CMA IADCOM - UBA CENTRO DE INVESTIGACIÓN EN METODOS CUANTITATIVOS APLICADOS A LA ECONOMÍA Y LA GESTIÓN

Universidad de Buenos Aires Facultad de Ciencias Económicas

http://www.economicas.uba.ar/institutos_y_centros/revista-modelos-matematicos/

GESTIÓN DE LA PRIVACIDAD DE DATOS PERSONALES: EL MODELO DE PRIVACIDAD DIFERENCIAL

Salaberry, Natalia

Universidad de Buenos Aires, Facultad de Ciencias Económicas, Av. Córdoba 2122 – 1120AAQ. Ciudad Autónoma de Buenos Aires, República Argentina

natalia.salaberry@economicas.uba.ar

Resumen

Recibido: 14-06-2021

Aceptado: 13-09-2021

Palabras clave

Gestión, Privacidad, Datos Personales, Privacidad Diferencial.

Este trabajo tiene por objetivo la implementación del modelo de Privacidad Diferencial para la protección de datos personales en el marco de un sistema de gestión responsable. La utilización de tecnología para el procesamiento de datos ha puesto de manifiesto la granularidad de la información recopilada lo que representa un riesgo en la privacidad de su titular. Esta es concebida como el derecho que tiene un individuo para poder controlar el manejo que se hace de sus propios datos. La concepción de la privacidad desde el diseño y por defecto, representa un marco acorde para la constitución de una gestión responsable de la privacidad de datos.

En este contexto, la implementación del modelo - mediante un caso de aplicación - resulta oportuno para poder ver la eficacia de este. En función de los resultados se logra obtener datos ruidosos o distorsionados sin que se afecte significativamente la distribución original de los mismos con el fin de evitar la re identificación del titular. De esta manera, se logra demostrar que la aplicación de privacidad diferencial cumple con el objetivo de protección de datos a la vez que no afectará un análisis posterior.

ISSN (En línea) 2362 3225

MANAGEMENT OF PERSONAL DATA PRIVACY: THE DIFFERENTIAL PRIVACY MODEL

Abstract

KEYWORDS

Management, Privacy, Personal Data, Differential Privacy.

This work aims to implement the Differential Privacy Model for the protection of personal data within the framework of a responsible management system. The use of technology for data processing has revealed the granularity of the information collected, which represents a risk to the privacy of its owner. This is conceived as an individual's right to control the handling of his or her own data. The conception of privacy from the design and by default, represents a suitable framework for the constitution of a responsible management of data privacy. In this context, the implementation of the model - through an application case - is timely to be able to see the effectiveness of the model. Based on the results, noisy or distorted data can be obtained without significantly affecting the original distribution of the data to avoid re-identification of the holder. In this way, it is possible to demonstrate that the differential privacy application meets the data protection objective while not affecting a subsequent analysis.

ISSN (En línea) 2362 3225

INTRODUCCIÓN

La explosión del desarrollo tecnológico durante el presente siglo ha facilitado la posibilidad de contar con grandes volúmenes de datos. Esto dio lugar a la conformación de un ecosistema de datos, sobre el que las diferentes organizaciones se encuentran cada vez más interesadas, en la medida que les brinda la oportunidad de obtener más información para mejorar su estrategia de toma de decisiones. Entre los diversos datos recolectados, los datos sensibles o personales han acaparado el centro de la escena.

La utilización de tecnologías para el procesamiento de datos ha puesto de manifiesto la granularidad de la información recopilada ya que facilitan datos útiles (personales entre otros) sobre la población objetivo. Surge, entonces, un primer riesgo asociado a la privacidad de su titular y con ello la necesidad de establecer metodologías que permitan garantizar su protección.

Es en este sentido que el presente trabajo persigue el objetivo de evaluar la eficacia del modelo de Privacidad Diferencial como metodología a ser aplicada para una eficiente protección de datos. De esta manera, se demuestra la potencialidad que posee este tipo de técnica, al mismo tiempo que se pone de relieve el marco bajo el cual debe llevarse a cabo. En este contexto, surge un interrogante principal a responder: ¿Es la Privacidad Diferencial una metodología que permite dar garantía de protección en el marco de una gestión responsable de datos?

Para poder brindar respuesta al interrogante planteado, en primer lugar, se aborda el concepto de privacidad desde dos enfoques: uno ético y otro normativo. En el primer apartado se desarrollan ambos, arribando a la conclusión de que si bien estos son necesarios no resultan suficientes a la hora de llevar a cabo una implementación dado que solo se limitan a establecer el deber ser, pero no el cómo debe llevarse a cabo. Entonces, la concepción de la privacidad desde el diseño y por defecto, representa un marco acorde para la constitución de una gestión responsable de la privacidad de datos.

En el segundo apartado, se introduce el concepto de Privacidad Diferencial. La privacidad diferencial es una metodología que consiste en aplicar ruido a los datos con el objetivo de generar cierta distorsión en los mismos como forma de protección. Su modelo matemático probabilístico, brinda la posibilidad de evitar la reidentificación del titular. No obstante, requerirá de la evaluación de un costo, que vendrá dado por cuanta información real se está dispuesto a no publicar o utilizar a cambio de obtener privacidad. De esta manera, el fin último es obtener información sobre un conjunto de individuos no así sobre uno en particular.

Finalmente, se lleva a cabo la implementación del modelo mediante un ejemplo de aplicación. Como resultado se logra dar cuenta de que, a medida que disminuye el valor de pérdida de privacidad, la variabilidad media que existe entre el verdadero valor y el valor obtenido de aplicar

privacidad diferencial crece. Pero a su vez, a mayor tamaño de muestra, e igual valor de pérdida de privacidad, el error disminuye. De esta manera, se logra observar la efectividad de la aplicación del modelo de Privacidad Diferencial para la protección de datos.

1. GOBERNANZA DE DATOS

Dado el gran volumen de datos (*Big Data*) disponibles, comienza**n** a surgir inconvenientes para poder administrarlos afectando la toma de decisiones en las organizaciones. Esto se debe a la inexistencia de procesos y políticas que permitan garantizar la confiabilidad en los datos. Es entonces como el tratamiento de grandes volúmenes de datos a puesto de manifiesto que un plan de gobernanza ya no es suficiente para garantizar la calidad de los datos, sino que, resulta necesario que los datos estén disponibles en tiempo adecuado, sean confiables, significativos y suficientes (Kim y Cho, 2018).

Un sistema de gobierno de datos operativo permite establecer un marco para la transparencia y usabilidad de los datos, a partir de determinar las estrategias de recolección, procesos de integración y, finalmente, la gestión de la información obtenida. Permite llevar adelante una comunicación sobre conceptos complejos y ambiguos dentro de una organización (Martínez, 2012). Esto se debe a que a menudo los datos no son verificados, son redundantes, incompletos y están peligrosamente desactualizados. En respuesta a esta problemática una puesta en práctica de un gobierno de datos surge como necesaria para controlarlos.

El principal objetivo, entonces, de un sistema de gobierno es la creación de nuevo valor a partir de los datos en línea con los objetivos de la organización. Para lograr este objetivo, el sistema debe contar con cuatro aspectos principales en su diseño: estructura, responsabilidad de vigilancia, talento y cultura e infraestructura (Deloitte, 2011). Por estructura se hace referencia al establecimiento de reglamentos claros que permitan establecer un diseño de gobierno adecuado a los objetivos de la organización. En cuanto la funcionalidad de vigilancia refiere a delinear las políticas que determina la junta directiva para especificar las funcionalidades administrativas que hacen a las prácticas para llevar adelante los objetivos establecidos. En cuanto al talento y la cultura, busca alinear los principios con las creencias centrales para establecer una cultura de trabajo. Y, por último, en cuanto a la infraestructura, se hace referencia al establecimiento manuales de políticas y procedimientos que permitan alinear la tecnología con el sistema de gobierno de datos.

En este sentido, un sistema de gobierno de datos podría decirse que posee tres pilares fundamentales: las personas, los procesos y la tecnología. A partir de ello, se establece un ciclo funcional que comienza por la identificación de desafíos, con el fin de establecer el impacto sobre

los objetivos permitiendo definir prioridades. En una segunda etapa, se desarrollan políticas y procedimientos alineados con las necesidades asociadas. Luego se procede con la ejecución de tales políticas con el fin de lograr los objetivos planteados. Y, finalmente, se realiza monitoreo y medición de resultados. Este ciclo, es de naturaleza continua en función de la dinámica del ciclo de vida de los datos para la toma de decisiones.

Es así como, el dato se convierte en un activo que debe ser gestionado formalmente en toda la organización. Ello implica que las personas asuman responsabilidad frente a cualquier situación adversa en su calidad, al mismo tiempo que deben brindar confiabilidad sobre estos. Es en este sentido que la tecnología colabora en el proceso de forma tal que ayuda a manejar la información para que pueda ser utilizada por toda la organización. Busca garantizar la integridad y calidad con el fin de generar nuevo conocimiento al mismo tiempo que resulta un desafío para toda la organización.

A su vez, busca prevenir efectos secundarios como ser la fuga de información privada o la violación en la privacidad de la información (Kim y Cho, 2018). Este tipo de información refiere a aquella que surge de los datos personales siendo estos los que describen atributos de un individuo identificado, o que permiten realizar una individualización e identificación única de este. De este modo, cuando se incorporan datos de tipo personal, debe existir una estrategia de protección de la información para no incurrir en violación de la privacidad. Es en este sentido, la manipulación de datos personales conlleva asociado un riesgo que, mediante el diseño de un adecuado plan de gobierno de datos, puede ser controlado.

2. ENFOQUE ÉTICO Y NORMATIVO SOBRE PRIVACIDAD DE DATOS

En este contexto, la privacidad es definida como el interés que tiene un individuo en poder controlar sus propios datos (Clarke, 1999), o al menos ejercer su derecho sobre el manejo que se hace de los mismos. Si bien en ocasiones hay razones suficientes para ponerla en segundo lugar -como es el contexto actual de pandemia de coronavirus-, surge la necesidad de establecer una metodología que garantice su protección cuando los mismos son incorporados en las bases de datos de una organización y luego utilizados para extraer información.

Su incorporación puede crear información sin el consentimiento de sus titulares, aunque la intención de su uso no sea mal intencionado. En tales circunstancias, los riesgos asociados pueden derivarse de cuatro acciones específicas: reciclar, reutilizar, recombinar y reanalizar (Steinmann, Matei y Collmann, 2016). La conjunción de estas cuatro acciones puede llevar a la generación de nueva información que excede al objetivo primario para el cual el individuo cedió sus datos. Por esta razón, la gobernanza de datos requiere de un marco estratégico para la transparencia y el uso

general de los estos. En dicho marco, deben definirse los accesos, controles y responsabilidad sobre las personas que los accedan.

A partir de ello, y desde un punto de vista ético, la manipulación de datos personales puede entrar en conflicto con la libertad de expresión, reunión y manifestación o a la presunción de inocencia; en valores como la confianza y la cohesión social; y en procesos humanos importantes como el desarrollo de la identidad (Buenadicha, Galdon, Hermosilla, Loewe, y Pombo, 2019). Un correcto sistema de gobierno de datos puede surtir un efecto positivo en la medida que establezca condiciones claras sobre la seguridad de la información y su almacenamiento. Prácticas como implementación de antivirus, el cifrado de datos, la anonimización y la codificación de datos suelen ser modalidades de protección de los datos personales que resultan efectivas, aunque no suficientes.

También se deberá respetar la autonomía de los individuos. Esto es, si el uso de los datos tuviera un fin diferente al que fue solicitado, deberá consultarse a su titular si está dispuesto a facilitar sus datos a tal fin. En general, los sistemas informáticos utilizados para procesar datos son percibidos por la mayoría de los individuos como cajas negras (Buenadicha, Galdon, Hermosilla, Loewe, y Pombo, 2019), es decir, como mecanismos incomprensibles. Dado que no se puede pretender que todos los individuos tengan el suficiente conocimiento respecto de su funcionalidad, se busca establecer un sistema que este fundamentado en la transparencia de la información. Para lograrlo la información personal debe ser protegida a la vez que accesible para quien la solicite. Así mismo debe garantizar a través de sus políticas la utilización que se hará de ellos, con el fin de obtener un consentimiento claro por parte de quien los otorga. De esta manera, se logrará evitar una crisis de confianza entre el usuario y quien manipula su información.

Dado todos los riesgos asociados a la manipulación de información personal, en función de tener en cuenta los aspectos éticos mencionados, es que ha surgido diversa regulación. Un estándar de referencia global es el Reglamento General de Protección de Datos (RGPD)[1] de la Unión Europea, con entrada en vigor en el año 2018 (Buenadicha, Galdon, Hermosilla, Loewe, y Pombo, 2019). Sin entrar en mayores detalles sobre el RGDP, uno de los elementos esenciales que se establece es que el consentimiento es la base de la gestión de datos personales. Esto implica que, además de las cuestiones entorno a la seguridad, quienes recolecten y gestionen los datos deberán asegurarse siempre de haber informado a sus propietarios (los individuos) y obtener su consentimiento tantas veces como sea necesario si la finalidad del uso que se le dará a sus datos cambia. Además de respetar los derechos de los individuos, esta noción se constituye en un eje

[1] El Reglamento General de Protección de Datos (RGPD) entra en vigencia en mayo de 2018 en la Unión Europea. Disponible en https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32016R0679

fundamental de un contrato social que permite generar confianza entre quienes proporcionan los datos y quienes los manejan.

En esta misma línea, la Organización para la Cooperación y el Desarrollo Económicos (OCDE), de la cual forma parte la República Argentina, establece unos principios generales, vigentes desde el año 2013, sobre la protección de datos personales, que constituyen un marco general para delinear unas políticas claras sobre su resguardo. En este sentido postula que se debe establecer límites claros para la obtención de los datos, así como determinar la relevancia de estos para el uso previsto. En cuanto a la recolección, establece definir con claridad el uso que se dará a los datos antes de solicitarlos. También se propone abstenerse de utilizar los datos para usos distintos al determinado originalmente sin el consentimiento de las personas afectadas, y que se deberá proteger los datos contra el acceso ilícito o piratería. Así mismo se deberá garantizar que las personas cuyos datos se han recolectado tengan acceso a los mismos y puedan solicitar modificaciones o su eliminación definitiva.

Desde un enfoque legislativo, se deberá tenerse en cuenta la regulación vigente en cada país a la hora de diseñar un sistema de gobierno de datos. En el caso particular de Argentina existe la ley 25.326[2] de Protección de Datos Personales en cuyo objeto establece la necesidad de establecer "…la protección integral de los datos personales asentados en archivos, registros, bancos de datos, u otros medios técnicos de tratamiento de datos, sean estos públicos, o privados destinados a dar informes, para garantizar el derecho al honor y a la intimidad de las personas…" (Art. 1°).

Particularmente, por tratamiento de datos personales refiere a "… procedimientos sistemáticos, electrónicos o no, que permitan la recolección, conservación, ordenación, almacenamiento, modificación, relacionamiento, evaluación, bloqueo, destrucción…así como también su cesión a terceros a través de comunicaciones, consultas, interconexiones o transferencias." Por último, se considera de vital importancia lo establecido en su artículo 4°- "Los datos objeto de tratamiento no pueden ser utilizados para finalidades distintas o incompatibles con aquellas que motivaron su obtención."- ya que expresa el respeto por los derechos de los individuos.

De esta manera, la gestión en torno de la privacidad de datos se convierte en un punto central que deberá ser diseñada bajo consideraciones éticas y dentro del contexto de la normativa vigente. Teniendo en cuenta ambos enfoques, la protección de datos personales deberá ser construida desde el diseño y por defecto, de forma tal que se garantice la privacidad de los datos durante todo el ciclo de vida en una organización.

[2] Ley 25.326 sancionada en octubre de 2000 por el Honorable Congreso de la Nación Argentina, disponible en http://servicios.infoleg.gob.ar/infolegInternet/verNorma.do?id=64790

3. PRIVACIDAD DESDE EL DISEÑO Y POR DEFECTO

Partiendo del paradigma europeo, la privacidad desde el diseño, y por defecto, significa gestionar el riesgo asociado de manera proactiva para establecer una estrategia que incorpore la protección de datos durante todas las etapas de su procesamiento (Agencia Española de Protección de Datos, 2019). Esto permitirá que las responsabilidades sean asignadas por defecto, informando y formando a quienes se encuentren involucrados en cada una de estas. De esta manera, y teniendo en cuenta la regulación vigente en cada país, permitirá determinar qué datos personales son los necesarios para cada fin específico dentro de un marco de gestión responsable.

Esta metodología deberá considerar la proactividad en la gestión del riesgo. Esta viene dada por la anticipación a las amenazas que pudiesen existir a partir de la identificación de las debilidades de los sistemas contenedores de datos con el objetivo de minimizar los riesgos en lugar de aplicar medidas correctivas una vez sucedido un incidente. Implicará la definición y asignación de responsabilidades concretas, así como el desarrollo de métodos sistemáticos para detección temprana. A su vez, requerirá del desarrollo de una cultura del compromiso, impulsada desde la organización y asumida por los trabajadores.

También, resulta necesario establecer la privacidad por defecto. Esto implica que desde la etapa de diseño del procesamiento deberá quedar configurada para que sea parte integral de los sistemas, así como de todos los procesos de la organización. Para ello será necesario considerarla un requisito necesario durante el ciclo de vida del dato. Implicará establecer evaluaciones de impacto sobre datos personales, así como evaluación de riesgo sobre los derechos y libertades de los individuos.

Un proceso frecuente como aproximación a garantizar la privacidad, es la anonimización de datos. Este permite eliminar o reducir al mínimo el riesgo de reidentificación de su titular, sin que afecte la veracidad de estos (Agencia Española de Protección de Datos 2016). Para ello, se requerirá de la ruptura de la cadena de identificación de un individuo.

La identificación de un individuo puede ser directa o indirecta. El primer caso involucra a datos que identifican de manera unívoca y directa a su titular -como por ejemplo el número de documento único-. En el último caso, se encuentran los datos que son obtenidos de otras fuentes y al combinarlos permiten la re identificación de un individuo, es decir, son datos cruzados. Por ejemplo, la combinación de género, edad, lugar de nacimiento y padecimiento de una determinada enfermedad pueden permitir la identificación indirecta de un individuo específico. En particular, una modalidad frecuente de identificación indirecta se da cuando los datos son transferidos. El

receptor puede contar con datos de su propia fuente (o ajena) que le facilitaría la recombinación con los datos recibidos.

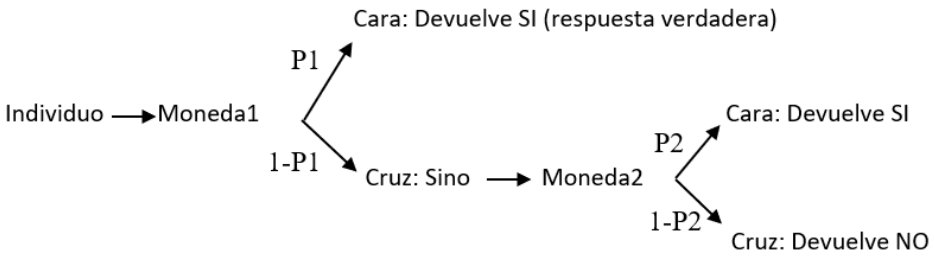
Frente a esta última situación, se pone de relieve que con tan solo una anonimización de datos resulta no ser suficiente para garantizar su protección en una organización. Y esto se debe a que no es posible anonimizar la totalidad de los datos. En consecuencia, se requerirá de la aplicación de una metodología adicional. La Privacidad Diferencial se presenta como una posible que contribuirá a llevar adelante el objetivo planteado.

4. LA PRIVACIDAD DIFERENCIAL

La privacidad diferencial es una metodología que consiste en aplicar aleatoriedad (ruido) sobre los datos al momento de implementar técnicas para su procesamiento (Dwork, 2008). Como resultado se obtendrá una cierta distorsión en los mismos. Esta, permitirá que un análisis posterior no se vea sustancialmente afectado. Si no, brinda la posibilidad de contar con un método matemático probabilístico riguroso para hacer frente a la posibilidad de reidentificación del titular. No obstante, requerirá de la evaluación de un costo, que vendrá dado por cuanta información real se está dispuesto a no publicar o utilizar a cambio de obtener privacidad. De esta manera, el fin último es obtener información sobre un conjunto de individuos no así sobre uno en particular.

Una definición técnica de la privacidad diferencial se puede encontrar en la respuesta aleatoria. Fue desarrollada por las ciencias sociales (Dwork, 2008) para recopilar información estadística sobre alguna temática tabú. Por ejemplo, supongamos que se quiere analizar una encuesta sobre una temática sensible (como el cáncer), donde las respuestas son sí o no, considerando que los encuestados (titular de los datos) no permiten el acceso directo a la información real brindada.

Frente a esta situación es posible aplicar privacidad diferencial para eliminar cierta información privada. Supongamos que el proceso contiene la siguiente lógica: se lanza una moneda y si sale cara devuelve la respuesta verdadera SI. Si no, se lanza una segunda moneda, y si sale cara devuelve como respuesta SI y si sale cruz devuelve NO. Esquemáticamente esto es:



En consecuencia, implicará que cada individuo está protegido con una negación plausible mediante el lanzamiento de una moneda (aleatoriedad). De esta forma, si se quiere conocer la

probabilidad de aquellos que tienen respuesta SI y están siendo protegidos por privacidad diferencial (PD), se deberán establecer las siguientes probabilidades condicionales:

P(SI/no está protegido por PD)= P1+ (1- P1)*P2

P(SI/ está protegido por PD) = (1- P1)*P2

Finalmente, la probabilidad de estar protegido por privacidad diferencial viene dada por:

$$\text{P(Estar protegido por PD)} = \frac{\text{P(SI)} - \text{P(SI/no está protegido por PD)}}{\text{P(SI/ está protegido por PD)} - \text{P(SI/ no está protegido por PD)}}$$

$$\frac{\text{P(SI)} - [\,\text{P1} + (1 - \text{P1}) * \text{P2}]}{(1 - \text{P1}) * \text{P2} - [\,\text{P1} + (1 - \text{P1}) * \text{P2}]} = \frac{\text{P(SI)} - [\,\text{P1} + (1 - \text{P1}) * \text{P2}]}{\text{P2} - \text{P1P2} - \text{P1} - \text{P2} + \text{P1P2}}$$

$$\frac{\text{P(SI)} - [\,\text{P1} + (1 - \text{P1}) * \text{P2}]}{-\text{P1}} = \frac{-\text{P1}}{-\text{P1}} + \frac{\text{P(SI)} - (1 - \text{P1}) * \text{P2}]}{-\text{P1}}$$

Por lo tanto:

$$\text{P(Estar protegido por PD)} = 1 - \frac{[\,\text{P(SI)} - (1 - \text{P1}) * \text{P2}]}{\text{P1}} \qquad (1)$$

La ecuación (1) implica que la varianza de la distribución de probabilidad de estar protegido por privacidad diferencial crece tendiendo a infinito a medida que P1 se acerca a 0, mostrando una reducción de la pérdida de la privacidad. Esto es, aplicar privacidad diferencial a los datos implica reducir el riesgo de pérdida de privacidad.

Al incorporar aleatoriedad en los datos arrojará como resultado una nueva base, que a diferencia de la original contendrá registros modificados como consecuencia de haber aplicado privacidad diferencial. Cuando un tercero consulte los datos no podrá notar la diferencia entre ambos en función de los resultados obtenidos, ya que estas no cambiarán sustancialmente el resultado probabilístico. La aproximación a este resultado requiere de la modelización matemática de la privacidad diferencial, la cual se realiza a continuación.

5. MODELO DE PRIVACIDAD DIFERENCIAL

Formalmente, dado dos conjuntos de datos $(\boldsymbol{D_1}, \boldsymbol{D_2})$, que difieren como máximo en un registro, pero uno está incluido en el otro, Dwork (2008) define a la privacidad diferencial como:

$$\boldsymbol{Pr}[\text{K}(\boldsymbol{D_1}) \in \boldsymbol{S}] \leq \boldsymbol{exp}(\boldsymbol{\varepsilon}) * \boldsymbol{Pr}[\text{K}(\boldsymbol{D_2}) \in \boldsymbol{S}] \quad (2)$$

donde K es una función aleatoria que brinda privacidad diferencial y S está incluido en todo el rango de K.

De la ecuación (2) se puede observar que la probabilidad de respuesta en $\boldsymbol{D_1}$ difiere en $\boldsymbol{exp}(\boldsymbol{\varepsilon})$ veces la probabilidad de respuesta en $\boldsymbol{D_2}$. Dicha diferencia es consecuencia de haber incorporado ruido en los datos. A su vez, la pérdida de privacidad viene dada por el valor de $\boldsymbol{\varepsilon}$. Si $\boldsymbol{\varepsilon = 0}$,

entonces se obtiene una privacidad completa, mientras que si $\boldsymbol{\varepsilon > 0}$ existirá una pérdida de privacidad.

Para poder determinar cuánto ruido se debe incorporar en los datos de forma tal de equilibrar la pérdida de privacidad, se debe medir la sensibilidad del peso de los datos de un individuo en los cálculos que se realizan (Dwork, 2008). Esto es, para *f*: D→$\boldsymbol{R^k}$

$$\Delta \boldsymbol{f} = \max_{\boldsymbol{D_1, D_2}} \| \boldsymbol{f(D_1)} - \boldsymbol{f(D_2)} \| \quad (3)$$

donde f es la obtención de la respuesta verdadera.

Si en la ecuación (3), $\boldsymbol{k = 1}$ entonces f es la diferencia máxima entre los valores que puede tomar en dos bases de datos que tan solo difieren en un solo registro. Por lo tanto, cuanto menor sea el valor de $\boldsymbol{\varepsilon}$, se deberá incorporar más ruido para garantizar un incremento en la privacidad de datos.

Llegado a este punto es importante tener en cuenta que a medida que se aumenta la cantidad de aplicaciones de ruido sobre la base de datos se debilitará la privacidad. Esto es, dada la sensibilidad de $\boldsymbol{f(X)}$, la aleatoriedad comienza a perder fuerza. Un mecanismo para sortear esta dificultad consiste en utilizar la distribución de Laplace para la generación de ruido aleatorio. De esta manera, Dwork (2008) propone:

$$\boldsymbol{f(X)} + (\boldsymbol{Lap}(\Delta \boldsymbol{f}/\boldsymbol{\varepsilon}))^{\boldsymbol{k}} \quad (4)$$

En la ecuación (4) es posible notar que la incorporación de ruido se independiza de la cantidad de los $\boldsymbol{k}$ componentes de $\boldsymbol{f(X)}$. Es decir, la forma de incorporar ruido en la base de datos ya no depende de la secuencialidad de aplicaciones que se haga sobre la misma sino del valor de $\boldsymbol{\varepsilon}$. A medida que $\boldsymbol{\varepsilon}$ disminuye, el ruido esperado será mayor -la curva de la distribución de Laplace será más aplanada (ver anexo)-. De esta manera, se observa que la privacidad queda garantizada a partir de que la privacidad diferencial solo depende de la sensibilidad de $\boldsymbol{f(X)}$ y del parámetro $\boldsymbol{\varepsilon}$.

Con el fin de poner a prueba el modelo de privacidad diferencial presentado en este apartado, a continuación, se desarrolla un caso de aplicación mediante la utilización del software Python[3].

6. APLICACIÓN DEL MODELO DE PRIVACIDAD DIFERENCIAL

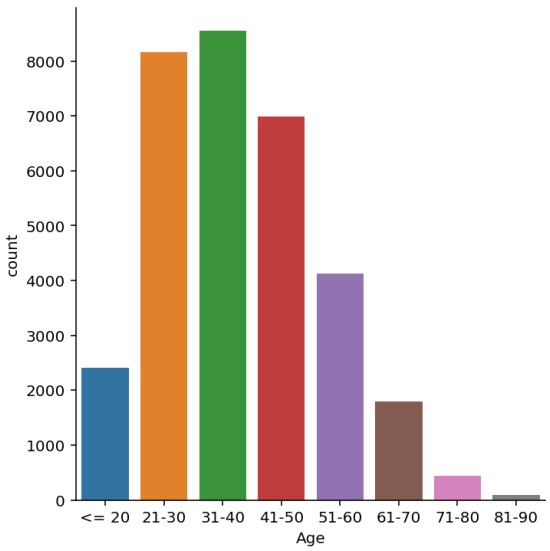
Llevar a cabo una implementación de privacidad de datos resulta de interés para poder evaluar la eficacia del modelo dentro del marco de una gestión responsable de datos. Para ello, se toma una muestra de datos y se realiza la aplicación mediante la utilización de Python. Seleccionando una

[3] Acerca de Python https://www.python.org/

variable se presenta la propuesta de aplicación del modelo. Para esto se utiliza la base datos "Adults"[4]. Este es una extracción del censo de 1994 de EE. UU. realizada por Barry Becker. Si bien contiene 32.561 registros en un subconjunto con 14 atributos más una clase que consiste en el nivel de ingreso anual (superior a 50 mil dólares o inferior a tal valor), se analizará la distribución de la variable "Age" (edad).

La elección de esta variable se debe a que es un atributo que caracteriza a un individuo y resulta sensible en la medida que combinado con otros atributos puede derivar en la identificación de su titular. Además, cualquier distorsión que se incorpore en la misma puede variar significativamente su distribución, afectando un análisis posterior. Los datos contenidos muestran que posee un valor mínimo de 17 años y un valor máximo de 90 años. Frente a esto y para facilitar su visualización, se la agrupa en rangos de edad, y se realiza un gráfico de barras que es representativo de la distribución de la variable.

Gráfico 1: Distribución de los individuos según rango de edad



Elaboración propia con Python

A partir del gráfico 1 se puede observar que la mayor cantidad de individuos contenidos en el conjunto de datos se encuentran entre las edades 21 y 50 años. Dada esta distribución original de la variable "Age" en el siguiente apartado se muestran los resultados obtenidos de la aplicación de privacidad diferencial.

RESULTADOS

A partir del modelo de Privacidad Diferencial descripto, en el presente apartado se establece la metodología que utilizará el algoritmo a implementar. Si bien existen diferentes alternativas, se opta por utilizar la metodología del histograma. Dado que este último, consiste en representar la frecuencia (absoluta o relativa) de los casos en función de las categorías (intervalos de clase) de la

[4] Obtenido de https://archive.ics.uci.edu/ml/datasets/adult

variable, cualquier afectación en los valores puede cambiar drásticamente su distribución. Ello se debe a que son muy sensibles a cambios en los valores que contiene cada intervalo de clase.

Considerando diferentes tamaños de muestra (5000, 10000, 15000, 20000) y teniendo en cuenta distintos valores del parámetro $\boldsymbol{\varepsilon}$ (0.5, 0.1, 0.05, 0.025), mediante la utilización de la librería OpenDP[5] en Pyhton, se realiza la aplicación de privacidad diferencial sobre la variable Age. A continuación, en el Gráfico 2 se muestran los resultados obtenidos.

Gráfico 2: Comparación de la distribución de los individuos según rango de edad considerando los valores originales y los valores obtenidos de aplicar privacidad diferencial

Gráfico 2.1

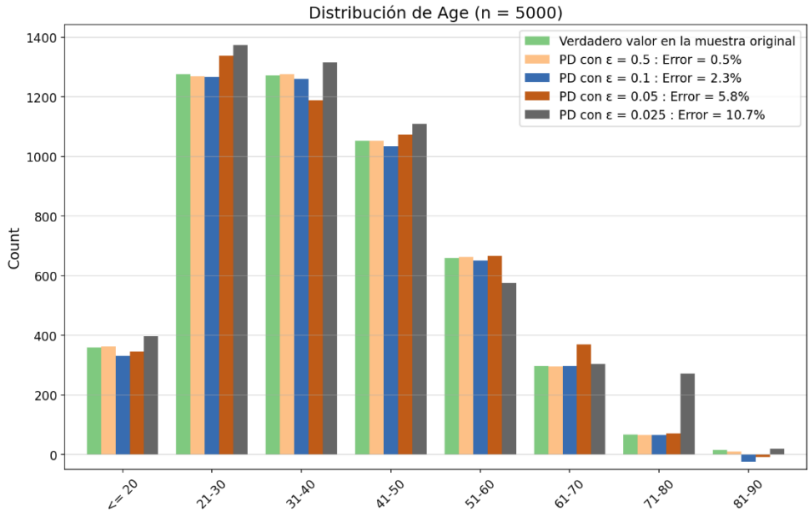
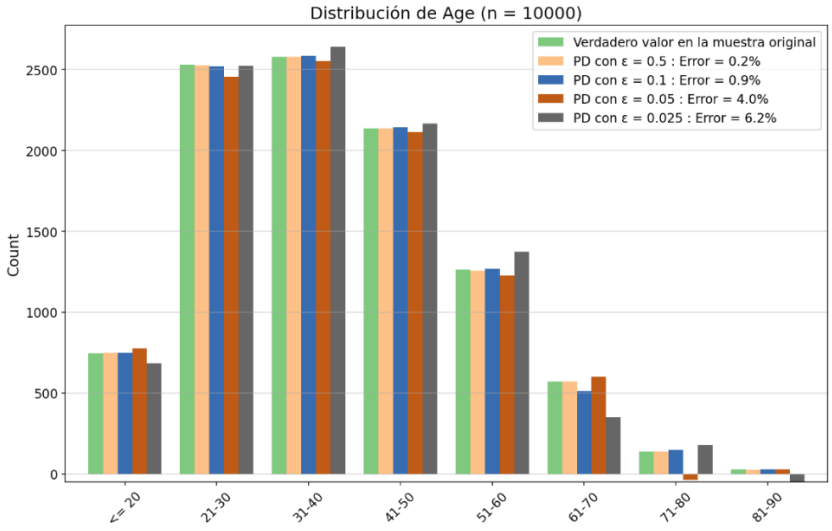


Gráfico 2.2



[5] Acerca de OpenDP https://pypi.org/project/opendp-smartnoise-core/

Gráfico 2.3

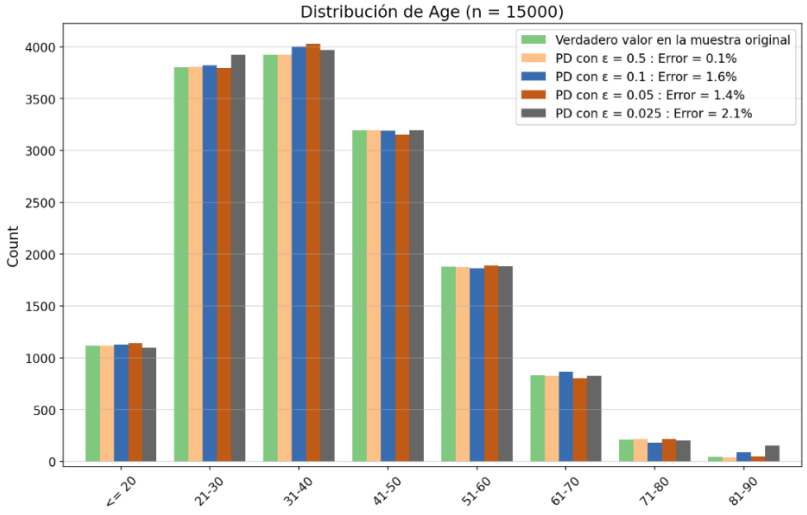
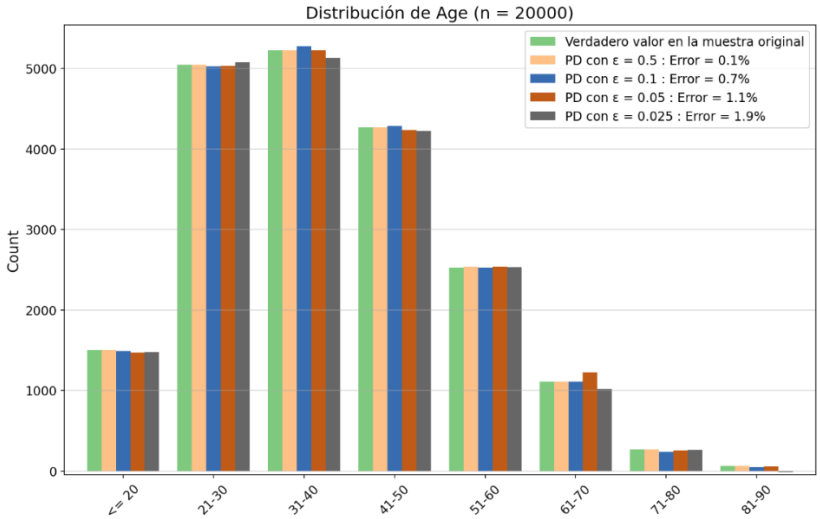


Gráfico 2.4



Elaboración propia con Python

Es de destacar que, para cada tamaño de muestra, a medida que disminuye el valor de $\boldsymbol{\varepsilon}$, es decir, a medida que se opta por reducir la pérdida de privacidad, el error aumenta. Este error representa en términos porcentuales, la variabilidad media que existe entre el verdadero valor en la muestra y el valor obtenido de aplicar privacidad diferencial. A su vez, se observa que a medida que aumenta el tamaño de la muestra, a igual valor del parámetro $\boldsymbol{\varepsilon}$, el error disminuye.

En función de los resultados obtenidos se logra obtener datos ruidosos o distorsionados sin que se afecte significativamente la distribución de la variable en cuestión con el fin de evitar la re-identificación del titular. De esta manera, se logra demostrar que la aplicación de privacidad diferencial cumple con el objetivo de protección de datos a la vez que no afectará un análisis posterior.

CONCLUSIÓN

A lo largo del presente trabajo se ha demostrado que la aplicación de Privacidad Diferencial sobre datos permite generar privacidad sin que se afecte significativamente la distribución original de los datos. Al mismo tiempo, se logra dar cuenta de que la incorporación de este tipo

de técnicas, en un contexto de grandes volúmenes de datos, contribuye significativamente a la protección de datos.

Además, se pone de relieve que los datos personales desafían a las organizaciones a considerar a la privacidad desde el diseño y por defecto para lograr con éxito una gestión responsable de datos. Si bien la regulación y legislación son parte necesaria no resultan suficientes, en la medida que estas solo establecen el deber ser y no el cómo llevarlo a cabo.

El modelo de Privacidad Diferencial permite observar que la forma de incorporar ruido en datos ya no depende de la secuencialidad de aplicaciones que se haga sobre estos sino de cuanta privacidad se está dispuesto a resignar. De esta manera, se observa que la privacidad queda garantizada a partir de que la privacidad diferencial solo depende de la sensibilidad de respuesta.

Finalmente, de la implementación del modelo sobre una variable (edad de los individuos tomados como muestra) se observa que, a medida que disminuye el valor de pérdida de privacidad, el error aumenta. Es decir, la variabilidad media que existe entre el verdadero valor en la muestra y el valor obtenido de aplicar privacidad diferencial se agranda. Pero a su vez, se observó que, a mayor tamaño de muestra, e igual valor de pérdida de privacidad, el error disminuye. De esta manera, se logra observar la efectividad de la aplicación del modelo para grandes volúmenes de datos.

Para finalizar, resulta de interés llevar adelante una ampliación de este trabajo, que permita realizar una implementación más profunda de la metodología de Privacidad Diferencial, tomando otros criterios para evaluar la distribución de los datos. También, realizar la implementación de algún modelo de predicción o clasificación sobre los datos protegidos que permitiría evaluar el impacto en la performance de estos.

REFERENCIAS BIBLIOGRÁFICAS

Agencia Española de Protección de Datos (2019). Guía de Privacidad desde el Diseño. Sede Electrónica. https://www.aepd.es/es

Buenadicha, C., Galdon, G., Hermosilla, M. P., Loewe, D., & Pombo, C. (2019). La Gestión Ética de los Datos. Por qué importa y cómo hacer un uso justo de los datos en un mundo digital BID, editor.

Chan, J., Gollakota, S., Horvitz, E., Jaeger, J., Kakade, S., Kohno, T., ... & Tessaro, S. (2020). PACT: Privacy Sensitive Protocols and Mechanisms for Mobile Contact Tracing. arXiv preprint arXiv:2004.03544.

Clarke, R. (1999). Internet privacy concerns confirm the case for intervention. Communications of the ACM, 42(2), 60-67.

Cleven, A., & Wortmann, F. (2010, January). Uncovering four strategies to approach master data management. In 2010 43rd Hawaii International Conference on System Sciences (pp. 1-10). IEEE.

Deloitte (2011). Desarrollo de un modelo operativo de gobierno que sea efectivo. Una guía para las juntas y los equipos de administración de servicios financieros.

Dwork, C. (2008, April). Differential privacy: A survey of results. In International conference on theory and applications of models of computation (pp. 1-19). Springer, Berlin, Heidelberg.

Hekmati, A., Ramachandran, G., & Krishnamachari, B. (2020). CONTAIN: privacy-oriented contact tracing protocols for epidemics. arXiv preprint arXiv:2004.05251.

Jones, D. T. (2018). Data Governance Framework Implementation Plan. Philadelphia: DBHIDS.

Kim, H. Y., & Cho, J. S. (2018). Data governance framework for big data implementation with NPS Case Analysis in Korea. Journal of Business and Retail Management Research, 12(3).

Martínez, J. (2012). Seis pasos para el Gobierno de Datos. ¿Qué es y cómo se implementa un programa de Gobierno de Datos? IBM, Developer Works.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. Harvard business review, 90(10), 60-68.

McKeen, J. D., & Smith, H. A. (2007). Developments in practice XXIV: information management: the nexus of business and IT. Communications of the Association for Information Systems, 19(1), 3.

Schmarzo, B. (2013). Big Data: Understanding how data powers big business. John Wiley & Sons.

Steinmann, M., Matei, S. A., & Collmann, J. (2016). A theoretical framework for ethical reflection in big data research. In Ethical Reasoning in Big Data (pp. 11-27). Springer, Cham.

ANEXO

Distribución de Laplace

La distribución de Laplace es la distribución de la diferencia de dos variables aleatorias e independientes, donde cada una de ellas tiene una distribución exponencial.

Su función de densidad de probabilidad viene dada por

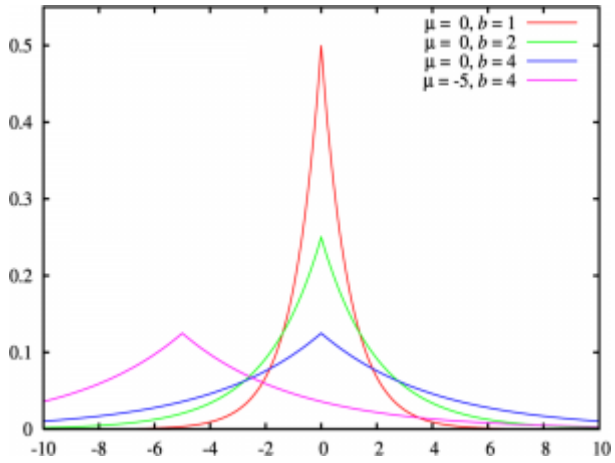
$$P(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}$$

donde b es el parámetro de escala.

La función de distribución acumulada viene dada por

$$F(x) = \frac{1}{2}[1 + signo(x-\mu)(1 - e^{-\frac{|x-\mu|}{b}})]$$

En el modelo de privacidad diferencial, la escala de la distribución de probabilidad viene dada b=$\Delta f/\varepsilon$ (Dwork, 2008). Por tanto, a medida que disminuye el valor de ε, el parámetro de escala b aumenta. Gráficamente:



Tomado de https://www.statisticshowto.com/laplace-distribution-double-exponential/ en donde se puede observar que, a mayor valor de b, se obtiene una curva más aplanada de la distribución.