuntos de datos analizados anteriormente, y es que no hay un clasificador

único en común para los mejores valores obtenidos, sino que se obtienen a partir de 2 clasificadores como SMO y SVM, donde destaca principalmente el de SMO por ser la técnica de anonimización que obtiene en la mayoría de los casos el mejor valor de exactitud para las técnicas comparadas.

Por lo tanto, al igual que con los conjuntos de datos estudiados anteriormente, para el conjunto de datos de *Breast Cancer Wisconsin* no es posible establecer una relación directa entre los mejores y peores valores de exactitud obtenidos para cada conjunto de datos y clasificador usado.