risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Revista Ibérica de Sistemas y Tecnologías de Información

Recebido/Submission: 20/11/2018
Aceitação/Acceptance: 14/01/2019

Marco de medición de la privacidad

Luis Gustavo Esquivel-Quirós[1], Elena Gabriela Barrantes[2], Fernando Esponda Darlington[3].

luis.esquivel@ucr.ac.cr, gabriela.barrantes@ecci.ucr.ac.cr, fernando.esponda@itam.mx

[1, 2] Universidad de Costa Rica, Escuela de Ciencias de la Computación e Informática, Sede "Rodrigo Facio Brenes" Montes de Oca, 11501-2060, San José, Costa Rica.

[3] Ciencias en Computación, Instituto Tecnológico Autónomo de México, Río Hondo #1 Colonia Progreso Tizapán Álvaro Obregón, 01080, Ciudad de México, México.

DOI: 10.17013/risti.31.66–81

Resumen: El aumento de sanciones por violaciones de la privacidad motiva la definición de una metodología de evaluación de la utilidad de la información y de la preservación de la privacidad de datos a publicar. Al desarrollar un caso de estudio se provee un marco de trabajo para la medición de la preservación de la privacidad. Se exponen problemas en la medición de la utilidad de los datos y se relacionan con la preservación de la privacidad en daos a publicar. Se desarrollan modelos de aprendizaje máquina para determinar el riesgo de predicción de atributos sensibles y como medio de verificación de la utilidad de los datos. Los hallazgos motivan la necesidad de adecuar la medición de la preservación de la privacidad a los requerimientos actuales y a medios de ataque sofisticados como el aprendizaje máquina.

Palabras-clave: aprendizaje máquina; preservación de la privacidad; publicación de datos; medición de la privacidad.

Privacy measurement framework

Abstract: The grown penalties for privacy violations motivate the definition of a methodology for evaluating the usefulness of information and the privacy-preserving data publishing. We developing a case study and we provided a framework for measuring the privacy-preserving. Problems are exposed in the measurement of the usefulness of the data and relate to privacy-preserving data publishing. Machine learning models are developed to determine the risk of predicting sensitive attributes and as a means of verifying the usefulness of the data. The findings motivate the need to adapt the privacy measures to current requirements and sophisticated attacks as the machine learning.

Keywords*:* machine learning; privacy-preserving; data publishing; privacy measurement.

1. Introducción

El uso exponencial de Internet en el mundo ha modificado la forma de compilar, intercambiar y manipular datos. Las velocidades de procesamiento, los volúmenes de información y las relevancias de los contenidos han cambiado rápidamente hasta niveles cada vez menos imaginables. Esto se ve asociado al interés en crear servicios personalizados basados en la información disponible. Diferentes organizaciones tienen datos sobre las personas, estos datos son un elemento vital y se estima que hasta el 80% de todos los datos almacenados en las organizaciones, pueden clasificarse como grandes datos (big data) (Khan et al., 2014).

Los datos que almacenan las organizaciones por lo general provienen de múltiples fuentes (dispositivos, personas, organizaciones, entre otros) y estas fuentes se convierten en los productores de los datos. Estos datos pueden ser publicados o compartidos con otras organizaciones o individuos, de esta forma los receptores de estas publicaciones se convierten en los consumidores de la información. Una publicación de datos en la mayoría de los casos contiene información de múltiples fuentes y muchas organizaciones utilizan análisis de datos (*"data mining"*) para extraer conocimiento relevante (Castillo-Rojas, Medina-Quispe, & Vega-Damke, 2017; Norambuena & Zepeda, 2017). Desde una perspectiva individual esto plantea preguntas acerca de cuánto conocimiento puede ser recabado sobre la vida de una persona, por ejemplo, sobre su situación actual o su paradero (Ohm, 2010). Una vez realizada la publicación de datos, las consecuencias para el dueño o productor de los datos pueden ser muy variadas, exponiendo al individuo a discriminación o escarnio público, entre otros (Martinez et al., 2017).

Las leyes y estándares apenas se mantienen al día con el potencial de invasión a la privacidad. La integración económica y social resultante del funcionamiento de los mercados ha llevado a un aumento sustancial de los flujos transfronterizos de datos y esto converge en el surgimiento de nuevas regulaciones, como por ejemplo, el Reglamento General de Protección de Datos de la Unión Europea (GDPR, por sus siglas en inglés) (El Parlamento Europeo y el Consejo de la Unión Europea, 2016). En consecuencia a estas regulaciones, la Preservación de la Privacidad en Datos a Publicar (PPDP, por sus siglas en inglés) se ha convertido en un área de interés para los investigadores y profesionales.

La PPDP supone que quienes intentan descubrir información confidencial sobre las personas, se pueden encontrar entre los destinatarios de los datos. Por lo tanto, el objetivo de las técnicas de PPDP es modificar los datos haciéndolos menos específicos, de modo que la privacidad de los dueños de los datos esté protegida y a la vez se mantenga la utilidad de los datos tratados. La preservación de la privacidad de los datos requiere del estudio del equilibrio entre el respeto a los deseos o preferencias de privacidad de múltiples dueños de datos, ante una posible inferencia autorizada o no, y la posibilidad de la reidentificación de cada dueño o el enlace de información sensible sobre el conjunto de datos publicado.

Regulaciones como la GDPR definen que las tareas deben estar acotadas, por ejemplo al solicitar el consentimiento informado de todos los posibles escenarios de uso para los datos recopilados. Es importante destacar que esto resulta sumamente difícil, por ejemplo, en iniciativas de datos abiertos, es casi imposible identificar todos los destinatarios y los posibles usos que les den a los datos (Conradie & Choenni, 2012).

Por lo tanto, cualquier publicador de datos necesita aplicar mecanismos de preservación de la privacidad (Ayala-Rivera, McDonagh, Cerqueus, & Murphy, 2014).

La aplicación de formas para medir el potencial abuso y pérdida de la información, mediante la experimentación sobre conjuntos de datos reales, permite brindar garantías científicas sobre la preservación de la privacidad. En este artículo, se estudia el trabajo de la profesora Latanya Sweeney, de la Universidad Carnegie Mellon, y su modelo de preservación de la privacidad conocido como k-anonimato (Sweeney, 2002a). Se aplica de forma práctica una variante multidimensional conocida como algoritmo de Mondrian (LeFevre, DeWitt, & Ramakrishnan, 2006a, 2006b) y se analiza una metodología de evaluación de la utilidad de la información y de la preservación de la privacidad de la información resultante.

La sección 2 describe formalmente la preservación de la privacidad en datos a publicar, así como algunos algoritmos y métricas alrededor de la preservación de la privacidad de los datos. En la sección 3 se describe trabajo relacionado y se enfatizan las contribuciones respecto a iniciativas similares. La evaluación experimental se describe en la sección 4, que incluye la metodología utilizada y el análisis de resultados. Las conclusiones se presentan en la sección 5.

2. Preservación de la privacidad de datos a publicar

La preservación de la privacidad de datos a publicar o PPDP requiere una definición clara de la preservación de la privacidad. En 1977, Dalenius proporcionó una definición muy estricta, donde específica, que el acceso a los datos publicados, no debe permitir que un atacante aprenda algo adicional sobre cualquier víctima objetivo, en comparación con la información que obtendría si no contara con el acceso a la base de datos publicadora, incluso con la presencia de conocimiento previo obtenido de otras fuentes (Dalenius, 1977).

Con base en esta definición, la principal motivación alrededor de la creación de modelos de preservación de la privacidad de datos a publicar es la de proveer garantías científicas. Donde estas permitan asegurar una utilidad práctica de uso sobre el conjunto de datos publicado y un grado de dificultad al intentar realizar una identificación de los dueños de los datos a partir de la información provista. La forma más básica e intuitiva de proveer garantías científicas de preservación de la privacidad es la desidentificación. La cual consiste en eliminar los datos que permiten identificar o relacionar directamente al dueño original de los datos, como por ejemplo eliminar el número de identificación (cédula en el caso de Costa Rica), número de teléfono o el nombre completo. Al proceso mediante el cual se identifican los datos desidentificados es conocido como reidentificación (Bayardo & Agrawal, 2005).

La desidentificación dentro de la PPDP se refiere al proceso mediante el cual un administrador de datos tiene una tabla T, donde cada columna se puede categorizar como identificador directo, cuasi-identificador, sensible o no sensible. Cada columna solo puede pertenecer a una de estas categorías. Las columnas definidas como identificadores directos permiten identificar explícitamente a los dueños de los datos; las columnas el cuasi-identificadoras contienen valores que pueden identificar a los dueños de los datos,

por medio de la vinculación a información externa que permita reidentificar los dueños originales. Si un atributo es cuasi-identificador su característica más importante es qué tan disponibles estén datos externos con una variable que corresponda al potencial valor de la columna cuasi-identificadora. Las columnas sensibles contienen información susceptible a crear algún perjuicio o discriminación específica hacía el dueño de los datos, como por ejemplo la religión, una enfermedad, el salario o el estado de discapacidad, entre otras. Las columnas no sensibles son todas aquellas que no se inscriben en las tres categorías anteriores (Fung, Wang, Fu, & Yu, 2010).

Dado esto, la PPDP toma la tabla *T* (*ID*: identificadores, *CID*: cuasi-identificadores, *S*: Columnas sensibles, *NS*: Columnas no sensibles) y la lleva a un estado *T'* (*CID', S', NSS*). Donde *CID'* y *S'* es una versión anónima del *CID* y *S* originales. Esto al aplicar operaciones de preservación de la privacidad (anonimización) a las columnas en *CID* y *S* de la tabla original *T*. Uno de los modelos más conocidos para preservar la privacidad de datos es el k-anonimato, el cual fue propuesto por primera vez por Samarati y Sweeney en (Samarati & Sweeney, 1998) y extendido por Sweeney en (Sweeney, 2002a). El aporte del trabajo de Sweeney es definir una propiedad de un grupo de datos tal que si es cumplida, dichos datos se pueden publicar con una reducción significativa en la posibilidad de realizar una reidentificación a partir de ellos.

2.1. K-anonimato

El k-anonimato es un marco de desarrollo que trabaja sobre un conjunto de columnas cuasi-identificadoras con un objetivo de privacidad definido por un parámetro *k* (Pierangela Samarati, 2001). El k-anonimato se puede ver como una propiedad de una tabla *T*. Se dice que una tabla *T* es k-anónima si, para cualquier combinación de valores de los campos cuasi-identificadores de la tabla, hay al menos *k* tuplas con la misma combinación (Sweeney, 2002b).

Las tablas no son k-anónimas de forma natural. Por lo tanto, se debe aplicar un proceso para forzar la propiedad sobre ellas. La forma más común de lograr esto es mediante el uso de la generalización y/o la supresión. Para la generalización, los valores se agrupan en clases de equivalencia de acuerdo con algún principio organizativo que depende de las particularidades de la información.

En la Fig. 1 se puede observar una estructura árbol que representa una distribución de clases de equivalencia disponibles para números de 0 a 99. La distribución de clases de equivalencia para una columna *i*, dado su papel en el proceso de generalización de los datos, se conoce como jerarquía de generalización. En la jerarquía de generalización, cada nivel proporciona un conjunto de clases de equivalencia. Estas clases permiten agrupar más o menos registros para obtener el valor *k* solicitado y cada clase de equivalencia no comparte elementos en común con otra clase.

Las operaciones de k-anonimato ocultan información detallada para que los múltiples registros se vuelvan indistinguibles con respecto a los valores en el *CID*. En consecuencia, si el dueño de los datos está vinculado a un registro a través de un valor en el *CID*, el dueño de los datos también está vinculado a todos los demás registros que tienen el mismo valor para el *CID*, lo que hace al enlace ambiguo con los demás dueños de los datos. De esta forma, podemos entender que el problema del k-anonimato es producir

una *T* anónima (*T'*) y que la misma satisfaga un requisito de privacidad determinado por el nivel del valor del *k* en el modelo de privacidad, de forma que retenga la mayor utilidad de datos posible.

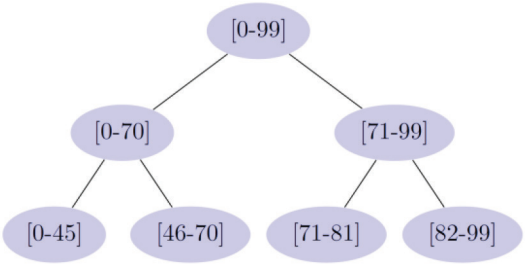

Figura 1 – Jerarquía de generalización con las clases de equivalencia para números de 0 a 99

En este trabajo utilizaremos una versión multidimensional del k-anonimato conocida como Mondrian Básico, propuesta por Kristen LeFevre en (LeFevre et al., 2006a) y modificada en (LeFevre et al., 2006b).

2.2. K-anonimato multidimensional – Mondrian Básico

Mondrian es un algoritmo de preservación de la privacidad de datos voraz descendente que permite la generalización de atributos *CID* de forma multidimensional. La versión inicial del algoritmo está diseñada para trabajar solo con atributos numéricos (LeFevre et al., 2006a). En un trabajo posterior LeFevre propuso una modificación a su algoritmo para que aceptara atributos categóricos, además de numéricos (LeFevre et al., 2006b).

Para satisfacer el requerimiento del k-anonimato sobre la tabla *T*, se debe validar que para cada valor distinto en una columna CID, este valor se encuentra *k* veces. En muchos casos esto no se cumple en la tabla original y mucho menos si se cuenta con varias columnas *CID*. Por este motivo la visión multidimensional de los datos ayuda a satisfacer el requerimiento de forma más rápida. Permitiendo que si tenemos como *CID* a varias columnas, y si por cada tupla única al seleccionar estas columnas sobre *T* se producen al menos *k* tuplas iguales, se dice que cumple el principio de k-anonimato en el grado del *k* reportado. Es decir, el tamaño de cada clase de equivalencia en *T* con respecto a las columnas es al menos de tamaño *k* o en otras palabras el tamaño de la clase de equivalencia se determina contando el número de tuplas que son agrupadas mediante ella (Fung et al., 2010).

La utilidad de datos de Mondrian se puede ver afectada por la distribución de los datos y el mecanismo utilizado para la creación de particiones (Ayala-Rivera et al., 2014). Por este motivo es importante contar con alguna métrica que verifique la utilidad de los datos y su preservación de la privacidad. En general se crea una pérdida de información cada vez que un valor en una columna *CID* se generaliza a una clase de equivalencia mayor en la jerarquía de generalización, esto debido a que cada clase de equivalencia es más genérica que la anterior. El objetivo de un buen algoritmo de PPDP es encontrar

una transformación de los datos originales, de forma que satisfaga un requerimiento de privacidad al tiempo que minimice la pérdida de información y maximice la utilidad de los datos resultantes. Por lo tanto, una medida es necesaria para indicar la calidad y la preservación de la privacidad de los datos.

2.3. Métricas de privacidad

En la literatura se presentan varias métricas que evalúan la calidad de los datos utilizando medidas simples basadas en el tamaño de la clase de equivalencia o el número total de generalizaciones (Aggarwal et al., 2005; Bayardo & Agrawal, 2005; LeFevre et al., 2006a; Pierangela Samarati, 2001; Sweeney, 2002a). Dado el efecto del uso de jerarquías de generalización, en este trabajo es de interés la métrica de *Penalización de Certeza Normalizada* (*NCP*, por sus siglas en inglés) (Xu et al., 2006). Está métrica intenta capturar la incertidumbre causada por la generalización de cada clase de equivalencia en el espacio de columnas *CID*. Para un valor continuo u ordinal en una columna *CID*, el NCP en una clase de equivalencia *G* se define como la división de los rangos de valores de la columna *CID* en los que se agrupa el valor del campo en la clase de equivalencia *G, entre* el dominio completo de la columna *CID*. En el caso de los valores de columnas categóricas, donde no exista un orden total o alguna función de distancia, el NCP se define con respecto a la jerarquía de generalización de la columna *CID*. Podemos expresar el NCP como cero cuando el número de hojas (es decir, el número de valores agrupados de la columna) en el subárbol que contiene el valor del campo es igual a uno y en caso contrario el NCP es el resultado de la división del número de hojas entre el número total de valores distintos en la columna categórica.

El valor NCP de la clase de equivalencia sobre todas las columnas *CID* es el resultado de la sumatoria de *i* igual a 1 hasta el número de columnas continuas, ordinales o categóricas del NCP de la clase de equivalencia. NCP mide la pérdida de información para una única clase de equivalencia y caracteriza la pérdida de información de una partición completa al sumar el NCP de todas las tuplas en cada clase de equivalencia. Para este trabajo adoptamos la formulación normalizada de la versión agregada de NCP, llamada Penalización de Certeza Global (Global Certainty Penalty - GCP) (Ghinita, Karras, Kalnis, & Mamoulis, 2009). Está mide la pérdida de información de toda la tabla anonimizada, tomando *P* como el conjunto de todas las clases de equivalencia en la tabla anonimizada y se define como:

$$GCP(P) = \frac{\sum_{G \in P} |G| * NCP(G)}{d * N}$$

Figura 2 – Ecuación GCP

En la Fig. 2 tenemos que *d* representa todas las columnas *CID* y *N* el número de registros en la tabla original *T*, |*G*| es la cardinalidad de la clase de equivalencia *G*. La ventaja de esta ecuación es la capacidad de medir la pérdida de información entre tablas con cardinalidad y dimensionalidad variable. El rango de valores de GCP está entre 0 y 1, donde 0 significa que no existe pérdida de información y 1 corresponde a la pérdida

total de información. Por facilidad de uso, en este trabajo el GCP se calcula dividiendo el valor de GCP original con el número de valores en el conjunto de datos para pasarlo a porcentaje.

La importancia de una métrica de privacidad está vinculada de manera innata a los cálculos que se pueden realizar sobre los datos y el objetivo de la PPDP es producir conjuntos de datos que tengan una "buena" utilidad para una gran variedad de trabajos.

El que esa variedad de trabajos sea desconocida es una premisa esencial para muchos investigadores, pero regulaciones como la GDPR (El Parlamento Europeo y el Consejo de la Unión Europea, 2016) se traen abajo esta premisa al solicitar informarle al usuario los procesos que se realizarán con sus datos. Al conocer de antemano los trabajos a realizar, el publicador de los datos simplemente puede ejecutar los trabajos en los datos originales y publicar solo los resultados o publicar una versión anonimizada que no permita realizar los trabajos que el dueño de los datos no autorizó. Otros investigadores han utilizado el aprendizaje máquina para medir la utilidad de los datos (Chen, LeFevre, & Ramakrishnan, 2008; LeFevre et al., 2006b). En este documento se propone utilizar el aprendizaje máquina no solo para medir la utilidad, sino también la preservación de la privacidad.

2.4. Aprendizaje máquina

El aprendizaje máquina permite en muchos casos extraer información útil, interesante y previamente desconocida de grandes conjuntos de datos. El éxito siempre se basa en la disponibilidad de datos de alta calidad y el intercambio efectivo de estos. Se ha logrado un impulso en la publicación de datos tanto por beneficio mutuo, como por regulaciones que obligan a la publicación de ciertos datos (Conradie & Choenni, 2012). La publicación de datos es omnipresente en muchos dominios. Por ejemplo, en el 2006, el proveedor de Internet AOL lanzó un conjunto de datos que contenían 3 meses de búsquedas de 650 000 usuarios. Los nombres fueron enmascarados con identificadores aleatorios y aun así, en cuestión de días, un reportero del New York Times identificó a Thelma Arnold, una viuda de 62 años como el usuario 4417749 (Barbaro & Zeller, 2006). Situaciones como esta han impulsado a los investigadores en PPDP a utilizar algoritmos de aprendizaje de clasificación y regresión en sus investigaciones (Chen et al., 2008; LeFevre et al., 2006b; Machanavajjhala et al., 2007).

En estos modelos de aprendizaje máquina, los atributos normalmente se caracterizan en al menos uno de los siguientes tipos (Han, Kamber, & Pei, 2012):

- *Atributo objetivo o de interés*: es el atributo nominal cuyo valor busca predecir con precisión el modelo de clasificación construido. En el caso de la regresión, es el atributo numérico cuyo valor tiene como propósito predecir el modelo de regresión construido.
- *Atributos de predicción*: son el conjunto de atributos (discretos o continuos, según el algoritmo) que se utilizan como entradas para construir el modelo que intenta predecir el atributo objetivo.

La utilidad requiere potenciar el uso de la información disponible y por ello se debe tomar en cuenta que los atributos no sensibles se publican si son importantes para la tarea de

minería de datos. El atributo objetivo en la minería es definido como no sensible en el proceso de preservación de la privacidad para que conserve su semántica. Por lo general cuando se considera un algoritmo de clasificación o de regresión junto con la PPDP, cada atributo tiene solo dos caracterizaciones (atributo de predicción y atributo objetivo). En el resto de este documento, se asumirá que el conjunto de atributos de predicción es un conjunto de cuasi-identificadores (*CID*). Bajo esta suposición, puede parecer contradictorio mantener la categoría de atributo sensible, pero para efectos de este trabajo se utiliza como atributo objetivo cuándo se evalúa la preservación de la privacidad.

El transformar un atributo sensible en un atributo objetivo permite medir la precisión del modelo de aprendizaje máquina creado para predecirlo y por lo tanto medir la preservación de la privacidad en el conjunto de datos en estudio. De esta manera en unos casos se toma como atributo objetivo el atributo sensible y en otros como atributo objetivo un atributo de interés hipotético. La evaluación del uso de modelos de aprendizaje máquina como métrica de preservación de la privacidad se llevara a cabo en la siguiente sección.

3. Trabajo relacionado

La selección de un algoritmo apropiado para proteger la privacidad cuando se difunden datos es una preocupación general para la PPDP. Como resultado, la comparación de múltiples algoritmos de preservación de la privacidad desde la perspectiva de la utilidad de los datos y la efectividad en la preservación de la privacidad, representa un importante trabajo de investigación. Algunos autores discuten que la eficacia de la preservación de la privacidad se evalúa mejor por la utilidad que proporciona a las aplicaciones de destino y que las métricas pueden comportarse de manera diferente con diferentes algoritmos de preservación de la privacidad (Nergiz & Clifton, 2007). Además de esto, se presenta un marco de evaluación en (Bertino, Fovino, & Provenza, 2005) para estimar y comparar diferentes tipos de algoritmos de preservación de la privacidad específicamente en la minería de datos. Aun cuando estos estudios están estrechamente relacionados con la PPDP, las tareas de minería de datos que se consideran están estrechamente relacionadas con las soluciones propuestas.

La PPDP publica los datos modificados mediante algoritmos de preservación de la privacidad a múltiples destinatarios que pueden usar los datos de muchas maneras diferentes. Por lo tanto, no sería adecuado evaluar los métodos de preservación de la privacidad utilizando estudios comparativos que solo tengan en cuenta métricas con fines específicos (es decir, dependientes de la aplicación). Esto se debe a que las métricas que tienen en cuenta un escenario de uso particular solo pueden capturar la utilidad de los datos protegidos según los requisitos para ese escenario. En cambio, en (Ayala-Rivera et al., 2014) se indica que un conjunto de métricas que pueden aplicarse a la mayoría de los escenarios de publicación proporciona un mejor enfoque para realizar una comparación sistemática. Aunque este mismo trabajo hace énfasis en las variaciones del rendimiento de los algoritmos, su aporte es importante, pero escaso de discusión sobre la incompetencia de algunas métricas en los distintos contextos.

El objetivo de las métricas de preservación de la privacidad es medir el grado y la cantidad de protección que ofrecen las tecnologías de preservación de la privacidad.

De esta forma, las métricas de preservación de la privacidad contribuyen a mejorar la privacidad de los dueños de los datos. La diversidad y la complejidad de las métricas de preservación de la privacidad en la literatura hacen que una elección informada de métricas sea desafiante. Como resultado, en lugar de utilizar métricas existentes, se proponen nuevas métricas con frecuencia, y los estudios de privacidad a menudo son incomparables (Wagner & Eckhoff, 2018).

En (Fung et al., 2010) se afirma que para abordar el objetivo de clasificación, la distorsión debe medirse por error de clasificación en casos futuros y sugieren que el conocimiento de clasificación útil es capturado por diferentes combinaciones de atributos. La generalización y la supresión pueden destruir algunas de estas "estructuras de clasificación" útiles, pero pueden surgir otras estructuras útiles para ayudar. En algunos casos, la generalización y la supresión pueden incluso mejorar la precisión de la clasificación, al eliminar ruido en los datos. También se afirma que es esencial evaluar experimentalmente el impacto de la preservación de la privacidad mediante la construcción de un clasificador a partir de los datos protegidos y ver cómo funciona en los casos de prueba. Algunos trabajos (Iyengar, 2002; LeFevre et al., 2006b) han llevado a cabo o analizan tales experimentos, aunque en la práctica general se utiliza alguna métrica especifica (Ghinita et al., 2009; Gong, Luo, Yang, Ni, & Li, 2017).

Este enfoque no es abordado al analizar métricas y modelos de privacidad, por ejemplo en (Wagner & Eckhoff, 2018) analizan una selección de más de ochenta métricas de privacidad e introducen caracterizaciones basadas en los aspectos de la privacidad que miden, sus requerimientos de entrada y el tipo de datos que protegen. Además, presentan un método sobre cómo elegir las métricas de privacidad basadas en nueve preguntas que ayudan a identificar las métricas de privacidad adecuadas según un escenario dado. Pero rescatan que se necesita trabajo adicional sobre métricas de privacidad. Esto fundamenta la necesidad de una metodología de evaluación de la preservación de la privacidad como la presentada en este artículo.

4. Evaluación experimental

La evaluación experimental tiene como objetivo proporcionar una metodología de evaluación de la utilidad y la preservación de la privacidad. Se describen los pasos de un protocolo experimental para evaluar un algoritmo de preservación de la privacidad con respecto a un conjunto de datos y se desarrolla un caso de estudio que permite comparar los resultados de varios modelos de aprendizaje máquina creados a partir de publicaciones construidas con diferentes niveles de privacidad y distintas configuraciones de cuasi-identificadores.

4.1. Metodología

Una forma de evaluar la calidad de la preservación de la privacidad es creando un modelo de predicción utilizando los datos anonimizados y luego evaluar el modelo según la precisión en la predicción tanto del atributo sensible, como del atributo de interés para el estudio. Para lograr este propósito se presenta la siguiente metodología de evaluación de la utilidad y la preservación de la privacidad:

1. Preprocesado de los datos.
2. Categorización de los datos en atributos identificadores, cuasi-identificadores, sensibles y no sensibles.
3. Definición de niveles de preservación de la privacidad e identificación de los cuasi-identificadores a utilizar.
4. Anonimización de los datos por medio del algoritmo de preservación de la privacidad (Mondrian Básico).
5. Construcción de conjunto de datos que comprende los datos anonimizados, los datos no sensibles, el atributo de interes y el atributo sensible.
6. Preprocesado y recodificación del conjunto de datos generado para poder ser utilizado en el algoritmo de aprendizaje máquina.
7. Categorización de los atributos como atributos de predicción y atributos objetivo.
8. Segmentación del conjunto en datos de entrenamiento y de prueba.
9. Generación del modelo de predicción mediante el algoritmo de aprendizaje máquina seleccionado.
10. Evaluación de la precisión de la predicción del modelo generado.

Está metodología propuesta para la evaluación de la utilidad de los datos y la preservación de la privacidad de los datos se utiliza en un caso de estudio práctico para demostrar su usabilidad. El caso de estudio desarrollado mantiene el atributo de interés y el atributo sensible en los datos a publicar durante las fases de preservación de la privacidad y durante la creación del modelo de aprendizaje máquina.

Se utiliza el algoritmo Mondrian Básico implementado en el lenguaje Python como algoritmo de preservación de la privacidad, así como implementaciones de los algoritmos de aprendizaje máquina proporcionados por el paquete de software "scikit-learn" (Pedregosa et al., 2011). Específicamente arboles de decisión, bosques aleatorios y regresión logística. El conjunto de datos de prueba es el AdultDatabase (Dheeru & Karra Taniskidou, 2017). Este conjunto contiene 32 561 filas con 15 atributos, una vez finalizado el preprocesado quedaron 30 162 filas. Se definió como atributo sensible (no cuasi-identificador): "race", como atributo de interés (no cuasi-identificador): "income" y se define como atributo no sensible: "education" (para asegurar un grado de utilidad constante, esto aunque es equivalente a el atributo "education-num"). Los demás atributos se utilizan como posibles cuasi-identificadores.

En los experimentos, se varió el tamaño del conjunto de atributos cuasi-identificadores en un rango de 1, 6 y 12. Cuando se modifica el número de atributos cuasi-identificadores, el atributo que deja de ser cuasi-identificador automáticamente pasa a ser no sensible para mantener un mismo número de entradas en los algoritmos de aprendizaje máquina. Las jerarquías de generalización para los atributos categóricos están disponibles en desarrollos de código abierto de GitHub. También, se definió como nivel de privacidad k el rango de valores: 2, 5, 10, 25, 50 y 100.

4.2. Análisis de resultados

El caso de estudio sobre la metodología propuesta incluye evaluar la utilidad de los datos y la preservación de la privacidad de los datos a publicar, para esto se crearon modelos de predicción utilizando los algoritmos de aprendizaje máquina seleccionados.

Se distinguen dos tipos de modelos creados, los que realizan predicción de la variable de interés para comprobar la utilidad de los datos y los que realizan predicción de la variable sensible para comprobar la preservación de la privacidad. Como se observa en la Fig. 3, los árboles de decisión tienen una pérdida en su precisión en la utilidad de los datos (predicción de variable de interes) al aumentar el nivel de privacidad k con doce cuasi-identificadores. Este comportamiento resulta normal si se considera que los datos son generalizados a partir de alguna de las clases de equivalencia. Lo que no resulta consecuente es que la pérdida de precisión no continua aumentando. Podemos observar que en general la precisión de la predicción del árbol de decisión no sufre variaciones conforme se cambia el nivel de privacidad.

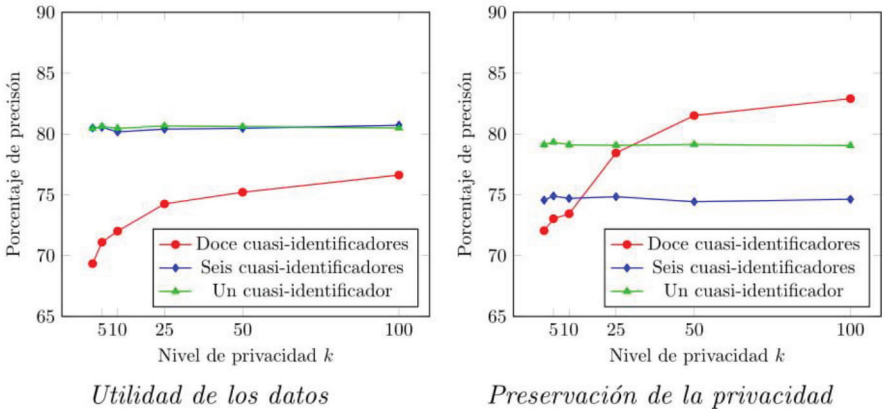

Figura 3 – Evaluación de modelos creados con árboles de decisión

También se puede observar en la Fig. 3 que los modelos que miden la preservación de la privacidad por medio de árboles de decisión mantienen una precisión superior al 70%. Si bien la medición de la precisión demuestra cambios respecto a los modelos que miden la utilidad, se puede observar que ambos obtienen comportamientos similares y con resultados superiores al 70%.

El comportamiento de la precisión en la predicción de los modelos creados con el algoritmo de bosques aleatorios se presenta en la Fig. 4. Se observa que la variación de la precisión para la medición de la utilidad de los datos es similar a la de los modelos creados con el algoritmo de árboles de decisión.

Los modelos creados con el algoritmo de regresión logística comparten similitud con los modelos evaluados anteriormente. Se observa en la Fig. 5 que todos los modelos creados tienen una precisión superior al 75%. Este resultado es consecuente con los arrojados por los modelos evaluados anteriormente y permiten teorizar que se mantiene una utilidad buena de los datos. Al mismo tiempo se puede teorizar que la preservación de la privacidad es muy baja. Esto al destacar que ni el número de cuasi-identificadores utilizado, ni el nivel de preservación de la privacidad parece influir en el resultado de la predicción para el dato sensible.

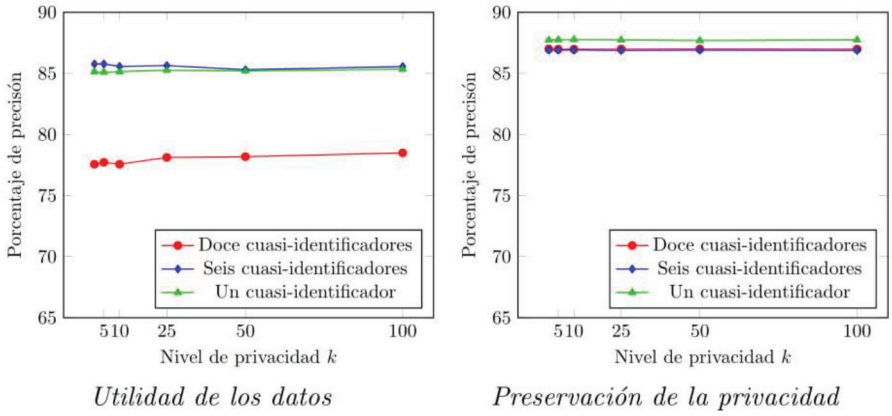

Figura 4 – Evaluación de modelos creados con bosques aleatorios

Destaca que la preservación de la privacidad no cambia significativamente al variar el nivel de privacidad k o el número de cuasi-identificadores. Al tomar como métrica de privacidad los resultados obtenidos por los algoritmos de aprendizaje máquina, podemos indicar que la precisión en la predicción sobre el atributo sensible es muy alta. Dados los resultados obtenidos se puede indicar que existe una preservación de la privacidad muy débil.

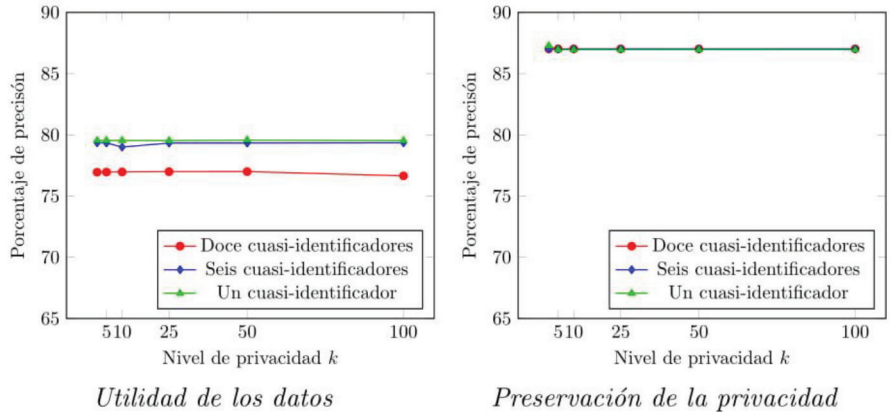

Figura 5 – Evaluación de modelos creados con regresión logística

En particular, como se explicó en la sección 2, es importante contar con una métrica de preservación de la privacidad. Para contrastar los resultados obtenidos por los modelos de aprendizaje máquina se evaluó la métrica GCP. Los resultados de la métrica GCP en la Fig. 6 indican que la pérdida de información puede llegar a 40% y este resultado no es consecuente con los resultados de los modelos creados con los algoritmos de aprendizaje máquina.

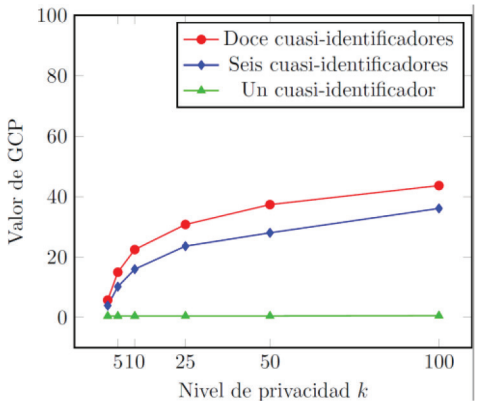

Figura 6 – Relación entre el nivel de privacidad k, el número de cuasi-identificadores y el valor obtenido por la métrica GCP

La métrica GCP reporta un incremento en la pérdida de información al modificar el nivel de privacidad k, situación que no se ve reflejada en la precisión de la predicción de los algoritmos de aprendizaje máquina. Es probable que la reducción de los dominios de los atributos influyera en la conservación de la precisión en las predicciones. Otro aspecto es que la métrica GCP no toma en cuenta la diversidad de valores en el atributo sensible y eso disminuye la calidad de la medición.

Dados estos resultados se puede afirmar que los modelos creados mediante los algoritmos de aprendizaje máquina mantienen el mismo porcentaje de éxito en la predicción aun cuando los datos han sido procesados por un modelo de preservación de la privacidad como Mondrian Básico.

5. Conclusiones

La utilidad de los datos es un aspecto fundamental para motivar el uso de técnicas de preservación de la privacidad, pero al mismo tiempo es posible que sean la llave para respetar los deseos o derechos de los dueños de los datos. Los algoritmos de preservación de la privacidad en datos a publicar son ampliamente estudiados en la búsqueda de proteger la privacidad pero resulta complicado indicar que datos corren más riesgo de reidentificación. Este artículo proporciona indicios sobre la existencia de dificultades para establecer buenos parámetros en algoritmos de privacidad como Mondrian Básico. El estudio experimental genera evidencia sobre la poca efectividad de la métrica GCP respecto a la utilidad de los datos anonimizados en modelos de aprendizaje máquina que efectúen predicciones. Además, la poca efectividad del GCP está presente tanto en la predicción del atributo de interés (midiendo la utilidad), como sobre el atributo sensible (midiendo la preservación de la privacidad).Los algoritmos de aprendizaje máquina se plantean como un instrumento para asegurar la privacidad y se espera que la metodología sea base para un aseguramiento de la privacidad más riguroso.

Agradecimientos

Un agradecimiento especial a la Profesora Ileana Castillo Arias de la UCR por todas sus recomendaciones y revisiones. Este trabajo fue apoyado por el Programa de Posgrado en Computación e Informática (PCI), la Escuela de Ciencias de la Computación e Informática (ECCI), el Centro de Investigaciones en Tecnología de la Información y la Comunicación (CITIC)), y el Sistema de Estudios de Posgrado (SEP) todos en la Universidad de Costa Rica (UCR). Así como también por el Ministerio de Ciencia, Tecnología y Telecomunicaciones (MICITT), y por el Consejo Nacional para Investigaciones Científicas y Tecnológicas (CONICIT) del Gobierno de Costa Rica.

Referencias

Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D., & Zhu, A. (2005). Anonymizing tables. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 3363, 246–258. Doi: 10.1007/978-3-540-30570-5_17.

Ayala-Rivera, V., McDonagh, P., Cerqueus, T., & Murphy, L. (2014). A Systematic Comparison and Evaluation of k -Anonymization Algorithms for Practitioners. Transactions on Data Privacy, 7(3), 337–370.

Barbaro, M., & Zeller, T. (2006). A Face Is Exposed for AOL Searcher No. 4417749. New York Times, (4417749), 1–3. Doi: 4417749.

Bayardo, R. J., & Agrawal, R. (2005). Data privacy through optimal k-anonymization. In: Proceedings - International Conference on Data Engineering (pp. 217–228). Doi: 10.1109/ICDE.2005.42.

Bertino, E., Fovino, I. N., & Provenza, L. P. (2005). A Framework for Evaluating Privacy Preserving Data Mining Algorithms. Data Mining and Knowledge Discovery, 11(2), 121–154. Doi: 10.1007/s10618-005-0006-6.

Brickell, J., & Shmatikov, V. (2008). The cost of privacy: destruction of data-mining utility in anonymized data publishing. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 70–78. Doi: 10.1145/1401890.1401904.

Castillo-Rojas, W., Medina-Quispe, F., & Vega-Damke, J. (2017). Esquema de visualización para modelos de clústeres en minería de datos. RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao, (21), 67–84. Doi: 10.17013/risti.21.67-84.

Chen, B.-C., LeFevre, K., & Ramakrishnan, R. (2008). Adversarial-knowledge dimensions in data privacy. The VLDB Journal, 18(2), 429–467. Doi: 10.1007/s00778-008-0118-x.

Conradie, P., & Choenni, S. (2012). Exploring process barriers to release public sector information in local government. In: Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance - ICEGOV '12 (p. 5). New York, USA: ACM Press. Doi: 10.1145/2463728.2463731.

Dalenius, T. (1977). Towards a methodology for statistical disclosure control. Statistik Tidskrift, 15, 429–444. Doi: 10.1145/320613.320616.

Dheeru, D., & Karra Taniskidou, E. (2017). UCI Machine Learning Repository. Retrieved from: http://archive.ics.uci.edu/ml.

El Parlamento Europeo y el Consejo de la Unión Europea. (2016). Reglamento (UE) 2016/679 del parlamento europeo y del consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la D. Diario Oficial de La Unión Europea, 2014(119), 1–88.

Fung, B., Wang, K., Fu, A., & Yu, P. (2010). Introduction to Privacy-Preserving Data Publishing (Vol. 17). CRC Press. Doi: 10.1201/9781420091502.

Ghinita, G., Karras, P., Kalnis, P., & Mamoulis, N. (2009). A framework for efficient data anonymization under privacy and accuracy constraints. ACM Transactions on Database Systems, 34(2), 1–47. Doi: 10.1145/1538909.1538911

Gong, Q., Luo, J., Yang, M., Ni, W., & Li, X. B. (2017). Anonymizing 1:M microdata with high utility. Knowledge-Based Systems, 115, 15–26. Doi: 10.1016/j.knosys.2016.10.012.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. San Francisco, CA: Morgan Kaufmann Publishers. Doi: 10.1016/B978-0-12-381479-1.00001-0.

Iyengar, V. S. (2002). Transforming data to satisfy privacy constraints. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02 (p. 279). New York, USA: ACM Press. Doi: 10.1145/775047.775089.

Khan, N., Yaqoob, I., Hashem, I. A. T., Inayat, Z., Mahmoud Ali, W. K., Alam, M., … Gani, A. (2014). Big Data: Survey, Technologies, Opportunities, and Challenges. The Scientific World Journal, 2014, 1–18. Doi: 10.1155/2014/712826.

LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2006a). Mondrian multidimensional K-anonymity. In: Proceedings - International Conference on Data Engineering (Vol. 2006, p. 25). IEEE.

LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2006b). Workload-aware anonymization. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06 (p. 277). New York, USA: ACM Press. Doi: 10.1145/1150402.1150435.

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). L -diversity. ACM Transactions on Knowledge Discovery from Data, 1(1), 3–es. Doi: 10.1145/1217299.1217302.

Martinez, F. R. C., Candelaria, A. D. H., Lozano, M. A. R., Zúñiga, A. R. R., Peláez, R. M., & Michel, J. R. P. (2017). Después de presionar el botón enviar, se pierde el control sobre la información personal y la privacidad: Un caso de estudio en México. RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao, (21), 115–128. Doi: 10.17013/risti.21.115-128.

Nergiz, M. E., & Clifton, C. (2007). Thoughts on k-anonymization. Data & Knowledge Engineering, 63(3), 622–645. Doi: 10.1016/J.DATAK.2007.03.009

Norambuena, B. K., & Zepeda, V. V. (2017). Minería de procesos de software: Una revisión de experiencias de aplicación. RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao, 21(21), 51–66. Doi: 10.17013/risti.21.51-66

Ohm, P. (2010). Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. UCLA Law Review, 57, 1701. Retrieved from: http://www.uclalawreview.org/?p=1353.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res., 12, 2825–2830. Retrieved from: http://dl.acm.org/citation.cfm?id=1953048.2078195.

Samarati, P., (2001). Protecting respondents' identities in micro- data release. IEEE Transactions on Knowledge and Data Engineering, 13(6), 1010–1027.

Samarati, P., & Sweeney, L. (1998). Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Suppresion. In: Proceedings of the IEEE Symposium on Research in Security and Privacy, (pp. 384–393). Doi: 10.1145/1150402.1150499.

Sweeney, L. (2002a). Achieving k-anonymity Privacy Protection using Generalization and Suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 571–588. Doi: 10.1142/S021848850200165X.

Sweeney, L. (2002b). k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst., 10(5), 557--570. Doi: 10.1142/S0218488502001648.

Wagner, I., & Eckhoff, D. (2018). Technical Privacy Metrics: a Systematic Survey. ACM Computing Surveys, 51(3), 1–38. Doi: 10.1145/3168389.

Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., & Fu, A. W.-C. (2006). Utility-based anonymization using local recoding. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06 (Vol. 18, p. 785). JMLR.org. Doi: 10.1145/1150402.1150504.