.UBAeconómicas posgrado

ENAP Escuela de Negocios y Administración Pública

Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Negocios y Administración pública

# MAESTRÍA EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

## TRABAJO FINAL DE MAESTRÍA

Gestión responsable de datos en el Estado argentino. La Privacidad Diferencial como propuesta para la protección de datos personales.

Responsible management of data in Argentine State. Differential Privacy as a proposal for the protection of personal data.

AUTORA: NATALIA ROMINA SALABERRY

DIRECTOR: DR. JAVIER IGNACIO GARCÍA FRONTI

SEPTIEMBRE 2022

## Agradecimientos

Mi agradecimiento a mi director de tesis, Dr. Javier García Fronti, y a la Dr. María José Bianco, por el acompañamiento, confianza y apoyo incondicional, así como todos los aportes fundamentales realizados para que pueda llevar adelante este trabajo. Sin ellos, este trabajo final de maestría no hubiese sido posible.

También quisiera mencionar mi gratitud hacia la Profesora Emérita Dra. María Teresa Casparri, ya que, sin su trabajo incansable y ejemplar, esta maestría no hubiese existido. Gracias a ella, tuve la posibilidad de profesionalizar aún más mis conocimientos a través de realizarla.

Agradezco al Dr. Pablo Matías Herrera, por todos los comentarios realizados en diferentes instancias de cursada como parte de esta maestría y de doctorado, pero también por ser un compañero de trabajo de investigación en alguna oportunidad, que me ha permitido adquirir diversos conocimientos para mi desarrollo académico.

Agradezco a la Universidad de Buenos Aires, a la Facultad de Ciencias Económicas, mi segundo hogar y desde hace un tiempo mi lugar de trabajo permitiéndome ejercer la docencia, función que con tanta pasión tengo el honor de poder realizar. Y al Centro de Métodos Cuantitativos Aplicados a la Economía y a la Gestión perteneciente al IADCOM, por permitirme comenzar a dar los primeros pasos para desarrollarme como investigadora académica.

Finalmente, agradecer a mi familia, pilar fundamental en mi vida. En especial a mi madre, que con su amor incondicional soportó los tiempos de ausencia. Pero sobre todo por haberme enseñado que, con perseverancia, compromiso y trabajo los sueños se pueden cumplir.

## Resumen

El Estado argentino recopila grandes volúmenes de datos con el fin de realizar políticas públicas. La utilización de las tecnologías para su procesamiento le permite ser más eficiente y eficaz en el diseño de las políticas, cuyo resultado la sociedad lo recibe a través de diversos servicios públicos. Sin embargo, las tecnologías le otorgan un poder superior. Este reside en las personas que la utilizan ya que les permite obtener más información y conocimiento sobre los individuos. Si bien, cierta parte de la población lo acepta; Otra parte, considera que trae aparejado consecuencias sobre su vida privada.

La preocupación acerca de cómo los datos personales son tratados y divulgados por parte de las organizaciones públicas del Estado argentino comienza a tomar mayor relevancia. En este contexto, el presente trabajo final de maestría tiene por objetivo general aplicar y evaluar la efectividad de la metodología de Privacidad Diferencial para evitar la identificación de ciertos individuos a partir de datos personales de 2015 obtenidos de organizaciones estatales argentinas y establecer los aspectos relevantes que se derivan para el diseño de una estrategia de gestión responsable de la privacidad en el Estado argentino. Para poder cumplimentar con el objetivo planteado, se utiliza un enfoque mixto que combina un método cualitativo y cuantitativo de investigación.

El trabajo se estructura en cuatro capítulos, agrupados en dos partes. En los capítulos uno y dos, se presenta el corpus teórico. En el capítulo tres, se lleva a cabo la implementación técnico-cuantitativa de la Privacidad Diferencial para la protección de datos personales. Por último, en el capítulo cuatro se realiza la generalización de resultados obtenidos a través de establecer los principales puntos relevantes a ser tenidos en cuenta a la hora de diseñar una estrategia de privacidad en un marco de gestión responsable en organizaciones públicas.

De este modo, el trabajo aporta una metodología para el tratamiento de la privacidad de datos en organizaciones públicas estatales, constituyendo una herramienta de gestión. La identificación de riesgos en el procesamiento de datos personales y la implementación de una técnica para mitigarlos permite anticipar los principales resultados potenciales de este tipo de desarrollo. Además, permite mostrar la importancia de elaborar un marco de gestión responsable para el tratamiento de la privacidad.

**Palabras claves:** Gestión responsable, Datos, Estado argentino, Privacidad Diferencial, Protección de datos personales.

# Índice

**Agradecimientos** ..... **2**

**Resumen** ..... **3**

**Introducción** ..... **6**

**Planteamiento del tema** ..... **12**

**Capítulo 1: Organizaciones y el mercado de datos personales** ..... **13**

Introducción ..... 13

1.1 Las organizaciones basadas en datos ..... 14

1.1.1 El papel de la tecnología de la información ..... 16

1.1.2 Configuración de la capacidad de acción ..... 18

1.1.3 El rol de la gestión de datos ..... 19

1.1.4 La toma de decisiones en base a datos ..... 20

1.2 Constitución del mercado de datos personales ..... 21

1.2.1 El mercado como una red sociotécnica ..... 22

1.2.2 Los datos personales conductuales ..... 24

1.2.3 La creación de valor ..... 25

Conclusión ..... 26

**Capítulo 2: El modelo de Privacidad Diferencial y los conceptos involucrados** ..... **28**

Introducción ..... 28

2.1 El concepto de privacidad ..... 29

2.1.1. La privacidad de datos personales ..... 30

2.1.2. La privacidad desde el diseño y por defecto ..... 32

2.2 La Privacidad Diferencial ..... 33

2.3 La modelización de la Privacidad Diferencial ..... 36

2.3.1 Función o algoritmo aleatorio ..... 36

2.3.2 Distribución de Laplace ..... 37

2.3.3 El mecanismo de Laplace ..... 39

2.3.4 El modelo de Privacidad Diferencial ..... 39

Conclusión ..... 41

**Capítulo 3: Aplicación y evaluación de la Privacidad Diferencial para evitar la identificación de ciertos individuos a partir de datos personales obtenidos de organizaciones estatales argentinas** ..... **42**

Introducción ..... 42

3.1 Obtención de datos y construcción de base de datos personales ..... 43

3.1.1 Análisis descriptivo de la base de datos personales ..... 46

3.1.2 Identificación de problemáticas ..... 49
3.1.3 Identificación indirecta de individuos ..... 50
3.2 Aplicación de Privacidad Diferencial y evaluación del modelo ..... 53
3.3 Aplicación de ruido aleatorio por mecanismo de Laplace y resultados ..... 57
Conclusión ..... 60
**Capítulo 4: Aspectos relevantes para el diseño de una gestión responsable de la privacidad en el Estado argentino ..... 61**
Introducción ..... 61
4.1 La estrategia del diseño de privacidad ..... 62
4.2 Aspectos principales en la implementación de privacidad sobre datos personales ..... 65
4.2.1. Acceso y disponibilidad de datos ..... 65
4.2.2. Detección de riesgos para la privacidad ..... 67
4.2.3. Selección de metodología para la protección de datos personales ..... 68
4.3 La responsabilidad por defecto ..... 70
Conclusión ..... 71
**Conclusión ..... 73**
**Referencias bibliográficas ..... 78**
**Apéndice ..... 81**
A1. Simulación de distribución de Laplace con Python ..... 81
A2. Implementación de Privacidad Diferencial con Python ..... 82

## Introducción

La privacidad de datos personales es definida como el interés que tiene un individuo en poder controlar sus propios datos (Clarke, 1999), o al menos ejercer su derecho de poder conocer que se hace con los mismos. Las organizaciones públicas reconocen el valor de los datos como generadores de información para la realización de políticas públicas eficientes y eficaces (Bryson et al., 2015). Su recopilación y procesamiento mediante el uso de tecnología también involucra a la estrategia necesaria a llevarse a cabo (Constantiou y Kallinikos, 2015). A su vez, el desarrollo tecnológico de las últimas décadas ha facilitado la creación de redes online que conforman un espacio digital de interacciones (Zysman y Kenney, 2018). Estas, favorecen el intercambio entre organizaciones e individuos, conformando un mercado de datos (Täuscher y Laudien, 2018). En particular, las organizaciones públicas se han visto beneficiadas ya que les ha permitido crear su propia red tecnológica denominada gobierno electrónico (OCDE, 2014). De la interacción entre los actores en este nuevo espacio, surgen grandes volúmenes de datos (Schmarzo, 2013).

Con los grandes volúmenes de datos las organizaciones públicas realizan cálculos frente a la necesidad de poder hacerlos inteligible (Rivoir y Morales, 2019). Dado que los individuos son interpretados a través de los datos que generan a partir de su interacción con medios digitales en el ámbito social, resulta necesario convertirlos en información. Y son las tecnologías de *Big Data* las que permiten llevar adelante este proceso generándose conocimiento para la toma de decisiones (Constantiou y Kallinikos, 2015). De esta manera, permiten convertir lo intangible (datos) en objetos materiales (Haskel y Westlake, 2017) –informes de resultados, gráficos, tablas de datos, entre otros–. Pero, a su vez, las tecnologías otorgan un poder superior a las organizaciones públicas. Este reside en las personas que la utilizan ya que les permite obtener más información y conocimiento sobre los individuos (Gandy Jr, 2021). Si bien cierta parte de la población acepta que se produzca la recopilación de sus datos – incluyendo los personales -; Otra parte, considera que trae aparejado consecuencias sobre su vida privada.

De esta manera, comienza a surgir una disyuntiva entre las potenciales promesas derivadas del uso de la tecnología para el procesamiento de datos y los riesgos asociado a su uso para la privacidad de los individuos. En este sentido, por un lado, se encuentran los países que pregonan el máximo aprovechamiento de las capacidades tecnológicas en un contexto de datos, entre los que se encuentran Estado Unidos de América. Y por el otro, principalmente

los países nucleados en la Unión Europea que advierten sobre los riesgos asociados y deciden establecer una regulación a través de un Reglamento que entra en vigor en el año 2018 que se ha convertido en un estándar de referencia mundial. Argentina, puede considerarse que toma de referencia ambas miradas ya que en los últimos años ha realizado políticas de impulso al desarrollo tecnológico en diferentes sectores que involucran a los datos, pero también cuenta con una ley de Protección de Datos Personales promulgada en el año 2000. No obstante, ante diferentes hechos de conocimiento público, la preocupación en la sociedad argentina acerca de cómo los datos personales son tratados y divulgados por parte de las organizaciones públicas comienza a tomar mayor importancia.

El tema seleccionado se considera de suma relevancia, ya que aún en el año 2022 no existe un consenso acerca del tratamiento de privacidad de datos personales en el mundo. Pero al mismo tiempo, con el continuo avance tecnológico en un contexto completamente digitalizado, surgen cada vez más diferentes hechos derivados del uso de datos personales que ponen en jaque el derecho a la privacidad que poseen los individuos. En este sentido, la fundamentación de la elección del tema a tratar para la elaboración del trabajo final de maestría se considera de interés para las diferentes organizaciones, así como para la sociedad argentina en su conjunto. Esto se debe a que se busca establecer los aspectos relevantes que permitan dar garantía sobre la efectividad de la protección de datos personales al momento de implementar técnicas para su procesamiento. Como resultado, las organizaciones públicas podrán contar con mayor cantidad de información sobre la población siendo un impacto positivo. Además, permitirá generar un mayor grado de confianza de los individuos para con estas lo que se traducirá en más información disponible para la toma de decisiones.

Por otra parte, se considera que el tema se encuentra contenido dentro del área de conocimiento que aborda la maestría en Métodos Cuantitativos para la Gestión y Análisis de datos en Organizaciones, en la medida que se encuadra dentro de las problemáticas asociadas al procesamiento automático de datos con estrecha relación a su gestión y gobernanza. A su vez, se considera que brinda un aporte a la academia mediante la implementación y evaluación de un método para la protección de datos personales.

En este contexto, el presente trabajo final de maestría tiene por objetivo general aplicar y evaluar la efectividad de la metodología de Privacidad Diferencial para evitar la identificación de ciertos individuos a partir de datos personales de 2015 obtenidos de organizaciones estatales argentinas y establecer los aspectos relevantes que se derivan para

el diseño de una estrategia de gestión responsable de la privacidad en el Estado argentino. Este objetivo general se aborda a través de cuatro objetivos específicos. El primero, es describir a una organización basada en datos, cómo se lleva a cabo el proceso de toma de decisiones en base a estos y cómo se produce su interacción en el mercado de donde los obtiene. El segundo, presentar el modelo de Privacidad Diferencial, así como todos los conceptos involucrados. El tercero, es construir una base de datos personales partiendo de un padrón de CUIT e identificar los riesgos de privacidad que surgen de esta y, finalmente, aplicar el modelo de privacidad diferencial para evaluar su efectividad en términos de mitigar tales riegos. Finalmente, el cuarto objetivo es especificar los puntos relevantes que surgen de los anteriores para ser tenidos en cuenta en el diseño de una estrategia de privacidad en organizaciones públicas del Estado. Para el cumplimiento de cada uno de ellos, la realización del trabajo se organiza en cuatro capítulos, agrupados en dos partes.

La primera parte se integra por los capítulos uno y dos, a través de los cuales se presenta el corpus teórico. Mientras en el primer capítulo se desarrolla la conceptualización de una organización basada en datos y sus implicancias; en el segundo se aborda el concepto de privacidad, de privacidad de datos personales y de privacidad diferencial. Se explica entonces porqué es necesario establecer un marco de responsabilidad para mitigar los riesgos asociados que surgen del procesamiento de datos personales en organizaciones públicas.

En el primer capítulo, el objetivo a desarrollar consiste en describir a las organizaciones públicas estatales como aquellas basadas en datos y su interacción con los individuos en el mercado de datos para la generación de valor. La hipótesis asociada es que el procesamiento de datos personales a través de tecnología por parte de las organizaciones públicas del estado les otorga un poder superior que pone en riesgo la privacidad de los individuos. A partir de la explicación realizada en las diferentes instancias se logra comprender el riesgo de privacidad de datos que se deriva del proceso de interacción entre organizaciones e individuos en el mercado de datos a través del uso de tecnología. En este este sentido, una mitigación de los riesgos requiere de un compromiso por parte de las organizaciones haciendo hincapié en los individuos.

Para argumentar por qué la tecnología es esencialmente un instrumento de poder, se desarrolla la conceptualización de una organización basada en datos a partir de tres dimensiones claves: las tecnologías de *Big Data*, la capacidad de generación de información y su interacción con los individuos para la obtención de datos. Para ello, se analizan las

principales particularidades de cómo es llevado a cabo el procesamiento de datos obtenidos en el mercado para la toma de decisiones y se remarca como la tecnología es un instrumento utilizado no solo para generar valor. Finalmente, se analiza como los datos recolectados a través de estas son utilizados para generar conductas humanas futuras identificando los riesgos asociados a privacidad y se expone la necesidad de la responsabilidad que ello conlleva.

El segundo capítulo, tiene por objetivo presentar el modelo de Privacidad Diferencial, así como todos los conceptos de privacidad involucrados. La hipótesis aquí planteada es que la privacidad de datos personales concebida desde el diseño y por defecto resulta necesaria para establecer un marco responsable de gestión de la privacidad. En el capítulo se analiza la necesidad de definir una estrategia que incorpore la protección de datos personales durante todas las etapas de su procesamiento ya que permitirá que las responsabilidades sean asignadas por defecto, informando y formando a quienes se encuentren involucrados. Si bien esto es posible de llevarse a cabo en organizaciones públicas argentinas con la ayuda de una metodología como la privacidad diferencial, no se verifica si está siendo implementada actualmente.

En este segundo capítulo para probar que la privacidad diferencial es una metodología que permite proteger datos personales desde el diseño y por defecto, en primer lugar, se propone una definición de privacidad y de la privacidad de datos personales. Luego, se aborda la presentación de la metodología mencionada fundamentando que otorga cierta garantía de privacidad. Para ello, se define su concepto, se ejemplifica su lógica de operatoria y se expone la necesidad de su modelización. Finalmente, se desarrolla el modelo de Privacidad Diferencial presentando los principales conceptos estadísticos y matemáticos involucrados. De esta manera, el modelo que se propone sería una herramienta de gestión que permite mitigar los riesgos de privacidad de datos personales.

De la articulación de las dos hipótesis planteadas en el corpus teórico, abordado en los capítulos uno y dos, es que se establece la hipótesis general de este trabajo. En ambas se hace referencia a los riesgos derivados del procesamiento de datos en organizaciones públicas y la necesidad de establecer responsabilidad en los procesos involucrados. De la vinculación entre ambas, la hipótesis general del trabajo final de maestría es: el uso y procesamiento de datos personales en organizaciones públicas argentinas vulnera la privacidad de los individuos.

Para llevar a cabo la verificación de esta, resulta necesario obtener datos personales para construir una base, identificar los riegos de privacidad derivados y finalmente realizar la implementación de Privacidad Diferencial para la evaluación de su efectividad en términos de protección de datos personales. En base a las etapas de implementación que se realizan y teniendo en cuenta los principios fundacionales propuestos por Cavoukian (2011), se establecen los aspectos principales para tener en cuenta a la hora de diseñar una estrategia responsable de privacidad en organizaciones públicas.

La segunda parte de este trabajo, desarrollada en los capítulos tres y cuatro, se lleva a cabo en primer lugar mediante un enfoque técnico cuantitativo realizando la aplicación de la Privacidad Diferencial sobre una base de datos personales que debió construirse manualmente y analizarse para la identificación de los riesgos de privacidad derivados en términos de identificación de individuos. En segundo lugar, mediante un enfoque cualitativo se establecen los principales aspectos a ser tenidos en cuenta en el diseño de una estrategia de privacidad en base a lo realizado a lo largo del trabajo. Mientras en el capítulo tres se lleva a cabo la aplicación, en el capítulo cuatro se realiza la generalización de resultados obtenidos a través de lo cual se establecen los aspectos principales.

En el tercer capítulo, el objetivo es, por un lado, construir una base de datos personales. Por el otro identificar los riesgos de privacidad que surgen de esta y, finalmente, aplicar el modelo de privacidad diferencial para evaluar su efectividad en términos de si es posible evitar la identificación de un individuo. La hipótesis asociada es que la aplicación de privacidad diferencial resulta efectiva para la protección de datos personales. Mediante la implementación que se realiza, se logra obtener un resultado inicialmente satisfactorio de la implementación de la metodología de privacidad diferencial sobre una base de datos personales en la medida que permitió mitigar el riesgo de privacidad detectado.

Para llevar adelante el capítulo tres, en primer lugar, se explicita como es construida la base de datos personales utilizada. Luego, se realiza una descripción de los datos finalmente obtenidos. A continuación, se exponen las problemáticas que se detectan en la base de datos en términos de riesgos de violación de privacidad. Finalmente, se desarrolla la implementación del modelo de privacidad diferencial y se realiza la evaluación de la efectividad del modelo en términos de los resultados obtenidos.

En el cuarto capítulo, el objetivo es establecer los principales puntos a ser tenidos en cuenta para llevar adelante el diseño de una estrategia de la gestión de la privacidad de datos personales en organizaciones públicas estatales. Si bien no cuenta con una hipótesis específica planteada, al utilizarse un enfoque cualitativo se deriva de los resultados alcanzados. Para lograrlo, en una primera parte, se explicitan los principios sobre los cuales es necesario sustentar el diseño de la privacidad en organizaciones. Luego se exponen los puntos principales a ser tenidos en cuenta en base a los resultados obtenidos previamente. Finalmente, se especifican los factores relevantes para definir la responsabilidad por defecto en la gestión de la privacidad de datos personales.

Tras el cumplimiento de cada uno de los objetivos mencionados, el trabajo aporta una primera aproximación para el diseño de una estrategia de gestión de la privacidad de datos personales en organizaciones públicas, siendo extensivo a las organizaciones privadas. En este sentido, constituye inicialmente una herramienta fundamental para dar garantías en el diseño de un sistema de gestión responsable de datos. Si bien resulta necesario ampliarse o profundizarse, los aspectos mencionados no dejan de constituir un punto de partida esencial que toda organización pública debería tener en cuenta para lograr con mayor eficiencia la protección de datos personales. Además, brinda la posibilidad de dar garantías de responsabilidad en el tratamiento de los datos para generar confianza con los individuos de la sociedad argentina.

## Planteamiento del tema

El Estado argentino recopila grandes volúmenes de datos a través de su plataforma de gobierno digital para realizar políticas públicas. Con el desarrollo y utilización de tecnologías, se ha visto favorecido para llevar adelante su procesamiento. La evolución de estas incluye también la estrategia necesaria para la extracción de información para la toma de decisiones, dando origen al concepto de *Big Data*. De esta manera, les permite ser más eficiente y eficaz en el diseño de las políticas, cuyo resultado la sociedad lo recibe a través de diversos servicios públicos.

Pero, a su vez, las tecnologías de *Big Data* otorgan un poder superior al Estado. Este reside en las personas que la utilizan ya que les permite obtener más información y conocimiento sobre los individuos. Si bien cierta parte de la población acepta que se produzca la recopilación de sus datos – incluyendo los personales -; Otra parte, considera que trae aparejado consecuencias sobre su vida privada. Frente a esta situación, se comienza a cuestionar como los datos son tratados y divulgados por parte de estas organizaciones públicas estatales. Y en consecuencia la privacidad comienza a ocupar el centro de discusión.

En este contexto, surge el siguiente interrogante: ¿Cómo pueden protegerse los datos personales recopilados y publicados por organizaciones del Estado argentino garantizando la privacidad de los individuos? Para poder responder este interrogante, el objetivo general de la tesis es aplicar y evaluar la efectividad de la metodología de Privacidad Diferencial para evitar la identificación de ciertos individuos a partir de datos personales de 2015 obtenidos de organizaciones estatales argentinas y establecer los aspectos relevantes que se derivan para el diseño de una estrategia de gestión responsable de la privacidad en el Estado argentino. La hipótesis general asociada es que el uso y procesamiento de datos personales para la elaboración de políticas públicas en organizaciones estatales argentinas vulnera la privacidad de los individuos.

Como objetivos específicos se considera realizar una concepción del Estado como una organización basada en datos que interactúa en el mercado de datos, identificar el riesgo de privacidad para los individuos derivado del procesamiento de datos personales que llevan a cabo las organizaciones estatales y realizar una aplicación de la metodología de Privacidad Diferencial para evaluar su efectividad en la protección de este tipo de datos. Finalmente, establecer los aspectos principales para el diseño de una estrategia responsable de privacidad en el Estado argentino.

# Capítulo 1: Organizaciones y el mercado de datos personales

## Introducción

El desarrollo tecnológico de las últimas décadas ha facilitado la creación de redes online que conforman un espacio digital de interacciones (Zysman y Kenney, 2018). Estas, facilitan el intercambio entre organizaciones e individuos, conformando un mercado de datos (Täuscher y Laudien, 2018). En particular, el Estado se ha visto beneficiado ya que le ha permitido crear su propia red tecnológica. Esta se denomina gobierno electrónico, caracterizada por facilitar el encuentro entre los propios actores estatales, organizaciones no gubernamentales, empresas, asociaciones de ciudadanos y los ciudadanos individuales (OCDE, 2014).

De la interacción entre los actores en este nuevo espacio, surgen grandes volúmenes de datos (Schmarzo, 2013). Si bien en un principio han impactado en los procesos productivos tradicionales del Estado, la posibilidad de contar con estos de manera dinámica e inmediata para la toma de decisiones le ha facilitado la generación de valor público (Naser, 2021). Es decir, a través de su procesamiento le ha permitido obtener más información para la realización de políticas públicas eficientes y eficaces (Bryson et al., 2015). Esto a su vez ha favorecido a los individuos ya que reciben servicios o bienes públicos de mayor calidad.

En este contexto, surge el siguiente interrogante: ¿cómo se produce la toma de decisiones en organizaciones del Estado en base a la información generada a partir de datos recolectados en el mercado de datos? Para responder esta pregunta el objetivo a desarrollar en el presente capítulo consiste en describir que es una organización basada en datos y cómo se lleva a cabo el proceso de toma de decisiones. Pero también, resulta imprescindible comprender el proceso constitutivo del mercado de datos y como se crea valor público en este.

A fin de realizar una comprensión integral del Estado como una organización basada en datos, el presente capítulo se divide en dos secciones. En la primera sección se desarrolla la conceptualización de una organización basada en datos en el contexto de la economía de la información. Luego, se especifica su definición a partir de tres dimensiones claves: las tecnologías de *Big Data*, la capacidad de generación de información y la gestión de datos. Finalmente se analiza como la conjunción de estas dimensiones se materializa en la toma de decisiones.

En la segunda sección, se analizan las principales particularidades del mercado de datos en el cual participa el Estado. En primer lugar, se describe la constitución del mercado de datos donde la tecnología es el factor esencial de poder. Luego, se analiza como los datos capturados a través de esta por parte del Estado son utilizados para generar conductas humanas futuras identificando los riesgos asociados. Por último, se especifica sobre la generación de valor público en base a los datos y se expone la responsabilidad que ello conlleva.

### 1.1 Las organizaciones basadas en datos

En el contexto de la economía de la información (Seidl, 2021), surge un nuevo tipo de organización cuya característica definitoria es que la toma de decisiones se realiza en base a datos (McAfee y Brynjolfsson, 2012), facilitada por el uso intensivo de tecnología. A diferencia de las organizaciones definidas en términos clásicos (Haskel y Westlake, 2017), estas conciben al dato como un bien primario que debe ser transformado para convertirlo en valor agregado. Es en este sentido que el dato es concebido como la unidad mínima de información. Siguiendo a autores como Javier Gil Flores et. al. (1994), el dato es el valor que toma una variable bajo una unidad de análisis, generados por la acción humana de forma individual o colectiva en el entramado.

A diferencia de los bienes primarios de producción – que en términos clásicos son los obtenidos de la naturaleza principalmente–, el dato posee la particularidad de ser un bien intangible (Haskel y Westlake, 2017). Pero también se presenta la particularidad de que el capital esencial utilizado para llevar a cabo la generación de valor también es intangible. Como sostienen Haskel y Westlake (2017), además de bienes de capital físico (computadoras, servidores, recursos humanos especializados), la inversión en capital en este tipo de organizaciones refiere principalmente a *software*, datos, procesos, conocimiento, ideas. Estos resultan esenciales ya que los datos se encuentran dispersos en la sociedad por lo que requieren de una coordinación adecuada para ser tratados eficientemente en pos de cumplir con un objetivo concreto.

Este último punto resulta de importancia, ya que expone la complejidad a la cual deben enfrentarse las organizaciones a la hora de trabajar con datos. Deberán atravesar por un proceso de adaptación que les permita determinar el mejor uso de los recursos (Milgrom, 1992) en este nuevo contexto. De esta manera, y con la llegada del siglo XXI, se comienza a transitar una revolución organizacional impulsada por los datos (Schmarzo, 2013). Esta ha

sido potenciada por el uso masivo de internet el cual se ha convertido en el canal facilitador de acceso a datos voluminosos. Si bien, en un comienzo resultó en una problemática para aquellas organizaciones que poseían estructuras tradicionales, con el transcurso del tiempo la innovación tecnológica facilitó herramientas para acceder a los datos.

Ahora bien, el impacto causado por la revolución organizacional ha alcanzado tanto a aquellas privadas como públicas. En las primeras ha dado origen al surgimiento de lo que se conoce como organizaciones de plataformas que se caracterizan por ser una red online de interacciones (Zysman y Kenney, 2018) – por ejemplo, Mercado Libre, Amazon, Google, entre muchas otras–. Estas actúan de intermediarias entre agentes de la sociedad poniendo a su servicio una infraestructura digital para el intercambio haciendo uso intensivo de la tecnología. Específicamente, son negocios basados en información que es generada en la red a costo cero (Moazed y Johnson, 2016).

Las organizaciones públicas estatales, en cambio, no tienen un objetivo comercial. Estas, históricamente, han recopilado datos para realizar políticas públicas cuyo resultado la sociedad lo recibe a través de diversos servicios (o bienes) públicos (Bryson et al., 2015). Es decir, son generadoras de valor público. Por este, se entiende a todo asunto de interés público (Naser, 2021) abarcando desde bienes y servicios que satisfagan un deseo o necesidad de los diferentes actores de la sociedad hasta el uso legítimo de recursos para lograr diversos propósitos públicos. Pero, aun así, también han sido impactadas por dicha transformación dando lugar al surgimiento de lo que se conoce como gobierno digital.

Siguiendo las recomendaciones del consejo de La Organización de Cooperación y Desarrollo Económicos (OCDE, 2014), se define a un gobierno digital como:

> "*Digital Government refers to the use of digital technologies, as an integrated part of governments' modernisation strategies, to create public value. It relies on a digital government ecosystem comprised of government actors, non-governmental organisations, businesses, citizens' associations and individuals which supports the production of and access to data, services and content through interactions with the government*" (pág. 7)

De esta manera, se considera al Estado como una organización basada en datos que requiere gestionarlos a través del uso de tecnología. Esto le permitirá obtener información que será utilizada como evidencia para la toma de decisiones. Con esta podrá diseñar políticas públicas eficientes y eficaces para la generación de valor público.

En este contexto organizacional, resulta fundamental comprender que se entiende por tecnología y cuál es el papel que desempeña a la hora de trabajar con datos. Para ello, en la primera sección, se desarrolla este concepto; luego, se propone un abordaje conceptual de la capacidad de cálculo derivada consecuentemente del uso tecnológico siendo lo que configura a este tipo de organización dentro de la figura de centro de cálculo. Se continúa planteando el rol necesario de la gestión de datos que da forma acabada a la mencionada figura. Para finalizar, teniendo en cuenta lo anterior, se establece como se desarrolla la toma de decisiones.

### 1.1.1 El papel de la tecnología de la información

A diferencia de la era de la industrialización, desde fines del siglo XX, comienza a referirse a Tecnologías de la Información y Comunicaciones (TIC)[1]. Con la llegada del siglo XXI, caracterizado por el uso masivo de Internet[2], se incorpora el término Tecnología Digital. Esta se caracteriza por ser facilitadora de "la aplicación de conocimiento e información a aparatos de generación de conocimiento y procesamiento de la información/comunicación" (Castells, 2004). En este sentido, podría decirse que es una tecnología de propósito general (Seidl, 2021) –como lo fue la máquina de vapor en su momento– que requiere de cambios organizativos y conceptuales para lograr la generación de valor.

El impacto causado por la tecnología digital en las organizaciones ha dado lugar al concepto de *Big Data*. En un inicio, asociado a la disponibilidad de grandes volúmenes de datos, se basa en la idea de que no es posible manejar aquello que no se puede medir (McAfee et al., 2012). En consecuencia, la organización requiere de un plan para el manejo de recursos frente a la posibilidad de obtener mejores resultados que derivan en un mayor conocimiento. En este mismo sentido, Constantiou y Kallinikos (2015), consideran que se trata de la estrategia necesaria a tomar para llevar a cabo el procesamiento de datos que permitan la extracción de información. Y resulta fundamental en la medida que su crecimiento a velocidad exponencial y su diversidad dificulta el descubrimiento de patrones para su generación.

[1] La UNESCO define a las TIC como "Conjunto diverso de herramientas y recursos tecnológicos utilizados para transmitir, almacenar, crear, compartir o intercambiar información"
[2] Internet es un mecanismo para la difusión de información y un medio para la colaboración e interacción entre las personas y sus computadoras sin importar su ubicación geográfica (Leiner, B. et al., 2009).

Con el correr de los años, el concepto evoluciona hacia un sistema de mayor complejidad, abarcando un conjunto de tecnologías que fueron transformando de manera disruptiva a la sociedad en su conjunto. En este sentido es posible hablar de un ecosistema de *Big Data* (Kolanovic y Krishnamachari, 2017) que ha tenido un alto impacto generando desde nuevas formas comunicacionales hasta cambios de hábitos en los individuos. Es decir, se configura una nueva cultura tecnológica[3] en la sociedad moderna. De esta manera, el *Big Data*, es el proceso por medio del cual las organizaciones convierten los datos generados por los individuos en información.

Sin embargo, este sistema tecnológico otorga un poder superior a la organización constituyéndose en un "imperativo tecnológico" como lo define Oscar H. Gandy Jr (2021). Este poder no reside en la tecnología en si misma sino en las personas que la utilizan ya que obtienen más información y conocimiento. En este sentido, es notable la cantidad de personas –como nunca en la historia de la humanidad– que conocen tanto sobre los individuos debido a la recopilación masiva de datos a través de medios digitales. Estos son utilizados en la mayoría de las ocasiones para favorecerlos, pero también pueden ser manipulados en su perjuicio, incluso violando el derecho a su privacidad (AEPD, 2020). Este riesgo asociado será abordado en el capítulo 2.

Las tecnologías del *Big Data*, sin duda, le permiten al Estado ofrecer servicios con eficacia, eficiencia, disponibilidad, interoperabilidad y con racionalización de los recursos, entre otros beneficios (AEPD, 2020) en la medida que facilitan la ejecución de cálculos en base a datos. Por ejemplo, en el contexto de la pandemia COVID-19[4], la facilidad y rapidez de recopilación de datos para realizar proyecciones confluyó en medidas de confinamiento como fue el caso de Argentina en 2020[5]. Es así como, le permite concretar la toma de decisiones basada en evidencia mediante la realización de cálculos, incluso de manera anticipada.

[3] Como sostiene Manuel Castells (2002, Sesión 1), "se trata de una sociedad (cultural moderna) en la que las condiciones de generación de conocimiento y procesamiento de información han sido sustancialmente alteradas por una revolución tecnológica centrada sobre el procesamiento de información, la generación del conocimiento y las tecnologías de la información"

[4] El 12 de febrero de 2020, un virus de rápida expansión se lo denomina oficialmente enfermedad infecciosa bajo el nombre de Coronavirus 2019 (COVID-19). A partir de entonces, el virus comienza a ser denominado COVID-19 alcanzando a diferentes ciudades del mundo. El 11 de marzo de 2020 la OMS lo declara una pandemia.

[5] Mediante el decreto 297/2020 emitido el 19 de marzo de 2020, las Autoridades Nacionales de Argentina, decretan el Aislamiento Social Preventivo y Obligatorio a fin de proteger la salud pública.

### 1.1.2 Configuración de la capacidad de acción

Al decir que las organizaciones realizan cálculos, implica que existe una necesidad preexistente. Esta radica en poder hacer inteligible los datos (Rivoir y Morales, 2019). Dado que los individuos son interpretados a través de los datos que generan a partir de su interacción con medios digitales en el ámbito social, resulta necesario convertirlos en información. Y son las tecnologías de *Big Data* las que permiten llevar adelante este proceso generándose conocimiento para la toma de decisiones (Constantiou y Kallinikos, 2015). Por lo tanto, la realización de cálculos puede ser concebida como una configuración algorítmica –secuencias de pasos lógicas– a través de la cual estos se materializan.

Tal materialización, en definitiva, permite convertir lo intangible (datos) en objetos materiales (Haskel y Westlake, 2017) –informes de resultados, gráficos, tablas de datos, entre otros–. Este proceso de cálculo puede concebirse de modo general en tres pasos (Callon y Muniesa, 2003). En primer lugar, las entidades deben ser aislados en un único espacio –es decir, los datos deben ser recolectados y almacenados–. Luego deberán asociarse con otras para ser transformados siendo posible a través de aplicar conocimiento a la tecnología. Finalmente, se deberá extraer un resultado, siendo el valor agregado generado para la toma de decisiones.

El proceso de cálculo es llevado a cabo dentro de la organización, lo que permite interpretarla como un centro de cálculo. Bruno Latour (1987) –quien ha desarrollado este concepto – expone como el conocimiento de diferentes disciplinas, que es propio de cada científico especializado, y la tecnología facilitadora de los cálculos son reunidos en este. Lo importante aquí, es que de esta interacción surge una capacidad de acción que es superior a la del individuo, otorgándole un poder superior. Pero, además, este poder es producido (Latour, 2005) siendo que ya no se deriva de la clásica acumulación de capital – es decir, los objetos pasan a un segundo plano- sino de dicha capacidad.

En este sentido, la configuración de la capacidad de cálculo en la organización resulta en un híbrido colectivo (Callon y Muniesa, 2003) en la medida que no solo se trata de cálculos mentales realizados por individuos especializados, sino que son integrados con el uso de la tecnología para generar información. De esta manera, el Estado en su conjunto es un agente con capacidad de acción que se apropia de los datos generados por los individuos a cambio de brindarles un servicio. Para que este proceso se lleve a cabo de manera eficiente,

necesariamente deberá implementar un sistema de gestión de datos por medio del cual podrá finalmente tratar a la información como un recurso integrado de calidad.

### 1.1.3 El rol de la gestión de datos

En el proceso de cálculo descripto en el apartado anterior, se establece la necesidad de que los datos recolectados deben asociarse con otros de forma tal que sea posible extraer resultados. Estos otros datos son definiciones centrales transversales a toda la organización, que facilitan la convivencia de diferentes sistemas y procesos para la construcción de información. La claridad de su definición es un objetivo central de un plan de gestión de datos, ya que permitirán integrar, analizar y extraer el valor que contienen los datos (Cleven y Wortmann, 2010).

De esta manera la gestión de datos toma un rol central, dado que se trata de un proceso a partir del cual la organización gestionará la información como un recurso integrado (McKeen y Smith, 2007), frente a la necesidad de establecer reglas y políticas que permitan garantizar la confiabilidad en los datos por parte de los usuarios finales. En este, las entidades centrales (o datos maestros) se convierten en un activo clave dentro una organización. Por consiguiente, la falta de una adecuada gestión de estos puede derivar en el surgimiento de diferentes problemas, como fallas en el funcionamiento de procesos hasta una mala toma de decisiones por generación de resultados erróneos. En este sentido, el rol principal de la gestión de datos es la gestión del riesgo asociado al uso de datos.

Frente a este riesgo, la organización deberá tener en cuenta cinco aspectos principales: la estructura, la arquitectura, el gobierno, el proceso y la calidad de definiciones centrales (Cleven y Wortmann, 2010). La estructura implica establecer un acuerdo sobre cada definición para establecer una estructura relacional entre las diferentes entidades. La arquitectura implica establecer el diseño de sistemas adecuados que den soporte al ciclo de vida de estos. En cuanto a la gobernanza de datos, requiere establecer una articulación entre todas las áreas de la organización para delinear adecuadamente una estructura de datos sólida, definiendo roles y responsabilidades sobre los datos en cada área. En línea con este punto, en el proceso de datos centrales se determinará la manera en que estos deben ser creados, usados, mantenidos y almacenados. Finalmente, la calidad de las definiciones centrales resultará como consecuencia de la implementación correcta de las cuatro fases anteriores.

Ahora bien, la responsabilidad sobre el tratamiento de los datos recolectados por el Estado se vuelve fundamental. Esto se debe a que no solo posee estadísticas gubernamentales – información agregada sobre grupos sociales – sino que también posee datos personales sobre cada individuo (Gandy Jr, 2021), existiendo un riesgo asociado como se esbozó en el aparatado 1.1.1. Ante esto, se requerirá del diseño de un sistema de gestión responsable que trate este riesgo de manera proactiva – asignando responsabilidades por defecto a quienes se encuentren involucrados – para establecer una estrategia que incorpore a la protección de datos personales durante todas las etapas de su procesamiento (AEPD, 2019). Sobre este punto se profundizará en el capítulo 2.

De esta manera, la organización logrará establecer un sistema de datos consolidado que le permitirá – en tanto agente de cálculo – reducir los riesgos asociados. Esto le permitirá llevar adelante una gestión de datos eficiente mediante la aplicación de tecnológicas de *Big Data.* En consecuencia, la extracción de resultados será utilizado como información para la toma de decisiones.

### 1.1.4 La toma de decisiones en base a datos

En los centros de cálculos, definidos en el aparatado 1.1.2, se produce conocimiento a través de la interacción de los individuos especializados (analista de datos, científicos de datos, especialistas en *software*, entre otros) con la tecnología. A diferencia del pasado, cuando los científicos elaboraban teoría de la ciencia prácticamente en abstracto, ahora cuentan con enormes volúmenes de datos y las herramientas para procesarlos facilitándoles hacer ciencia empírica (Rhodes y Lancaster, 2019). De esta manera, la ciencia deja de ser un conjunto de meras posibilidades teóricas para materializarse en conocimiento basado en evidencia, es decir, en información.

Como sostienen Rhodes, T., Lancaster, K., & Rosengarten, M. (2020), los modelos matemáticos y sus proyecciones proveen control basada en evidencia. Es decir, las proyecciones producen conocimiento, ofreciendo una sensación de seguridad a través del cálculo. De esta manera, dejan de ser meras posibilidades teóricas para concretarse en toma de decisiones que afectan las acciones en el presente. Por lo tanto, podría decirse que, en la sociedad moderna, la ciencia es vista como la ciencia que participa en la materialidad relacional de implementar y evidenciar (Rhodes y Lancaster, 2019).

La medida de aislamiento social impuesta en Argentina en pandemia mencionada con anterioridad constituye un claro ejemplo. Permite de manera clara notar como la capacidad de cálculo es ejercida por el Estado para obtener información que es utilizada para la toma de una decisión que modificó el accionar social de los argentinos en aquel momento. Pero, sin lugar a duda, también permite ver como esta capacidad es utilizada para el ejercicio del poder de control anticipatorio basado en evidencia (Rhodes, et al., 2020). Es en este sentido que en el centro de cálculo se produce la integración de tecnología e individuos con conocimiento para la generación de información.

Sin ir más lejos, en el mismo escenario de pandemia en Argentina, las Autoridades Nacionales pusieron a disposición de la sociedad un aplicativo móvil denominado CuidAR[6] con el fin de obtener más datos para realizar políticas públicas de control de la pandemia (Salaberry y Herrera, 2021). Lo interesante de este es que recopila datos personales –incluso de geolocalización– de los usuarios, que son utilizados para ejercer un poder necesario de control, pero a su vez permite conocer y determinar comportamientos humanos a través de su procesamiento. Esto pone en evidencia el poder de instrumentación que brinda el uso de tecnología (Zuboff, 2019) así como la necesidad de una gestión de datos basada en responsabilidad como se mencionó en el apartado anterior.

### 1.2 Constitución del mercado de datos personales

La organización descripta en el apartado anterior interactúa con los individuos, lo cual comienza a dar forma a un mercado de datos, en particular de datos personales. Estos, que les pertenecen a los individuos, son apropiados por este centro de cálculo ya que poseen un poder superior. El objetivo de esta sección es describir su constitución para comprender su funcionamiento.

Los datos personales refieren a un atributo o conjunto de atributos que permiten realizar un perfilamiento del individuo (Custers, 2013), es decir, una identificación unívoca de este. Por ejemplo, el nombre y apellido, número de documento, edad, dirección postal, entre otros; pero también los identificadores únicos de los dispositivos móviles, datos de geolocalización, así como datos biométricos, también conforman parte de estos. Todos estos

[6] Acerca de la App CuidAR: https://www.argentina.gob.ar/jefatura/innovacion-publica/acciones-coronavirus/aplicacion-y-tableros-de-gestion

en conjunto permiten conocer el comportamiento de un individuo. Por ello resultan muy atractivos para ser recolectados por las organizaciones.

A diferencia de la concepción económica clásica, este mercado es concebido como una red sociotécnica (Caliskan y Callon, 2010), concepto que se ampliará en el aparatado 1.2.1. En esta los agentes que la conforman ya no son meros agentes de intercambio, sino agentes sociotécnicos dotados de capacidad de cálculo. Esta propuesta no se trata de una oposición a la concepción económica clásica, sino más bien, dado que son los agentes quienes están dotados de esta capacidad es que pueden materializar el valor para el intercambio – es decir, pueden convertir un activo intangible (los datos) en algo objetivo (información)–.

Ahora bien, en este escenario el poder es producido (Latour, 2005) y ya no se deriva de la clásica acumulación de capital – es decir, los objetos pasan a un segundo plano-. Sino que la interacción humana en la red genera un flujo constante de datos. Los agentes dotados de capacidad de cálculo se apropian de estos y los utilizan como conductos para modificar conductas humanas. Es así como, esta red puede ser interpretada como un mercado de datos conductuales (Zuboff, 2019), siendo la tecnología inteligente el instrumento que brinda la posibilidad de convertir al dato en un predictivo de conductas humanas futuras.

De esta manera, la organización pública del Estado – en tanto agente con capacidad de cálculo superior– interactúa con los individuos ya que posee un incentivo para la extracción de datos (Mazzucato, Entsminger y Kattel, 2020) dado que reconoce su cualidad de ser generador de valor. Para ello pone a disposición de la sociedad una red digital (gobierno electrónico). Y los individuos, que poseen necesidades particulares (Zuboff, 2015), están dispuestos a ceder sus datos (principalmente los personales) a través de esta para poder satisfacerlas. A cambio reciben una contraprestación de servicio público (o un bien público).

Frente a la conformación de este mercado como una red sociotécnica, resulta necesario profundizar acerca de su configuración. Para ello, en la primera sección, se lleva adelante este desarrollo. Luego, teniendo en cuenta lo anterior, se pone de relieve como los datos personales son usados como conductos para ejercer el poder de control del Estado. Finalmente, se amplía acerca del proceso de generación de valor en la red.

### 1.2.1 El mercado como una red sociotécnica

Desde una perspectiva clásica, y siguiendo a los autores Carl Shapiro y Hall Varían (1998), el mercado de datos es concebido como un espacio de intercambio entre oferentes y

demandantes, donde el bien principal es la información. Esta es básicamente cualquier cosa que pueda ser digitalizada. En este sentido, toda página web, base de datos, libros, entre muchos otros, constituyen bienes de información. Y lo son, porque existe una demanda que está dispuesta a pagar un precio por esta. Además, desde un enfoque del lado de la oferta, existe una estructura de costos para producirla, aunque un muy bajo costo para reproducirla.

Esta última característica particular de la información digitalizada, no se trata de otra cosa que de la existencia de costos marginales tendientes a cero. De esta manera, existen costos fijos de producción inicial de información y costos marginales insignificantes. Una implicancia directa es que el precio, entonces, no está determinado en base a los costos – en términos de la teoría clásica de formación de precios[7]– sino de cuanto está dispuesto a pagar el consumidor (Shapiro y Varían, 1998). Dado que cada consumidor tendrá un mayor o menor interés por este bien, el precio será diferente para un mismo bien. Por lo tanto, existirán precios diferenciales.

A su vez, el gran flujo de información que circula rápidamente en internet sobre un producto requiere que el oferente logre que le llegue al consumidor interesado solo la información necesaria y clave. Pero también le permite captar información sobre los consumidores que puede ser utilizada para generar nuevas estrategias de ventas, siendo posible gracias a la estructura tecnológica disponible. En consecuencia, la economía de la información está caracterizada por la información como un bien y su asociado que es la tecnología (Shapiro y Varían, 1998).

Pero esta concepción resulta escasa para poder explicar cuál es el rol del poder de cálculo que permite la apropiación de datos por parte de las organizaciones y como impacta sobre los individuos. Como sostienen Culpepper y Thelen (2020), esta visión de mercado no resulta suficiente para estudiar el poder de la organización dado que no solo se trata del dominio de un mercado en función de tamaño y escala. Sino del ejercicio de poder que ejercen sobre los individuos. Por esta razón, resulta necesario concebir al mercado de datos como una red ensamblada donde se disputa el poder.

En este sentido, la propuesta de Koray Caliskan y Michel Callon (2010) resulta más adecuada. Partiendo de la concepción de centro de cálculo de Bruno Latour presentada en el apartado 1.1.2., estos autores conciben al mercado de datos como una red ensamblada que

[7] Para la cual el mercado es un espacio abstracto en el que la demanda agregada y la oferta interactúan entre sí hasta producirse un ajuste que determina el precio (Coumot, 1838).

se enmarca en lo que denominan '*agencements'*. Esta propone la idea de una división entre los agentes que participan. Por un lado, están los que crean esta red, y por el otro los elementos que se ensamblan, generándose una lucha de poder. La capacidad de acción es lo que les permite actuar de diferente manera según la posición adquirida en la red. Pero también les permite adquirir nuevas capacidades, mientras que las cosas son tan solo un espejo que reflejan su valor.

De esta manera, todo el ensamblaje de la red sociotécnica se configura dentro de una figura denominada agencia, cuya clave definitoria es su capacidad de acción. Se trata entonces de una red ensamblada entre organizaciones e individuos. Pero también con dispositivos (tecnologías) y humanos (especializados) sin los cuales la ejecución del poder de cálculo por parte de la organización es imposible. Los individuos que interactúan también son agentes de cálculo pero que utilizan la red para satisfacer necesidades. A diferencia de la organización, su capacidad de cálculo es muy inferior ya que no poseen el herramental necesario para poder ejecutarla.

A su vez, como la tecnología – en tanto dispositivo para realizar cálculos en base a datos– la posee la organización, entonces el Estado puede conocer y determinar comportamientos humanos (Zuboff, 2019) en esta red mediante el procesamiento de datos personales y aplicando conocimiento científico para obtener resultado que serán utilizados para cumplir un fin determinado. Esto lo posiciona con un poder de instrumentación superior dentro de la red frente a los individuos, generándose una asimetría. Particularmente, esta asimetría es una asimetría de información, es decir, el Estado posee más información que los individuos que la que estos poseen de sí mismos. Este punto será especificado con mayor detalle en el capítulo 2.

### 1.2.2 Los datos personales conductuales

Ante la insuficiencia de la explicación de la concepción económica clásica del mercado, Zuboff (2015) realiza una crítica de esta. La misma consiste en plantear que la necesidad que poseen los individuos y buscan satisfacerla interactuando con el Estado a través de un contexto digital en red (gobierno digital), genera una dependencia que le otorga poder de control. Esta necesidad es la que los predispone para ceder sus datos, incluso los personales. De esta manera, en la red se producen estos, y la organización puede extraerlos y apropiárselos otorgándole mayor valor predictivo de información.

Pero tal apropiación de los datos suele realizarse de manera silenciosa, en una especie de acuerdo tácito. En este sentido, Zuboff, (2015) plantea la existencia de una relación subjetiva entre la red y los individuos que carece de consentimiento explícito. Desde esta perspectiva, entonces, las subjetividades son la fuente de donde los datos son extraídos y no ya por la mera existencia de la red. Para convertir estas subjetividades en resoluciones de objetivos concretos, la organización utiliza los dispositivos de cálculo. A su vez, esto genera un ciclo de retroalimentación donde ya no existe contrato alguno entre partes sino más bien un poder de control que Zudoff denomina poder de vigilancia.

En este contexto, el dato personal, es visto como un futuro conductual. Por conducto se entiende a aquel elemento por medio del cual las conductas humanas pueden ser modificadas[8] (Zuboff, 2019). Desde esta perspectiva es entonces como se configura un mercado de datos conductuales. En este, es el uso del herramental tecnológico por especialista con conocimiento superior en el centro de cálculo el que brinda la posibilidad de convertir al dato en un predictivo de conductas humanas futuras. De esta manera, se conforma un poder de vigilancia que opera a través de un poder superior de cálculo generando información que habilita el saber más sobre los individuos, mientras que estas operaciones son desconocidas (o difícil de comprender) para estos.

Cierto grupo de los individuos acepta que se produzca la recopilación de los datos personales asociados con su necesidad. Es decir, casi cualquier dato recopilado puede justificarse si la organización puede demostrar que es pertinente hacerlo para tomar una decisión en su favor (Gandy Jr, 2021). En cambio, otros, buscan encontrar el límite de su ignorancia. Esto es, algunos manifiestan que consideran que trae aparejado consecuencias sobre sus vidas en el sentido de que la utilización de la tecnología resulta cuanto menos invasiva, generándoles preocupación. En este sentido, es que se considera adecuado que el Estado implemente pautas claras y en la medida de lo posible, comprensibles para cualquier individuo de la sociedad.

### 1.2.3 La creación de valor

A pesar de las complejidades sociales que se surgen en la red, el valor se origina por la posibilidad de ejecutar cálculos (Çalışkan, y Callon, 2010). Los agentes dotados de habilidades e instrumentos desarrollan esta capacidad siendo la que les permite valorizar, es

[8] Un claro ejemplo es el expuesto en el apartado 1.1.4.

decir, los cálculos se materializan. De esta manera, la configuración algorítmica de la red facilita que los agentes pueden llevar a cabo el proceso de valorización. De aquí que sin la existencia de esta el mercado deja de existir, ya que sería imposible encontrar los ensambles sociotécnicos reales que en él se producen.

En este espacio de encuentro entre los agentes, el flujo de datos es continuo y brindan la posibilidad de conocer detalladamente a los individuos. Los datos personales – como fueron definidos anteriormente– que son posibles de extraer permiten generar un valor sin precedente. Esta acción permite construir lo que se conoce como perfilamiento. Este consiste en generar la determinación del comportamiento individual mediante la agregación de rastros de las actividades realizadas (Gandy Jr, 2021). En consecuencia, el poder de control que le brinda al Estado para la realización de políticas públicas, sin lugar a duda es de una importancia superior como por ejemplo en el caso de una pandemia. Además, la eficiencia de acción generada le permite afianzar la relación con los ciudadanos.

De esta manera, en la concepción moderna de la gestión pública el Estado tiene un papel especial como garante de los valores públicos (Bryson et al., 2015). Esto porque contribuye a generar el bienestar colectivo. Por tanto, el valor público creado en el escenario planteado se vuelve de vital importancia. Pero al mismo tiempo invita a exponer la responsabilidad que conlleva el contar con una capacidad de cálculo tan poderosa. Por esta razón, no debe perderse de vista los riesgos asociados a la privacidad que se fueron deslizando y que serán profundizados en el capítulo 2.

**Conclusión**

En el presente capítulo, se ha expuesto como el Estado es concebido como una organización basada en datos. Se ha mostrado como lleva adelante un proceso de creación de valor público en base a datos mediante el uso de tecnología. Estos son recolectados en el mercado y utilizados como conductos para la modificación de conductas humanas en pos de generar políticas públicas eficientes y eficaces que benefician al colectivo social. Pero también se explicitó como esta capacidad de acción superior – respecto del individuo– que es ejecutada a través del uso de tecnología le otorga poder de control.

La gestión pública del Estado, en su concepción moderna, pone al servicio de la sociedad una red de interacción (gobierno electrónico). De esta obtiene datos que mediante realización de cálculos le permite crear valor público. En este proceso las tecnologías de la información

tienen un rol central ya que son el dispositivo por medio del cual se llevan a cabo. Pero también le otorgan un poder de control superior en la medida que le permite obtener un mayor conocimiento sobre los individuos.

A su vez, la tecnología por si sola resulta insuficiente para convertir los datos en información. Es necesario combinarla con recursos humanos especializados para llevar a cabo su producción. De esta manera, la organización puede llevar adelante el proceso de cálculo necesario. En consecuencia, el Estado es concebido como un centro de cálculo. En este la gestión de datos toma un rol central, ya que le permite integrar, analizar y extraer el valor que contienen los datos. De esta manera, la organización logra establecer un sistema de datos consolidado que le facilita reducir los riesgos asociados. A su vez, las tecnológicas de *Big Data* le permite llevar adelante una gestión eficiente de datos para la extracción de resultados. Finalmente, los resultados son utilizados como información para la toma de decisiones.

A partir de esta consolidación de la organización, es que puede interactuar en el mercado de datos. De esta manera, se ensambla en la red sociotécnica tomando forma de agencia. Su clave definitoria es la capacidad de acción para obtener principalmente datos personales. Estos son utilizados como conductos, es decir, son convertidos en un predictivo de conductas humanas futuras para la realización de políticas públicas. De esta manera, obtiene un poder de vigilancia ya que la habilita a saber más sobre los individuos, mientras que las operaciones llevadas a cabo son difíciles de comprender por ellos.

En consecuencia, el valor se origina por la posibilidad de ejecutar cálculos siendo los agentes dotados de habilidades e instrumentos los que desarrollan esta capacidad para su materialización. De esta manera, la configuración algorítmica de la red facilita que los agentes pueden llevar a cabo este proceso de valorización. De aquí que sin su existencia el mercado deja de existir, ya que sería imposible encontrar los ensambles sociotécnicos reales que en él se producen.

Al ser las organizaciones del Estado las que poseen el herramental para llevar adelante los cálculos, obtienen un poder superior de acción. Pero este poder, expone la responsabilidad que ello conlleva. Por esta razón, no debe perderse de vista el riesgo sobre la privacidad de los individuos que fue deslizado a lo largo de este capítulo y que será profundizado en el capítulo 2 a continuación.

# Capítulo 2: El modelo de Privacidad Diferencial y los conceptos involucrados

## Introducción

El procesamiento, análisis e incluso publicación de datos personales por parte de las organizaciones públicas estatales conlleva un riesgo asociado para la privacidad de los individuos (AEPD, 2020). La privacidad refiere al control que un individuo tiene sobre el acceso que otros podrían tener a cuestiones personales como ser características propias que decide mantener como privadas (Gandy Jr, 2021). En este sentido, la privacidad de datos personales se define como el interés que tiene un individuo en poder controlar sus propios datos (Clarke, 1999).

La recolección de datos personales por parte de las organizaciones puede implicar que quien los manipule genere un perjuicio para los individuos titulares de estos. Como consecuencia puede incurrirse en una violación de su privacidad (AEPD, 2020). Frente a esta problemática, su gobernanza resulta necesaria. Ello implicará establecer metodologías que durante todo su procesamiento den cierta garantía de resguardo, pero también la asignación de responsabilidades sobre cada participante involucrado.

En este contexto, surge el siguiente interrogante: ¿qué metodología es necesarias para garantizar la privacidad de datos personales en organizaciones públicas estatales? Para responder esta pregunta el objetivo del presente capítulo consiste en presentar el modelo de Privacidad Diferencial, así como todos los conceptos involucrados. Pero para ello, resulta imprescindible comprender en primer lugar de que se trata la privacidad y, en particular, la privacidad de datos personales.

Para poder establecer una metodología que brinde privacidad de datos personales, el presente capítulo se divide en tres secciones. En la primera sección se desarrolla conceptualmente a la privacidad. En primer lugar, se explicita una concepción de esta. Luego, se aborda el concepto de privacidad de datos personales. Finalmente se especifica la necesidad de abordar este riesgo desde el diseño y por defecto.

En la segunda sección, se aborda la presentación la metodología de la Privacidad Diferencial que da cierta garantía sobre la privacidad de datos personales. Para ello, en primer lugar, se define su concepto. Luego, se ejemplifica su lógica de operatoria. Por último, se expone la necesidad de su modelización.

Finalmente, en la tercera sección, se desarrolla el modelo de Privacidad Diferencial presentando los principales conceptos involucrados. En primer lugar, se define que es un algoritmo aleatorio. Luego, se presenta la distribución de Laplace. A continuación, se aborda la explicación del mecanismo de Laplace. Por último, especifica el modelo de Privacidad Diferencial teniendo en cuenta los conceptos previamente presentados.

### 2.1 El concepto de privacidad

El puntapié inicial para concebir a la privacidad es que se trata de un derecho humano universal que se encuentra contenido en el artículo 12[9] de la Declaración Universal de los Derechos Humanos de 1948. En este se refiere a que todo individuo tiene derecho a vivir en privacidad (Salaberry y Herrera, 2021). Entonces, basándose en este, la privacidad puede ser interpretada como un valor social con alcance universal. Una consecuencia directa es que el individuo tiene el derecho a controlar quien puede acceder y quien no a las cuestiones de índole personal.

De esta manera, la privacidad refiere al control que un individuo tiene sobre el acceso que otros podrían tener a cuestiones personales como ser características propias que decide mantener como privadas (Gandy Jr, 2021). Cuando este control es puesto en duda, surge entonces una preocupación en los individuos. Particularmente, en un contexto de mercado de datos en el cual se produce la interacción de los individuos con las organizaciones estatales, comienza a cuestionarse cuál es el nivel de acceso a información personal que se produce y que implicancia tiene en términos de pérdida de control de la privacidad para el individuo.

En este contexto, la preocupación ha ido subiendo de nivel en la medida que se produce más y más recopilación de datos a través de nuevos desarrollos tecnológicos en un espacio digitalizado. De esta manera, comienza a tomar mayor relevancia la sensibilidad de la información recopilada (Gandy Jr, 2021). Esta sensibilidad ha dado lugar a incorporar dos nuevas dimensionalidades sobre la discusión de la privacidad.

Como sostiene Gandy Jr, 2021, la sensibilidad de la información tiene dos dimensiones relacionadas: los aspectos personales – a diferencia de los aspectos públicos del individuo–,

[9] Art. 12: “Nadie será objeto de injerencias arbitrarias en su vida privada, su familia, su domicilio o su correspondencia, ni de ataques a su honra o a su reputación. Toda persona tiene derecho a la protección de la ley contra tales injerencias o ataques.” Organización de las Naciones Unidas. Declaración Universal de Derechos Humanos.

y el potencial daño sobre este que podría resultar si se hace un uso no autorizado de la información. Esto a su vez, ha dado lugar a cuestionar acerca de si se realiza combinación de información personal proveniente de diferentes fuentes para producir nueva información sensible. Particularmente, en las esferas de las organizaciones públicas del Estado es donde más se observa que se pone en duda este punto.

Frente a los cuestionamientos entorno a la manipulación de información personal, ha surgido diversa regulación. Un estándar de referencia mundial es el Reglamento General de Protección de Datos (RGPD)[10] de la Unión Europea, con entrada en vigor en el año 2018. Uno de los elementos esenciales que se establece en este es que el consentimiento es la base para el tratamiento de datos personales. Esto implica que 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 cambia. Además de respetar el derecho a privacidad del individuo, esta noción se constituye en un eje fundamental de un contrato social que permite generar confianza entre quienes proporcionan los datos y quienes los manejan.

### 2.1.1. La privacidad de datos personales

En la definición de la privacidad realizada en el apartado anterior, se expuso como a partir de la recopilación y uso de datos personales por parte de las organizaciones estatales, la privacidad del individuo puede verse afectada. Ya no solo se trata de un paparazzi tomando fotos a escondida o de una escucha telefónica no autorizada, sino que el uso de los datos recolectados también puede derivar en un riesgo para la privacidad. En este contexto en particular, entonces, surge la noción de privacidad de datos personales.

La privacidad de datos personales es definida como el interés que tiene un individuo en poder controlar sus propios datos (Clarke, 1999), o al menos ejercer su derecho de poder conocer que se hace con los mismos. Si bien, en el caso particular de las organizaciones públicas del Estado, en la mayoría de las ocasiones son utilizados para favorecerlos, también pueden ser manipulados en su perjuicio por terceros cuando son publicados abiertamente, pudiendo incurrir en la violación de su privacidad (AEPD, 2020).

[10] Reglamento General de Protección de Datos (RGPD). Disponible en https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32016R0679

De esta manera, la incorporación de datos personales resulta un desafío para toda organización del Estado. 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 datos.

En dicho marco, deben definirse los accesos, controles y responsabilidad sobre las personas que accedan a estos. Pero también deberá considerase prevenir efectos secundarios como ser la fuga de información privada (Kim y Cho, 2018). 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. En este sentido, la manipulación de datos personales conlleva asociado un riesgo que, mediante el diseño de un adecuado plan de gobierno de datos, podría ser controlado.

En esta misma línea, la Organización para la Cooperación y el Desarrollo Económicos (OCDE)[11], de la cual forma parte la República Argentina, establece unos principios generales vigentes desde el año 2013. Constituyen un marco general para delinear políticas claras para el resguardo de datos personales. 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 se deberán proteger 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.

Por otra parte, en Argentina existe la ley 25.326[12] de Protección de Datos Personales en cuyo objeto determina la necesidad de establecer "…la protección integral de los datos personales

[11] Personal Data Protection at the OECD https://www.oecd.org/general/data-protection.htm#:~:text=The%20OECD's%20data%20protection%20rules&text=The%20rules%20require%20that%20personal,for%20no%20longer%20than%20necessary.

[12] 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

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."

De esta manera, la gestión en torno de la privacidad de datos personales en contextos organizacionales se convierte en un punto central. Esto implicará el diseño de un sistema de gestión bajo consideraciones éticas y dentro del marco de la normativa vigente. A su vez requerirá que la protección de datos personales sea construida también por defecto, de forma tal que se establezcan ciertas responsabilidades para dar garantías sobre la privacidad de los individuos.

### 2.1.2. La privacidad desde el diseño y por defecto

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 (AEPD, 2019). Esto permitirá que las responsabilidades sean asignadas por defecto, informando y formando a quienes se encuentren involucrados en cada una de estas. Además, teniendo en cuenta la regulación vigente, permitirá determinar qué datos personales son los necesarios para cada fin específico dentro de un marco de gestión responsable.

Considerar la proactividad en la gestión del riesgo requiere de implementar metodologías que permitan reducir el riesgo asociado a la violación de la privacidad. Por ejemplo, definir anticipadamente posibles amenazas a partir de identificar debilidades de los sistemas contenedores de datos. De esta manera, se busca minimizar los riesgos en lugar de aplicar medidas correctivas una vez sucedido un incidente. Por lo tanto, también es necesaria 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 sus trabajadores (AEPD, 2019)

De esta manera, la privacidad establecida desde el diseño deberá quedar configurada para que sea parte integral de los sistemas, así como de todos los procesos llevados a cabo por la

organización. Para ello resulta necesario establecer evaluaciones de impacto sobre datos personales, así como evaluación de riesgo sobre los derechos de los individuos. Un proceso frecuente como aproximación a garantizar la privacidad, es el de anonimización de datos. Este permite eliminar o reducir el riesgo de reidentificación de su titular, sin que afecte la veracidad de estos (AEPD, 2019).

Pero también, como fue mencionado en el apartado 2.1.1, se deberá tener en cuenta que la combinación de datos personales provenientes de diferentes fuentes puede generar la identificación indirecta de un individuo. En particular, una situación frecuente donde se puede originar resulta ser cuando los datos son transferidos entre organizaciones. El receptor puede contar con datos que le habilita la posibilidad de recombinación con los datos recibidos.

Por lo tanto, tan solo una anonimización de datos puede no resultar suficiente para garantizar la privacidad de datos por parte de una organización. Y esto se debe a que no es posible anonimizar la totalidad de los datos preservando su utilidad para ser publicados. En consecuencia, se requerirá de la aplicación de una metodología diferente que evite incurrir en esta situación. Una opción posible puede resultar en la aplicación de Privacidad Diferencial.

## 2.2 La Privacidad Diferencial

La privacidad diferencial es un concepto introducido por Cynthia Dwork en 2008 en su trabajo "*Differential Privacy: A Survey of Results*". Consiste en una redefinición de la privacidad adaptada al problema de preservar la privacidad cuando se realiza un análisis de datos (Dwork y Roth, 2014). Esto es, se trata de tener conocimiento sobre el comportamiento de una población y no así de un individuo. Y será independiente de si se cuenta con información adicional. Es decir, mediante recombinación de datos provenientes de diferentes fuentes tampoco se incurrirá en la reidentificación de un individuo.

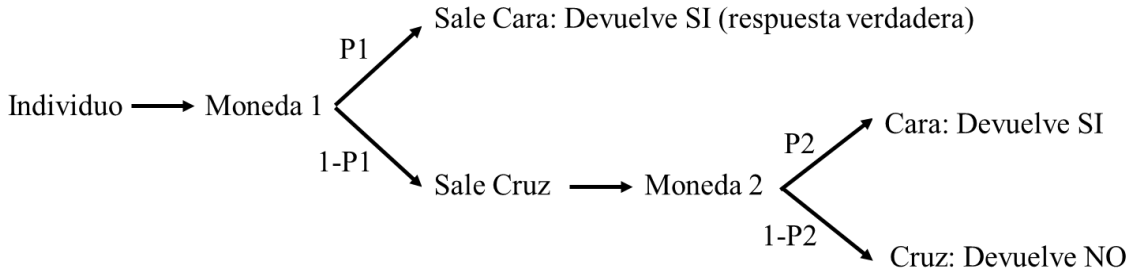
Siguiendo a la autora Cynthia Dwork (2008), metodológicamente la privacidad diferencial consiste en aplicar ruido aleatorio sobre los datos al momento de implementar técnicas para su procesamiento. Como resultado se obtendrá una cierta distorsión en los mismos sin afectar un análisis posterior. De esta manera, brinda la posibilidad de contar con un método matemático probabilístico riguroso para hacer frente a la posibilidad de re identificar al

individuo titular. Pero 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.

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. Las respuestas son del tipo sí o no, considerando que los encuestados (titulares de los datos) no permiten el acceso directo a la información real brindada.

Frente a esta situación es posible aplicar privacidad diferencial para evitar exponer cierta información privada (Dwork, 2008). 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. A continuación, se puede observar este proceso de manera esquemática.

Imagen 0: Esquema de obtención de respuesta aleatoria




Fuente: elaboración propia en base a Dwork (2008)

En este esquema, cada individuo estará protegido con una negación plausible mediante el lanzamiento de una moneda (aleatoriedad) (Dwork, 2008). Entonces, si se quiere conocer la probabilidad obtener respuesta SI y estar siendo protegido por privacidad diferencial (PD), se deberán considerar las siguientes probabilidades condicionales:

$$P(SI/\text{no está protegido por PD}) = P1 + [(1 - P1) * P2] \quad (1)$$

siendo la probabilidad de obtener como respuesta SI dado que no está protegido por PD.

$$P(SI/\text{ está protegido por PD}) = (1 - P1) * P2 \quad (2)$$

siendo la probabilidad de asignar la respuesta SI dado que está protegido por PD.

A partir de la ecuación (1) y (2), la probabilidad de estar protegido por privacidad diferencial vendrá dada por:

$$\text{P (Estar protegido por PD)} =$$

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

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

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

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

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

$$1 - \frac{[\,\text{P(SI)} - (1 - \text{P1}) * \text{P2}]}{\text{P1}} \qquad (3)$$

De la ecuación (3) puede interpretarse que mediante la aplicación de privacidad diferencial a los datos es posible obtener una reducción del riesgo de privacidad. Este punto será profundizado con el desarrollo del modelo más adelante.

Aplicar esta metodología a 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 los consulte no podrá notar la diferencia entre ambos en función de los resultados agregados obtenidos, ya que esta no cambiará sustancialmente el resultado probabilístico, su distribución. La aproximación a este resultado requiere de modelización matemática, la cual se expone en el apartado 2.3.3. Pero para ello previamente resulta necesario introducir dos conceptos: función aleatoria y distribución de probabilidad de Laplace.

## 2.3 La modelización de la Privacidad Diferencial

Para poder presentar el modelo de Privacidad Diferencial en el apartado 2.3.3, resulta necesario especificar tres conceptos. En primer lugar, el significado de una función o algoritmo aleatorio. Luego, cual es el comportamiento de la distribución de Laplace. Finalmente, y en base a esta última, que es el mecanismo de Laplace.

### 2.3.1 Función o algoritmo aleatorio

El esquema de obtención de respuesta aleatoria presentado en el apartado anterior puede ser modelizado mediante un algoritmo aleatorio. Un algoritmo es una secuencia mecánicamente ejecutable de instrucciones elementales, generalmente destinadas a lograr un propósito específico (Erickson, 1999). En particular, los algoritmos aleatorios son un caso especial de algoritmos estocásticos.

Un algoritmo estocástico puede ser visto como un algoritmo parcialmente controlado por un proceso aleatorio (Hromkovič, 2004). Por ejemplo, el control puede venir dado por el resultado obtenido del lanzamiento de una moneda. Este resultado será utilizado para decidir de qué manera el algoritmo continuará su proceso. La calidad de un algoritmo estocástico se mide generalmente en el tiempo de funcionamiento y en el grado de fiabilidad.

El grado de fiabilidad puede interpretarse como la probabilidad de obtener el resultado correcto (siendo P1 en el esquema planteado en el apartado 2.3). Los algoritmos aleatorios son un caso especial de algoritmos estocásticos en el sentido de que devuelven la respuesta correcta con una alta probabilidad (Hromkovič, 2004). A continuación, se ejemplifica su funcionamiento.

Considérese un algoritmo A como una distribución de probabilidad sobre un conjunto finito $\{A_1, A_2, \dots, A_n\}$ de estrategias deterministas. Para cualquier *input w*, A elige un $A_i$ al azar para operar sobre *w*. De esta manera habrá n cálculos (o en términos computacionales, n corridas) de A en todo el rango de $\{A_1, A_2, \dots, A_n\}$. Entonces, es posible modelar el proceso de trabajo de A como un espacio de probabilidad dado por:

$$(S_{A,w}, Prob)$$

donde $S_{A,w}$=$\{A_1, A_2, \dots, A_n\}$ y $Prob$ es la distribución de probabilidad en todo el rango de $\{A_1, A_2, \dots, A_n\}$.

Siguiendo a Hromkovič (2004), se considera a $Prob$ como $S_{A,w}$=$\{C_1, C_2, \dots, C_n\}$, donde $C_i$ es cada cálculo realizado por A sobre *w*. Por lo tanto, A elige de manera determinista cada paso i a seguir con una probabilidad $Prob(\{A_i\})$. Luego, si se quiere medir la fiabilidad del resultado devuelto de A sobre *w*, puede considerarse la variable X: $S_{A,w} \rightarrow \{0,1\}$ definida como:

$$X(C_i)^{13} = \begin{cases} 1 \; si \; C_i \text{ devuleve la respuesta correcta sobre } w \\ 0 \; si \; C_i \text{ devuleve la respuesta incorrecta sobre } w \end{cases}$$

Entonces, para todo $i = 1, \dots, n$, el valor esperado de X es la probabilidad de éxito de A sobre *w*. Esto es:

$$\mathrm{E(X)} = \sum_{i=1}^{n} X(C_i) * Prob(\{C_i\})$$

$$= \sum_{X(C_i)=1} 1 * Prob(\{C_i\}) + \sum_{X(C_i)=0} 0 * Prob(\{C_i\})$$

$$= Prob(\{C_i\})$$

siendo la probabilidad de que A devuelva la respuesta correcta, es decir, cuando X toma el valor 1.

### 2.3.2 Distribución de Laplace

La distribución de Laplace es una distribución de probabilidad continua que lleva el nombre de Pierre-Simon Laplace[14]. Su función de densidad de probabilidad (Dwork y Roth, 2014) está dada por la siguiente ecuación:

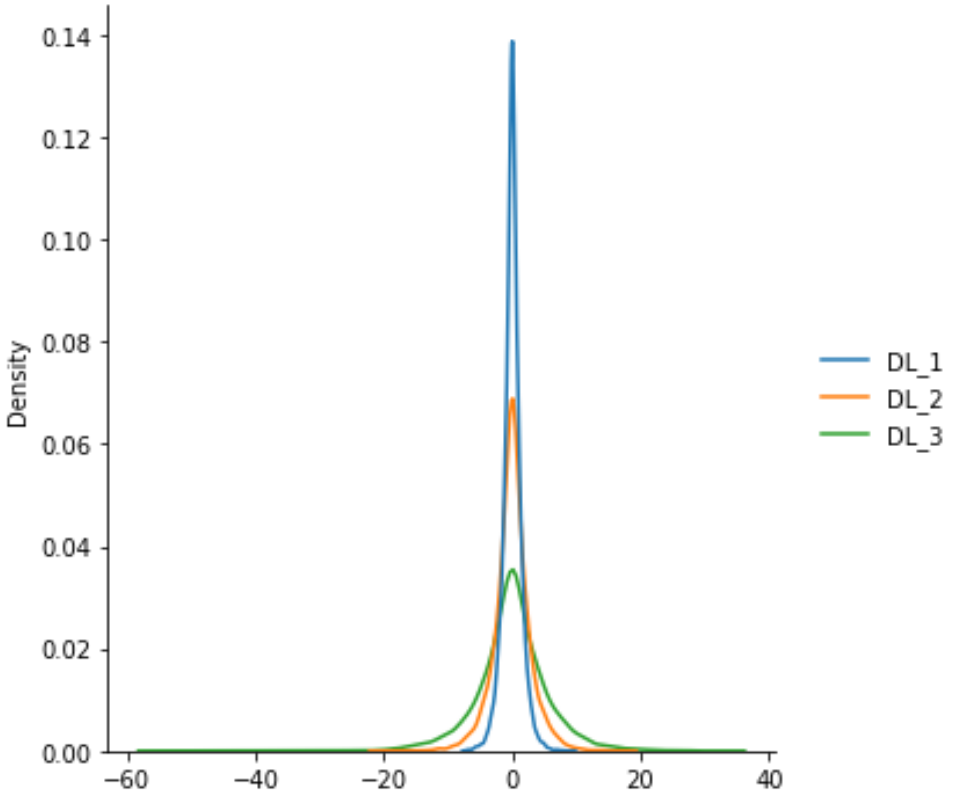
$$f(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}} \qquad (4)$$

donde b es el parámetro de escala, y μ es el parámetro de posición.

[13] Esta representación emula el lanzamiento de una moneda que solo tiene dos resultados posibles (1, si sale cara, 0 si sale cruz)
[14] Ver https://hmong.es/wiki/Pierre-Simon_Laplace

Para visualizar su comportamiento, se crea una función en *Python* (ver apéndice A.1) para simular aleatoriamente la densidad de probabilidad de Laplace con μ=0 y diferentes valores del parámetro b para 1000 valores. Gráficamente es posible observar lo siguiente:

Gráfico 0: Distribución de Laplace




Fuente: elaboración propia con *Python*

donde en el caso de la curva DL_1 el parámetro b vale 1, vale 2 para la curva DL_2 y vale 4 para la curva DL_3.

Del gráfico 0, se puede observar que a medida que aumenta el parámetro de escala (b), la distribución se vuelve más aplanada. Tal comportamiento tendrá como consecuencia la necesidad de aumentar la distorsión de datos para obtener mayor privacidad. Si bien este punto será explicitado en el apartado 2.3.4 con el desarrollo del modelo de Privacidad Diferencial, previamente resulta necesario presentar un mecanismo que permite distorsionar datos utilizando la distribución presentada.

### 2.3.3 El mecanismo de Laplace

El mecanismo de Laplace por medio del cual puede obtenerse distorsión en los datos (Near y Abuah, 2021), puede definirse como sigue. Dada una función $f(X)$ que devuelve un número, se define una función aleatoria $F(X)$ como:

$$F(X) = f(X) + Lap\left(\frac{s}{\varepsilon}\right) \qquad (5)$$

donde la sensibilidad de $f(X)$ será un valor s, y $Lap(\frac{S}{\varepsilon})$ representa una muestra de la distribución de Laplace con parámetros μ=0 y b=$\frac{s}{\varepsilon}$ – cuyo comportamiento pudo observarse en la ejemplificación realizada en el apartado 2.3.2–.

La sensibilidad de una función representa los cambios en la suma se los resultados que devuelve cuando el valor de entrada cambia en una unidad. De esta manera, en la ecuación (5), tal sensibilidad vendrá dada por la consulta que se realiza a una base de datos. Por lo tanto, si una consulta cuenta el número de filas de la base de datos, y se modifica exactamente una fila de esta base, entonces el resultado de la consulta puede cambiar en un máximo de 1.

### 2.3.4 El modelo de Privacidad Diferencial

Formalmente, dado dos bases de datos $(D_1, D_2)$, que difieren como máximo en una fila, pero una está incluida en la otra, Cynthia Dwork (2008) define a la privacidad diferencial como una función aleatoria Κ que permite distorsionar datos:

$$Pr[\text{Κ}(D_1) \in S] \leq exp(\varepsilon) * Pr[\text{Κ}(D_2) \in S] \quad (6)$$

donde S está incluido en todo el rango de Κ.

De la ecuación (6) se puede observar que la probabilidad de respuesta en $D_1$ difiere de la probabilidad de respuesta en $D_2$ al aplicar la función Κ. Dicha diferencia es consecuencia de haber incorporado ruido aleatorio a través de esta función en los datos de $D_2$. A su vez, la pérdida de privacidad vendrá dada por el valor de $\varepsilon$. Si $\varepsilon = 0$, entonces se obtiene una privacidad completa, mientras que si $\varepsilon > 0$ existirá una pérdida de privacidad.

Para poder determinar cuánto ruido aleatorio 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$\rightarrow R^k$ con k $\epsilon$ [0;1], la sensibilidad es definida como:

$$\Delta f = \max_{D_1, D_2} \| f(D_1) - f(D_2) \| \quad (7)$$

donde $f$ no es otra cosa que la obtención de la respuesta verdadera – es decir, P1 en el esquema planteado en el aparatado 2.2 –.

Si en la ecuación (7), $k = 1$ entonces $f$ es la diferencia máxima entre los valores que puede tomar en dos bases de datos que tan solo difieren en una sola fila. Por lo tanto, cuanto menor sea el valor de $\varepsilon$, se deberá incorporar más ruido aleatorio 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 consultas sobre la base de datos para incorporar ruido aleatorio se debilitará la privacidad. Esto es, dada la sensibilidad de $f(X)$ , la aleatoriedad comienza a perder fuerza (Dwork, 2008). Un mecanismo para sortear esta dificultad consiste en utilizar el mecanismo de Laplace para la generación de ruido aleatorio, presentado en el apartado 2.3.3.

De esta manera, Cynthia Dwork (2008) propone:

$$f(X) \; + \; [Lap(\Delta f/\varepsilon)]^k \quad (8)$$

donde el parámetro de escala de la distribución de Laplace presentada en el apartado 2.3.2 viene dado por $\frac{\Delta f}{\varepsilon}$

En la ecuación (8) es posible notar que la incorporación de ruido aleatorio mediante la función de Laplace independiza de la cantidad de los $k$ componentes de $f(X)$. Es decir, la forma de incorporar ruido en la base de datos ya no depende de la secuencialidad de consultas que se haga sobre la misma sino del valor de $\varepsilon$. A medida que $\varepsilon$ disminuye, el ruido esperado será mayor – la curva de la distribución de Laplace será más aplanada–. De esta manera, se observa que la privacidad queda garantizada a partir de que la privacidad diferencial solo depende de la sensibilidad de $f(X)$ y del parámetro $\varepsilon$.

Con el fin de evaluar la efectividad del modelo de privacidad diferencial presentado, en el siguiente capítulo se desarrolla una implementación de esta metodología mediante la utilización del software *Python*[15] en el entorno *Google Colaboratory*[16].

**Conclusión**

En el presente capítulo, se ha expuesto que la privacidad de los individuos puede ser puesta en riesgo a partir del procesamiento de datos personales por parte de las organizaciones estatales. En consecuencia, se presenta la necesidad de establecer metodologías que permitan dar cierta garantía para el tratamiento de este tipo de datos. A su vez, se requiere de incorporar a la privacidad durante el diseño de todo ciclo de procesamiento de datos personales, pero también asignando responsabilidades sobre quienes participan. De esta manera, se introdujo el modelo Privacidad Diferencial como aquella metodología que permite obtener protección de datos personales en organizaciones. Por lo tanto, se brindaron los fundamentos de un método que permite evitar la violación de la privacidad del individuo titular de los datos.

Para poder llevar a cabo la presentación de la metodología de Privacidad Diferencial, en primer lugar, se expuso la definición de la privacidad. Por esta, se definió a todo aquello que un individuo decide mantener en privado y, por ser un valor social, se encuentra regulado. Teniendo en cuenta un contexto organizacional estatal basado en datos, particularmente se hizo referencia a la privacidad de datos personales. El procesamiento de este tipo de datos requiere tener en cuenta el riesgo asociado a la violación de la privacidad del individuo.

Frente a este riesgo de privacidad, en las organizaciones públicas del Estado donde se recolectan y procesan datos personales, surge la necesidad de dar cierta garantía para su protección. Esto implica la necesidad de que se establezca un diseño de procesamiento de datos con eje en la privacidad. Pero también, surge la necesidad de establecer un marco de responsabilidad sobre quienes participan de este proceso. Ello como consecuencia de que en Argentina existe una regulación legal sobre el tratamiento de datos personales otorgándole derechos a los individuos titulares de los datos.

De esta manera, la utilización de métodos para la protección de datos personales en organizaciones estatales resulta necesario. En particular, la metodología de Privacidad

[15] Acerca de *Python* https://www.python.org/
[16] Acerca de *Google Colaboratory* https://colab.research.google.com/?utm_source=scs-index

Diferencial brinda un método que da garantías en este sentido. La posibilidad de aplicar ruido aleatorio sobre datos personales permite obtener datos distorsionados pero utilizables y, en consecuencia, facilita su publicación. Frente a esta tentadora posibilidad, surge la necesidad de realizar la demostración a través de una aplicación del método presentado para evaluar su efectividad. Por esta razón, en el siguiente capítulo se lleva a cabo su implementación y evaluación.

# Capítulo 3: Aplicación y evaluación de la Privacidad Diferencial para evitar la identificación de ciertos individuos a partir de datos personales obtenidos de organizaciones estatales argentinas

## Introducción

La publicación de datos personales conlleva un riesgo asociado para la privacidad de los individuos. Si bien, las organizaciones públicas del Estado los utilizan en la mayoría de las ocasiones para favorecerlos, también pueden ser manipulados en su perjuicio por terceros –cuando son publicados abiertamente–. De esta manera se puede incurrir en la violación de la privacidad de los individuos (AEPD, 2020).

Frente a esta problemática, la gobernanza de los datos personales resulta necesaria, ya que involucra la utilización de diferentes metodologías para la protección de datos. Una de estas es la anonimización de datos, pero resulta limitada ya que no es posible anonimizar una base de datos completa. Otra técnica es La Privacidad Diferencial (Dwork, 2008). Si bien es una técnica más compleja, resulta ser más robusta para la protección de datos personales permitiendo que los mismos sean utilizables para un análisis posterior.

En este contexto, surge el siguiente interrogante: ¿qué nivel de efectividad tiene la aplicación de privacidad diferencial para la protección de datos personales? Para responder esta pregunta el objetivo del presente capítulo consiste en, por un lado, construir una base de datos personales. Por el otro identificar los riesgos de privacidad que surgen de esta y, finalmente, aplicar el modelo de privacidad diferencial para evaluar su efectividad en términos de si es posible evitar la identificación de un individuo.

Para llevar adelante el desarrollo del capítulo, se divide en cinco secciones. En la primera sección se explicita como es construida la base de datos personales con la cual se trabajará. Para ello, en primer lugar, se describe cuáles datos serán tomados de partida para determinar una muestra. Luego, se explica como es el proceso de obtención de los datos personales que

permiten conformar la base de datos final. En la segunda sección, se realiza una descripción de los datos finalmente obtenidos. Esto permite conocer los atributos personales con los cuales se trabajará.

En la tercera sección se exponen las problemáticas que se detectan en una base de datos personales en términos de riesgos de violación de privacidad. En primer lugar, se expone el riesgo que se deriva a partir de la publicación de identificadores únicos. Luego, se muestra como otros atributos también pueden resultar en un identificador directo del individuo. Finalmente, se introduce la problemática que surge de la combinación de atributos personales.

En la cuarta sección se demuestra cómo es posible re identificar individuos a partir de recombinar atributos provenientes de bases de datos diferentes. En primer lugar, se particiona la base datos personales para emular una práctica habitual de trasferencia de datos entre las organizaciones. Luego, se muestra como la combinación de datos personales contenidos en una base con los de la otra, permite realizar la identificación de un individuo.

Finalmente, en la quinta sección, se desarrolla la aplicación del modelo de privacidad diferencial. En primer lugar, se realiza la implementación del modelo con el *software Python* en el entorno *Google Colaboratory*. A continuación, se realiza la evaluación de la efectividad del modelo. Por último, se implementa ruido aleatorio por el mecanismo de Laplace y se evalúan los resultados obtenidos.

### 3.1 Obtención de datos y construcción de base de datos personales

Para llevar adelante la aplicación del modelo de privacidad diferencial, se parten de datos publicados en el portal de datos abiertos del Gobierno de la Ciudad de Buenos Aires[17]. Consisten en un padrón de contribuyentes de alto riesgo, responsables del Impuesto sobre los Ingresos Brutos (IIBB) para el primer trimestre de 2015. Este consta de 487514 registros y una serie de atributos, entre ellos el CUIT[18] (Clave Única de Identificación Tributaria según la Administración Federal de Ingresos Públicos), la razón social y el tipo de contribuyente inscripto.

[17] Portal de datos abiertos BAData. Ver https://data.buenosaires.gob.ar/dataset/
[18] El CUIT es un código que identifica de manera unívoca a trabajadores autónomos, comercios y empresas.

El CUIT se conforma de 11 dígitos. Los dos primeros se corresponden a la identificación del género (masculino o femenino) según normativa vigente hasta junio de 2021[19]. Los ocho siguientes se corresponden al Documento Nacional de Identidad (DNI) y el último es un dígito verificador. Por tanto, cada individuo posee un número de CUIT único. En consecuencia, cumple la función de ser un identificar único de cada individuo de la República Argentina.

A partir del padrón de CUIT con el que se cuenta, se define una muestra. Esta se conforma de 266564 números de CUIT de ciudadanos argentinos. La selección de esta muestra consistió en tomar los casos cuyo CUIT comienzan con número 27 y con número 20. De esta manera, se consigue obtener individuos según género y cuya razón social es su nombre y apellido. Finalmente, la muestra queda conformada con 172161 CUIT pertenecientes a individuos de género masculino y 94403 CUIT pertenecientes a individuos de género femenino.

Con el objetivo de conformar una base de datos personales con más atributos, se seleccionan 1000 números de CUIT al azar. La mitad de estos corresponden al género femenino y la mitad restante al género masculino. Con estos se procede a buscar manualmente datos personales consultando distintos servicios *web* del Estado argentino.

En primer lugar, se obtiene la constancia de inscripción al Monotributo consultando gratuitamente el servicio *web* de AFIP[20] (Administración Federal de Ingresos Públicos). En este se debe ingresar de manera individual cada número de CUIT y un código de seguridad por cada consulta, como puede observarse en la imagen a continuación:

Imagen 1: Servicio *web* de AFIP para la obtención de Constancia de Inscripción a Monotributo

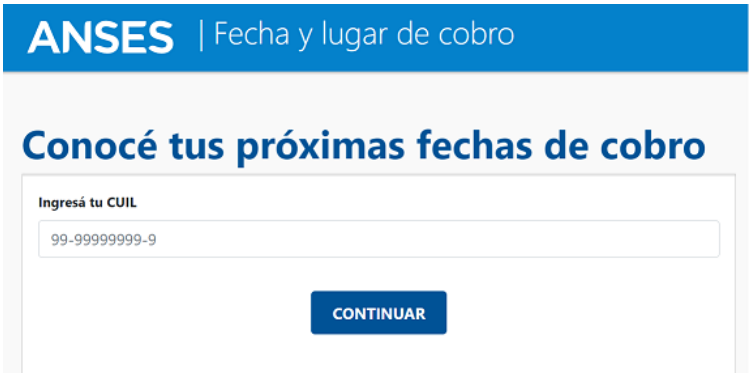
[19] El 8 de junio de 2021 se sanciona la normativa 5007/2021 a partir de la cual los primeros dos dígitos ya no identificaran a los individuos según sexo binario. https://www.argentina.gob.ar/normativa/nacional/resoluci%C3%B3n-5007-2021-350854/texto

[20] Portal AFIP en el siguiente link https://seti.afip.gob.ar/padron-puc-constancia-internet/ConsultaConstanciaAction.do

Una vez ingresados ambos, se obtiene en versión PDF (*Portable Document Format)* y de manera *online* la constancia de inscripción. A partir de esta se procede a copiar manualmente cada uno de los siguientes datos personales contenidos: el código postal, domicilio fiscal, jurisdicción, barrio y código y descripción de actividad que realiza. Pero también, se verifican la razón social (nombre y apellido) y el género que se encontraban entre los datos originales de partida. Respecto de la primera, en algunos casos solo se contaba con el apellido. Por esto, también se procedió a completar el nombre.

Finalizado lo anterior, se procede a consultar el portal *web* de ANSES[21] (La Administración Nacional de la Seguridad Social) denominado "Donde cobro". En este se debe ingresar de manera individual cada número de CUIT, sin necesidad de ingresar un código verificador, como puede observarse en la imagen a continuación:

Imagen 2: Portal *web* "Donde cobro" de ANSES




Al ingresar individualmente un CUIT, se logra verificar si el individuo es beneficiario social[22]. Y en los casos que lo es, se detallan, además del nombre y apellido y CUIT, nombre del tipo de beneficio que percibe, nombre de la organización bancaria donde percibe el cobro, pero también la dirección postal y barrio de la sucursal donde el pago es efectuado. Estos datos se copian manualmente para complementar a los anteriormente obtenidos.

[21] Portal "Donde cobro" de ANSES en el siguiente link https://servicioswww.anses.gob.ar/dondecobrov2/
[22] Un beneficiario social de ANSES es aquel individuo que percibe alguna de las siguientes prestaciones que brinda el Estado argentino: Jubilación, Pensión, AUH, AXE, SUAF, PROGRESAR, Prestación por Desempleo. Para más información ver https://www.anses.gob.ar/sites/default/files/archivo/2021-05/Bases%20y%20condiciones%20Beneficios%20ANSES.pdf#:~:text=%E2%80%9CBENEFICIARIOS%22%3A%20son%20aquellas%20personas,(PNC)%2C%20de%20la%20Asignaci%C3%B3n

Finalmente, la base de datos personales es denominada “INDIVIDUOS” y fue creada en formato Excel[23]. Para su almacenamiento se utilizó Google Drive[24]. Contiene los datos personales para 1000 individuos en base a los siguientes atributos:

Tabla 1: Atributos que contiene la base de datos personales

| Atributos | Descripción |
|---|---|
| cuit | Clave Única de Identificación Tributaria |
| tipo_contr_insc | Vale C o D según categoría del contirbuyente inscripto |
| raz_soc | Razón social (nombre y apellido del individuo inscripto) |
| GENERO | Vale M si es masculino y F si es femenino |
| CP | Código postal |
| Domicilio_fiscal | Domicilio fiscal |
| JURISDICCION | Provincia |
| BARRIO | Barrio |
| Código_actividad | Número identificatorio de actividad según nomenclador AFIP |
| Actividad | Nombre de la actividad que realiza según nomenclador AFIP |
| Beneficiario_social | Vale SI si es beneficiario social, NO en caso contrario |
| Nombre_beneficio_social | Nombre del beneficio social |
| Lugar_de_cobro | Nombre de la organización bancaria, dirección postal y barrio de la sucursal de percepción del cobro |

Fuente: elaboración propia

De esta manera, se logró identificar información personal de 1000 individuos argentinos, siendo 500 de género femenino y 500 de género masculino. La información de código postal, dirección postal, jurisdicción y barrio se obtuvo de manera completa en todos los casos. Respecto de la actividad que realizan, solo en 508 se logró identificar, ya que para los restantes no figuraba en su constancia de inscripción a Monotributo. En cuanto a los datos obtenidos de ANSES, 401 resultaron ser beneficiarios sociales, obteniéndose datos completos. El 59,9% restante resultó ser no beneficiario.

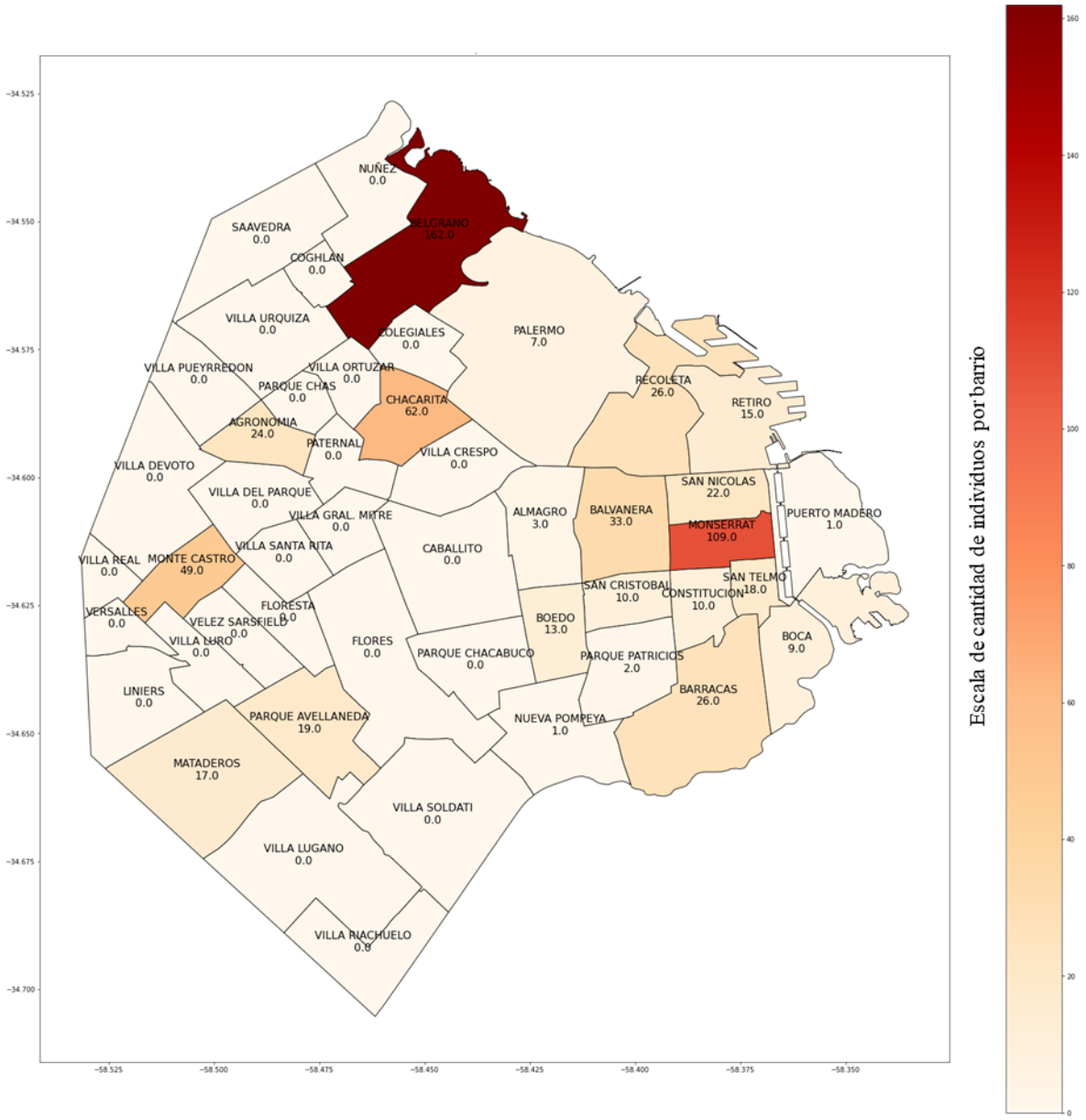
### 3.1.1 Análisis descriptivo de la base de datos personales

Entre las características principales que presenta la base construida se encuentra la distribución geográfica de los 1000 individuos según jurisdicción. El 63,8% de los individuos tiene domicilio fiscal en CABA (Ciudad Autónoma de Buenos Aires), el 31,8% en Provincia de Buenos Aires y el 4,4 % restante en diferentes provincias de Argentina. Entre los individuos de CABA la mayor cantidad se encuentran en el barrio de Belgrano

[23] Acerca de Excel https://www.microsoft.com/en-us/microsoft-365/excel
[24] Acerca de Google Drive https://www.google.com/drive/

(162). La distribución completa de individuos de CABA según barrio se puede observar en el mapa a continuación.

Mapa 1: Cantidad de individuos de CABA según barrio



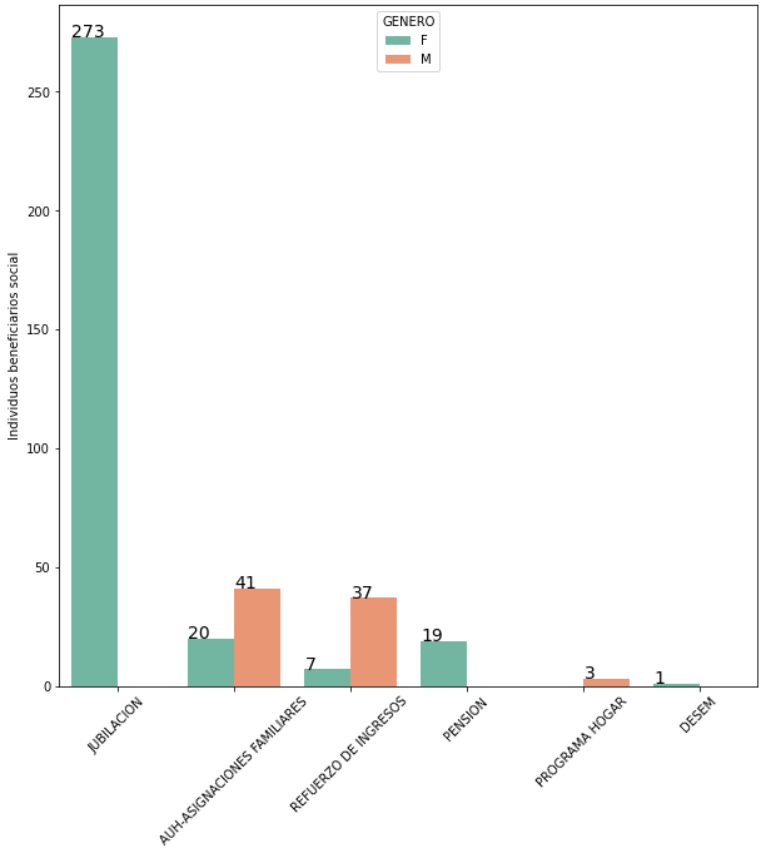

Fuente: elaboración propia con Python

La principal actividad que realizan los individuos es Servicios Personales según actividad declarada en la constancia de inscripción a Monotributo. En segundo lugar, se encuentra Servicios Inmobiliarios por cuenta propia y en tercer lugar Servicios Jurídicos. A su vez, en función de la categoría de Monotributo, 610 poseen categoría D y 390 categoría C. La categoría está relacionada con el monto bruto de facturación anual permitida según AFIP.

En el caso de la D implica una facturación bruta anual de como máximo $ 1.934.273,04, mientras que en el caso de la C es de como máximo $ 1.557.443,75[25].

En cuanto a los datos obtenidos de ANSES, 401 individuos resultaron tener algún beneficio social. De estos, el 68.08% corresponde a Jubilación mientras que el 15,2% corresponde a AUH (Asignación Universal por Hijo) y el resto a otros beneficios. A continuación, se puede observar la distribución de quienes poseen algún beneficio social según el beneficio que percibe y género.

Gráfico 1: Beneficiarios sociales, según tipo de beneficio y generó




Fuente: elaboración propia con Python

Del gráfico 1, se puede observar que todos los casos que poseen Jubilación son mujeres. Lo mismo sucede en el caso de Pensión. En cambio, donde existe mayor cantidad de individuos con género masculino es en el caso de AUH, Refuerzo de Ingresos y Programa Hogar. De esta manera, de un total de 500 individuos de género masculino, solo el 16,2% posee algún

[25] Datos según AFIP a julio de 2022.

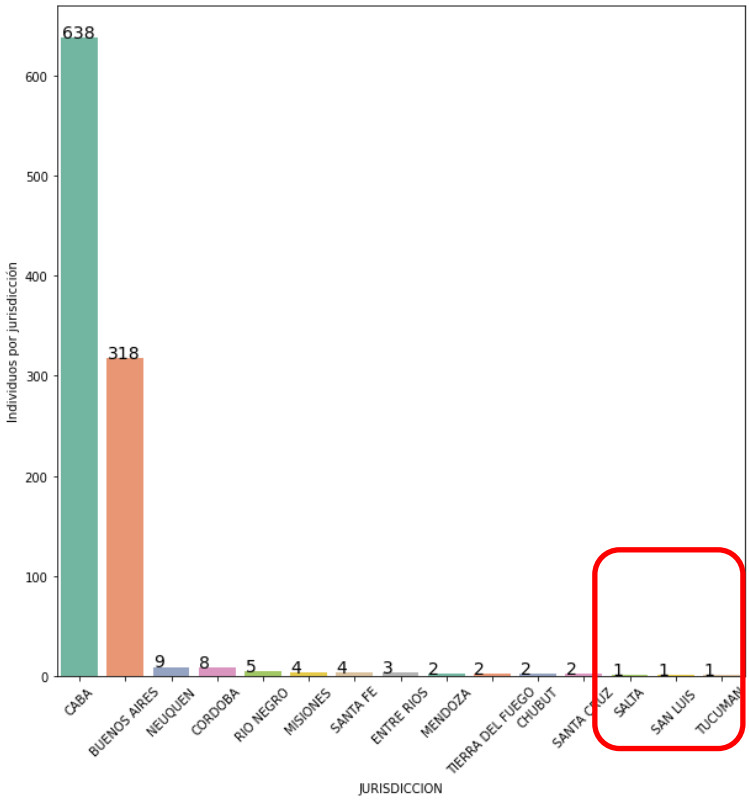
beneficio social. En el caso de los 500 individuos de género femenino esta cifra alcanza al 64%.

### 3.1.2 Identificación de problemáticas

Conformada la base de datos personales, la primera problemática que se logra identificar resulta ser que con tan solo el CUIT es posible obtener mucha más información personal de los individuos consultando diferentes servicios *web* gratuitos del Estado. De aquí que no debiera publicarse el mismo, o si se publica debiera ser en forma anonimizada. Esto debido a que la información posible de obtener es accesible por cualquier persona y puede ser utilizada en perjuicio de su titular. Por esta razón, se decide anonimizar el CUIT, respetando un principio ético del uso de datos para procesos de investigación (Saunders et al., 2016).

Mediante la anonimización del CUIT en *Python*, se crea una nueva variable en la base datos denominada "Fake_cuit" y se procede a eliminar la original ("cuit"), manteniendo todo el resto de los atributos. Pero, aun así, surge una problemática de identificación directa según cantidad de casos por jurisdicción. En el gráfico 2 a continuación es posible observar esta situación.

Gráfico 2: Cantidad de individuos identificables según Jurisdicción




Fuente: elaboración propia con Python

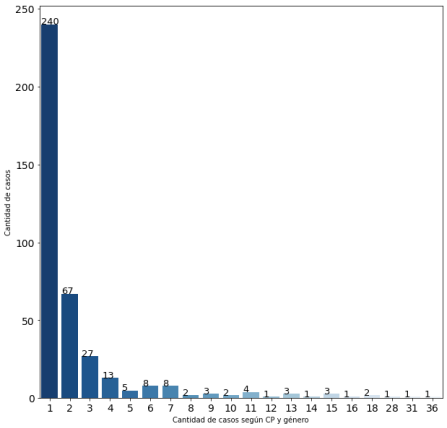
Como se observa en el gráfico 2, en el caso de Salta, San Luis y Tucumán, existe un solo individuo por cada una de estas jurisdicciones. Esto implica una identificación única de caso por jurisdicción. Ante esta situación, debería eliminarse los casos – o por lo menos anonimizar los datos completos de cada individuo involucrado– si se quiere evitar su identificación. Con esto se pretende exponer que no solo los identificadores únicos como el CUIT pueden facilitar la identificación unívoca de un individuo en una base datos. Otros atributos pueden resultar de este tipo según la cantidad de casos involucrados.

También se detectó una problemática a partir de la combinación de dos atributos. En el gráfico 1 del apartado 3.2, es posible observar que para el caso de género masculino y tipo de beneficio social "DESEM" existe un solo caso. De esta manera, y en una instancia inicial de descripción, queda expuesta una problemática frecuente de reidentificación de individuos a partir de la combinación de atributos en una base de datos. Además, esta puede escalar en riesgo si se cruzan datos provenientes de diferentes fuentes, mediante la transferencia de datos entre organizaciones. En el siguiente apartado se realiza una demostración de esta problemática.

### 3.1.3 Identificación indirecta de individuos

En el apartado 3.3 se expusieron diferentes problemáticas de identificación directa a partir de identificadores personales únicos que ponen en riesgo la privacidad de los individuos. Pero también, se introdujo una nueva problemática que surge de la recombinación de atributos en una misma base de datos. Para ejemplificar la cantidad de casos que tienen una combinación única de atributos en la base de datos bajo análisis, a continuación, se muestra gráficamente en base a tomar el código postal (CP) y el género.

Gráfico 3: Cantidad de individuos identificables según combinación de código postal y género

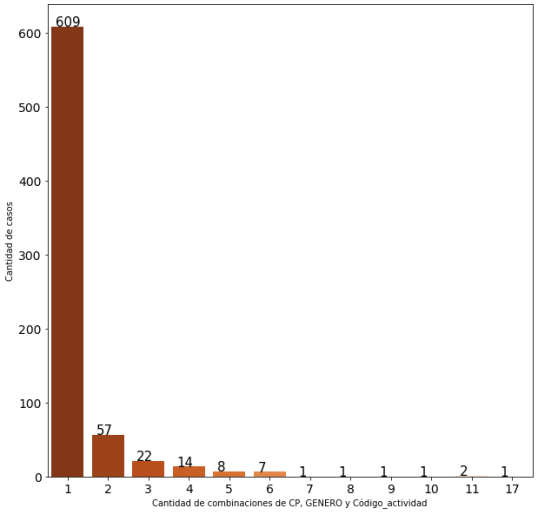


Fuente: elaboración propia con Python

El gráfico 3 muestra en el eje de abscisas la cantidad existente de combinaciones de código postal y género (vale 1 si es única); y en el eje de ordenadas, la cantidad de individuos. Su interpretación es como sigue: existen 240 individuos de un total de 1000 (24%) de la base de datos que poseen una combinación única (no repetida) de ambos atributos. Mientras que existen 67 casos donde la combinación de los atributos existe dos veces (esto involucra a un total de 134 individuos). Y así, es posible interpretar los restantes casos. Por lo tanto, en el primer caso, la identificación del individuo es unívoca, mientras que en el segundo caso mencionado no lo es, ya que existen dos individuos por cada combinación de valor de los atributos.

Si se incorpora un atributo más, por ejemplo, el código de actividad, entonces los casos obtenidos se muestran en el gráfico a continuación.

Gráfico 4: Cantidad de individuos identificables según combinación de código postal, género y actividad




Fuente: elaboración propia con Python

A diferencia de la situación del gráfico 3, en el gráfico 4 se observa que la cantidad de individuos identificables unívocamente a partir de combinar tres atributos aumenta a 609, alcanzando al 60.9% del total de la base de datos. De esta manera, a medida que se aumenta la cantidad de atributos combinados en una misma base de datos, aumenta el riesgo de reidentificación de un individuo y, por tanto, de violación de su privacidad. De esta manera, la problemática identificada invita a preguntarse qué atributos son publicables en conjunto y cuales no cuando se cuenta con una única base de datos.

Ahora bien, la recombinación de atributos no necesariamente surge de una misma base de datos. Del cruce de datos proveniente de diferentes bases, también surge riesgo de reidentificación de individuos ya que facilita la combinación de atributos. Teniendo en cuenta que entre organizaciones del Estado es frecuente la práctica de transferencia de datos para la toma de decisiones, también se debe tener en cuenta este riesgo. Este surge como consecuencia de la recombinación de los datos recibidos con los propios, lo que también pone en riesgo la privacidad del individuo.

Para llevar adelante la demostración de este caso, en primera instancia se particiona en dos la base de datos con el CUIT ya anonimizado. A partir de esto, se conforman dos bases de datos. Por un lado, una (individuos_1) que contiene 'tipo_contr_insc', 'raz_soc', 'GENERO', 'CP', 'Domicilio_fiscal', 'JURISDICCION', 'BARRIO', 'Código_actividad', 'Actividad' y 'Fake_cuit'. Por el otro, una segunda (individuos_2) con los atributos 'GENERO', 'CP', 'Beneficiario_social', 'Nombre_beneficio_social', 'Lugar_de_cobro' y 'Fake_cuit'. En cada uno de ellos se mantiene el total de individuos de 1000. De esta manera se emula una base datos posiblemente recibida y una propia en un contexto organizacional.

Entonces, si se agrupan los datos según código postal, género y código de actividad en la base de datos (individuos_1) se obtiene que hay 609 casos con combinación única, como se había mostrado en el gráfico 4. A modo de ejemplo, se toma un caso de estos 609, cuyo código postal es 1002, el género es femenino y el código de actividad es 702091. Si a continuación se cruza esta información con el segundo conjunto de datos (individuos_2) se obtiene unívocamente un caso con toda la información. Los datos completos del individuo involucrado son[26]:

| | tipo_contr_insc | raz_soc | GENERO | CP | Domicilio_fiscal | JURISDICCION | BARRIO | Código_actividad | Actividad | Fake_cuit | Beneficiario_social | Nombre_beneficio_social | Lugar_de_cobro |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | D | | F | 1002 | | CABA | MONSERRAT | 702091 | SERVICIOS DE ASESORAMIENTO, DIRECCIÓN Y GESTIÓ... | XNZF0716667456452 4 | SI | JUBILACION | BBVA- BANCO FRANCES S.A. RECONQUISTA 00199 - ... |

De esta manera se demuestra como mediante el cruce de datos de diferentes bases a partir de atributos personales en común es posible identificar unívocamente a los individuos (que para la base de datos bajo análisis son en total 609 casos) y ampliar la información sobre el mismo. Al mismo tiempo, también queda expuesto como la anonimización de algún identificador único puede resultar insuficiente para evitar esta problemática. Ahora bien,

[26] No se muestran el nombre y apellido ni dirección postal para proteger los datos del individuo involucrado.

podría pensarse en anonimizar todos los atributos. Pero si se realiza esto, se convertirían los datos en códigos sin sentido, como puede observarse en el caso del CUIT.

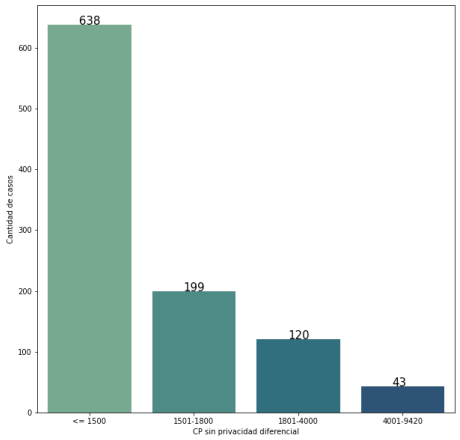
Por esta razón, la anonimización de datos resulta ser una técnica de uso limitado. Si se quiere conservar información útil para un análisis posterior, no es posible aplicarla a toda la base de datos. En consecuencia, se requiere de otras metodologías para sortear este problema. Una metodología posible es la aplicación de Privacidad Diferencial mediante el mecanismo de Laplace. Para poder evaluar la efectividad de esta metodología, en el siguiente apartado se lleva a cabo su implementación y evaluación.

### 3.2 Aplicación de Privacidad Diferencial y evaluación del modelo

Para llevar a cabo la implementación de Privacidad Diferencial sobre datos resulta necesario con el fin de evaluar la efectividad del modelo de privacidad diferencial. Esto es, si se aplica esta metodología a por ejemplo un atributo de entre los atributos personales utilizados en el apartado 3.4, se busca poder evaluar si su distribución se verá afectada. Para ello, se toma el atributo código postal de la base datos individuos_2 y se realiza la aplicación mediante la utilización de la librería OpenDP[27] en Python.

Previamente, se observa que el código postal de individuos_2 posee un valor mínimo de 1002 y un valor máximo de 9420. Frente a esto y para facilitar su visualización, se la agrupa en categorías, y se realiza un gráfico de barras que es representativo de su distribución. En el gráfico a continuación se visualiza la distribución del código postal según categorías.

Gráfico 5: Cantidad de individuos según rangos de código postal




Fuente: elaboración propia con Python

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

A partir del gráfico 5 se puede observar que la mayor cantidad de individuos contenidos en la base de datos individuos_2 se encuentran entre los códigos postales 1002 y 1500. Dada esta distribución original de la variable recategorizada a continuación se aplica privacidad diferencial sobre la misma. Para llevarlo a cabo de desarrollan dos funciones.

La primera función realiza la aplicación de ruido aleatorio por el mecanismo de Laplace según la lógica del modelo de privacidad diferencial explicado en el apartado 2.3.4, tomando una muestra. Para ello se opta por utilizar la metodología del histograma para representar el comportamiento de la distribución del código postal por categorías. Dado que este método consiste en representar la frecuencia (absoluta o relativa) de los casos en función de las categorías (intervalos de clase) de la 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.

El código implementado es el siguiente:

```
def dp_histo(epsilon=0.5):
    np.random.seed(2022)
    var_names = ['GENERO', 'CP', 'Beneficiario_social', 'Nombre_beneficio_
      social', 'Lugar_de_cobro', 'Fake_cuit', 'CP_2']

    with sn.Analysis(protect_floating_point=False) as analysis:
        data = sn.Dataset(path = '/content/samplefile', column_names =
               var_names)
        nsize = sample_size

        inc_prep = sn.histogram(data['CP_2'],
                categories = edad_categorias, null_value = "-1")

        inc_histogram2 = sn.laplace_mechanism(inc_prep, privacy_usage=
               {"epsilon": epsilon, "delta": .000001})

    analysis.release()

    return  inc_histogram2.value
```

La segunda función, evalúa el error en términos de diferencia porcentual promedio entre las clases del histograma del código postal original y la variable código postal con privacidad diferencial aplicada. El código implementado es el siguiente:

```
def perc_error(true_col, dp_col):
    return sum(abs(true_col - dp_col)) / sum(true_col) * 100
```

Considerando diferentes tamaños de muestra (250, 500, 750, 850, 950), teniendo en cuenta distintos valores del parámetro ε (0.8, 0.5, 0.25, 0.05), y considerando las categorías de código postal ['<= 1500', '1501-1800', '1801-4000','4001-9420'] se obtienen los resultados contenidos en la tabla 2 a continuación.

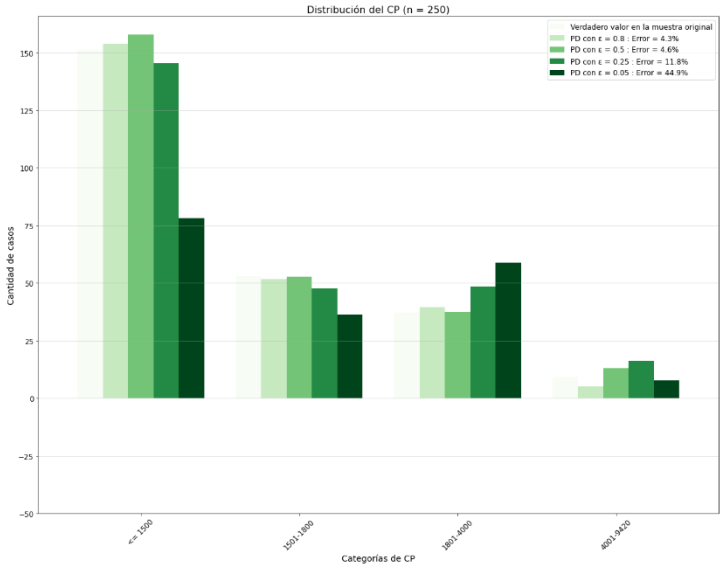
Tabla 2: Evaluación de resultados del modelo de privacidad diferencial

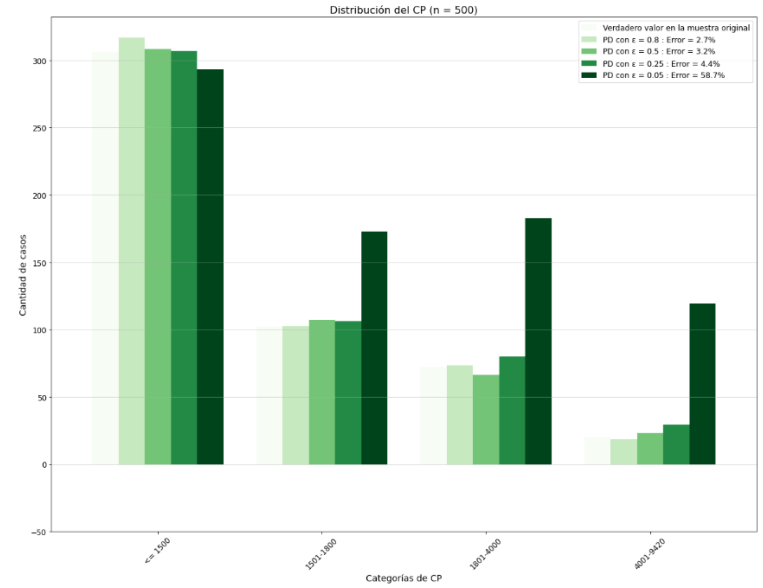
| Tamaño de muestra | 250 | 500 | 750 | 850 | 950 |
|---|---|---|---|---|---|
| | Error obtenido | | | | |
| PD con épsilon (ε=0.8) | 4,3% | 2,7% | 1,5% | 0,9% | 0,5% |
| PD con épsilon (ε=0.5) | 4,6% | 3,2% | 1,5% | 2,9% | 1,7% |
| PD con épsilon (ε=0.25) | 11,8% | 4,4% | 10,7% | 6,3% | 1,8% |
| PD con épsilon (ε=0.05) | 44,9% | 58,7% | 22,6% | 12,3% | 21,0% |

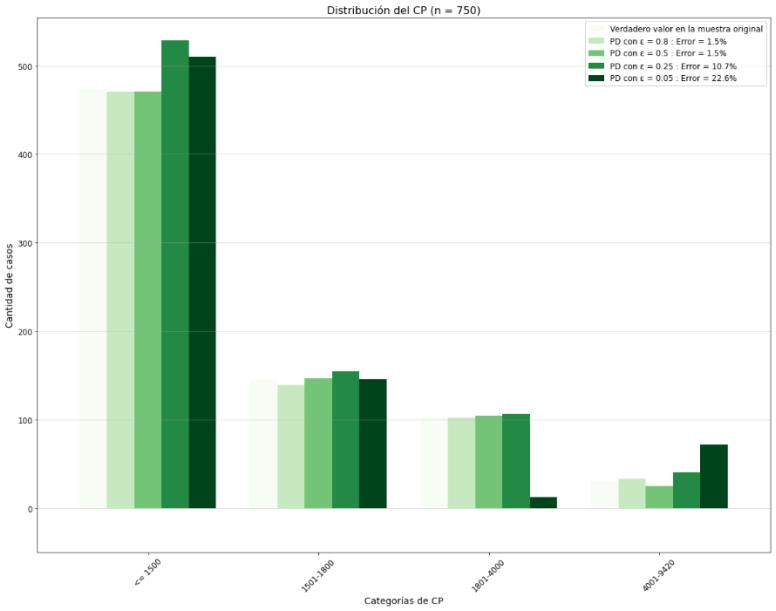
Fuente: elaboración propia

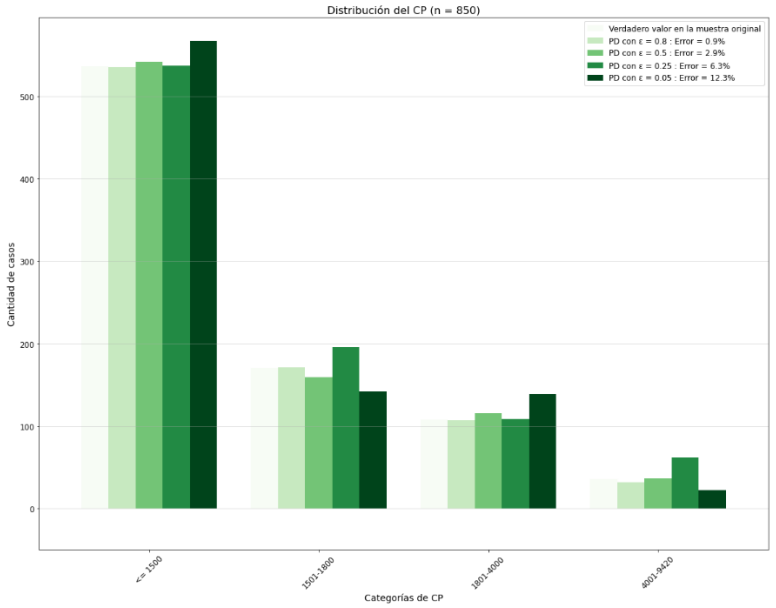
De la Tabla 2, se puede concluir que, para cada tamaño de muestra, a medida que disminuye el valor de ε – es decir, a medida que se opta por reducir la pérdida de privacidad–, el error aumenta. Este error representa en términos porcentuales, la variabilidad media que existe entre el verdadero valor de la clase en la muestra y el valor obtenido de aplicar privacidad diferencial. A su vez, es posible observar que a medida que se aumenta el tamaño de la muestra, a igual valor del parámetro ε, el error disminuye. Para exponer los resultados de una manera visual, a continuación, se muestra gráficamente los obtenidos para las distintas muestras:

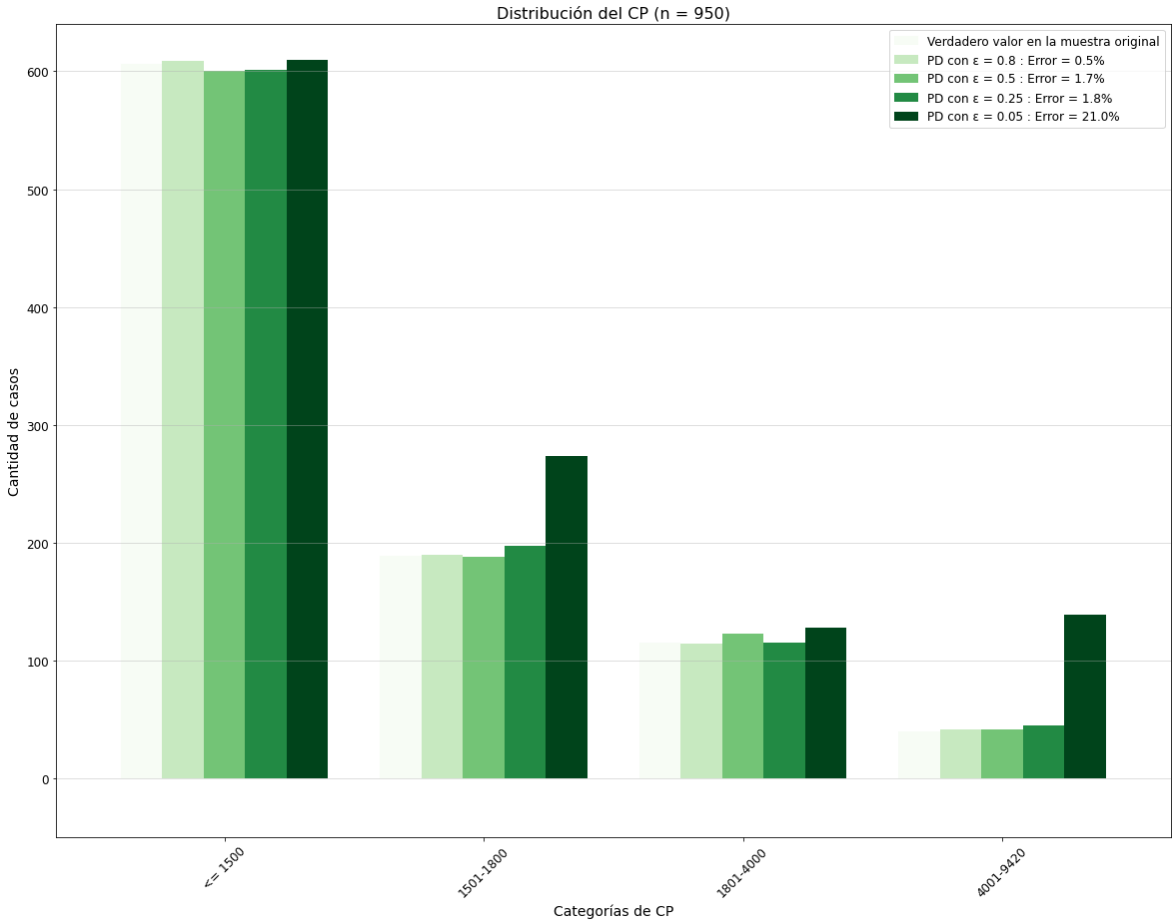
Gráfico 6: Cantidad de individuos según rangos de código postal con y sin privacidad diferencial (según diferente valor de épsilon) y para diferentes tamaños de muestra




Distribución del CP (n = 500)
Verdadero valor en la muestra original
PD con ε = 0.8 : Error = 2.7%
PD con ε = 0.5 : Error = 3.2%
PD con ε = 0.25 : Error = 4.4%
PD con ε = 0.05 : Error = 58.7%
Cantidad de casos
Categorías de CP
<= 1500
1501-1800
1801-4000
4001-9420


Distribución del CP (n = 750)
Verdadero valor en la muestra original
PD con ε = 0.8 : Error = 1.5%
PD con ε = 0.5 : Error = 1.5%
PD con ε = 0.25 : Error = 10.7%
PD con ε = 0.05 : Error = 22.6%
Cantidad de casos
Categorías de CP
<= 1500
1501-1800
1801-4000
4001-9420


Distribución del CP (n = 850)
Verdadero valor en la muestra original
PD con ε = 0.8 : Error = 0.9%
PD con ε = 0.5 : Error = 2.9%
PD con ε = 0.25 : Error = 6.3%
PD con ε = 0.05 : Error = 12.3%
Cantidad de casos
Categorías de CP
<= 1500
1501-1800
1801-4000
4001-9420




Fuente: elaboración propia con Python

De los gráficos 6 puede observarse que la distribución de la variable código postal en categorías no se ve fuertemente modificada por haber aplicado privacidad diferencial, en particular a medida que se aumenta el tamaño de la muestra y para valores no tan bajos de ε. Por lo tanto, en función de los resultados obtenidos se logra obtener datos distorsionados sin que se afecte la distribución de una variable de manera significativa. Esto permitirá contar con información sobre el código postal en categorías siendo utilizable para cualquier análisis posterior. De esta manera, se logra demostrar la efectividad de aplicar privacidad diferencial.

### 3.3 Aplicación de ruido aleatorio por mecanismo de Laplace y resultados

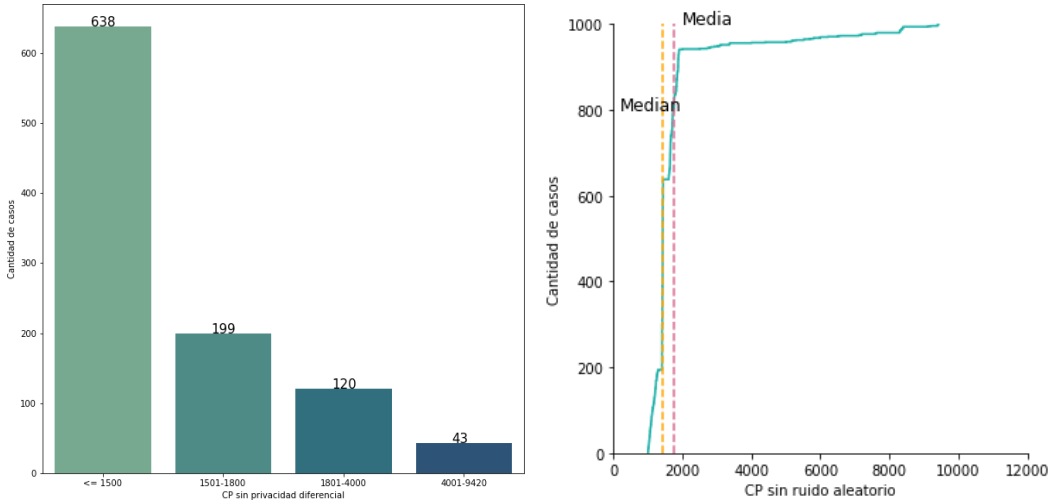
En base a la evaluación de la efectividad de aplicación de ruido aleatorio sobre la variable código postal realizada en el apartado anterior, a continuación, se implemente sobre la misma variable, pero en el conjunto de datos original (INDIVIDUOS) de manera completa. Para realizar esta implementación se crea la siguiente función que modela la ecuación (8) del apartado 2.3.4. El código de la función desarrollada es el siguiente:

```
def ruido(data, variable, variable_new, sensitivity, epsilon):
```

```
    np.random.seed(202206)
    data[variable_new]=
        data[variable]+np.random.laplace(loc=0, scale=sensitivity/
        epsilon)
    return data
```

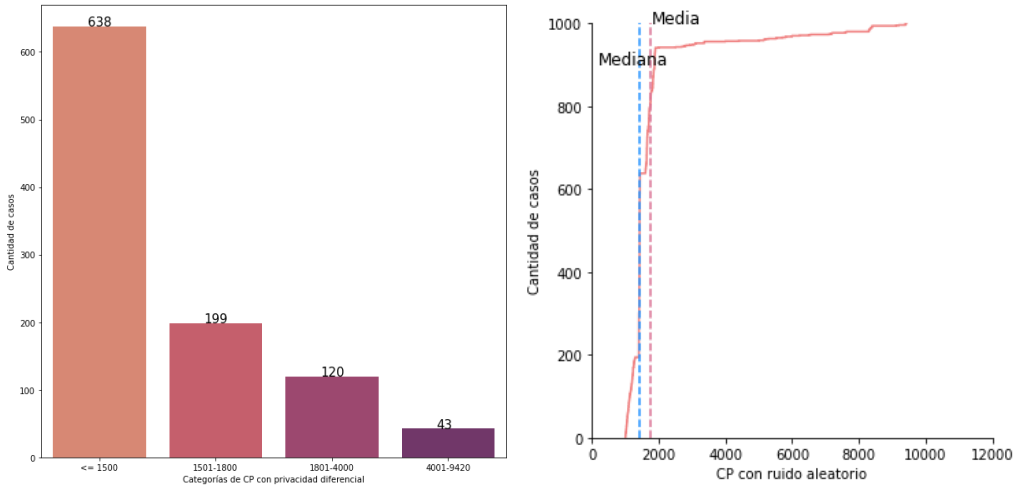
Se parametrizó con un valor épsilon (ε) igual a 0.5 porque en base a los resultados analizados en el apartado anterior, el error obtenido fue menor del 5% para un tamaño de muestra de 950. Llevando a cabo la implementación de esta función, se pueden observar gráficamente los resultados obtenidos a continuación.

Gráfico 7: Distribución y distribución acumulada de individuos según rangos de código postal sin ruido aleatorio (variable original)




Fuente: elaboración propia con Python

Gráfico 8: Distribución y distribución acumulada de individuos según rangos de código postal con ruido aleatorio (variable distorsionada)




Fuente: elaboración propia con Python

De la comparación entre los gráficos 7 y 8 puede observarse que la distribución de la variable código postal en categorías (gráficos de barras en cada caso) no se ve prácticamente afectada por haber aplicado ruido aleatorio. Y si se observa el comportamiento de la función de distribución (gráficos de la derecha en cada caso) puede notarse que es prácticamente igual. Observando las estadísticas resumen (con líneas punteadas en los gráficos) en la tabla a continuación, puede notarse como no se vieron afectados severamente sus valores.

Tabla 3: Medidas estadísticas resumen del código postal (CP) con y sin ruido aleatorio

| Medidas | CP sin ruido aleatorio | CP con ruido aleatorio |
|---|---|---|
| Moda | 1425 | 1424.206903 |
| Media | 1741.804 | 1741.010903 |
| Mediana | 1428 | 1427.206903 |

Fuente: elaboración propia

Finalmente, para poder comprobar si existe homogeneidad de las varianzas entre el código postal con y sin ruido aleatorio, se realiza una prueba de hipótesis de Levene[28]. Las hipótesis que se plantean son:

$$\begin{cases} H0: \textit{la varianza del código postal} \text{ con y sin } \textit{ruido aletorio son iguales} \\ H1: \textit{la varianza del código postal} \text{ con y sin } \textit{ruido aletorio son distintas} \end{cases}$$

De su implementación (ver apéndice A.2 - Evaluación de resultados) se obtiene un p-valor mayor a 0,99. De esta manera se concluye que no se rechaza la hipótesis nula (H0). Por lo tanto, hay evidencia suficiente para considerar que las varianzas de la distribución de la variable con y sin ruido aleatorio son iguales. En consecuencia, puede concluirse que la distribución de código postal con ruido aleatorio no se vio significativamente modificada con un 95% de confianza.

Finalmente, si se toma la variable código postal con ruido aleatorio y se intenta realizar el cruce de datos mostrado en el ejemplo del apartado 3.4, ya no es posible obtener los datos completos del individuo. Esto se debe a que ahora el código postal vale 1001.206903 siendo distinto al valor original que era 1002. De esta manera no es posible realizar al reidentificación del individuo a partir de cruzar dos bases de datos en base a este atributo, ya sea individualmente o combinado con otros.

[28] Acerca de la prueba de hipótesis de Levene ver Bisquerra, R. (1987) "La prueba de Levene para la homogeneidad de varianzas en el BMDP" Revista de investigación educativa. 1987, v. 5, n. 9; p. 79-85.

Por lo tanto, en función de los resultados obtenidos se logra obtener datos distorsionados sin que se afecte significativamente la distribución de la variable código postal. De esta manera, es posible contar con información sobre el código postal utilizable y publicable para cualquier análisis posterior. A su vez, se logra evitar la identificación de un individuo dando cierta garantía sobre su privacidad. De esta manera, se logra demostrar la efectividad de aplicar ruido aleatorio por el mecanismo de Laplace propuesto por la metodología de Privacidad Diferencial.

**Conclusión**

En el presente capítulo, se ha expuesto como a partir de una base de datos personales se pone en riesgo la privacidad de un individuo. Sin lugar a duda, para evitar la exposición de identificadores únicos como el CUIT, es viable aplicar una técnica de anonimización. Pero, aun así, de la combinación de otros atributos contenidos en la base de datos, continúa existiendo el riesgo de reidentificación. Frente a esta situación se aplicó la metodología de privacidad diferencial, resultando efectivo para evitar la identificación de un individuo a partir de combinar sus atributos personales. En consecuencia, se ha alcanzado una aproximación a mostrar que es posible obtener protección de datos personales – y evitar la violación de la privacidad del individuo titular de los datos – mediante la aplicación de ruido aleatorio sin que ello impida su validez para un análisis posterior.

Para poder llevar a cabo la implementación realizada, en primer lugar, debió construir manualmente una base de datos personales. Esto fue posible partiendo de un padrón de CUIT tomado de un repositorio online de datos abiertos del Gobierno de la ciudad de Buenos Aires. Con este identificador único, se identificó individuos según género y luego se tomó una muestra aleatoria de 1000 individuos. A partir de estos casos, se consultaron dos servicios *web* gratuitos que permitieron obtener varios atributos personales, conformando así la base de datos personales trabajada.

A partir de la conformación de la base datos personales, en una primera instancia se anonimizó el CUIT. Demostrada la insuficiencia de esta técnica para garantizar la privacidad del individuo, se aplicó la metodología de privacidad diferencial para distorsionar un atributo (el código postal) que previamente se demostró que combinado con otros permitía identificar a un individuo. Realizando su evaluación se pudo establecer su efectividad en términos de que es posible contar con el código postal distorsionado para ser publicado y realizar

cualquier análisis posterior. Y a su vez, se logra romper la identificación de un individuo otorgando cierta garantía sobre su privacidad.

Pero al mismo tiempo, en base a la demostración realizada se expone como la falta de una gobernanza de datos personales puede poner en riesgo la privacidad de los individuos. Y no solo surge como consecuencia de exponer identificadores únicos como el CUIT como es el caso presentado inicialmente. Sino también de la recombinación de atributos personales. En consecuencia, el manejo de datos personales por parte de una organización públicas estatales invita a cuestionar cual es el grado de responsabilidad que ello conlleva. Por esta razón, surge la necesidad de conocer cuál es el marco de gestión de los datos que está siendo implementado en las diferentes organizaciones públicas del Estado argentino. Pero, además, invita a cuestionar si realmente la privacidad de datos está siendo considerada como eje central con el fin de contar con una gestión responsable de datos.

# Capítulo 4: Aspectos relevantes para el diseño de una gestión responsable de la privacidad en el Estado argentino

## Introducción

En los dos primeros capítulos de este trabajo, se ha presentado el corpus teórico sobre el cual se sentaron las bases del desarrollo de la investigación. Partiendo de la concepción del Estado como una organización basada en datos, se ha afirmado que el procesamiento de datos involucrado implica riesgos de privacidad. En este centro de cálculo coexisten cuatro elementos claves: los datos, la tecnología, los especialistas y los individuos. De la interacción entre ellos en el mercado de datos, se deriva el riesgo de violación de la privacidad de los individuos. La definición de privacidad fue presentada en el capítulo 2, particularmente definiendo a la privacidad de datos personales como el derecho que tiene el individuo a conocer que se hace con sus datos. Frente a esto, se expuso que un marco de responsabilidad es necesario llevar adelante por parte de las organizaciones públicas estatales.

Desde un enfoque cuantitativo, en el capítulo tres fue desarrollada la aplicación práctica de privacidad de datos personales. Comenzando por la creación de una base de datos, continuando con la identificación de riesgos asociados, finalmente se seleccionó y aplicó el método de Privacidad Diferencial para la protección de datos personales. Evaluados los resultados obtenidos, se obtuvo un resultado satisfactorio de su efectividad, así como

también se sentaron las bases para determinar los aspectos principales a tenerse en cuenta para definir una estrategia de privacidad en el Estado argentino.

En el presente capítulo, se generalizan los resultados desde un enfoque cualitativo mediante la especificación de los puntos relevantes a ser tenidos en cuenta para llevar adelante el diseño de una estrategia de privacidad en organizaciones públicas estatales. Esto permitirá aproximar a la implementación de una gestión responsable de la privacidad de datos personales. Para ello se utilizan los resultados obtenidos en los capítulos previos. De este modo, se determinarán los primeros aspectos para la elaboración de un marco de gestión responsable de la privacidad en el Estado argentino.

Para llevar adelante el desarrollo del capítulo, se divide en tres secciones. En la primera, se explicitan los principios sobre los cuales es necesario sustentar el diseño de la privacidad en organizaciones estatales. Para ello, en primer lugar, se describe el significado de cada uno. Finalmente, se expone como la responsabilidad para su cumplimiento requiere de responsables a cargo y de técnicas para su implementación. De este modo se establecen las bases para el diseño de un sistema de gestión responsable de la privacidad.

En la segunda sección se exponen los puntos relevantes a ser tenidos en cuenta en base a los resultados obtenidos en el capítulo 3. En primer lugar, se especifica acerca de las consecuencias del acceso y disponibilidad de datos personales. Luego, se relevan los principales riesgos que se derivan de la manipulación de los datos que deben ser tenidos en cuenta. Finalmente, se establecen los criterios para tener en cuenta a la hora de seleccionar una metodología para la protección de datos personales.

Finalmente, en la tercera sección, se definen los factores relevantes en la definición de la responsabilidad de la privacidad. El primero de ellos involucra la dimensión del volumen de datos personales. A continuación, se menciona acerca del alcance de su tratamiento. Luego, se ahonda sobre el tiempo de conservación de los datos. Por último, se amplía acerca de la responsabilidad sobre la definición de accesibilidad a datos personales.

### 4.1 La estrategia del diseño de privacidad

Como se mencionó en el apartado 2.1.2, la privacidad desde el diseño implica que deba ser parte integral de los sistemas de datos, así como de todos los procesos llevados a cabo por la organización del Estado. Una manera de alcanzarlo es a través de definir una estrategia que

incorpore siete principios fundacionales definidos por Cavoukian (2011). El primero de ellos, implica considerar ser proactivo y preventivo antes que reactivo y correctivo. Para lograrlo se deberá tener en cuenta implementar metodologías de detección temprana de riesgos con el fin de corregirlos o minimizarlos.

También se requerirá que la privacidad quede establecida como una configuración predeterminada. Esto implica que los datos personales sean protegidos de manera automática a través de implementar aplicaciones o software. Pero además será necesario que se establezcan criterios de limitación de uso, así como de conservación, y también restricciones de acceso acorde a la necesidad de cada función del personal de la organización.

A su vez, la privacidad deberá ser incorporada desde la fase de diseño de los sistemas. No alcanza con añadir un simple módulo a un sistema existente, sino que resulta necesario integrarla como un requisito más en el diseño de cada sistema. A su vez ello requerirá del diseño de procesos de evaluación continua que den garantía sobre el correcto funcionamiento del proceso.

Para evitar caer en una discusión de privacidad versus necesidad (u oportunidad) de uso, resulta fundamental tener en claro que existen intereses contrapuestos entre la necesidad de la organización y los derechos de los individuos. Frente a ello, resulta necesario buscar un balance con un pensamiento de “todos ganan”. Posiblemente surja la necesidad de buscar nuevas soluciones frente a la evolución continúa de la tecnología como facilitadora de datos. En este sentido, resulta necesario considerar que la obtención de datos se encuentra en constante evolución. Y ello trae aparejado la necesidad de un continúo proceso de revisión y control.

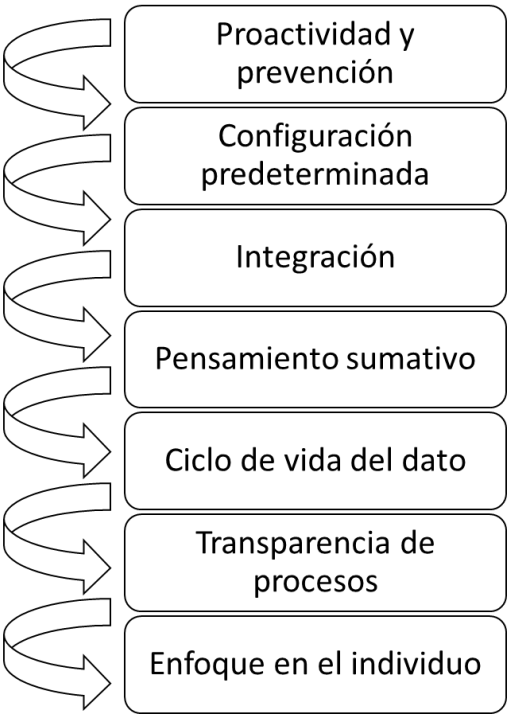
Teniendo en cuenta que el ciclo de vida del dato en la organización atraviesa por diferentes procesos y usos, se deberá establecer garantía de privacidad durante todo este. No solo alcanza con definirla en el diseño previo a la implementación de un sistema, sino que también deberá garantizarse durante todas las etapas posteriores. Ello requerirá de un análisis de todas las etapas que atraviesa el dato, implementando en cada una las medidas adecuadas.

La transparencia resulta un principio fundamental. Con este se busca demostrar la responsabilidad en el tratamiento de los datos personales de forma tal de generar confianza con el individuo. Para lograrlo será necesario hacer públicas las políticas de privacidad que establece la organización y que deben ser definidas en base a todos los principios antes

mencionados. Pero también deberá tener en cuenta la regulación vigente y la comprensión del derecho a la privacidad que tienen los individuos. La comunicación transparente y de fácil entendimiento debe ser la clave en este proceso.

Finalmente, teniendo en cuenta que la privacidad de datos personales fue definida en el aparatado 2.1.1. como el interés que tiene un individuo en poder controlar sus propios datos, resulta necesario tener al mismo "en mente" a la hora de establecer el diseño. De esta manera, se podrá establecer el desarrollo de procesos y aplicaciones que estén focalizadas en el individuo para el respeto de su derecho a la privacidad. Pero también resulta necesario informar a los individuos de las consecuencias en su privacidad de cualquier cambio implementado. Para ello será necesario poner a su disposición toda la información necesaria de manera clara y transparente para su fácil comprensión. A modo de resumen a continuación se listan los siete principios mencionados:

Esquema 2: Principios fundacionales para el diseño de la privacidad




Fuente: elaboración propia

Sin lugar a duda, todos estos principios requieren de un compromiso por parte de las organizaciones estatales para poder ser llevados a cabo. Desde concebir a la privacidad de datos personales en los altos mandos hasta el seguimiento y evaluación continua de todos los procesos. A su vez, se requerirá de la formación de los trabajadores de la organización

en este sentido, de forma tal que cada uno pueda conocer su grado de responsabilidad en el tratamiento de la privacidad.

La disponibilidad de metodologías para el tratamiento de privacidad de datos personales, constituyen una herramienta que facilita la implementación de protección de datos personales. La elección de estas deberá ser de un trabajo conjunto entre los diferentes especialistas buscando establecer un equilibro en base a las diversas necesidades. Por esta razón, la conformación de grupos interdisciplinarios requiere de un líder gestor que pueda llevar adelante el proceso.

El gestor se requerirá que sea un funcionario con formación multidisciplinaria entorno a datos sin necesidad de ser un especialista en cada área, pero si con un conocimiento mínimo que le permita dialogar con otros. Esto es, deberá ser quien brinde la posibilidad de generar una comunicación adecuada para el entendimiento entre partes especializadas de cada área. De esta manera, deberá buscar el consenso para lograr la aplicación efectiva de todos los principios detallados.

### 4.2 Aspectos principales en la implementación de privacidad sobre datos personales

En base a los principios mencionados en el apartado anterior, y teniendo en cuenta la aplicación realizada en el capítulo 3, en este apartado se definen los puntos relevantes que no deben omitirse a la hora de implementar privacidad de datos personales. Llevarlos a cabo requerirá tener en cuenta diversas dimensiones de áreas de trabajo. En estas intervendrán diferentes especialistas según los requisitos de la tarea a realizar en cada área de la organización estatal.

#### 4.2.1. Acceso y disponibilidad de datos

La obtención de datos requiere considerar al menos dos situaciones principales. Por un lado, la restricción de acceso a datos dentro de la organización. Por el otro, la publicación de datos para que terceros accedan a estos.

En cuanto al acceso a datos personales dentro de la organización, resulta necesario establecer restricciones acordes a la necesidad de cada área en particular. La administración de permisos frecuentemente suele estar a cargo de un área de seguridad informática con un responsable designado. Este deberá contar con ciertos parámetros que definen diferentes perfiles de usuarios que estarán relacionados con la funcionalidad ejercida.

Ahora bien, que los datos personales sean accesibles por ciertos perfiles, no implica que todos deben estar disponibles. Por ejemplo, si la demanda de acceso es para un analista de datos, probablemente resulte necesario un acceso a ciertos datos personales (como el género, la edad, entre otros) frente a otros especialistas. En este caso, un identificar único como el CUIT, no debe ser accesible por este. No solo que no le aporta ningún valor para un análisis de comportamiento, sino que puede existir el riesgo de uso mal intencionado o de fuga.

Si la necesidad es la identificación de caso, esto debe realizarse a través de un identificador interno (un número aleatorio ordinal único) de la organización que no involucre la identificación del individuo real. En este sentido, ciertos datos deben procurar contar con algún proceso de ocultamiento, exclusión o de anonimización. Pero también, estar almacenados en una base de datos particular que solo contenga este tipo de datos. Así se reducirá la posibilidad de obtener nuevos datos o de cruzamiento de información.

A su vez, se deberá tener en cuenta que la utilización de ciertos datos debe ser temporal. Es decir, una vez que la finalidad del uso fue completada, deben eliminarse por completo. Pero también, deberán eliminarse aquellos nuevos datos generados. En diversas ocasiones, pueden crearse nuevos atributos derivados que si bien son utilizados para un fin específico no deben quedar disponibles para su reutilización. Por ejemplo, si se cuenta con la edad, es posible determinar el año de nacimiento del individuo de una manera sencilla[29]. Este junto con otros atributos puede conducir a la obtención de otros datos o a la identificación directa de un individuo.

Por otra parte, la decisión de publicar datos en un portal propio o ajeno o la transferencia de datos a otras organizaciones requiere de establecer un método que evite que ciertos datos sean públicos. Como fue posible determinar en la obtención de datos realizada en el apartado 3.1, exponer identificadores únicos como el CUIT permite la identificación directa de un individuo. A su vez, esto habilita la posibilidad de obtener mayor cantidad de datos a través de diferentes portales que están disponibles para brindar servicio a sus titulares.

La obtención de nuevos datos abre la posibilidad de incurrir en una recombinación de atributos que deriven en la identificación ciertos individuos. Frente a esta posibilidad, una opción es la publicación de datos agregados. Por ejemplo, en vez de publicar

[29] Al año vigente se le resta la edad y se obtiene como resultado el año de nacimiento.

individualmente un código postal, en su lugar publicar rangos de valores. Similar sería en el caso de la edad o el año de nacimiento, entre otros.

Resulta necesario que especialistas idóneos tengan intervención para definir los criterios (o formas) de publicación de datos. Para ello se requerirá de realizar evaluaciones previas, que permitan identificar los riesgos asociados. Los procesos de preevaluación o puesta a prueba deben ser una constante frente a la diversidad de datos existentes y del surgimiento de nuevos. Valerse de herramientas tecnológicas para la automatización de tareas, puede resultar en una necesidad o en una práctica valiosa. Pero no debe perderse de vista el monitoreo continúo para ajuste y validación.

### 4.2.2. Detección de riesgos para la privacidad

Una de las tareas fundamentales para lograr la proactividad y prevención frente a la privacidad, es la detección de riesgos. En el caso de datos personales, el principal riesgo de violación de privacidad es la posibilidad de identificación o reidentificación de un individuo. Con excepción de los casos solicitados por orden judicial en términos de la legislación argentina, todos los datos personales no deben estar accesibles para cualquier persona.

Existen diferentes maneras de realizar evaluaciones de riesgo en el sentido planteado o, como fue definido en el apartado 3.1.2, identificar problemáticas sobre una base de datos. Para llevarlo a cabo se requerirá de un analista que tenga la capacidad de definir criterios para determinar que atributos son los que conllevan una mayor exposición al riesgo. Una primera aproximación para lograrlo es a través de un análisis descriptivo.

Durante el proceso descriptivo se debe considerar realizar, en primer lugar, una búsqueda de identificadores únicos. En caso de detectarse alguno, se deberá evaluar si es posible anonimizarse o si se debe quitar de la base de datos. En ocasiones, ciertos atributos son necesarios conservarlos para poder llevar a cabo un análisis posterior ya que aportan cierto valor agregado en la construcción de información. Pero si se detecta un caso único en base a un solo atributo, este deberá ser quitado de la base de datos.

Para lograr la detección de casos únicos por atributo resulta necesario realizar diferentes conteos. Esta tarea puede resultar compleja si la base de datos con la que se cuenta contine gran cantidad de estos. La visualización de resultados mediante la utilización de herramientas tecnologías debe ser considerada ya que facilita la detección automática, como lo expuesto en el mapa 1 del apartado 3.1.2. Pero requerirá de tiempo de formación para su

utilización. No obstante, el criterio para la construcción gráfica debe ser determinado por un especialista idóneo.

En segundo lugar, debe considerarse el conteo de casos en base a la recombinación de atributos dentro de una misma base de datos. En este caso valerse también de visualizaciones de resultados puede facilitar la detección de riesgos. Un ejemplo lo constituye el resultado mostrado en el gráfico 2 del apartado 3.1.2. En este se muestra que de la combinación de dos atributos se identifican casos únicos. Frente a esta posibilidad deberá evaluarse que atributo se quitará de la base de datos y cual se dejará, de forma tal de romper con la posibilidad de identificación.

En ocasiones, quitar atributos o anonimizarlos en una base de datos disminuye el potencial valor que los datos pueden generar. Ante esto, se debe evaluar la posibilidad de aplicar otras metodologías que permitan dejar disponibles ciertos atributos, pero con alguna transformación de su valor real. En particular, esto resulta necesario cuando los datos se van a publicar o transferir.

### 4.2.3. Selección de metodología para la protección de datos personales

Si bien existen diferentes métodos para poder llevar a cabo la protección de datos personales, en este trabajo se propuso la utilización de la Privacidad Diferencial ya que permite obtener datos distorsionados sin afectar un análisis posterior. Esto brinda la posibilidad de utilización o publicación de datos que resultan valiosos para la generación de información. Su rigurosidad matemática probabilística permite obtener garantías de efectividad en términos de privacidad. A su vez, 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.

Dada la configuración de su método, sin lugar a duda requiere de un especialista que tenga el suficiente conocimiento formal para poder comprenderlo. Superada la designación del personal idóneo, como todo modelo, requiere de puesta a prueba y evaluación. Si bien, en el apartado 3.2 se realizó una aplicación y evaluación para demostrar su efectividad, la complejidad de esta tarea puede escalar. Esto se debe a que existe la posibilidad de aplicar diferentes métodos para realizar su aplicación.

Un punto de no menor importancia es que requiere de cierta programación para utilizarse sobre una base de datos, como pudo observarse en la aplicación realizada. Ello conlleva a la necesidad de que el responsable a cargo tenga un cierto nivel de formación en este aspecto.

Esto le permitirá llevar acabo esta tarea o al menos poder comunicarla adecuadamente para que sea implementada por un programador.

Frente a la complejidad mencionada, resulta recomendable desarrollar una API (*Application Programming Interface*) a través de la cual se realicen reiteradas aplicaciones de manera automatizada a través de variar los parámetros involucrados. Esto permitirá realizar tantas pruebas como sea posible, con eficiencia, de forma tal de minimizar el riesgo de privacidad. En consecuencia, su diseño e implementación requerirá de la intervención de diferentes especialistas.

Superada la fase de desarrollo e implementación productiva, surge la necesidad de contar con un especialista que finalmente tome la decisión de cuál es el nivel del parámetro épsilon adecuado. Este parámetro será el que brinde la medida de costo en términos de privacidad, como fue mencionado. Las pruebas realizadas para su determinación en el apartado 3.2, muestran la lógica de su determinación. Aquí pudo observarse que a medida que se aumenta el tamaño de la muestra, para un mismo valor de épsilon, el error disminuye. Pero también, se observó que a medida que su valor es más chico (se aumenta la privacidad), se requiere de mayor distorsión de los datos.

A su vez expone la potencial complejidad que esto puede implicar, asociada a la necesidad de contar con un conocimiento de base especializado. Pero también, demuestra el nivel de responsabilidad involucrado. En consecuencia, el diseño de una API para la evaluación contribuye a minimizar el margen de error. Al mismo tiempo también debe considerarse programar optimizaciones que devuelvan como resultado final el valor buscado. Es decir, obtener una definición de manera automática de forma tal de evitar cualquier sesgo que un analista pueda involucrar ya se manera consciente o inconsciente.

Ahora bien, previo a la determinación de la metodología a aplicar, deberá considerarse aplicar otras metodologías de forma tal de realizar una comparación. Esto permitirá hacer una selección de cuál es el mejor método para la protección de los datos personales. Podría resultar necesario aplicar más de uno o de varios combinados para lograr el objetivo de obtener la mayor privacidad posible.

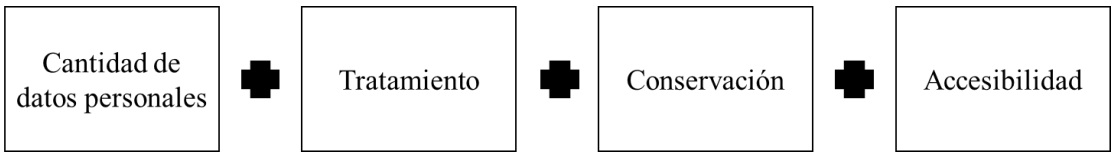
Por otra parte, una de las funciones principales de una API, es que actúa de intermediaria entre la base de datos y el usuario. De este modo, se garantiza que el usuario no acceda directamente a esta. Ello contribuye aún más a disminuir el riesgo de fuga de datos. Por lo

tanto, el diseño de una interfaz a través de la cual operar brinda múltiples beneficios a la hora de operar con datos. Por esta razón se recomienda su implementación.

### 4.3 La responsabilidad por defecto

La responsabilidad es un factor clave a la hora de diseñar la estrategia de un sistema de protección de datos personales en una organización estatal. Esta requiere de considerar cuatro aspectos fundamentales que se encuentran involucrados en los puntos mencionados en el apartado anterior. En el siguiente esquema se enuncian cada uno de ellos:

Esquema 3: Factores para la determinación de responsabilidades




Fuente: elaboración propia en base a AEPD (2020)

La cantidad de datos personales que son recolectados o generados implica aspectos tanto cuantitativos como cualitativos. El responsable involucrado deberá tener en cuenta el volumen de datos tratados discerniendo entre los realmente necesarios y los que no. Pero también deberá considerar el tipo de dato requerido. No todos los datos personales necesariamente deben ser utilizados o expuestos, algunos pueden ser anonimizados ya que no aportan valor de análisis.

En cuanto al tratamiento de datos personales, implica que el responsable deba limitarse a los estrictamente necesario en base a su objetivo. Es decir, dada las diferentes etapas por las que atraviesan, deberá tenerse en cuenta que en cada una de ellas se realicen las operaciones únicamente necesarias para el cumplimiento de cada una. Pero también, podría resultar necesario extender el uso según la necesidad del problema a resolver. En ese caso, el responsable y el especialista deberán definir la estrategia necesaria a llevarse a cabo.

La conservación de datos personales, ya sean estos originales o derivados, requiere de una atención particular. Su perdurabilidad debe ser limitada para lograr el principio de minimizar riesgos. En ocasiones, durante una fase de análisis, se desarrollan procesos para la generación de nuevos datos. Cuando esto sucede, los datos son derivados o generados de una recombinación. Entonces, se deberá tener en cuenta no conservarlos indefinidamente.

En su lugar es recomendable resguardar los procesos generadores para que puedan ser repetidos en caso de ser necesario.

Finalmente, uno de los puntos centrales, es la accesibilidad a datos personales en cada una de las fases del ciclo de vida en la organización. Será necesario establecer criterios de perfiles minuciosos según la necesidad. Pero también deberá prestarse especial atención cuando los datos son publicados o transferidos a terceros o, incluso, frente a la posibilidad de almacenamiento en nube o búsquedas automatizadas a través de aplicativos que acceden a datos.

Los frentes de accesibilidad son amplios en un contexto tecnológico y digitalizado. Las tareas de control de acceso a datos deben basarse en políticas internas definidas claramente en la organización, así como la definición de roles y responsabilidades. En este sentido, se requiere fuertemente de tecnología en la cual apoyarse para lograr establecer la mayor cantidad de controles posibles. El diseño de una interfaz con amplios componentes de privacidad puede resultar en una manera eficiente de lograrlo.

De esta manera, la responsabilidad de privacidad por defecto quedará establecida y será transversal a toda la organización. Se requerirá informar y formar a todos aquellos que se encuentren involucrados. Esto, a su vez, implicará la definición y asignación de responsabilidades concretas. En definitiva, se requiere del desarrollo de una cultura del compromiso, impulsada desde la organización y asumida por sus trabajadores.

**Conclusión**

En el presente capítulo, se realizó la generalización de resultados para la determinación de los aspectos principales de un proceso responsable de gestión de datos personales en organizaciones estatales. Para ello, la privacidad de datos es considerada la unidad de análisis. Si bien resulta inevitable considerar los diversos riesgos involucrados en el procesamiento de los datos, la responsabilidad juega un rol central. De esta manera, el aporte principal del capítulo es que, sin la correcta definición de responsabilidades basadas en un compromiso por parte de las organizaciones, no es posible lograr un sistema de gestión responsable de la privacidad.

Para poder llevarlo a cabo, en primer lugar, se tomaron siete principios fundacionales de la privacidad. El cumplimiento de todos ellos en conjunto resulta necesario como base del

diseño de la estrategia de un sistema de gestión responsable de la privacidad de datos personales. La consecuencia principal resultó ser que la responsabilidad de su cumplimento se constituye en el factor clave que es necesario tener en cuenta.

A partir de los principios fundacionales, y en base a los resultados cuantitativos obtenidos en el capítulo 3, se establecieron los puntos relevantes a ser tenidos en cuenta en el diseño de un sistema de gestión de la privacidad. Un punto principal que surge de lo establecido es la necesidad de contar con tecnología para llevar a cabo de manera eficiente los controles de cumplimientos necesarios. Para ello, el diseño de una API se presenta con una posible solución eficiente ya que permite evitar el acceso directo a una base de datos, así como diseñar todos los métodos y controles necesarios para una implementación automática. De esta manera se podrían minimizar los problemas de sesgo y de fuga de información.

Finalmente, se estableció como la privacidad por defecto queda configurada. La consecuencia directa es que la responsabilidad en el tratamiento de datos personales requiere de una definición acabada de pautas y políticas de privacidad definidas dentro de la organización. Para ello resulta necesario no solo un compromiso de esta, sino también la elección de un responsable que pueda llevar adelante el cumplimiento de lo establecido en dialogo con todas las áreas de la organización.

Tener en cuenta los puntos principales establecidos, permite concluir que la gestión de la privacidad de datos personales en organizaciones estatales resulta necesaria para lograr con eficiencia un sistema responsable. Si bien pueden ampliarse o profundizarse, los aspectos mencionados no dejan de constituir un punto de partida esencial que toda organización pública debería tener en cuenta. Esto le permitirá lograr con mayor eficiencia la mitigación de riesgos de privacidad derivados de los datos. Además, brinda la posibilidad de garantizar la responsabilidad involucrada basada en el respeto del derecho de la privacidad que poseen los individuos. En consecuencia, evidenciar la falta de responsabilidad en el tratamiento de datos publicados y establecer los criterios para evitar este tipo de problemáticas se considera un aporte fundamental para el manejo de datos personales en el sector público.

## Conclusión

El principal resultado alcanzado en el trabajo final de maestría es la identificación de los aspectos relevantes para el diseño de una estrategia responsable de la privacidad de datos personales en organizaciones públicas estatales en base a la generalización de resultados obtenidos de la implementación de la Privacidad Diferencial. Esta también puede hacerse extensiva a organizaciones privadas. En este sentido, constituye una herramienta fundamental como primera aproximación a establecer garantías de privacidad de datos en el diseño de una estrategia en un marco de gestión responsable de datos en el Estado. La identificación de riesgos de privacidad, la implementación de la privacidad diferencial para la protección de datos personales y la definición de responsabilidades asignadas por defecto, resultaron ser los aspectos relevantes.

El objetivo general que guio la elaboración de este trabajo fue que la aplicación de la metodología de Privacidad Diferencial es efectiva para la protección de datos personales obtenidos de organizaciones públicas estatales lo que a su vez permitió establecer los aspectos relevantes para aproximar a una definición de una gestión responsable de la privacidad de datos en el Estado argentino. Su desarrollo fue posible comenzando por la elaboración manual de una base de datos a partir de seleccionar una muestra de un padrón de CUIT de 2015 publicados por una organización pública del Gobierno de la Ciudad de Buenos Aires. Luego, se identificaron los riesgos derivados en torno a la privacidad de los individuos. Finalmente, se llevó a cabo la aplicación de la metodología seleccionada obteniendo resultados satisfactorios que luego permitieron establecer los aspectos principales para la definición de una estrategia de gestión responsable de la privacidad. Este procedimiento permitió mostrar que el uso y procesamiento de datos personales en las organizaciones públicas vulnera la privacidad de los individuos. Si bien se advierten las limitantes de lo realizado en términos del alcance, se considera que constituye un aporte inicial significativo para las organizaciones, así como para la sociedad argentina.

La realización de la tesis se estructuró en cuatro capítulos, agrupados en dos partes. En los capítulos uno y dos, se presentó el corpus teórico. Mientras en el primero, se definió al Estado como una organización basada en datos para la realización de políticas públicas y su interacción en el mercado de datos; En el segundo se conceptualizó a la privacidad y, particularmente, a la privacidad de datos personales, que derivó en la definición de la Privacidad Diferencial. Los capítulos tres y cuatro, presentaron la aplicación técnico-

cuantitativa y cualitativa de protección de datos personales desde un marco de responsabilidad. En base a la generalización de resultados obtenidos, se identificaron los puntos relevantes para el diseño de una estrategia de privacidad en un marco de gestión responsable en organizaciones públicas del Estado.

El objetivo del primer capítulo fue describir a las organizaciones públicas estatales como una organización basada en datos y su interacción con los individuos en el mercado para la generación de valor. La hipótesis asociada fue que el procesamiento de datos personales a través de tecnología por parte de las organizaciones públicas del estado les otorga un poder superior que pone en riesgo la privacidad de los individuos. Mediante el desarrollo de las diferentes instancias, se pudo establecer los riesgos derivados que vulneran el derecho a la privacidad que poseen los individuos. En este sentido, el aporte principal del capítulo es la necesidad de diseñar un marco de gestión responsable de datos personales para dar garantías de privacidad en organizaciones estatales.

La contribución de este primer capítulo fue argumentar por qué la tecnología es esencialmente utilizada como un instrumento de poder en las organizaciones públicas estatales. En primer lugar, se desarrolló la conceptualización de una organización basada en datos a partir de tres dimensiones claves: las tecnologías de *Big Data*, la capacidad de generación de información y su interacción con los individuos para la obtención de datos. Para ello, se analizaron las principales particularidades de cómo se lleva a cabo el procesamiento de datos obtenidos en el mercado para la toma de decisiones remarcando como la tecnología es un instrumento utilizado no solo para generar valor. Finalmente, se analizó como los datos recolectados a través de estas son utilizados para generar conductas humanas futuras identificando los riesgos asociados a privacidad y se expuso la necesidad de la responsabilidad que ello conlleva. A su vez, se determinó la necesidad de implementar una metodología que brinde garantías de privacidad en el procesamiento de datos.

El segundo capítulo, tuvo por objetivo presentar el modelo de Privacidad Diferencial, así como todos los conceptos de privacidad involucrados. La hipótesis planteada fue que la privacidad de datos personales concebida desde el diseño y por defecto resulta necesaria para establecer un marco responsable de gestión de la privacidad. A través de la exposición de los diferentes conceptos, se pudo establecer la necesidad de establecer un diseño de procesamiento de datos con eje en la privacidad teniendo en cuenta también el marco

legislativo vigente en Argentina. La responsabilidad de quienes participan se constituyó en un punto central a ser tenido en cuenta.

En el capítulo se analiza la necesidad de definir una estrategia que incorpore la protección de datos personales durante todas las etapas de su procesamiento ya que permite asignar las responsabilidades por defecto, informando y formando a quienes se encuentren involucrados. Si bien esto es posible de llevarse a cabo en organizaciones públicas estatales argentinas con la ayuda de una metodología como la privacidad diferencial, no se verifica si está siendo implementada actualmente. No obstante, el modelo propuesto constituye una herramienta de gestión que permite mitigar los riesgos de privacidad derivados del procesamiento de datos personales.

En el tercer capítulo, el objetivo fue, por un lado, construir una base de datos personales. Por el otro identificar los riesgos de privacidad que surgen de esta y, finalmente, aplicar el modelo de privacidad diferencial para evaluar su efectividad en términos de si es posible evitar la identificación de un individuo para proteger su privacidad. La hipótesis asociada fue que la aplicación de privacidad diferencial resulta efectiva para la protección de datos personales. Mediante la aplicación que se realizó, se logra obtener un resultado satisfactorio sobre una base de datos personales en la medida que permitió mitigar el riesgo de privacidad detectado.

Para llevar adelante el capítulo tres, en primer lugar, se explicitó como fue construida la base de datos personales utilizada. Luego, se realizó una descripción de los datos finalmente obtenidos. A continuación, se identificaron las problemáticas asociadas en torno a privacidad. Finalmente, se desarrolló la aplicación del modelo de privacidad diferencial y se realizó la evaluación de su efectividad en términos de los resultados obtenidos.

Si bien se advierte la limitación de este capítulo en cuanto al alcance, el aporte generado mediante la aplicación llevada a cabo es la demostración de como una organización pública estatal a través de publicar un identificador único como el CUIT pone en riesgo la privacidad de los individuos. Esto porque habilita a que cualquier individuo u organización pueda recopilar otros datos personales pudiendo hacer un uso mal intencionado. A partir de ello, se aplicó ruido aleatorio a través de la Privacidad Diferencial a los datos que permitieron distorsionarlos obteniendo como resultado la mitigación del riesgo involucrado.

Por último, el cuarto capítulo tuvo como objetivo especificar los principales puntos a ser tenidos en cuenta en base a lo realizado para llevar adelante el diseño de una estrategia de la gestión de la privacidad de datos personales en organizaciones públicas del Estado. Para lograrlo se utilizaron los resultados obtenidos en los capítulos previos, a partir de los cuales desde un enfoque cualitativo se derivaron los puntos relevantes a ser tenidos en cuenta. De este modo, se logró aproximar a una primera elaboración de un marco de gestión responsable para la definición de una estrategia del diseño de la gestión de la privacidad. Si bien no cuenta con una hipótesis específica planteada, al ser un enfoque cualitativo se deriva de los resultados alcanzados.

El desarrollo de este último capítulo consistió, en una primera parte, en explicitar los principios sobre los cuales es necesario sustentar el diseño de la privacidad en organizaciones. Luego se expusieron los puntos relevantes a ser tenidos en cuenta en base a los resultados obtenidos en los capítulos previos. Finalmente, se definieron de manera aproximada los factores relevantes para determinar la responsabilidad por defecto en la gestión de la privacidad de datos personales.

Tras el desarrollo de cada uno de los objetivos mencionados, el trabajo final de maestría aporta un resultado que se considera de especial interés para las organizaciones, pero también para la sociedad argentina en su conjunto. Una primera aproximación para el diseño de un sistema de gestión de la privacidad de datos personales en una actualidad de discusiones antagónicas y sin consenso sobre la privacidad de datos, puede considerarse una herramienta que aporta valor. Si bien resulta necesario profundizarse en los puntos implementados y destacados, no dejan de constituir un punto de partida esencial que toda organización pública debería tener en cuenta para lograr con mayor eficiencia el diseño de una estrategia de privacidad. Además, brinda la posibilidad de garantizar el respeto del derecho de la privacidad que poseen los individuos.

El trabajo presenta diversas posibilidades para continuar ampliando su alcance. Una línea de trabajo en particular que surge es el desarrollo de la interfaz propuesta en el capítulo cuatro de forma tal de mostrar como la tecnología puede ayudar a lograr una gestión eficiente y responsable de la privacidad y no solo ser utilizada como un instrumento de poder. También, es posible extender el caso de estudio hacia otros datos personales y evaluar la aplicación de diferentes metodologías para lograr determinar cuál es la más efectiva en términos de protección de datos personales.

Una vez realizado el análisis de la privacidad desde un marco de responsabilidad, explicada e implementada la metodología de la Privacidad Diferencial a partir del cual se construyó y se derivaron los aspectos relevantes, y teniendo en cuenta las posibilidades de ampliar el alcance de lo realizado, el trabajo a futuro consiste en realizar una propuesta formal para implementar esta metodología a través del diseño de una interfaz en organizaciones públicas estatales. Posiblemente algunas de ellas cuenten con métodos para la protección de datos personales, pero son desconocidos ya que no son informados a la sociedad. Sin embargo, el foco está puesto en la potencialidad de la tecnología para el procesamiento de datos vinculada con la idea realizar políticas públicas eficientes, y no con la idea de incorporar un marco de responsabilidad sobre ese desarrollo.

## Referencias bibliográficas


AEPD (Agencia Española de Protección de Datos) [en línea] https://www.aepd.es/es/guias-y-herramientas/guias

——— (2019), Guía de Privacidad desde el Diseño. [en línea] https://www.aepd.es/es/documento/guia-privacidad-desde-diseno.pdf

——— (2020), Guía de Protección de Datos por Defecto. [en línea] https://www.aepd.es/sites/default/files/2020-10/guia-proteccion-datos-por-defecto.pdf

——— (2020), Tecnologías y Protección de Datos en las AA.PP. [en línea] https://www.aepd.es/es/documento/guia-tecnologias-admin-digital.pdf

Bryson, J. M., Crosby, B. C., & Bloomberg, L., (2015). *Introduction*. In J. M. Bryson, B. C. Crosby & L. Bloomberg (Eds.), Creating public value in practice: Advancing the common good in a multi-sector, shared-power, no-one-wholly-in-charge world (pp. 128). CRC Press. ISBN 9781482214604

Callon, M. & Muniesa, F. (2003). Les marchés économiques comme dispositifs collectifs de calcul. Réseaux, vol. no 122, no. 6, 2003, pp. 189-233. https://www.cairn.info/revue--2003-6-page-189.htm.

Castells, M. (2004). La era de la información: economía, sociedad y cultura (Vol. 1). siglo XXI.

Cavoukian, A. (2011), Privacy by Design: The 7 Foundational Principles. Ph.D. Information & Privacy Commissioner Ontario, Canada. https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf

Clarke, R. (1999). Internet privacy concerns confirm the case for intervention. Communications of the ACM, Volume 42, Issue 2, Feb. 1999, pp 60–67 https://doi.org/10.1145/293411.293475

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. DOI: https://doi.org/10.1109/HICSS.2010.488

Constantiou, I., Kallinikos, J. (2015) New games, new rules: big data and the changing context of strategy. J Inf Technol 30, 44–57. https://doi.org/10.1057/jit.2014.17

Culpepper, P. D., & Thelen, K. (2020). Are we all Amazon primed? Consumers and the politics of platform power. Comparative Political Studies, 53(2), 288-318. https://doi.org/10.1177/0010414019852687

Custers, B. (2013). Data Dilemmas in the Information Society: Introduction and Overview. In Calders, T., Schermer, B., Zarsky, T., & Custers, B. (Eds.), Discrimination and Privacy in the Information Society (pp. 3-26). Springer. ISBN: 978-3-642-30487-3

Dwork, C. (2008). *Differential Privacy: A Survey of Results*. In: Agrawal, M., Du, D., Duan, Z., Li, A. (eds) Theory and Applications of Models of Computation. TAMC 2008. Lecture Notes in Computer Science, vol 4978. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-79228-4_1

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci., 9(3-4), 211-407. https://doi.org/10.1561/0400000042

Erickson, J. (1999). Algorithms. Creative Commons Attribution 4.0 International license. https://jeffe.cs.illinois.edu/teaching/algorithms/book/Algorithms-JeffE.pdf

Gandy Jr, Oscar H. (2021). Information and Power. In O. Gandy Jr, The panoptic sort: A political economy of personal information (pp. 29-50). Oxford University Press, 2021. https://doi.org/10.1093/oso/9780197579411.001.0001

García Jiménez, E., Gil Flores, J., & Rodríguez Gómez, G. (1994). Análisis de datos cualitativos en la investigación sobre la diferenciación educativa. Revista de investigación educativa, 23, 179-213. URI: https://hdl.handle.net/11441/77867

Haskel, J., & Westlake, S. (2017). Capitalism without capital. Princeton University Press. ISBN: 9780691175034

HromkoviČ, J. (2005). *Fundamentals*. In HromkoviČ, J, Design and Analysis of Randomized Algorithms. Introduction to Design Paradigms (pp. 19-99). Springer Berlin, Heidelberg. ISBN: 978-3-540-27903-7

Kolanovic, M., & Krishnamachari, R. T. (2017). Big data and AI strategies: Machine learning and alternative data approach to investing. JP Morgan Global Quantitative & Derivatives Strategy Report. https://cpb-us-e2.wpmucdn.com/faculty.sites.uci.edu/dist/2/51/files/2018/05/JPM-2017-MachineLearningInvestments.pdf

Latour, B. (1987). Science in action: How to follow scientists and engineers through society. Harvard university press. ISBN 9780674792913

Latour, B. (2005). Reassembling the social: An introduction to actor-network-theory. Oxford university press.

Mazzucato, M., Entsminger, J., & Kattel, R. (2020). Public Value and Platform Governance. UCL Institute for Innovation and Public Purpose WP 2020-11. http://dx.doi.org/10.2139/ssrn.3741641

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. https://hbr.org/2012/10/big-data-the-management-revolution

McKeen, J., & Smith, H. (2007). Developments in Practice XXIV: Information Management: The Nexus of Business and IT. Communications of the Association for Information Systems, 19, pp-pp. https://doi.org/10.17705/1CAIS.01903

Milgrom, P. (1992). Economics, organization and management. Prentice-Hall, Inc. Pag. 43. ISBN: 9780132246507

Moazed, A., & Johnson, N. L. (2016). Modern monopolies: what it takes to dominate the 21st century economy. St. Martin's Press. ISBN: 9781250091895

Naser, A. (2021). Gobernanza digital e interoperabilidad gubernamental: una guía para su implementación. Documentos de Proyectos (LC/TS.2021/80), Santiago, Comisión Económica para América Latina y el Caribe (CEPAL). URI https://hdl.handle.net/11362/47018

Near, J. P., & Abuah, C. (2021). Programming Differential Privacy. Vol 1. [en línea] https://uvm-plaid.github.io/programming-dp/

OCDE (Organización de Cooperación y Desarrollo Económicos) (2019), Índice de gobierno digital OCDE 2019: resultados y mensajes clave. [en línea] https://www.oecd.org/gov/recomendacion-del-consejo-sobre-gobierno-abierto-141217.pdf

——— (2014), Recommendation of the Council on Digital Government Strategies. [en línea] https://one.oecd.org/document/C(2014)88/en/pdf

Rivoir, A. L., & Morales, M. J. (2019). *Derechos digitales y regulación de Internet. Aspectos claves de la apropiación de tecnologías digitales*. En A. L. Rivoir, & M. J. Morales (Eds.). Tecnologías digitales: Miradas críticas de la apropiación en América Latina (pp. 35-49). CLACSO. http://biblioteca.clacso.edu.ar/clacso/se/20191128031455/Tecnologias-digitales.pdf

Rhodes, T., & Lancaster, K. (2019). Evidence-making interventions in health: A conceptual framing. Social Science & Medicine, 238, 112488. https://doi.org/10.1016/j.socscimed.2019.112488

Salaberry, N. R., & Herrera, P. M. (2021). Gestión de la privacidad de datos sensibles.: El aplicativo CuidAR para el control del COVID-19. Gestión Joven, 22(2), 48-58. ISSN-e 1988-9011

Saunders, M., Lewis, P., & Thornhill, A. (2016). Research methods for business students (7th ed.). Pearson Education. ISBN: 0273701487, 9780273701484

Schmarzo, B. (2013). Big Data. Understanding How Data Powers Big Business. John Wiley & Sons, Inc. ISBN: 978-1-118-73957-0

Seidl, T. (2021). Commodification and Disruption: Theorizing Digital Capitalism. Centre for European Integration Research. WP, 02/2021. https://eif.univie.ac.at/workingpapers/index.php

Shapiro, C., Carl, S., & Varian, H. R. (1998). Information rules: a strategic guide to the network economy. Harvard Business Press. https://doi.org/10.2307/1183273

Täuscher, K., & Laudien, S. M. (2018). Understanding platform business models: A mixed methods study of marketplaces. European Management Journal, 36(3), 319-329. https://doi.org/10.1016/j.emj.2017.06.005

Zuboff, S. (2015). Big other: surveillance capitalism and the prospects of an information civilization. Journal of information technology, *30(1), 75-89*. https://doi.org/10.1057/jit.2015.5

Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. Profile books. ISBN-13 9781610395694

Zysman, J., & Kenney, M. (2018). The next phase in the digital revolution: Intelligent tools, platforms, growth, employment. Communications of the ACM, 61(2), 54–63. https://doi.org/10.1145/3173550

## Apéndice

### A1. Simulación de distribución de Laplace con Python

Se crea una función en Python para simular la densidad de probabilidad de Laplace con diferentes combinaciones de parámetros:

```
#Función para simulación de la densidad de probabilidad de Laplace
import numpy as np
import pandas as pd
import random as rd

def simula_laplace(b1, b2, b3, n):
  np.random.seed(2022)

  DL_1 = pd.DataFrame(np.random.laplace(0, b1, n))
  DL_2 = pd.DataFrame(np.random.laplace(0, b2, n))
  DL_3 = pd.DataFrame(np.random.laplace(0, b3, n))

 laplace_data=pd.concat([DL_1, DL_2, DL_3], axis=1)
 laplace_data.columns=["DL_1", "DL_2", "DL_3"]
 return laplace_data
```

```
#se aplica la función para simular 10000 valores:
 laplace_data=simula_laplace(1,2,4, 10000)

#visualización de la densidad de probabilidad de Laplace
import seaborn as sns
sns.displot(laplace_data, kind="kde")
```

## A2. Implementación de Privacidad Diferencial con Python

Implementaciones de Privacidad Diferencial: Código implementado en Python en el entorno Google Colaboratory

- Obtención de datos:

```
#descarga del conjunto de datos desde el portal de datos abiertos del Gobierno
 de la Ciudad de Buenos Aires
!wget https://cdn.buenosaires.gob.ar/datosabiertos/datasets/administracion-
gubernamental-de-ingresos-publicos/alto-riesgo-fiscal/alto-riesgo-fiscal-
primer-trimestre-2015.csv
#importación de librería pandas
import pandas as pd
#lectura del conjunto de datos
data=pd.read_csv("/content/alto-riesgo-fiscal-primer-trimestre-
2015.csv", sep=";")
#Se seleccionan las columnas "cuit", "tipo_contr_insc" y "raz_soc"
data_1=data.iloc[:,[3,4,11]]
#selección de cuits que comienzan con 20 para género masculino
data_1_M=pd.DataFrame(data_1[(data_1['cuit']>=20000000000) & (data_1['cuit']<2
1000000000)])
# se verifican los casos que contienen S.A. o S.A en el nombre ya que no son p
ersonas físicas
data_1_M["raz_soc"][data_1_M.raz_soc.str.contains('S.A.|S.A',na=False)].count(
)
#se eliminan los casos identificados anteriormente
data_1_M.drop(data_1_M[data_1_M.raz_soc.str.contains('S.A.|S.A',na=False)].ind
ex , inplace=True)
#se resetea el indice
data_1_M=data_1_M.reset_index(0)
```

- Selección de muestra en base al CUIT

```
#selección de cuits que comienzan con 27 para género femenino
data_1_F=data_1[(data_1['cuit']>=27000000000) & (data_1['cuit']<28000000000)]
# se verifican los casos que contienen S.A. o S.A en el nombre ya que no son p
ersonas físicas
data_1_F["raz_soc"][data_1_F.raz_soc.str.contains('S.A.|S.A',na=False)].count(
)
#se eliminan los casos identificados anteriormente
```

```
data_1_F.drop(data_1_F[data_1_F.raz_soc.str.contains('S.A.|S.A',na=False)].ind
ex , inplace=True)
#se resetea el índice del conjunto de datos
data_1_F=data_1_F.reset_index(0)
```

- Creación de la variable género

```
#se crea una variable "GENERO" que contiene el valor M=masculino
data_1_M['GENERO']="M"
#se crea una variable "GENERO" que contiene el valor F=femenino
data_1_F['GENERO']="F"
```

- Conjunto de datos inicial

```
individuos = pd.concat([data_1_F, data_1_M], axis=0, ignore_index=True)
#se cuentan casos donde no existe el nombre
individuos[individuos['raz_soc']=='S/DATOS TITULARIDAD'].count()
#se elimina el caso sin nombre
individuos.drop(individuos[individuos.raz_soc.str.contains('S/DATOS TITULARIDA
D',na=False)].index , inplace=True)
#se elimina la columna index
individuos=individuos.drop(['index'], axis=1)
#se exportan los datos a un archivo csv
individuos.to_excel('/content/sample_data/individuos.xlsx', index=False, engin
e='xlsxwriter')
```

- Importación de la nueva base de datos creada con datos reales obtenidos en forma manual en base a una muestra de 1000 CUITS al azar de la base anteriormente exportada.

```
import pandas as pd
df = pd.read_excel('/content/INDIVIDUOS.xlsx')
df.head(5)
```

- Descriptiva de la base de datos

```
#se verifican cuantos casos no poseen código de actividad
df['Código_actividad'].isnull().sum()
#se cuentan la cantidad de casos con y sin beneficio social
df.groupby(['Beneficiario_social']).Beneficiario_social.agg('count')
#se cuentan la cantidad de casos por jurisdiccion
df.groupby(['JURISDICCION']).JURISDICCION.agg('count')
#se visualiza la cantidad de casos por jurisdiccion
plt.figure(figsize=(10,10))

a=sns.countplot(x=df['JURISDICCION'], data=df,  order = df['JURISDICCION'].val
ue_counts().index, palette="Set2" )
a.tick_params(labelsize=10)

for p in a.patches:
   a.annotate('{:.0f}'.format(p.get_height()), (p.get_x()+0.01, p.get_height()
+0.01), fontsize=14)
```

```
plt.xticks(rotation=+45)
plt.ylabel("Individuos por jurisdicción")

#se verifican cantidad de individuos según categoría de monotributo
df.groupby(['tipo_contr_insc']).tipo_contr_insc.agg('count')

#se verifican las actividades principales que realizan los individuos
df.groupby(['Actividad']).Actividad.agg('count').sort_values()

#se verifican cantidad de individuos según tipo de beneficio y género
df.groupby(['Beneficiario_social', 'Nombre_beneficio_social']).Beneficiario_so
cial.agg('count')
#se visualiza la cantidad de casos por tipo de beneficio y género
plt.figure(figsize=(10,10))

a=sns.countplot(x=df['Nombre_beneficio_social'], data=df[df['Beneficiario_soci
al']=='SI'], hue=df['GENERO'], order = df['Nombre_beneficio_social'].value_cou
nts().index, palette="Set2" )
a.tick_params(labelsize=10)

for p in a.patches:
   a.annotate('{:.0f}'.format(p.get_height()), (p.get_x()+0.01, p.get_height()
+0.01), fontsize=14)

plt.xticks(rotation=+45)
plt.ylabel("Individuos beneficiarios social")

#instalación de módulo
!pip install geopandas
#importación de librería para manipular dataframes con datos geográficos
import geopandas as gpd
#obtención de archivo shp de barrios de CABA
!wget https://cdn.buenosaires.gob.ar/datosabiertos/datasets/barrios/barrios-
zip.zip
#se descomprime el archivo
! unzip /content/barrios-zip.zip
#lectura del archivo shp de barrios de CABA
import geopandas as gpd
barrios = gpd.read_file("/content/barrios_badata_wgs84.shp")
print(barrios)
#se normalizan dos nombres de barrios en el base de datos
df['BARRIO']=df['BARRIO'].replace({'MONTSERRAT':'MONSERRAT', 'LA BOCA':'BOCA'}
)
#se generan la cantidad de individuos por barrio de CABA
BARRIOS=pd.DataFrame(df[df['JURISDICCION']=='CABA'].groupby(['BARRIO']).BARRIO
.agg('count'))
BARRIOS=BARRIOS.add_suffix('_Count').reset_index()
BARRIOS
#al archivo shape se le agregan los nuevos datos
```

```
barrios_1=pd.merge(barrios, BARRIOS, on="BARRIO", how='outer' )
#se observa
barrios_1.head(5)
#se completa la variable con valor cero en caso de ser null
barrios_1=barrios_1.fillna(0)
#obtención del centroide del polígono
barrios_1['coords'] = barrios_1['geometry'].apply(lambda x: x.representative_p
oint().coords[:])
barrios_1['coords'] = [coords[0] for coords in barrios_1['coords']]
#visualización del mapa
barrios_1.plot(column='BARRIO_Count', cmap='OrRd', edgecolor='k', legend=True,
  figsize=(30, 30))
for idx, row in barrios_1.iterrows():
    plt.annotate(s=row['BARRIO']+'\n'+str(row['BARRIO_Count']), xy=row['coords
'],horizontalalignment='center', fontsize=16)

plt.title("Individuos por Barrio de CABA", fontsize=14)
```

- Anonimización del CUIT

```
#se intala el módulo de anonimización
!pip install anonymizedf
#se importa la librería para anonimización
from anonymizedf.anonymizedf import anonymize
#se anonimiza el CUIT
an = anonymize(individuos)
an.fake_ids("cuit")
#los casos con código actividad null se completan con 0
individuos['Código_actividad']=individuos['Código_actividad'].fillna(0)
#se observa
individuos['Código_actividad'].unique()
#se cambia a tipo entero el dato de código de actividad
individuos['Código_actividad']=individuos['Código_actividad'].astype('int')
#se verifica
individuos.dtypes
```

- Cantidad de combinaciones para identificación indirecta

```
import seaborn as sns
from matplotlib import pyplot as plt
#Se agrupan los casos por CP y GENERO para contarlos y determinar cuantos caso
s combinando ambas variables son únicos
data_1=individuos.groupby(['GENERO', 'CP'],as_index=False).CP.count()
#se visualiza cantidad de casos según CP y género
plt.figure(figsize=(10,10))
a=sns.countplot(x='CP',data=data_1, palette="Blues_r")
a.tick_params(labelsize=14)
for p in a.patches:
```

```
   a.annotate('{:.0f}'.format(p.get_height()), (p.get_x(), p.get_height()), fo
ntsize=13)
plt.xlabel("Cantidad de casos según CP y género")
plt.ylabel("Cantidad de casos")
#porcentaje de casos posibles de identificar de manera unívoca combinando CP y
 GÉNERO
'el {:.2f}% de individuos son posibles de identificar a partir de combinar las
 variables código postal y género'.format(round(240/len(individuos),4)*100)
#Se agrupan los casos por CP, EDAD y GÉNERO para contarlos y determinar cuanto
s casos combinando las tres variables son únicos
data_2=individuos.groupby(['GENERO', 'Código_actividad', 'CP'],as_index=False
).CP.count()
#Se agrupan los casos por CP, GÉNERO y Código_actividad para contarlos y deter
minar cuantos casos combinando las tres variables son únicos
data_2=individuos.groupby(['GENERO', 'Código_actividad', 'CP'],as_index=False
).CP.count()
#porcentaje de casos posibles de identificar de manera unívoca
'El {:.2f}% de individuos son posibles de identificar a partir de combinar las
 variables CP, GENERO y JURISDICCION'.format(round(609/len(individuos),4)*100)
```

- Generación de conjuntos de datos identificar indirectamente individuos

```
#Conjunto de datos que contiene 'tipo_contr_insc', 'raz_soc', 'GENERO', 'CP',
'Domicilio_fiscal','JURISDICCION', 'BARRIO', 'Código_actividad', 'Actividad',
el cuit anonimizado
individuos_1= individuos.iloc[:,[0, 1, 2, 3, 4, 5, 6, 7, 8, 12]]
#Conjunto de datos que contiene el 'GENERO', 'CP', 'Beneficiario_social', 'Nom
bre_beneficio_social', 'Lugar_de_cobro','Fake_cuit'
individuos_2= individuos.iloc[:,[2, 3, 9, 10, 11,12]]
```

- Reidentificación indirecta

```
#se cuentan los casos agrupados por CP, GÉNERO Y Código_actividad
individuos_1.groupby(['CP','GENERO','Código_actividad']).agg('count')
#se individualiza un caso de los 609 como ejemplo a partir de los tres atribut
os que combinados permiten  identificar a un individuo
caso1 = individuos_1[(individuos_1['CP'] == 1002) & (individuos_1['GENERO'] ==
'F') & (individuos_1['Código_actividad'] ==702091)]
#se busca el caso1 en el conjunto de datos individuos_2
pd.merge(caso1, individuos_2, how='inner')
```

- Aplicación de privacidad diferencial

```
#se instala el módulo OpenDP
!pip install opendp-smartnoise seaborn
# se carga la librería
import opendp.smartnoise.core as sn

 #se genera una función que sobre la muestra aplica privacidad diferencial a l
 a variable CP, y otra que calcula el error
```

```
def dp_histo(epsilon=0.5): #la elección del valor de epsilon no es determinan
te (sino un valor fracuente de usar)
    np.random.seed(2022) #se setea una semilla
    var_names = ['GENERO', 'CP', 'Beneficiario_social', 'Nombre_beneficio_soc
ial', 'Lugar_de_cobro', 'Fake_cuit', 'CP_2']

    with sn.Analysis(protect_floating_point=False) as analysis:
        data = sn.Dataset(path = '/content/samplefile', column_names = var_na
mes)
        nsize = sample_size

#se utiliza el método del histograma debido a que es muy sensible ante cambio
s (ruido) en los datos
        inc_prep = sn.histogram(data['CP_2'], #histograma sin privacidad dife
rencial
                categories = cp_categorias, null_value = "-1")
        #se aplica privacidad diferencial por el mecanismo de Laplace
        inc_histogram2 = sn.laplace_mechanism(inc_prep, privacy_usage={"epsil
on": epsilon, "delta": .000001})

    analysis.release()

    return  inc_histogram2.value

#se genera una función que calcula el error (diferencia porcentual promedio e
ntre las clases del histograma)
def perc_error(true_col, dp_col):
    return sum(abs(true_col - dp_col)) / sum(true_col) * 100

# se aplica la función creada para generar privacidad diferencial
epsilon = 0.5
np.random.seed(2022)
laplace_counts = dp_histo(epsilon)
print("Histrograma de la EDAD aplicando PD-
Laplace:     " + str(laplace_counts.astype(int)))

# se calcula el error
error_lapl = perc_error(results['Valores_originales'], results['PD-Laplace'])
print ('PD -
Laplace con ε = ' + str(epsilon) + ' : Desviación respecto de los datos sin P
D = ' + str(round(error_lapl,1)) +'%')

#se crea uan función para evaluar el modelo para diferentes tamaños de muestr
as y valores de epsilon(privacidad)

def evaulacion(datos, sample_sizes,epsilons, edad_categorias):
  np.random.seed(2022)
  for sample_size in sample_sizes:
      np.random.seed(2022)
```

```
        sample = datos.sample(n = sample_size, random_state=np.random.seed(2022
))
        sample.to_csv('samplefile', index=False)

        results = pd.DataFrame(index=cp_categorias)
        private_df = pd.read_csv('/content/samplefile')
        results['True'] = private_df['CP_2'].value_counts()

        labels = ['Verdadero valor en la muestra original']

        for epsilon in epsilons:
            np.random.seed(2022)
            laplace_counts = dp_histo(epsilon)

            results['Laplace-PD-
'+str(epsilon)] = laplace_counts[:len(cp_categorias)]

            error = perc_error(results['True'], results['Laplace-PD-
'+str(epsilon)])

            labels.append('PD con ε = ' + str(epsilon) + ' : Error = ' + str(ro
und(error,1)) + '%')

        ax = results.plot.bar(rot=45, figsize=(20, 15), width=0.8, fontsize=12,
 colormap = 'Greens')
        plt.title('Distribución del CP (n = '+str(sample_size)+')', fontsize=16
)
        plt.legend(fontsize = 12, labels=labels)
        plt.grid(axis='y', alpha=0.5)
        plt.ylabel('Cantidad de casos', fontsize = 14)
        plt.xlabel('Categorías de CP', fontsize = 14)
        plt.ylim(-50)
    return ax

#se aplica la función
np.random.seed(2022)
evaulacion(individuos_3, [250, 500, 750, 850, 950], [0.8, 0.5, 0.25, 0.05], [
'<= 1500' , '1501-1800', '1801-4000','4001-9420'])
```

- Implementación de ruido aleatorio por mecanismo de Laplace a la variable CP

```
import numpy as np
#se define una función que aplica ruido aletorio según mecanismo de Lapalce
def ruido(data, variable, variable_new, sensitivity, epsilon):
  np.random.seed(202206)
  data[variable_new]=data[variable] + np.random.laplace(loc=0, scale=sensitivi
ty/epsilon)
  return data
#se aplica la función, eligiendo un epsilon de 0.5 ya de la evaluación se obtu
vo un error de 0%.
```

```
dapar=ruido(individuos_2.copy(), 'CP', 'CP_2', 1, 0.5)
```

- Evaluación de resultados

```
# Se crea una nueva variable CP_3 recategorizando la variable CP_2 y se visual
iza
edad_categorias_2 = ['<= 1500' , '1501-1800', '1801-4000','4001-9420']
cut_bins_2 =  [-1, 1500, 1800, 4000, 9420] # 10 categories
dapar['CP_3'] = pd.cut(dapar['CP_2'], bins=cut_bins_2, labels=edad_categorias_
2)
plt.figure(figsize=(10,10))
dp3=sns.countplot(x="CP_3", data=dapar, palette="crest")
for p in dp3.patches:
   dp3.annotate('{:.0f}'.format(p.get_height()), (p.get_x()+0.3, p.get_height(
)+0.01), fontsize=15)
plt.ylabel("Cantidad de casos")
plt.xlabel("CP sin privacidad diferencial")

# Se visualiza la variable CP original pero recategorizada para comparar
dapar['CP_4'] = pd.cut(dapar['CP'], bins=cut_bins_2, labels=edad_categorias_2)


plt.figure(figsize=(10,10))
dp4=sns.countplot(x="CP_4", data=dapar, palette="flare")
for p in dp4.patches:
   dp4.annotate('{:.0f}'.format(p.get_height()), (p.get_x()+0.3, p.get_height(
)), fontsize=15)
plt.ylabel("Cantidad de casos")
plt.xlabel("Categorías de CP con privacidad diferencial")

import statistics as stat

#visualización de estadísticas y distribución acumulada de CP
m1 = [dapar['CP'].median()]
med=[stat.mean(dapar['CP'])]
plt.figure(figsize=(10,10))
g = sns.displot(data=dapar,x='CP',kind='ecdf', color='lightseagreen', stat='co
unt')
axes = g.axes.flatten()
for ax in axes:
    ax.axvline(x=m1[0], c='orange', ls='--', lw=1.5)
    ax.axvline(x=med[0], c='palevioletred', ls='--', lw=1.5)
plt.text(200,800,"Median",fontsize=12)
plt.text(2000,1000,"Media",fontsize=12)
plt.xlim(0,12000, 100)
plt.ylabel("Cantidad de casos")
plt.xlabel("CP sin ruido aleatorio")

#visualización de estadísticas y distribución acumulada de CP_2
```

```
m1 = [dapar['CP_2'].median()]
med=[stat.mean(dapar['CP_2'])]
g = sns.displot(data=dapar,x='CP_2',kind='ecdf', color='lightcoral', stat='cou
nt')
axes = g.axes.flatten()
for ax in axes:
    ax.axvline(x=m1[0], c='dodgerblue', ls='--', lw=1.5)
    ax.axvline(x=med[0], c='palevioletred', ls='--', lw=1.5)
plt.text(200,900,"Mediana",fontsize=12)
plt.text(1800,1000,"Media",fontsize=12)
plt.ylabel("Cantidad de casos")
plt.xlabel("CP con ruido aleatorio")
plt.xlim(0,12000, 100)

#test de hipótesis de Levene
import scipy.stats as stats
stats.levene( dapar['CP'],dapar['CP_2'],center= 'mean')

#agrupación mediante combinación de variables
dapar.groupby(['GENERO','CP_2']).agg('count')

#se individualiza un caso a partir de tres atributos que combinados permiten
identificar a un individuo
caso2 = dapar[(dapar['GENERO'] =='F') & (dapar['CP_2'] ==1001.206903)]
#se busca el caso en el conjunto de datos individuos_1
pd.merge(caso2, individuos_1, how='inner')
```