¿Cómo compartir datos de manera efectiva y privada?

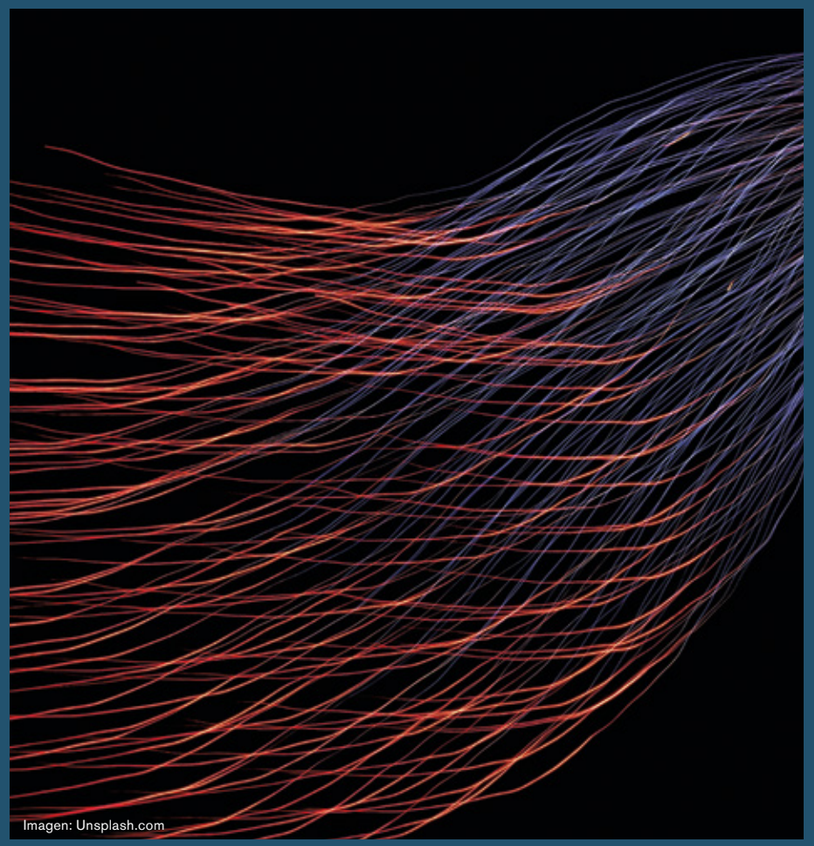


Imagen: Unsplash.com

Autores:
Sergio Yovine,
Facultad de Ingeniería,
Universidad ORT Uruguay
Franz Mayr,
Facultad de Ingeniería,
Universidad ORT Uruguay

Utilidad de los datos y protección de la privacidad

El crecimiento acelerado de los datos disponibles es uno de los factores centrales que motorizan los avances en el campo de la inteligencia artificial. Estos están dando lugar a mejoras significativas en la capacidad de resolver una gran variedad de tareas con la ayuda de algoritmos de *machine learning*. Este es el caso en áreas críticas como salud y ciberseguridad, donde se trabaja activamente en el desarrollo de algoritmos cada vez más precisos para abordar problemas como por ejemplo, el diagnóstico de enfermedades [1] y la detección de intrusiones [2], respectivamente.

Para las organizaciones, la oportunidad de construir sistemas predictivos inteligentes a partir de datos conlleva grandes desafíos. Por lo general, los algoritmos de *machine learning* necesitan cantidades considerables de datos de entrenamiento que permitan alcanzar niveles de exactitud satisfactorios, pero este requisito no suele estar al alcance de una organización sola. Esta carencia podría superarse si las organizaciones compartieran sus datos o los modelos predictivos entrenados con ellos.

Para paliar este déficit, en la última década se han impulsado políticas de datos públicos abiertos [3]. En el caso de Europa, por ejemplo, el acceso a los datos públicos está legislado por la Directiva (UE) 2019/1024 sobre datos abiertos y reutilización de la información del sector público. En Uruguay, la iniciativa de llevar adelante políticas públicas de datos abiertos está legislada en la Ley Sobre el Derecho de Acceso a la Información Pública (Ley N° 18.381), cuya finalidad es fomentar y prescribir la disponibilización de los datos producidos, obtenidos, en poder y/o bajo control de organismos públicos.

Ciertamente, la necesidad de compartir datos no está solamente motivada por la legislación sobre datos abiertos, sino también por el interés de ponerlos a libre disposición de organizaciones públicas y/o privadas que tienen la capacidad técnica (humana y material) de utilizar efectivamente estos datos para la innovación científica y productiva.

Sin embargo, a pesar de las ventajas de compartirlos, en la mayoría de los casos los datos no pueden publicarse ni transferirse fácilmente [4]. De hecho, la mayoría de los datos recogidos por las organizaciones, ya sean públicas o privadas, contienen información sobre individuos (ciudadanos, clientes, usuarios, pacientes, etc.), como por ejemplo, números de documento de identidad, de seguridad social, de cuentas bancarias y de tarjetas de crédito, contraseñas, etc.

Por otro lado, más allá del valor que la propia organización (u otra entidad externa) pueda extraer de los datos, es evidente que estos no les pertenecen, dado que sus verdaderos propietarios son los individuos. Esto ha motivado la aparición de legislación para la protección de la privacidad de los datos. Por ejemplo, el Reglamento General de Protección de Datos (GDPR) de Europa define un marco normativo que se aplica a todas las organizaciones de la UE, independientemente de su ubicación. En Uruguay, los datos publicados están sujetos al cumplimiento de la legislación vigente contemplada en la Ley N° 18.331 Protección de Datos Personales y acción de Habeas Data. Por lo tanto, existe una clara tensión entre la capacidad de proporcionar acceso a los datos y mantener la privacidad.

Para ilustrar la situación, consideremos el siguiente caso de uso de inteligencia artificial como servicio. Una organización (mutualista médica, empresa de comercio electrónico, banco, etc.) tiene almacenados en su base de datos información (historias clínicas, perfiles de compra, ingresos y gastos, etc.) de individuos (pacientes, usuarios, clientes, etc.) con el objetivo de extraer valor de estos datos, por ejemplo, entrenar un algoritmo de *machine learning* para detectar una patología, hacer recomendaciones de compra, calcular el riesgo crediticio, la probabilidad de fuga, o realizar estadísticas sobre enfermedades, consumos de productos, etc.

Con este fin, la organización contrata a un proveedor externo de servicios de análisis de datos y para que este pueda cumplir su cometido debe consultar la base de datos. Sin las medidas de protección adecuadas, estas consultas son factibles de revelar, inclusive de manera no intencional, información de carácter sensible para los propietarios, como por ejemplo la propia identidad de la persona, si tiene o no una cierta enfermedad, o particularidades de sus hábitos de consumo.

Para evitar esta pérdida de privacidad, toda organización debe tomar las medidas de protección adecuadas al otorgar acceso a su base de datos. Necesariamente, estas medidas tienen que ir más

allá del uso de técnicas de anonimización que eliminen cualquier dato de identificación personal. Esto se debe a que existen diversas maneras de re-identificar a los individuos cuyos datos se encuentran en bases de datos anonimizadas con dichas técnicas [5]. Además, la pérdida de privacidad no solo ocurre a través de la publicación de datos, sino que la información sensible puede quedar expuesta a ser revelada al permitir que terceros consulten los modelos predictivos aprendidos a partir de esos datos por algoritmos de *machine learning*. En efecto, esto es posible mediante la utilización de los llamados ataques de inversión de modelos [6].

En suma, el intercambio de información, ya sea en forma de datos brutos o de modelos aprendidos usando algoritmos de *machine learning*, debe garantizar niveles apropiados de privacidad. Esta cuestión no es sólo técnica, sino también jurídica, ya que existen leyes relativas a la salvaguarda de la privacidad. Por lo expuesto, queda claro que es esencial contar con mecanismos para proteger la información privada contenida en los datos que se ponen a disposición de terceros. Dichos mecanismos deben aplicarse independientemente de la forma en que se compartan los datos, ya sea publicando abiertamente un conjunto de datos o permitiendo que interesados externos consulten una base de datos o un modelo predictivo. Además, los mecanismos de protección de datos deben ser capaces de conservar suficiente información útil para resolver las tareas para las cuales se los necesita [7].

Propuesta de una solución

Como mencionamos anteriormente, dos áreas en las cuales la utilización de datos para la construcción de modelos predictivos mediante algoritmos de *machine learning* es de gran interés, son ciberseguridad y salud. Ambas están particularmente expuestas a la problemática analizada dado que los datos contienen indudablemente información sensible. Para ilustrar la situación, consideremos por ejemplo el caso de la detección de intrusiones a servicios web. Un enfoque exitoso para abordar el problema es entrenar modelos de *machine learning* con logs de acceso [8]. Estos datos son de naturaleza secuencial (cadenas de caracteres) y contienen información sensible de los usuarios del servicio en cuestión (claves de acceso, números de tarjeta, etc.). Un ejemplo del ámbito de la salud es el análisis de electrocardiogramas para predecir cardiopatías [9]. En este caso, los datos son secuencias temporales y forman parte de la historia clínica de los pacientes. Estos ejemplos son ilustrativos de la necesidad de desarrollar herramientas para garantizar la privacidad no solamente en el caso de datos tabulares [10] sino también secuenciales.

Esta observación motivó nuestro interés en investigar el problema de dar garantías de privacidad manteniendo niveles aceptables de utilidad predictiva en el contexto de datos secuenciales sensibles. Para abordar el tema, distinguimos dos escenarios de interés práctico. El primer escenario considera el caso de una organización que desea disponibilizar públicamente o a un proveedor de servicios de inteligencia artificial un conjunto de datos. El segundo escenario analiza la situación en la cual varias organizaciones (públicas y/o privadas) comparten entre ellas y/o con terceros modelos entrenados con datos privados de cada una de ellas.

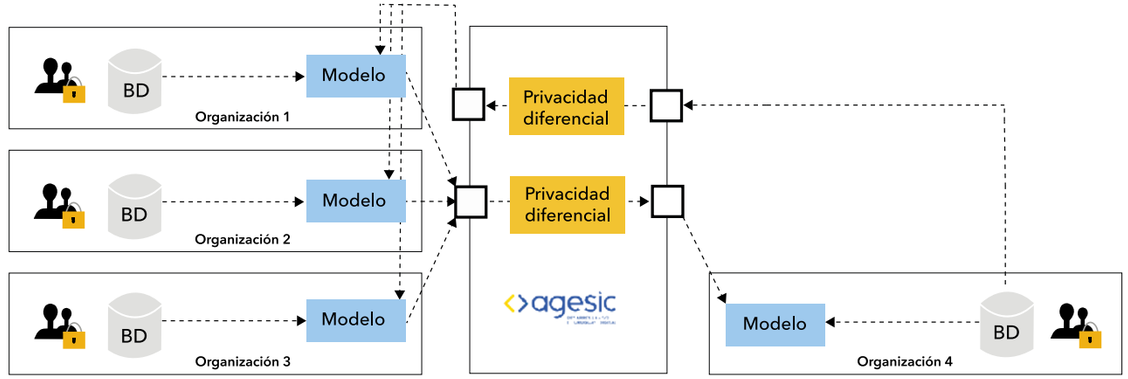
A fin de formalizar la noción de "garantía de privacidad" nos cerniremos al enfoque denominado "privacidad diferencial" [11]. Este formalismo propone un marco matemático general que consiste en cuantificar la pérdida de privacidad como una variable aleatoria. El objetivo de este marco es permitir el diseño de mecanismos específicos que proporcionan privacidad limitando de manera matemáticamente demostrable el valor de esa variable por una cantidad deseada y con una confianza dada. La cuantificación de la pérdida de privacidad es una de las características salientes y distintivas de este marco dado que permite comparar mecanismos y controlar la pérdida de privacidad acumulada por la composición de varios de ellos. Otra propiedad fundamental de la privacidad diferencial es la inmunidad al post-procesamiento. Esto es, un dato obtenido de un mecanismo diferencialmente privado no puede ser procesado a fin de obtener mayor pérdida de privacidad. Los mecanismos de privacidad diferencial, a diferencia de otros, proveen protección contra riesgos arbitrarios.

En el primer escenario el enfoque tratado consiste en generar una nueva base de datos sintética construida a partir de la aplicación de un mecanismo de privacidad diferencial a los datos originales. Esto implica generar nuevas secuencias en sustitución de las originales. Para esto, dado que los datos son secuencias, el enfoque privilegiado consistió en el entrenamiento de modelos de deep learning conocidos como redes neuronales recurrentes (RNN) [12]. Específicamente se entrenó una RNN generativa conjuntamente con un modelo predictivo dedicado a la tarea específica, por ejemplo, una clasificación binaria (normal vs anormal) de las secuencias. En este método se observó que la exactitud del aprendizaje realizado sobre los datos sintéticos aumenta a medida que aumenta el límite de privacidad diferencial. Los resultados observados experimentalmente fueron interesantes ya que los

modelos obtenidos aprendieron a generar secuencias normales y anormales que son muy diferentes a las originales pero manteniendo altos niveles de utilidad predictiva [13].

Para el segundo escenario, propusimos una solución que consiste en integrar dos mecanismos de privacidad diferencial a través de dos "curadores confiables" [14]. El rol del primer curador es proteger los datos privados de entrenamiento de las organizaciones que comparten sus modelos. Para conseguir este objetivo, el curador construye un "ensamble" con los modelos de las organizaciones participantes, aplicando un mecanismo llamado PATE [15]. El segundo curador aplica un mecanismo de privacidad diferencial con el propósito de proteger los datos privados de quien consulta dicho ensamble para hacer uso de sus predicciones. Esta técnica fue implementada y evaluada experimentalmente con datos de ciberseguridad y salud, habiéndose obtenido excelentes compromisos de privacidad y exactitud predictiva.

consultas a esos modelos a través de un mecanismo de privacidad diferencial que pone a resguardo los datos de las organizaciones. La organización 4 desea construir un modelo pero no tiene suficientes datos propios para hacerlo y no quiere compartir sus datos privados con el resto de las organizaciones. Esta organización envía sus datos a Agesic. Antes de hacer la consulta a los modelos de las otras organizaciones, Agesic aplica un mecanismo de privacidad diferencial para proteger los datos de la organización 4. De esta manera, ésta puede acceder al conocimiento de aquellas, sin que la privacidad de la información de ninguna de las organizaciones participantes sea vulnerada. Sin lugar a dudas, la implementación de este esquema requiere definir importantes aspectos de gobernanza de datos que deberán ser establecidos por todas las organizaciones participantes.



Conclusión

El análisis comparativo de ambos escenarios permitió observar que en el primero suele ser necesario aceptar altas cotas de pérdida de privacidad al momento de la generación de la base de datos sintética para luego conseguir modelos con métricas de predicción aceptables, mientras que en el segundo esto puede ser mejor controlado, en particular si el ensamble está compuesto por una cantidad importante de modelos. En consecuencia, consideramos que es razonable pensar en implementar este enfoque en la práctica. La Figura 1 muestra un despliegue esquemático de esta solución en Uruguay en la cual Agesic actúa como curador confiable en ambos sentidos. Las organizaciones 1 a 3 (organismos públicos, mutualistas, etc.) ponen a disposición de Agesic sus modelos entrenados con datos privados. Agesic disponibiliza

Agradecimientos

Este trabajo ha sido parcialmente financiado por ICT4V—Information and Communication Technologies for Verticals (POS_ICT4V_2016_1_15) y ANII—Agencia Nacional de Investigación e Innovación (FSDA_1_2018_1_154419 y FMV_1_2019_1_155913).

Referencias bibliográficas

[1] M. J. Iqbal, Z. Javed, H. Sadia et al. Clinical applications of artificial intelligence and *machine learning* in cancer diagnosis: Looking into the future. Cancer Cell Int 21, 270 (2021).

[2] A. Thakkar, R. Lohiya. A survey on intrusion detection system: feature selection, model, performance measures, application perspective, challenges, and future

research directions. Artif Intell Rev 55, 453–563 (2022).

[3] E. Ruijer, F. Détienne, M. Baker, J. Groff, A. Meijer. The Politics of Open Government Data: Understanding Organizational Responses to Pressure for More Transparency. Am. Rev. Public Adm. 2020, 50, 260–274.

[4] N. Gruschka, V. Mavroeidis, K. Vishi, M. Jensen. Privacy Issues and Data Protection in Big Data: A Case Study Analysis under GDPR. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 5027–5033.

[5] L. Rocher, J. Hendrickx, Y. de Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. Nat. Commun. 2019, 10, 3069.

[6] M. Fredrikson, S. Jha, T. Ristenpart. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, 12–16 October 2015.

[7] B. Chen, D. Kifer, K. LeFevre, A. Machanavajjhala. Privacy-Preserving Data Publishing. Found. Trends Databases 2009, 2, 1–167.

[8] C. Yin, Y. Zhu, J. Fei, X. He. A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks. IEEE Access 2017, 5, 21954–21961.

[9] Moody, G.; Mark, R. The impact of the MIT-BIH Arrhythmia Database. IEEE Eng. Med. Biol. Mag. 2001, 20, 45–50.

[10] S. Abdelhameed, S. Moussa, M. Khalifa. Privacy-preserving tabular data publishing: A comprehensive evaluation from web to cloud. Computers & Security 2018, 72, 74-95.

[11] C. Dwork, A. Roth. The Algorithmic Foundations of Differential Privacy. Found. Trends Theor. Comput. Sci. 2014, 9, 211–407.

[12] I. Goodfellow, Y. Bengio, A. Courville. Deep Learning. MIT Press. 2016.

[13] R. Visca. Estudio de modelos de privacidad de datos. Tesis de Maestría, Universidad ORT Uruguay, 2021.

[14] S. Yovine, F. Mayr, S. Sosa, R. Visca. An Assessment of the Application of Private Aggregation of Ensemble Models to Sensible Data. Mach. Learn. Knowl. Extr. 2021, 3, 788-801

[15] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, K. Talwar. Semi-supervised knowledge transfer for deep learning from private training data. arXiv 2016, arXiv:1610.05755.

Ingeniería y diseño: compromiso por la evolución sostenible

PAUTAS METODOLÓGICAS PARA EL CAMBIO

Imagen: Unsplash.com