AGENCIA NACIONAL
DE INVESTIGACIÓN
E INNOVACIÓN

# Informe final publicable de proyecto
# Herramienta de privatización de datos secuenciales para su aplicación en la detección de anomalías colectivas en ciberseguridad

**Código de proyecto ANII: FMV_1_2019_1_155913**

**23/12/2022**

**Resumen del proyecto**

Los beneficios globales de los sistemas de inteligencia artificial, como por ejemplo aquellos construidos para la de detección de ciberataques y de enfermedades, deberían superar con creces los riesgos individuales previsibles. Claramente, estos sistemas pueden contribuir a reducir la ciberdelincuencia y a mejorar la salud pero de un modo que puede atentar contra la privacidad de los individuos. Esto se debe a que el entrenamiento de estas inteligencias artificiales requiere la inspección de datos que contienen información sensible para las personas, como claves de acceso, números de tarjetas, resultados de análisis clínicos, etc. Por otro lado, los algoritmos de aprendizaje necesitan cantidades considerables de datos de entrenamiento con la finalidad de alcanzar niveles satisfactorios de eficacia. Sin embargo, tal volumen de datos no suele estar al alcance de una única organización, por lo que en muchos casos, notablemente en el ámbito de la salud, resulta necesario que múltiples organizaciones (públicas y/o privadas) compartan sus datos y los modelos predictivos entrenados con ellos, en pos de conseguir de manera conjunta beneficios sustanciales para la sociedad. Por esta razón, el objetivo de este proyecto es proponer una solución que permita a organizaciones compartir datos y modelos garantizando la privacidad de la información sensible a la vez que se preserva el valor de los datos para construir inteligencias artificiales beneficiosas. El resultado del proyecto fue el diseño, implementación y validación de una plataforma colaborativa que reposa sobre la integración de dos mecanismos de privacidad diferencial. El primero permite la disponibilización de modelos mediante la construcción de un ensemble que protege los datos de entrenamiento. El segundo permite consultar el ensemble protegiendo los datos de quien efectúa la consulta. La evaluación experimental de la solución propuesta con datos de dos áreas de aplicación, concretamente ciberseguridad y salud, fue exitosa.

**Ciencias Naturales y Exactas / Ciencias de la Computación e Información / Ciencias de la Computación / Inteligencia artificial aplicada a cibersegur**

**Palabras clave: aprendizaje profundo / redes generativas antagónicas / privacidad de la información /**

## Introducción

El acceso a la información pública juega un rol fundamental para desarrollar sociedades más transparentes e inclusivas. 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 Nº 18.381), cuya finalidad es fomentar y prescribir la disponibilización de los datos producidos, obtenidos, en poder y/o bajo control de organismos públicos.

Al mismo tiempo, los datos disponibilizados están sujetos al cumplimiento de la legislación vigente sobre protección de la privacidad de los datos. En el caso de Uruguay, estos aspectos están contemplados en la Ley Nº 18.331 Protección de Datos Personales y acción de Habeas Data. En particular, esta ley define expresamente el proceso de disociación como todo tratamiento de datos personales de manera que la información obtenida no pueda vincularse a una persona determinada o determinable.

Además, la publicación de datos de cualquier organización de Uruguay, pública o privada, podría también estar alcanzada por el marco normativo de la GDPR (General Data Protection Regulation) [1] que define el conjunto de regulaciones de protección de datos que rige sobre todas las organizaciones que operan en la Unión Europea, independientemente de donde estén ubicadas, en la medida que se trate de datos personales de ciudadanos o residentes de la Unión Europea.

Por otro lado, el acceso a los datos no está solo motivado por la legislación sobre datos abiertos, sino también por la necesidad de hacerlos disponibles a actores determinados, públicos o privados, con la capacidad técnica para analizarlos y/o para usarlos como insumo esencial en proyectos de investigación e innovación científico-tecnológica.

Brindar acceso a datos no es solo obligación de los entes estatales sino también una necesidad para cualquier organización, debido a que estos se están transformando en su activo más valioso [2,3,4]. De hecho, los grandes volúmenes de datos disponibles, junto con el aumento de la capacidad de cómputo, han habilitado la investigación y el desarrollo de algoritmos de machine learning que aprenden de los datos con la finalidad de construir modelos predictivos claves para la toma de decisiones inteligentes. Sin embargo, a pesar de que los datos constituyen un activo evidente para las organizaciones, estas enfrentan un problema mayor al intentar extraer valor de los mismos, y es que los datos no son fáciles de publicar o transferir, en la medida que, en muchos casos, contienen información de índole personal y privada propia o de terceros [5].

Existe entonces una tensión entre la necesidad de habilitar el acceso a datos y la privacidad de los mismos [6]. Por consiguiente, es esencial brindar mecanismos que permitan la protección de la privacidad de datos disponibilizados para

garantizar el cumplimiento de la legislación en cuanto a la privacidad de la información confidencial, pero al mismo tiempo, es necesario que los datos privatizados contengan información útil para llevar adelante con éxito las tareas de análisis a las que están destinados [7].
Por otro lado, los algoritmos de aprendizaje necesitan cantidades considerables de datos de entrenamiento con la finalidad de alcanzar niveles satisfactorios de eficacia. Sin embargo, tal volumen de datos no suele estar al alcance de una única organización, por lo que en muchos casos, notablemente en el ámbito de la salud, resulta necesario que múltiples organizaciones (públicas y privadas) compartan sus datos y los modelos predictivos entrenados con ellos, en pos de conseguir de manera conjunta beneficios sustanciales para la sociedad.
Para ilustrar una situación realista, consideremos el siguiente escenario. Una mutualista médica, que tiene guardadas en su base de datos las historias clínicas de sus asociados, decide entrenar un algoritmo de aprendizaje para detectar una determinada patología a partir de los resultados de análisis clínicos (hemogramas, biopsias, etc.). El desarrollo del sistema es encargado a una empresa experta en inteligencia artificial. Para que esta pueda realizar su trabajo debe consultar la base de datos. Es evidente que estas consultas son susceptibles de revelar información sensible, atentando contra la privacidad de los pacientes. Por esto se deben tomar medidas de protección que deben ir más allá del uso de técnicas de anonimización tradicionales basadas en la eliminación de datos de identificación personal. Esto se debe a que estas no protegen eficazmente contra la re-identificación [8]. Por otro lado, la privacidad no solo se puede vulnerar a través de la publicación de datos. En efecto, la información de carácter sensible puede ser revelada por medio de las consultas a los modelos entrenados con los datos [9].
En suma, queda claro que es esencial contar con mecanismos para proteger adecuadamente la información privada contenida en los datos que se ponen a disposición de terceros, ya sea directamente mediante mecanismos de disponibilización pública de datos o indirectamente a través de artefactos de inteligencia artificial entrenados con esos datos. Dichos mecanismos deben ser capaces de conservar suficiente información útil para que los sistemas de inteligencia artificial que los necesitan puedan cumplir satisfactoriamente su cometido.
Por esta razón, el objetivo de este proyecto es proponer una solución que permita a organizaciones compartir datos y modelos garantizando la privacidad de la información sensible a la vez que se preserva el valor de los datos para construir inteligencias artificiales beneficiosas.

**Metodología/diseño del estudio**

Este proyecto se enfocó en un escenario en el que varias organizaciones comparten datos y sistemas inteligentes entrenados con datos privados de cada una de ellas. El objetivo que se persiguió fue diseñar, implementar y validar experimentalmente una plataforma de software colaborativa que proveyera mecanismos adecuados de privacidad y permitiera obtener métricas satisfactorias de utilidad. La evaluación empírica se realizó en el ámbito de dos áreas de aplicación en las que privacidad y eficacia predictiva son vitales: ciberseguridad y salud.
El trabajo se organizó alrededor de tres líneas de investigación directoras, a saber: el estudio de mecanismos de privacidad aplicables al escenario definido, el desarrollo de software vinculado con la implementación de los mecanismos estudiados y con la generación de casos de prueba, y la validación experimental.
Para instanciar cuantitativamente el concepto de "garantía de privacidad", el proyecto adoptó el enfoque denominado "privacidad diferencial" (DP) [10]. DP es un marco probabilístico general que provee una definición abstracta de "mecanismo de privacidad" basada en la cuantificación de la "pérdida de privacidad" como una variable aleatoria. Todo mecanismo de DP tiene dos propiedades fundamentales, a saber, la imposibilidad de la re-identificación y la inmunidad al post-procesamiento. Además, una característica distintiva del enfoque DP es que permite componer y comparar mecanismos en función de dos parámetros centrales: la cota máxima de pérdida de privacidad deseada (epsilon) y la confianza estadística en su cumplimiento (delta). La puesta en práctica de DP consiste en diseñar e implementar mecanismos específicos para casos de uso concretos.
Durante el desarrollo del proyecto se estudiaron diferentes mecanismos de DP que condujeran a definir procesos y herramientas que permitieran la liberación de los datos y la realización de consultas a través de modelos de inteligencia artificial garantizando privacidad [11]. Para cada mecanismo se analizó la pérdida de información y la utilidad predictiva. Atendiendo al escenario planteado, se investigaron dos tipos de mecanismos. El primer tipo, denominado "local", consiste en generar nuevos datos a partir del agregado de ruido aleatorio a los datos originales. El segundo, denominado "centralizado", consiste en proteger las consultas realizadas a los modelos de aprendizaje automático lo que conduce in fine a garantizar la privacidad de los datos utilizados para entrenar dichos modelos.
En una primera etapa, los diferentes mecanismos de DP estudiados fueron implementados con el objetivo de estudiar el trade-off entre privacidad y eficacia predictiva. La complejidad de la validación experimental de los mecanismos analizados motivó el diseño y la implementación de la herramienta de software DP-GEM con el objetivo de automatizar

dicho proceso. Paralelamente, se implementó una batería de herramientas para la generación de datos de prueba, en particular logs de ataques y requests normales a servicios web. Durante la segunda etapa se realizó un estudio detallado de los aspectos arquitecturales y de gobernanza del mecanismo seleccionado que condujo a dos propuestas de implementación concretas.

## Resultados, análisis y discusión

La primera etapa del trabajo de investigación consistió en el estudio de diferentes mecanismos de DP con la finalidad de estudiar la relación entre capacidad predictiva y privacidad [11]. Esta investigación se apoyó en el análisis comparativo de diversos modelos de redes neuronales [11,12,13, 14]. Los resultados observados experimentalmente condujeron a concluir que el mecanismo más efectivo, es decir el que tenía mejor trade-off entre utilidad y privacidad, era el basado en el enfoque denominado "Private Aggregation of Teacher Ensembles" (PATE) [15,16]. El objetivo de PATE es garantizar la privacidad de los datos de entrenamiento de los modelos que conforman el ensemble. Sin embargo, PATE no ofrece ninguna forma de proteger los datos enviados en las consultas al ensemble. Esto es así dado que PATE asume que estos datos son públicos. Claramente, esto no aplica en nuestro escenario donde todos los datos son privados. En consecuencia, para resolver este problema, propusimos una solución que compone dos mecanismos. El primero está basado en PATE y su rol es permitir la disponibilización pública de dispositivos de inteligencia artificial a través de un ensemble. El segundo consiste en la aplicación de un mecanismo de DP a los datos enviados en las consultas. Este esquema teórico fue analizado y validado experimentalmente de manera exitosa sobre datos provenientes de dos ámbitos de aplicación de interés: ciberseguridad y salud [17].

A partir de ese trabajo, se propuso el siguiente escenario de despliegue "tipo". Las organizaciones O1 a On (organismos públicos, mutualistas, etc.) ponen sus modelos de inteligencia artificial entrenados con sus datos privados a disposición de Agesic, quien juega el rol de "curador confiable". Agesic disponibiliza como servicio las consultas a esos modelos a través del mecanismo de ensemble con DP que pone a resguardo de terceros no confiables los datos sensibles de las N organizaciones participantes. La organización S que desea usar el servicio, por ejemplo para realizar un diagnóstico médico puntual o para entrenar su propio modelo predictivo, envía sus datos a Agesic, en quien confía. Antes de hacer la consulta al ensemble de modelos de las otras organizaciones O1 a On, Agesic aplica el mecanismo local de DP para proteger los datos de S. De esta manera, la privacidad de la información de todas de las organizaciones participantes es garantizada.

Sobre la base de esta propuesta, se realizaron estudios preliminares de plataformas de software [18] y de gobernanza [19] que condujeron a definir una arquitectura de software de referencia. Estos trabajos permitieron identificar un punto mejorable del esquema de despliegue propuesto: el hecho que los participantes deban confiar en el curador (prestador del servicio) enviándole sus datos no privatizados, ya sea las predicciones de los modelos o los datos para la consulta, es en efecto un punto débil que puede conducir a la vulnerabilidad del sistema. Para remediar este inconveniente, se hicieron dos mejoras al esquema inicial de [17]. La primera consiste en que los datos sean enviados ya privatizados por la organización S al prestador del servicio. La segunda es que las respuestas de los modelos participantes del ensemble sean enviados bajo encriptado homomórfico (EH), permitiendo la realización de las operaciones matemáticas requeridas por PATE sin la necesidad de conocer los valores de las respuestas. En definitiva, por medio de este esquema que combina DP con EH es posible garantizar la privacidad de los datos de forma "end-to-end", sin el requerimiento de tener que confiar en ninguno de los actores participantes. Este trabajo de análisis arquitectural y de gobernanza nos llevó a diseñar y desarrollar dos implementaciones concretas basadas en código abierto [20, 21].

## Conclusiones y recomendaciones

El trabajo llevado adelante en el marco de este proyecto produjo resultados a nivel de investigación y de desarrollo.

En términos de investigación, podemos decir que se validó la tesis de que es factible construir inteligencias artificiales útiles sin vulnerar la privacidad de los individuos cuyos datos son utilizados en algún momento del proceso, contribuyendo así al desarrollo de inteligencias artificiales responsables [6]. Los estudios realizados revelan que es necesario identificar claramente el escenario de uso y definir apropiadamente los mecanismos adecuados, recurriendo a la integración de privacidad diferencial y encriptado homomórfico.

A nivel de desarrollo, la contribución del proyecto es doble. Por un lado, esta consistió en la propuesta de un esquema concreto para un caso de uso realista garantizando la privacidad de los datos de forma "end-to-end", en un contexto de inteligencia artificial como servicio en el que ninguno de los participantes involucrados es confiable. En segundo lugar, se desarrollaron dos implementaciones totalmente basadas en software de código abierto, que constituyen sendas pruebas de concepto de la viabilidad práctica del enfoque propuesto.

## Referencias bibliográficas


[1] EUGDPR (2017). The EU general data protection regulation. http://www.eugdpr.org/.

[2] The Economist. The world's most valuable resource is no longer oil, but data. The Economist, 2017.

[3] Rose, M. (2018). Data: Your Most Ignored And Valuable Asset. Retrieved February 2019. https://www.forbes.com/sites/forbesagencycouncil/2018/02/12/data-your-most-ignored-and-valuable-asset/

[4] C. Pettey (2017, November 30). Treating Information as an Asset. Retrieved February 21, 2019, from https://www.gartner.com/smarterwithgartner/treating-information-as-an-asset/

[5] Gruschka N, Mavroeidis V, Vishi K, Jensen M. Privacy Issues and Data Protection in Big Data: A Case Study Analysis under GDPR. In2018 IEEE International Conference on Big Data (Big Data) 2018 Dec 10 (pp. 5027-5033). IEEE.

[6] High-Level Expert Group on Artificial Intelligence. European Comission. Ethics guidelines for trustworthy AI. https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

[7] Bee-Chung Chen, Daniel Kifer, Kristen LeFevre and Ashwin Machanavajjhala (2009), "Privacy-Preserving Data Publishing", Foundations and Trends® in Databases: Vol. 2: No. 1–2, pp 1-167.

[8] Y-A. de Montjoye, L. Radaelli, V. K. Singh , A. Pentland (2018). Unique in the shopping mall: On the reidentifiability of credit card. Science 347 (6221), 536-539.

[9] Fredrikson M, Jha S, Ristenpart T. Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security 2015 Oct 12 (pp. 1322-1333). ACM.

[10] Dwork C (2006) Differential privacy. In: Proceedings of the 33rd International colloquium on automata, languages and programming (ICALP)(2), Venice, pp 1–12.

[11] Visca, R. Estudio de modelos de privacidad de datos. Tesis de Maestría, Universidad ORT Uruguay, 2021. https://hdl.handle.net/20.500.12381/463

[12] Biardo, Deborah; González, Guzmán; Lanzotti, Sabrina. Análisis y desarrollo de modelos predictivos con redes neuronales para Web Application Firewall. Tesis de Maestría, Universidad ORT Uruguay, 2020. https://hdl.handle.net/20.500.12381/461

[13] N. Martínez Varsi. Comparison of LSTM and Transformer Neural Network on multiple approaches for weblogs attack detection. Tesis de Maestría, Universidad ORT Uruguay, 2022 https://hdl.handle.net/20.500.12381/2363.

[14] S. Sosa. Application of private aggregation of teacher ensembles framework for malicious web request detection. Trabajo final de Ingeniería en Sistemas, Universidad ORT Uruguay, 2021. https://hdl.handle.net/20.500.12381/459

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

[16] Papernot, N.; Song, S.; Mironov, I.; Raghunathan, A.; Talwar, K.; Erlingsson, Ú. Scalable private learning with pate. arXiv 2018, arXiv:1802.08908.

[17] Yovine, S.; Mayr, F.; Sosa, S.; Visca, R. An Assessment of the Application of Private Aggregation of Ensemble Models to Sensible Data. Mach. Learn. Knowl. Extr. 2021, 3, 788-801.https://hdl.handle.net/20.500.12381/456

[18] P. Ampuero, J. Sánchez. Prueba de concepto del framework de OpenMined para modelos de Machine Learning. Tesis de Maestría, Universidad ORT Uruguay, 2021. https://hdl.handle.net/20.500.12381/462

[19] J. Ramas, A. Rodríguez, S. Zanotta. Implementación de las prácticas de MLOps para PATE. Tesis de Maestría, Universidad ORT Uruguay, 2022. https://hdl.handle.net/20.500.12381/2362

[20] M. Pisani, S. Yovine. Prototipo de "Application of Private Aggregation of Ensemble Models to Sensible Data" en la plataforma PySyft de OpenMined. Documento de trabajo, Universidad ORT Uruguay, 2022. https://hdl.handle.net/20.500.12381/2374

[21] W. Imbert, S. Uriarte, G. Wagner. Software basado en Pyfhel para garantizar privacidad de datos en un contexto de machine learning as a service. Documento de trabajo (informe preliminar de Tesis de Maestría), Universidad ORT Uruguay, 2022. https://hdl.handle.net/20.500.12381/2375


## Licenciamiento