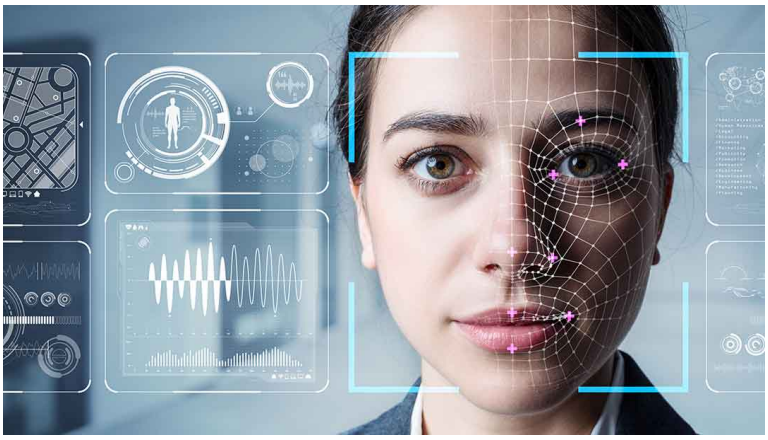
NOVEDADES

La técnica que permite saber todo sobre los clientes sin poner en riesgo su privacidad

23 set. 2021

Utilizada por empresas como Google y Apple para manejar el enorme flujo de información que les permite saber todo sobre sus clientes sin exponer datos sensibles, la privacidad diferencial fue abordada en Uruguay por la cátedra de Inteligencia Artificial y Big Data de la Facultad de Ingeniería de ORT, en una investigación que contó con el apoyo del CERTuy de Agesic, el Centro Tecnológico ICT4V y la Agencia Nacional de Investigación e Innovación.



El incremento de la disponibilidad de datos y de la potencia de cómputo, sumados a los sucesivos avances en el campo de la Inteligencia Artificial (IA),

permiten contar con mejoras significativas a la hora de resolver tareas partiendo del uso de artefactos inteligentes impulsados por algoritmos de aprendizaje automático.

Estos progresos son importantes cuando se trata, por ejemplo, de datos que hacen a aspectos críticos para las personas, como son la salud y la seguridad.

Hoy sabemos que la información y el conocimiento acumulado, resguardados por una sola organización, resultan ineficientes dado que, de ser compartidos, pueden ayudar a superar diferentes problemáticas.

De hecho, desde hace unos años diferentes ONG e institutos de investigación están impulsando la liberación mayor de datos partiendo del entendido de que si las organizaciones compartieran sus conocimientos mediante el intercambio de datos brutos o de modelos predictivos entrenados con dichos datos, el avance sería más abarcativo y eficiente.

Un posible abordaje para este problema es la denominada Privacidad Diferencial (DP), “un marco matemático general basado en la cuantificación de la pérdida de privacidad como una variable aleatoria, cuyo objetivo es permitir el diseño de mecanismos específicos que proporcionen protección de datos, limitando de forma demostrable la pérdida de privacidad por una cantidad deseada y con una confianza dada”, explica el Dr. Sergio Yovine, responsable de la cátedra de Inteligencia Artificial y Big Data de la Facultad de Ingeniería de Universidad ORT Uruguay.

Manejo común de datos y privacidad

Si bien las técnicas de anonimización de datos existen y se utilizan, el resultado no siempre es el esperado porque los peligros que representan los posibles ataques permanecen y se intensifican en todo el mundo.

Esta realidad lleva a seguir pensando en alternativas y así surge la DP, un marco de trabajo basado en garantías matemáticas que permite manejar distintos parámetros para crear una solución y validar su cumplimiento a partir de lo que exige el marco de trabajo matemático planteado. Es decir, que por detrás de la DP existe una teoría matemática que avala su funcionamiento.

Tal como indica el Ing. Ramiro Visca, integrante del equipo de investigación de la cátedra de Inteligencia Artificial y Big Data de ORT, y becario del centro tecnológico ICT4V, la clave está en que ese marco de trabajo propone una definición que fija pautas a cumplir de forma matemática y luego se deben diseñar mecanismos de privacidad que puedan demostrar que cumplen con el marco matemático establecido por lo que es la DP.

“Es algo difícil de hacer; no funciona de la noche a la mañana, porque la definición matemática es bastante engorrosa, la propuesta es compleja de entender y más difícil aún es encontrar un mecanismo que luego pueda ser demostrable y que cumpla con el marco matemático. No es una herramienta sencilla de usar”, alerta Visca.

Pese a las dificultades, la garantía matemática es la que valida y robustece la capacidad de la DP, “porque básicamente establece un parámetro llamado ε, el cual, si vale 0 indica que se están privatizando los datos al 100 %. En un conjunto de datos, si yo le hago una consulta, al tener esa ε en 0 está garantizado que no estoy filtrando información. Es decir, que tengo la utopía del mecanismo perfecto que me permite liberar el conjunto de datos sin preocuparme de que se viole la privacidad”, afirma el investigador.

Para Visca, más allá de los individuos a los que pertenece la información, para las organizaciones el rol de los datos “en el mejor escenario pasa por asistirlas en la toma de decisiones, o les permite encontrar una ventaja competitiva de cara a la innovación”.

“Creo que las organizaciones no son conscientes de la necesidad de administrar la privacidad de datos, más allá de que haya leyes. La gran mayoría de las organizaciones no piensa en la privacidad cuando está diseñando sus sistemas. De lo contrario no estaríamos hablando de este tema, porque se diseñaría todo alrededor de esta idea y básicamente lo que estamos haciendo ahora es tratar de emparchar el problema con mecanismos de privacidad posrecaudación de datos”, subraya Visca.

“Dentro de lo que se considera DP, existen dos clases: una es el modelo centralizado o global que trata de proteger toda una base de datos, y después está el modelo local que dice que a medida que se va armando esa base tengo

una consideración previa donde pongo mi filtro de privacidad. Entonces, cuando armo esa base de datos, ya sé que cumple con la definición de DP", agrega.

Pros y contras

Para el Ing. Eduardo Giménez, integrante del Equipo de Operaciones de ICT4V, la DP es un marco teórico que provee "una respuesta precisa y elegante" a la pregunta de qué es la privacidad de los datos.

"Acierta en ver ese tema como una propiedad de los programas, y no de los datos como se hace habitualmente. Persigue metas ambiciosas, como ser resistente a cualquier tipo de ataques sobre la privacidad", señala.

Sin embargo, explica Giménez, sin ser un experto en esa técnica, algunos aspectos de ella hacen difícil su utilización en la práctica. Uno de ellos es que introduce una noción de "presupuesto de privacidad", cuyo valor adecuado es muy difícil de estimar a priori. Cuando se realiza un experimento de ciencia de datos, no se sabe del todo lo que busca a priori y es necesario realizar muchas pruebas y ajustes antes de dar encontrar el modelo predictivo que necesita o ajustarlo al punto que se desea. En ese camino, es muy posible que agote su "presupuesto de privacidad".

Otra dificultad de la DP es que no permite combinar libremente desarrollos para obtener otros, sino que hay que ser cuidadoso en la manera en que estas combinaciones se realizan, para no perder las propiedades que se esperan.

Todo esto requiere técnicos especializados que conozcan muy bien la técnica de DP y sus definiciones, los cuales no abundan en el mercado.

"Veo más posible que esa técnica sea utilizada como parte de un servicio que la integre y empaquete. Por ejemplo, en PedidosYa el datalake que se utiliza para operaciones de inteligencia de negocios está en Google Cloud. Quizás hay una chance de que la DP sea utilizada si Google lo integra como parte de Big Query", concluye Giménez.

Marco legal de avanzada

En nuestro país, el manejo de datos personales está legislado por la Ley N.° 18.331, que reconoce el derecho de las personas a controlar el uso que se hace de la información propia registrada en cualquier soporte que permita tratarla y usarla posteriormente de diversos modos, tanto en el ámbito privado como público.

Esta ley fue actualizada en 2018 por los artículos 37 a 40 de la Ley N.° 19.670,

que introdujeron el principio de responsabilidad proactiva, la figura del delegado de protección de datos (Agesic), las notificaciones de vulneraciones de seguridad (CERTuy) y la ampliación del ámbito territorial.

Posteriormente, en diciembre del 2020, los artículos 86 y 87 de la Ley N.° 19.924 incluyeron una definición del dato biométrico y condiciones particulares para su tratamiento, tal como consignó Agesic.

El trabajo orientado a la protección de datos comenzó en 2018 con el desarrollo de la llamada Estrategia y Política Nacional de Datos, la cual sentó las bases a seguir por los organismos públicos y para la formulación de políticas públicas, a la vez que estableció los principios en el desarrollo de proyectos para el uso intensivo de datos, que incluyen expresamente la protección de datos personales y la seguridad.

ta con un marco normativo de protección de datos personales pot etros a los que maneja la Unión Europea.

El Ing. Javier Barreiro, gerente de IUGO Software & Design Studio y exdirector de Tecnología de Agesic, donde participó en la elaboración de las estrategias de implementación de IA en Uruguay, entiende que uno de los aspectos a destacar de la política implementada es que desde el vamos siguió un proceso de consulta pública, con un primer borrador elaborado por un equipo multidisciplinario.

Este equipo realizó su análisis en experiencias de otros países referentes en el tema, como Canadá e Italia, para luego adaptarlo a la realidad y normativa local.

“La consulta pública permitió obtener feedback de la sociedad civil, de organizaciones internacionales, organismos públicos y ciudadanos. Cada uno de los comentarios fue analizado, incorporándose a la estrategia o dando una devolución al respecto a quien lo realizó”, afirma.

Según Barreiro, desde el primer momento la estrategia trazada contempló la necesidad de preservar los datos, de modo que esta perspectiva fue plasmada en los principios generales y los distintos objetivos.

El perfil multidisciplinario del equipo de trabajo fue clave, asegura. “A nivel normativo, Uruguay cuenta con una legislación de avanzada, alineada con los estándares europeos a nivel de protección de datos. El mayor desafío seguramente esté en cómo hacer efectiva dicha regulación, tanto para asegurar su cumplimiento como para generar capacidades en los stakeholders para su consideración”, puntualiza.

Desde Agesic se destaca que Uruguay es el país con “la madurez más alta de la región en cuatro de las cinco dimensiones medidas”, según la segunda edición del “Reporte Ciberseguridad 2020: riesgos, avances y el camino a seguir en América Latina y el Caribe”, elaborado por el BID y la OEA.

Asimismo, el país se ubica entre los primeros puestos en las Américas según el Global Cybersecurity Index (GCI), elaborado por la International Telecommunication Union (ITU), tercero en 2018 detrás de Estados Unidos y Canadá.

Recursos humanos calificados

Uno de los mayores desafíos que se enfrentan a nivel mundial es la falta de profesionales y técnicos calificados en ciberseguridad.

En este sentido, Agesic está trabajando con distintos actores para apoyar el desarrollo de currículos de ciberseguridad en los diferentes centros educativos del país, así como para promover el desarrollo de una red de expertos a nivel nacional que incorpore a todos los involucrados.

yores desafíos que se enfrentan a nivel mundial es la falta de prof :ados en ciberseguridad.

En relación a la DP, desde Agesic se entiende que una de sus ventajas consiste en el hecho de que los conjuntos de datos se entregan a terceros autorizados como respuesta a una consulta concreta y no simplemente como consecuencia de la publicación de un único conjunto de datos.

Desde el punto de vista de la protección de datos, la mayor dificultad que existe es poseer la capacidad de generar la cantidad adecuada de ruido (es necesario hacer bastante ruido, ya que es un error frecuente no hacerlo), que se añade a las respuestas verdaderas a fin de proteger la privacidad de las personas y, al mismo tiempo, preservar la utilidad de las respuestas difundidas.

Además, es necesario tener cuidado de no caer en el error de pensar que los datos son anónimos para el tercero cuando el responsable del tratamiento todavía puede identificar al interesado en la base de datos original mediante el conjunto de medios que pueden ser razonablemente utilizados.

Los responsables de Agesic también señalan que esta técnica, junto con otras, forma parte de las determinadas por la Unidad Reguladora y de Control de Datos Personales en la Resolución N.° 68/017, que aprobó el documento "Criterios de Disociación de Datos Personales". Este documento fue elaborado en función a lo dispuesto por el Decreto N.° 54/017, reglamentario del artículo 82 de la Ley N.° 19.355, en el que se establece que las Entidades Públicas, sujetos obligados por la Ley N.° 18.381, N° 232/010, en formato de dato abierto.

Datos abiertos a la innovación

Volviendo a la privacidad y seguridad de los datos, desde el punto de vista de los investigadores y profesionales, la exigencia de compartirlos no solo está motivada por la legislación sino también por la necesidad de dejarlos a disposición de los actores públicos y privados con la suficiente capacidad técnica de utilizarlos con fines de innovación científico-tecnológica.

Para la Lic. Sabrina Lanzotti, una de las primeras egresadas del Master en Big

Data de Universidad ORT Uruguay y hoy jefa de Equipo Técnico DevSecOps de Atos, la gestión de la seguridad de la información es fundamental en un contexto como el actual, en el que muchas organizaciones reconocen el valor de una gestión adecuada de seguridad de la información.

Pese a este reconocimiento, el comportamiento es principalmente reactivo, porque “algunas organizaciones no han alcanzado la madurez en sus sistemas de gestión pese a haber implementado medidas técnicas referentes a la protección de sus activos”, asegura.

Señala también que ciertos marcos ampliamente utilizados en la industria son de la familia de normas ISO 27.000, NIST CyberSecurity Framework, Marco de Ciberseguridad de Agesic, entre otros, y todos ellos “se basan en conocer qué activos tengo, qué valor tienen para nuestra organización, cómo estamos expuestos, la implementación de medidas para mitigar el riesgo y la mejora continua”.

En este contexto, Lanzotti considera que la DP es una buena técnica para poder saber todo sobre los clientes sin ponerlos en riesgo.

“Permite conocer mucha información sobre un grupo pero no sobre sus individuos. En un primer intento por poder procesar grandes volúmenes de datos personales y que los individuos no sean reconocidos, se podrían eliminar datos que en primera instancia parecerían solucionar el problema. Por ejemplo, nombre, cédula, dirección, etc. pero técnicas de reidentificación han hecho que esto no sea suficiente. La privacidad diferencial asegura matemáticamente que esto no suceda, agregando aleatoriedad y ruido. Empresas como Google o Apple ya lo usan en sus productos”, explica Lanzotti.

Empresas locales

Empresas locales

Desde su experiencia como responsable de Seguridad en **PedidosYa**, Eduardo Giménez expresa que todas las soluciones que se toman en esta materia deben considerar el hecho de que año a año la empresa crece a un ritmo exponencial, no en sentido metafórico sino literal: cada año se duplica la cantidad de pedidos del año anterior, y esto ocurre desde hace ya varios años.

Esta realidad obliga a priorizar las soluciones automatizadas antes que manuales, dado que estas no escalan correctamente. “Al mismo tiempo, no podemos desplegar soluciones de seguridad que frenen ese avance. Tenemos que buscar siempre compromisos correctamente balanceados entre el riesgo y el crecimiento del negocio”, puntualiza Giménez.

Por este motivo, de momento la empresa no utiliza técnicas de DP. “Creo que en PedidosYa estamos muy lejos de hacerlo; nuestros desafíos en torno a la privacidad por el momento son mucho más básicos: cómo cumplir adecuadamente con la reglamentación, cómo asegurarnos cuando un usuario pide la baja de nuestros sistemas que es dado de baja de ‘todos’ los sistemas, cómo anonimizar sus datos sabiendo que tenemos que conservar otros relativos a sus compras por razones de negocio, cómo hacer que las técnicas de anonimización que usamos no puedan ser explotadas por atacantes que buscan borrar sus trazas luego de cometer un fraude, cómo limitar el acceso de los desarrolladores a los datos de los usuarios a la hora de dar soporte para resolver un bug”.

Para la Ing. Soledad Rivas, responsable de Aprendizaje Automático en **Tryolabs**, cuando se trabaja con conjuntos de datos que contienen información sensible “se debe ser consciente de que preservar la privacidad es tarea de quien está implementado la solución. Además, se deben tener en cuenta las consecuencias legales y éticas que podrían existir en caso de que

no se tengan las garantías de privacidad".

Rivas afirma que si bien en los últimos años se ha trabajado sobre el tema a nivel país, "la pandemia nos demostró que aún nos falta un largo camino por recorrer; es por eso que creemos fundamental que se continúe investigando sobre el problema para lograr mejores resultados en el futuro".

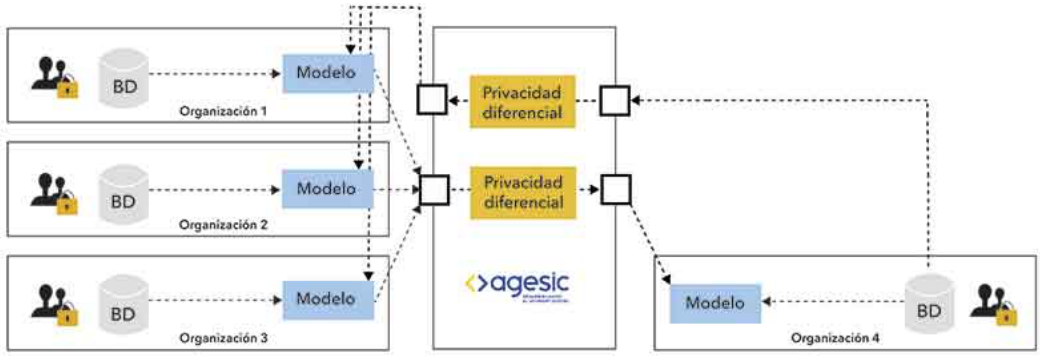
Esta situación motivó a Sebastián Sosa, Machine Learning Engineer en Tryolabs, a realizar su tesis de grado de Ingeniería en Sistemas de ORT, bajo la dirección de Ramiro Visca y Sergio Yovine. "Decidí comenzar una tesis de investigación para entender cómo aplicar técnicas que permitan entrenar modelos de aprendizaje automático a partir de datos sensibles con garantías de privacidad robustas", señala.

Las investigaciones

Las investigaciones realizadas en el ámbito de la cátedra de Inteligencia Artificial y Big Data de la Facultad de Ingeniería de ORT, consisten en explorar una técnica de entrenamiento de modelos de aprendizaje automático llamada PATE: Private Aggregation of Teacher Ensembles, para entender cómo se puede aplicar a un contexto distinto al cual plantean los autores originales.

La técnica PATE permite entrenar un modelo a partir de un conjunto de datos sensibles, y obtener garantías de privacidad robustas, expresadas en términos de DP.

De acuerdo con Sosa, "el aporte de la investigación fue extender la técnica PATE para demostrar que es posible entrenar al modelo en contextos donde sus datos de entrenamiento sean sensibles, en vez de públicos. Ello permite trabajar con esta técnica en contextos donde antes no era posible. En nuestro caso en particular, queríamos entrenar modelos predictivos a partir de datos sensibles que pertenecían a una organización gubernamental, con garantías rigurosas de que no se comprometería la privacidad de estos datos".



Ejemplo de aplicación esquemática de la solución propuesta.

Las organizaciones 1 a 3 (organismos públicos, mutualistas, etc.) ponen sus modelos de IA entrenados con sus datos a disposición de Agesic, en quien confían.

Agesic disponibiliza consultas a esos modelos a través de un mecanismo de DP que pone a resguardo los datos de las organizaciones 1 a 3. La organización 4 desea construir su IA 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, en quien sí confía. Antes de hacer la consulta a los modelos de IA de las otras organizaciones, Agesic aplica un mecanismo de DP para proteger los datos de la organización 4. De esta manera, puede acceder al conocimiento de aquellas, sin que la privacidad de la información de ninguna de las organizaciones participantes sea vulnerada.

“Lo que propusimos es un modelo de privacidad mixto. En lo que respecta a las organizaciones 1 a 3, esta técnica proporciona un modelo centralizado, mientras que para la organización 4 se trata de un modelo local”, explica el Mag. Franz Mayr, estudiante de doctorado e investigador de la cátedra de Inteligencia Artificial y Big Data de la Facultad de Ingeniería de ORT.

“Los resultados experimentales en aplicaciones de ciberseguridad (detección de ataques en weblogs) y salud (identificación de cardiopatías en electrocardiogramas) fueron muy positivos”, concluye Yovine.