TechDispatch

APRENDIZAJE FEDERADO — FEDERATED LEARNING

EDPS
European Data Protection Supervisor

aepd agencia española protección datos

Luxembourg: Publications Office of the European Union, 2025

PDF ISBN 978-92-9242-926-3 ISSN 2599-932X doi:10.2804/1422559 QT-01-24-000-ES-N

CONTENTS

1. Resumen ejecutivo

El aprendizaje federado (Federated Learning, FL de aquí en adelante) presenta un enfoque prometedor para el aprendizaje automático (Machine Learning, ML de aquí en adelante) al permitir que múltiples fuentes de datos (dispositivos o entidades) entrenen de forma colaborativa un modelo de inteligencia artificial compartido mientras se mantienen los datos descentralizados. Este enfoque mitiga riesgos de privacidad, ya que los datos sin procesar permanecen localmente en las fuentes, lo que es particularmente beneficioso en escenarios donde la confidencialidad de los datos o los requisitos regulatorios hacen que la centralización de datos[1] sea una solución poco práctica. Las aplicaciones del FL son diversas y abarcan recomendaciones personalizadas, análisis de datos sanitarios, espacios de datos y sistemas de transporte autónomos, donde garantizar la privacidad y la protección de datos es primordial.

Desde el punto de vista de la protección de datos personales, el FL ofrece importantes ventajas al minimizar el intercambio de datos. Este enfoque descentralizado se alinea con los principios básicos de la protección de datos, como la minimización de datos y la limitación de la finalidad, al garantizar que los datos personales permanecen bajo el control del responsable y no están expuestos a agentes externos. Además, el FL mejora la responsabilidad proactiva y la auditabilidad, ya que los responsables tienen una visión más clara de cómo se tratan sus datos personales. Además, al mantener los datos originales en dispositivos/servidores locales y compartir solo modelos o actualizaciones de modelos (gradientes o pesos), el FL puede mejorar la confidencialidad de los datos personales, limitando la necesidad de su centralización y reduciendo el impacto de las brechas de datos a gran escala.

A pesar de sus ventajas, EL FL presenta algunos desafíos que aún no se han resuelto por completo, uno de los principales es la posibilidad de fuga de datos a través de las actualizaciones de los modelos, ya que incluso sin acceso directo a los datos originales, un atacante podría inferir información confidencial mediante el análisis de los gradientes o pesos compartidos entre los dispositivos (y el servidor central en los casos en los que existe). Esta vulnerabilidad abre la puerta a los ataques de inferencia de membresía, en los que los adversarios pueden determinar si puntos de datos específicos formaban parte del conjunto de entrenamiento. Adicionalmente, la seguridad debe implementarse en todo el ecosistema porque si no, los atacantes tendrían la oportunidad de atacar el eslabón más débil y luego comprometer todo el sistema. Además, el FL debe poner en marcha medidas específicas de garantía de calidad de los datos de entrenamiento distribuido, y estar libre de sesgos, cuando los datos se tratan para un fin previsto. En comparación con las arquitecturas que no son

1 es decir, centralizar los datos en un único lugar

FL, el FL tiene vectores de amenaza distintos[2] que pueden afectar a la integridad de los datos, por lo que se deben contemplar e implementar las medidas adecuadas.

No se debe suponer que los datos intercambiados entre los dispositivos cliente y los modelos de ML resultantes se pueden tratar como datos anónimos; se debe realizar un análisis técnico y legal cuidadoso para analizar la naturaleza de los datos, los riesgos asociados con las actualizaciones del modelo y las medidas que se deben aplicar para mitigar dichos riesgos.

Para aprovechar al máximo los beneficios del FL y abordar sus desafíos, es esencial un enfoque holístico de cómo se procesan realmente los datos personales. Esto incluye la implementación de arquitecturas de sistemas que prioricen la protección de datos desde el diseño y por defecto, garantizando que el acceso a los datos entre las partes federadas se lleve a cabo equilibrando el nivel de riesgo del tratamiento, la precisión y la utilidad del modelo resultante.

Al centrarse en la privacidad y la protección de los datos personales, el FL puede utilizarse eficazmente para desarrollar sistemas de IA que sean potentes y respetuosos de los derechos y libertades de los usuarios.

2. Inteligencia artificial y las PETs (Privacy Enhancing Technologies)

En los últimos años, gracias a la disponibilidad de una enorme potencia de cálculo y el acceso a cantidades masivas de datos, el éxito de los sistemas de inteligencia artificial (IA)[3] ha aumentado con el uso de técnicas de desarrollo de aprendizaje automático (ML). Al igual que cualquier otro producto, los sistemas de IA deben seguir un proceso de diseño, desarrollo, validación y prueba que garantice los requisitos de rendimiento para un propósito y contexto específicos y ser acordes, entre otros, con la legislación de protección de datos personales.

El proceso de desarrollo de sistemas de ML incluye las siguientes fases (además del desarrollo

2 Varios métodos o vías que utilizan los atacantes para obtener acceso no autorizado a los datos

3 El Reglamento de IA [Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y se modifican los Reglamentos (CE) n.º 300/2008, (UE) n.º 167/2013, (UE) n.º 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial) (Texto pertinente a efectos del EEE)] define un sistema de IA como "un sistema basado en una máquina que está diseñado para funcionar con distintos niveles de autonomía y que puede mostrar capacidad de adaptación tras el despliegue, y que, para objetivos explícitos o implícitos, infiere de la información de entrada que recibe la manera de generar resultados de salida, como predicciones, contenidos, recomendaciones o decisiones, que pueden influir en entornos físicos o virtuales".

regular de sistemas que no son de ML[4]).

1. **El entrenamiento del sistema:** el entrenamiento de una IA requiere grandes cantidades de datos que sean relevantes para el propósito/objetivo de la IA. Por ejemplo, en el caso de un modelo de lenguaje de gran tamaño (LLM, IA diseñada para comprender y generar lenguaje humano), los datos de entrenamiento deben comprender un gran volumen de texto legible automáticamente[5] (actualmente el entrenamiento de algunos de estos sistemas implica cientos de miles de millones o billones de palabras[6]) para que la IA pueda proporcionar resultados que parezcan texto legible por humanos.
2. **Evolución automática del sistema:** Algunos[7] sistemas de IA incluyen algoritmos que permiten la evolución del sistema durante la fase de implementación[8]; es decir, la capacidad de utilizar los datos de operación para el entrenamiento continuo, lo que implica la necesidad de una validación y pruebas continuas.

Los avances en IA[9],[10] están cambiando rápidamente el estado del arte. La cantidad y diversidad de datos[11] utilizados para el proceso de aprendizaje de los sistemas de IA está creciendo a un ritmo acelerado para mantenerse al día con los diferentes y más sofisticados usos de esta nueva tecnología.

Muchas aplicaciones de IA utilizan datos personales[12]. Por lo tanto, para hacer frente a los riesgos legales y relacionados con la privacidad cada vez mayores que plantea la IA cuando utiliza datos personales, la aplicación de medidas adecuadas de protección de datos desde el diseño y por defecto[13] es esencial para proteger los derechos fundamentales de las personas. En este contexto, las tecnologías de mejora de la privacidad, comúnmente conocidas como PETs (Privacy Enhancing Technologies), podrían desempeñar un papel cada vez más importante.

Las PETs[14] son una variedad de técnicas para mejorar la privacidad y el control de los datos

4 Mas detalles se proporcionan en ISO 22989:2022
5 Información o datos que están en un formato que puede ser fácilmente procesado por una computadora sin intervención humana
6 How much data from the public Internet is used for training LLMs? , Michael Humour, 2023
7 A Comprehensive Survey of Continual Learning: Theory, Method and Application, Liyuan Wang et al., 2024
8 ISO/IEC/IEEE 15288:2023— Systems and software engineering — System life cycle processes
9 Defining Artificial Intelligence 2.0, Samoili, S. et al., 2021
10 The European's Commission's High-Level Expert Group on Artificial Intelligence - A Definition of AI: Main Capabilities and Scientific Disciplines, 2018
11 Por ejemplo, la clasificación de imágenes puede requerir millones de imágenes; Los grandes modelos de lenguaje normalmente se entrenan en miles de millones o billones de tokens...
12 The impact of the General Data Protection Regulation (GDPR) on artificial intelligence, Professor Sartor and Dr Francesca Lagioia, 2020
13 El artículo 27 del Reglamento (UE) 2018/1725 del Parlamento Europeo y del Consejo, de 23 de octubre de 2018, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales por las instituciones, órganos y organismos de la Unión, y a la libre circulación de dichos datos, y por el que se derogan el Reglamento (CE) n.o 45/2001 y la Decisión n.o 1247/2002/CE, y el artículo 25 del Reglamento (UE) 2016/679 del Parlamento Europeo y del Parlamento Europeo y del Parlamento Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos)
14 OECD (2023), Emerging privacy-enhancing technologies: Current regulatory and policy approaches, OECD Digital Economy Papers, No. 351, OECD Publishing, Paris, 2023

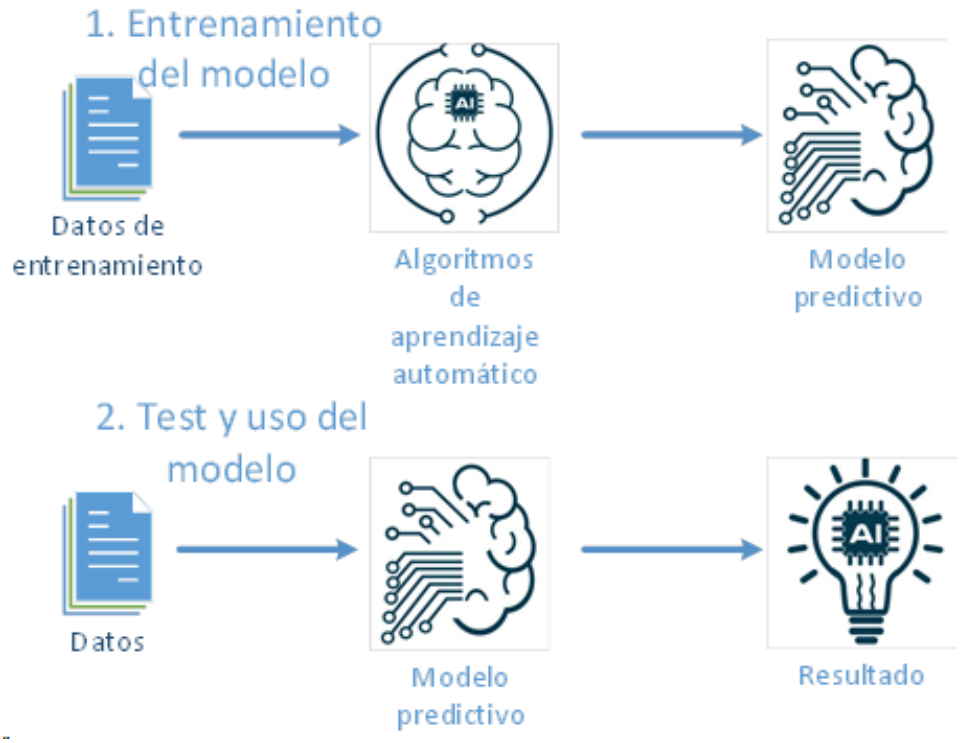
personales y se pueden implementar, entre otras, en la fase de entrenamiento del desarrollo de la IA. En este contexto, el aprendizaje federado podría considerarse como una forma de PET y, si se aplica correctamente, podría utilizarse en combinación con otros PETs para proporcionar una mayor protección cuando se procesen datos personales.

3. Aprendizaje federado

3.1 ¿Qué es el aprendizaje federado?

El aprendizaje federado (FL) es un tipo de aprendizaje automático (ML) en el que múltiples fuentes de datos (dispositivos o servidores) colaboran para entrenar un modelo compartido mientras se mantienen los datos descentralizados. En lugar de enviar datos originales a un servidor central (cuando lo hay), cada fuente trata sus propios datos localmente y solo comparte actualizaciones del modelo (por ejemplo, gradientes o pesos[15]).

Los datos y el proceso de aprendizaje son dos piezas clave en la construcción de sistemas de IA.



15 Los gradientes representan la dirección y la tasa de cambio de una función, específicamente relacionada con cómo un pequeño cambio en la entrada afecta a la salida. Los pesos son parámetros numéricos que determinan la fuerza de las conexiones entre las neuronas de una red neuronal.

Figura 1 Aprendizaje automático[16]

Originalmente, en ML, los datos y el proceso de aprendizaje estaban centralizados[17], es decir, los datos se encontraban en un centro de datos específico donde se produce el proceso de aprendizaje.

Con el tiempo, el proceso de aprendizaje se convirtió en un proceso demasiado costoso en tiempo, potencia de cálculo y almacenamiento para ser centralizado en una sola máquina. Por lo tanto, los desarrolladores terminaron cargando y almacenando sus conjuntos de datos en servicios en la nube. Estos servicios están optimizados para distribuir los datos y el tratamiento entre varias máquinas.

Aprendizaje federado con coordinación de servidores centrales

En un entorno de FL con servidor central:

1. Un servidor central o proveedor de servicios envía un modelo de ML inicial o previamente entrenado a cada dispositivo cliente federado.
2. Cada uno de estos dispositivos cliente entrena localmente su modelo de ML con sus propios datos, lo que da como resultado varios modelos de ML entrenados localmente.
3. Cada uno de estos dispositivos cliente envía de vuelta al servidor central sus parámetros o las actualizaciones de esos parámetros[18] para los modelos entrenados localmente.
4. El servidor central compila todos los datos recibidos de los modelos locales para producir un modelo combinado. En algunos casos de FL[19], se implementa un sistema de votación para seleccionar los modelos locales que se integrarán (los que producen los mejores resultados).
5. A continuación, el servidor central envía de nuevo este modelo combinado a todos los dispositivos cliente múltiples.
6. Este proceso se repite un número fijo de veces o hasta que el rendimiento del modelo central alcanza un umbral de rendimiento.

16 Iconos cortesía de https://vecteezy.com
17 Federated learning, Qiang Yang and al., Morgan and Claypool Publishers Preface (xiii) and Introduction (p1)
18 A systematic review of federated learning from clients' perspective: challenges and solutions, , Yashothara Shanmugarasa et al., 2023
19 Robust federated learning with voting and scaling, Xiang-Yu Liang et al., 2024

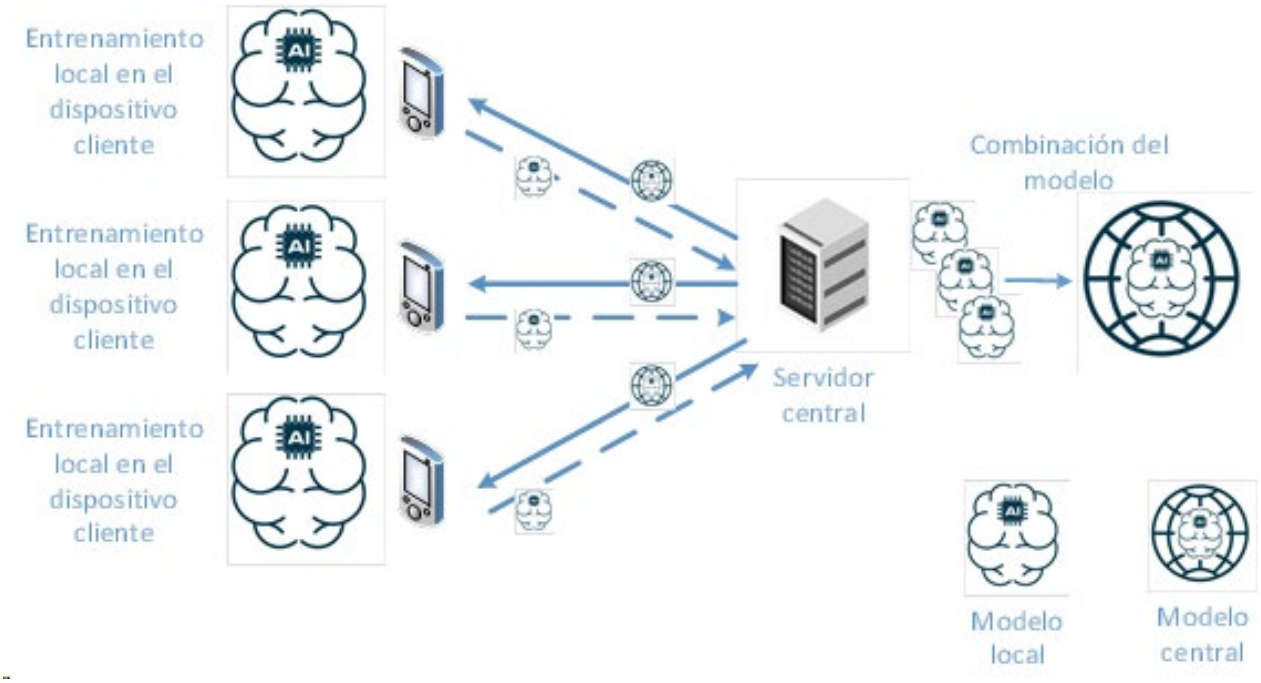


Figura 2: Aprendizaje Federado[20]

Es una aplicación de la estrategia compute-to-data, lo que significa llevar el proceso a los datos, en lugar de llevar (o mover) los datos al proceso, utilizando técnicas similares a las utilizadas en los grandes proveedores de servicios en la nube para aumentar la eficiencia.

Aprendizaje federado sin coordinación de servidores centrales

El FL también puede ser completamente descentralizado (Aprendizaje Federado Descentralizado (DFL)). En este caso, no hay un servidor central. Cada dispositivo cliente construye su modelo local e intercambia sus parámetros directamente con otros dispositivos cliente a través de una arquitectura de red punto a punto que no incluye un servidor central.

20 Iconos cortesía de https://vecteezy.com

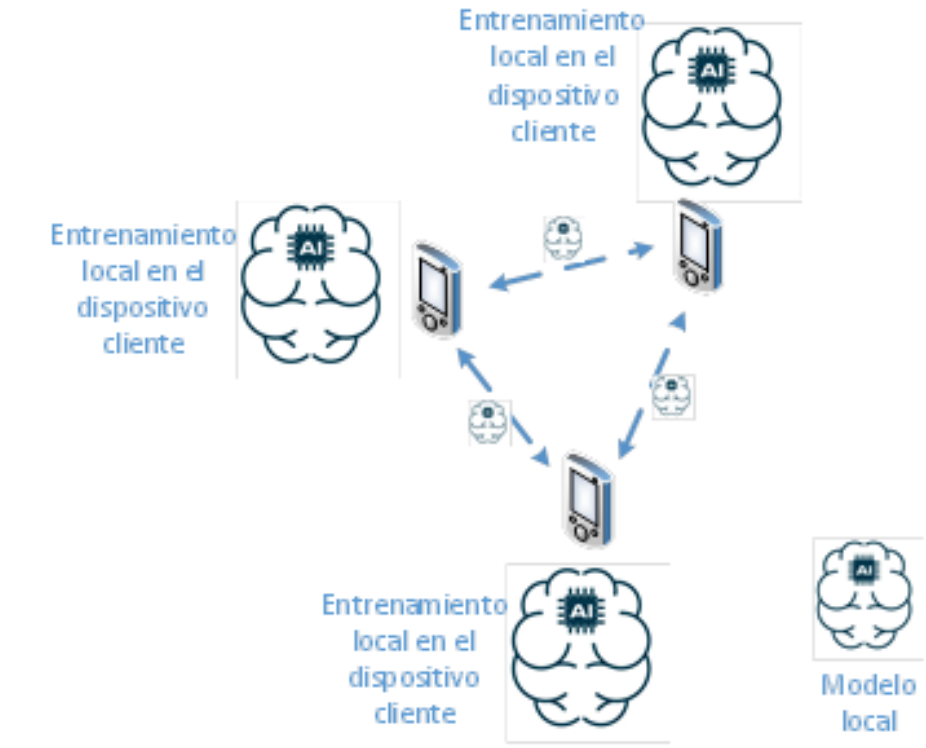


Figura 3: Aprendizaje federado descentralizado[21]

3.2 ¿Cómo se pueden clasificar los Modelos de Aprendizaje Federado?

3.2.1 Horizontal vs Vertical

Los modelos de FL se pueden clasificar de la siguiente manera[22].

- **Aprendizaje horizontal:** En el aprendizaje horizontal, los datos almacenados por los diferentes dispositivos cliente comparten las mismas características, es decir, todos los dispositivos cliente utilizan la misma estructura de datos (véase Figura 4: datos de aprendizaje horizontal).

21 Iconos cortesía de https://vecteezy.com
22 Understanding the types of Federated Learning, Openminded blog

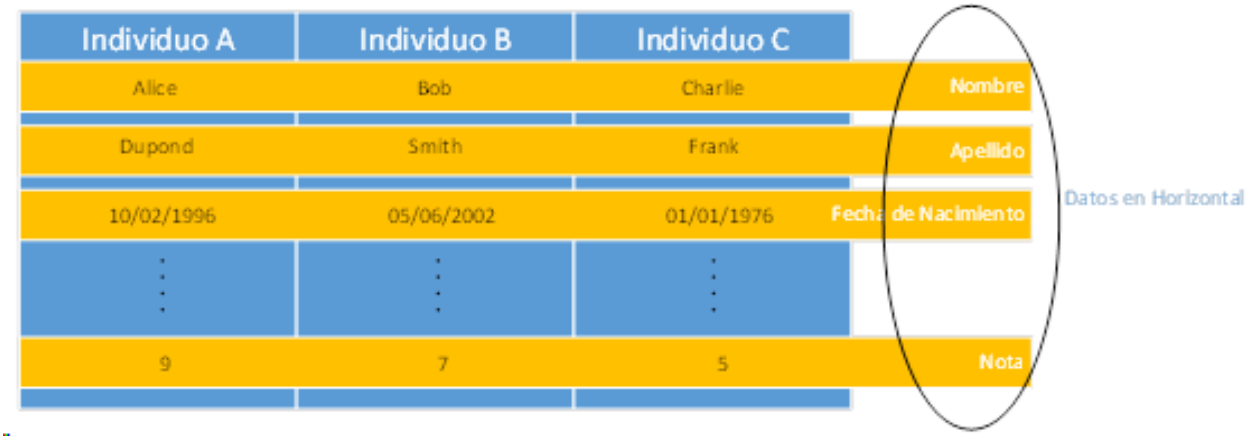


Figura 4: Datos de aprendizaje horizontal

- **Aprendizaje vertical:** En el aprendizaje vertical, los datos a través de dispositivos pueden contener datos sobre la misma entidad (por ejemplo, individuos), aunque con diferentes tipos de información, como se muestra en Figura 5: datos de aprendizaje vertical.

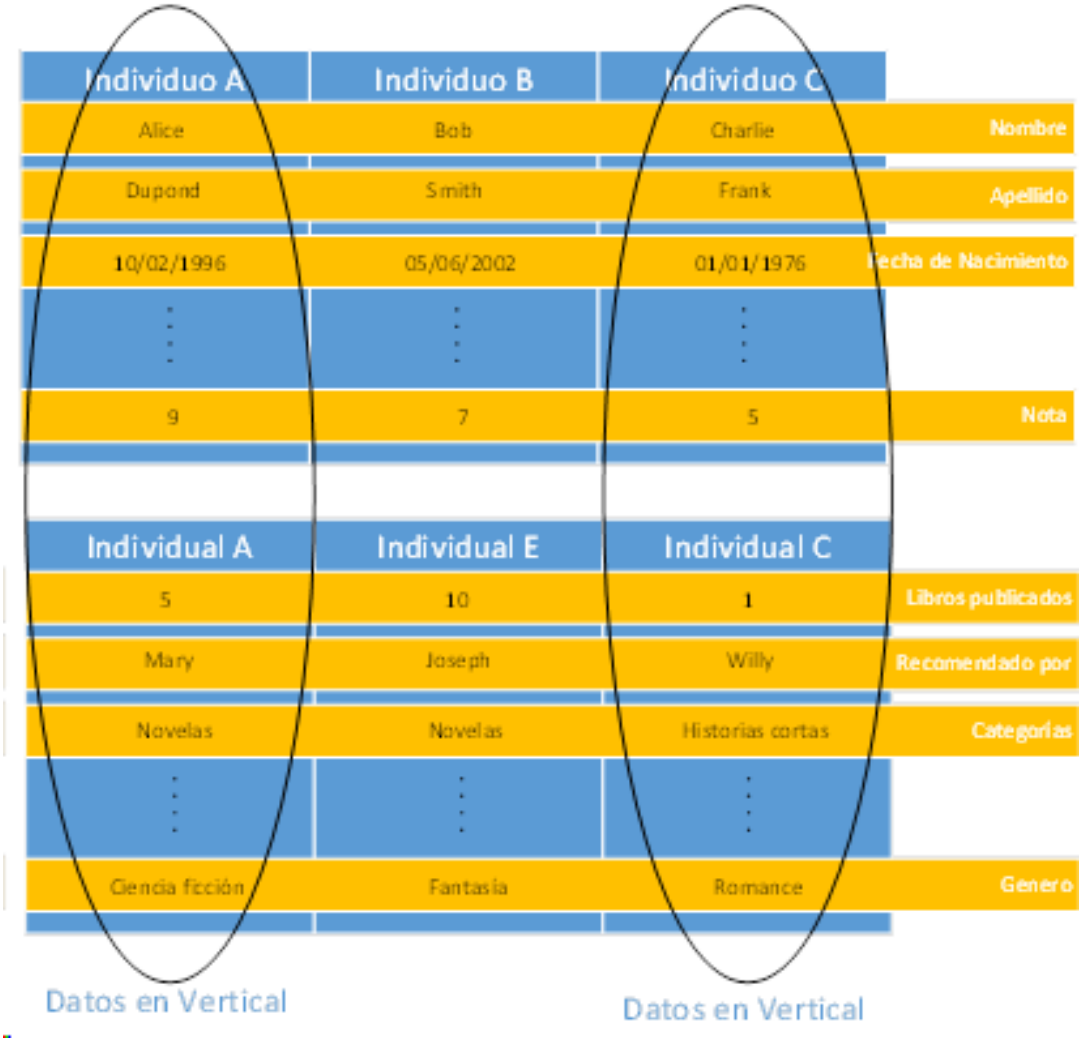


Figura 5: Datos de aprendizaje vertical

3.2.2 Dispositivos transversales vs silos transversales

Además, es posible clasificar los sistemas FL en función de su tipo de clientes.

- **Dispositivos** transversales: Los clientes son individuos con dispositivos personales (por ejemplo, teléfonos inteligentes, dispositivos portátiles...) en grandes cantidades. Los datos almacenados en cada dispositivo se limitan a los generados por su propio usuario. En algunos casos, los nuevos datos se generan dinámicamente mientras que los datos más antiguos se eliminan. Para algunos casos de uso, como el ajuste de un modelo a la especificidad de un individuo, el FL entre dispositivos encaja bien.

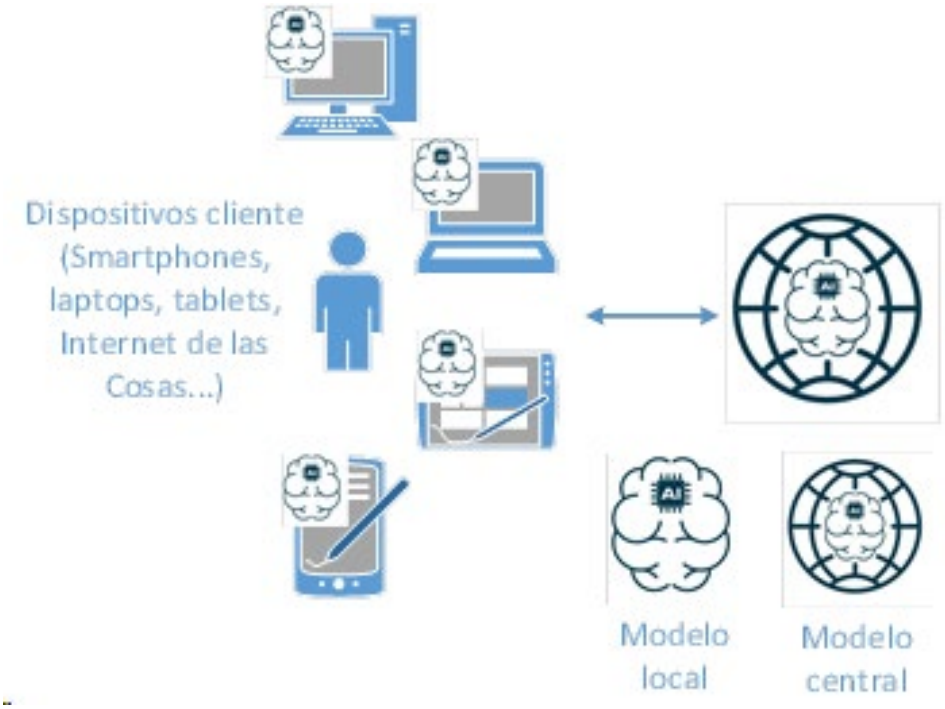


Figura 6: Dispositivos transversales en FL (los clientes son individuos dentro de una organización o no)[23]

- **Silos** transversales: Los clientes son organizaciones (por ejemplo, bancos, hospitales...) en un número reducido. Los datos que posee cada uno son grandes en cantidad y un conjunto de datos de una organización puede (pero no necesariamente) tener las mismas características generales que todos los datos de todas las organizaciones.

23 Iconos cortesía de https://vecteezy.com

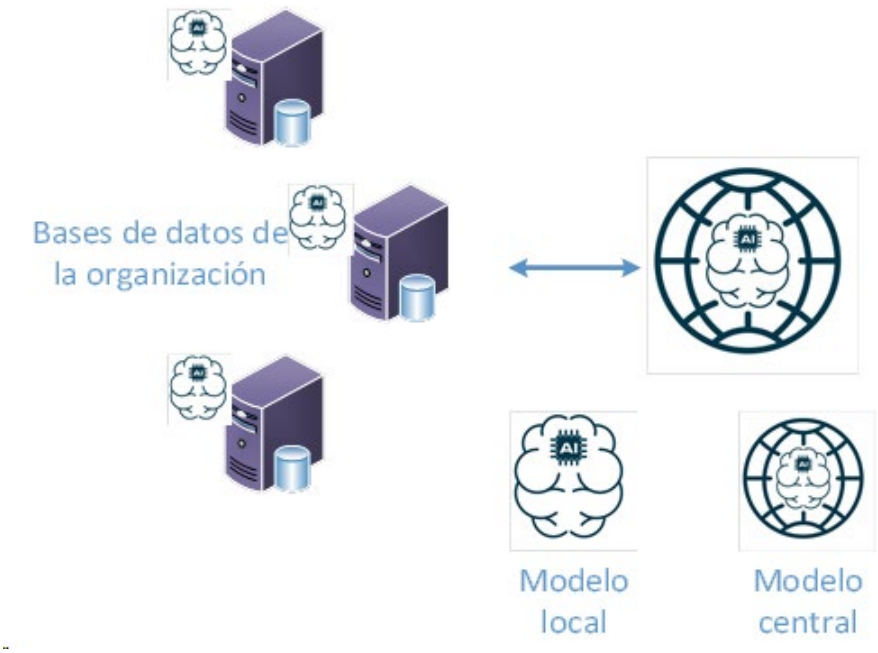


Figura 7: Silos transversales: los clientes son organizaciones y no individuos[24]

3.3 Ejemplos de casos de uso del Aprendizaje Federado

Como se mencionó anteriormente, el FL es una alternativa interesante al aprendizaje centralizado cuando el desarrollo de sistemas de IA requiere datos de diferentes fuentes, pero no es posible o deseable compartir estos datos. Las aplicaciones pueden ser diversas; Aquí se presentan algunos de los casos de uso más comunes para ejemplificar el potencial de la tecnología.

3.3.1 Modelos de IA para el sector sanitario

Las organizaciones en el campo de la salud (como hospitales, instituciones de investigación médica...) pueden crear un sistema de IA utilizando el FL para evitar compartir datos confidenciales (por ejemplo, datos médicos de pacientes) con terceros. En este contexto, obtener conjuntos de datos de entrenamiento lo suficientemente grandes puede ser difícil[25],[26],[27],[28] debido a razones de privacidad (los datos médicos son una categoría especial

24 Idem
25 Bridging federated learning theory and practice with real-world healthcare data, Jean Ogier du Terrail et al., 2022
26 A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applicationst, Prayitno et al., 2021
27 Federated learning project connects pharma with university to train AI model
28 Federated learning for medical imaging radiology, Muhammad Habib ur Rehman et al., 2023

de datos, sujeta a un régimen de protección específico) o al bajo número de muestras en cada organización (por ejemplo, ciertos hospitales pueden no tener muchos pacientes con una enfermedad concreta que deba estudiarse y para la que la IA podría ayudar). Cuando los sistemas de IA no pueden obtener suficientes datos de entrenamiento, el FL puede ayudar a mejorar el rendimiento de la IA, es decir, entrenar adecuadamente un sistema de IA con datos relacionados con una enfermedad concreta procedentes de diferentes servicios de salud. Por ejemplo, esto se logró para ayudar a luchar contra el cáncer[29] (en particular en casos de tumor cerebral) donde la heterogeneidad de la enfermedad, la preparación de los tejidos y los procesos de tinción hacían imposible encontrar una única ubicación central que contuviera datos suficientes para entrenar los sistemas de IA. La agrupación de recursos mediante el FL resolvió este problema sin necesidad de mover todos los datos a la ubicación central.

3.3.2 Modelos de voz

El FL también se ha utilizado para entrenar un modelo de voz[30,31] utilizando como dispositivos cliente smartphones y dispositivos portátiles (por ejemplo, relojes inteligentes) para evitar compartir sus datos de voz con el proveedor de servicios (por ejemplo, reconocimiento de voz o texto predictivo).

3.3.3 Sistemas de transporte autónomos

Otro caso es el de los sistemas de transporte autónomos (por ejemplo, los coches), en los que cada vehículo puede enviar los parámetros de los modelos locales a un modelo central sin revelar los datos personales del interesado (como los datos de ubicación), con el fin de mejorar un modelo de IA utilizado para fines como la detección de objetos y la planificación de rutas[32]. Esto se hizo para pronosticar las condiciones del tráfico, identificar el comportamiento de los peatones y ayudar a los conductores a tomar decisiones.

3.4 Desafíos técnicos para la implementación de sistemas FL

Recursos informáticos. Para alguna situación en el escenario entre dispositivos, el proceso de entrenamiento podría ser demasiado exigente en términos de capacidad de computación

29 A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applications, Prayitno et al., 2021
30 https://support.google.com/assistant/answer/11140942?hl=en#zippy=%2Cfederated-learning
31 Federated Learning for mobile keyboard prediction, A. Hard et al., 2019
32 Real-time End-to-End Federated Learning: An Automotive Case Study, Hongyi Zhang et al., 2021

para los dispositivos cliente si su potencia computacional es limitada. Por diseño, algunos dispositivos cliente pueden tener menos potencia de procesamiento que los servidores, que pueden integrar varias unidades de procesamiento (CPU, GPU y TPU) que realizan cálculos en un equipo[33].

Eficacia. Dependiendo del caso de uso de FL, el rendimiento de la comunicación podría ser una ventaja o una desventaja. Cuando el FL se utiliza en un escenario con varias bases de datos masivas en diferentes clientes (generalmente corporaciones u organismos públicos), y el modelo local es más pequeño que dichas bases de datos, la comunicación puede ser más eficiente porque no hay necesidad de comunicaciones masivas de datos. De lo contrario, en los casos de recopilación de datos en tiempo real para personas físicas, la comunicación puede ser menos eficiente dependiendo del tamaño del modelo. En ambos escenarios, el proceso de entrenamiento podría ser más eficiente porque no hay necesidad de una alta potencia de cómputo centralizada; el tratamiento se distribuye en varias máquinas (en el caso de los usuarios físicos, utilizando el paradigma de edge-computing).

Complejidad. Los distintos dispositivos cliente pueden variar en términos de tamaño, potencia de cálculo, medios de comunicación, arquitectura..., lo que hace que establecer un entrenamiento de FL sea una tarea compleja[34]. Además, el escalado de FL a un gran número de clientes puede introducir complejidades adicionales, incluida la gestión de la participación del cliente (por ejemplo, abandonos o errores de clientes durante el proceso de entrenamiento), la gestión de los rezagados (clientes lentos o poco fiables) y el equilibrio de la carga.

Convergencia. El logro de una convergencia rápida y estable en la configuración de FL debe administrarse debido a la posible naturaleza asincrónica de las actualizaciones y la distribución de datos no independiente e idéntica[35,36] (IID).

4. ¿En qué parte de una arquitectura FL puede haber un tratamiento de datos personales?

Dada la arquitectura de los sistemas de FL, hay tres etapas en las que pueden estar presentes datos personales.

33 A Survey of Three Types of Processing Units: CPU, GPU and TPU, Goran S. Nikolić et al., 2022
34 Heterogeneous Federated Learning: State-of-the-art and Research Challenges, Mang Ye et al., 2023
35 Es decir, no hay tendencias generales.
36 Independent and Identically Distributed (IID) Data Assessment in Federated Learning, Arafeh, Mohamad & Hammoud et al. 2022

- Los datos personales pueden tratarse **dentro de cada dispositivo** (los datos de entrenamiento pueden contener datos personales).
- Los datos personales pueden intercambiarse entre dispositivos (los datos compartidos entre dispositivos son los pesos y/o gradientes de los modelos de ML).
- Los datos personales pueden ser tratados **dentro de los modelos de ML** (tanto en el modelo central, cuando existe, como en los modelos locales).

4.1 ¿Los datos personales se tratan localmente en cada dispositivo?

En FL, los datos de entrenamiento se recopilan y tratan en cada dispositivo participante para entrenar un modelo local (cuando hay un servidor central, este servidor central proporciona primero un modelo de referencia). Si estos datos de entrenamiento incluyen datos personales, esos datos personales también podrían memorizarse parcialmente en el modelo local resultante.

La naturaleza descentralizada de FL no exime a los proveedores de sistemas de IA que utilizan FL de garantizar el cumplimiento de la legislación aplicable en materia de datos personales. En consecuencia, los proveedores que utilicen FL deberán considerar las salvaguardas adecuadas para proteger los datos personales en cada dispositivo.

Cualquier tratamiento local de datos personales en una configuración de FL debe realizarse, entre otras cosas, de tal manera que se garantice una base legal válida, se brinde transparencia, se minimicen los datos y se implementen medidas de seguridad sólidas para proteger los datos personales contra el acceso o las modificaciones no autorizados.

4.2 Posible intercambio de datos personales entre los dispositivos participantes

En FL, la comunicación entre dispositivos implica inherentemente la transmisión de conocimientos derivados de los datos de entrenamiento presentes en cada dispositivo. Los datos intercambiados consisten en gradientes y/o pesos que encapsulan el conocimiento y los patrones aprendidos de los datos de entrenamiento. Estos gradientes y/o pesos son necesarios para agregar actualizaciones locales en un modelo global completo. Sin embargo, la cuestión de si estos datos (gradientes y pesos) transmitidos entre los dispositivos permiten

el tratamiento de datos personales debe ser evaluada caso por caso por el responsable del tratamiento.

Dado que solo se comparten pesos y/o gradientes, el FL tiene menos riesgos desde el punto de vista de la protección de datos personales que el intercambio de los conjuntos de datos de entrenamiento completos. La reconstrucción de los datos de entrenamiento a partir de los datos intercambiados en un entorno de FL es complicada y solo funcionará para una fracción de los datos de entrenamiento (el nivel de dificultad y la tasa de éxito dependen de la configuración de FL).

Para verificar si los pesos y gradientes permitirían el tratamiento de datos personales, es necesario responder a la pregunta: "¿Se puede extraer información relacionada con un individuo cuyos datos personales se utilizaron durante el entrenamiento de la información intercambiada entre los dispositivos?" La respuesta no siempre es evidente. Reconstruir los datos de entrenamiento originales a partir de gradientes y pesos no será sencillo y, en la mayoría de los casos, puede que no sea posible.

Este riesgo debe determinarse al inicio del sistema. Esto es importante puesto que, si los pesos y gradientes permiten reconstruir los datos personales utilizados durante la fase de formación, se deben establecer salvaguardas adecuadas en cada caso[37].

4.3 ¿Se puede extraer información relacionada con un individuo del modelo resultante?

Al igual que con la información intercambiada, la respuesta no es sencilla y requiere un análisis caso por caso. Los modelos de ML pueden retener características y correlaciones de las muestras de datos de entrenamiento[38]; se pueden atacar para reconstruir los datos personales utilizados en la fase de entrenamiento (ataques de extracción) o para inferir si había muestras de datos específicas en el conjunto de datos de entrenamiento[39,40] (ataque de inferencia de pertenencia[41]). Por ejemplo, si le preguntamos a un modelo de lenguaje de gran tamaño (LLM) sobre una figura pública, podría ser posible recuperar información personal y, a veces, incluso

37 A review of federated learning: taxonomy, privacy and future directions, Ratnayake, H. et al., 2023

38 Wei, J., Zhang, Y., Zhang, L. Y., Ding, M., Chen, C., Ong, K. L., ... & Xiang, Y. (2024). Memorization in deep learning: A survey. arXiv preprint arXiv:2406.03880

39 Fang, H., Qiu, Y., Yu, H., Yu, W., Kong, J., Chong, B., ... & Xia, S. T. (2024). Privacy leakage on DNNs: A survey of model inversion attacks and defences. arXiv preprint arXiv:2402.04013.

40 Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., ... & Lee, K. (2023). Scalable extraction of training data from (production) language models. arXiv preprint arXiv:2311.17035.

41 Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., & Zhang, X. (2022). Membership inference attacks on machine learning: A survey. ACM Computing Surveys (CSUR), 54(11s), 1-37

producir una imagen de esa persona.

Dado que estas reconstrucciones y extracciones de datos personales de los modelos de ML son posibles, se puede concluir que existe el riesgo de que parte de los datos personales utilizados en el entrenamiento puedan extraerse de los modelos de ML resultantes. Por lo tanto, se debe realizar una evaluación caso por caso para determinar si los modelos de ML deben considerarse como datos personales.

Para más información sobre las circunstancias en las que los modelos de IA podrían considerarse anónimos y la demostración relacionada, véase la sección 3.2 del Dictamen 28/2024 del CEPD sobre determinados aspectos de protección de datos relacionados con el tratamiento de datos personales en el contexto de los modelos de IA[42].

5. ¿Cuáles son los beneficios y desafíos para la protección de datos del FL?

Cuando se utilizan datos personales para el entrenamiento, en los modelos no federados, cada dispositivo recogerá y transmitirá directamente estos datos, mientras que en los sistemas FL cada dispositivo entrenará un modelo local y solo transferirá el resultado del entrenamiento (pesos y parámetros). Esto evita tanto el intercambio como el tratamiento directo de datos personales por parte del sistema central u otros dispositivos federados y mitiga los riesgos de protección de datos relevantes.

Desde el punto de vista de la protección de datos personales, el FL puede aportar ciertas ventajas en comparación con los procesos centralizados de ML. Sin embargo, no debe darse por sentado que el FL resuelve todos los problemas, ya que persistirán algunos riesgos.

42 EDPB Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_en

5.1 Ventajas de FL frente a los sistemas ML centralizados desde el punto de vista de la protección de datos personales.

5.1.1 Transferencia/Minimización de datos

Debido a su naturaleza, el FL puede contribuir a la implementación del principio de minimización de datos (un principio básico de protección de datos) porque en lugar de enviar todo el conjunto de datos[43] a otras partes, solo se transmiten los parámetros del modelo o sus actualizaciones.

5.1.2 Responsabilidad proactiva o "accountability" mejorada

El FL puede ayudar a los responsables del tratamiento en la aplicación del principio de responsabilidad proactiva o "accountability": potencialmente podrían controlar mejor el acceso a los datos personales, evitando así también cualquier posible reutilización ilegal del tratamiento.

5.1.3 Tratamiento de datos confidenciales más seguro (incluidas categorías especiales de datos)

El FL permite el tratamiento de diferentes categorías de datos personales (incluidos datos sensibles) sin necesidad de compartir datos con las otras partes. Debido al tratamiento local y a la ausencia de intercambio de datos, el FL ayuda a reducir los riesgos para los derechos y libertades de las personas, principalmente en los casos de tratamiento masivo de categorías especiales de datos personales (artículo 9 del RGPD[44]), para obtener una evaluación más positiva del principio de proporcionalidad en el contexto de la EIPD que se llevará a cabo de conformidad con el artículo 35.7.b del RGPD y para garantizar la responsabilidad proactiva o "accountability". Sin embargo, para garantizar un tratamiento justo y evitar el sesgo de los patrones existentes en los datos de entrenamiento, se necesitan salvaguardas para detectar y

43 Sin embargo, los datos personales aún deben procesarse en los dispositivos del cliente. El FL no influye en la cantidad de datos de entrenamiento que se van a utilizar localmente.

44 Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos)

mitigar el sesgo presente en los datos de origen[45].

5.1.4 Gestión del consentimiento

El FL permite tener un mejor control de los datos personales del interesado, aumentando la transparencia (qué datos, para qué, cuándo). Idealmente, el interesado podrá verificar mejor qué uso se hace de sus datos personales, ya que el FL mejora el control y la soberanía sobre su propio entorno. Por lo tanto, como los datos de entrenamiento permanecen en los dispositivos, el FL simplifica la gestión del consentimiento.

5.1.5 Seguridad de los datos

En un escenario de silos cruzados, el FL podría ayudar a reducir la reticencia de las organizaciones con grandes volúmenes de datos a revelar información privada. Esto podría ayudar a implementar escenarios colaborativos de intercambio de datos, como por ejemplo en espacios de datos, reducir los riesgos y, por lo tanto, aprovechar mejor los beneficios potenciales. Además, dado que no existe un almacenamiento central de datos personales, es muy poco probable que una brecha de datos personales afecte a todos los datos personales utilizados en el entrenamiento del modelo.

Sin embargo, dado que existen riesgos de que los datos personales se reconstruyan utilizando los gradientes o pesos, o los modelos locales/centrales (véase el capítulo 4), se debe realizar un análisis y se debe informar adecuadamente al sujeto interesado de estos riesgos.

5.2 Desafíos del FL sobre sistemas ML centralizados desde el punto de vista de la protección de datos personales

5.2.1 Gestión de la calidad de los datos de entrenamiento

En ingeniería de software, la calidad de los datos podría definirse como el grado en que los datos satisfacen los requisitos de su propósito previsto[46]. Esto significa que un conjunto de datos de entrenamiento de ML tiene suficiente calidad si es posible desarrollar un sistema

45 EDPS's Generative AI and the EUDPR. First EDPS Orientations for ensuring data protection compliance when using Generative AI systems.

46 Modelo de calidad de datos ISO/IEC 25012

de ML que cumpla con sus requisitos de rendimiento. Es importante no confundir la calidad de los datos del conjunto de datos de entrenamiento (en particular, las características de exactitud y precisión) con el principio de exactitud del RGPD. Por ejemplo, el uso de técnicas de anonimización en un conjunto de datos de entrenamiento puede hacer que algunos de sus datos sean inexactos (desde el punto de vista de la protección de datos), pero el conjunto de datos aún puede tener la calidad suficiente para ser una entrada en el proceso de entrenamiento de ML.

En un entorno no FL (entorno en el que los datos están centralizados), antes de iniciar el proceso de entrenamiento, se debe realizar una evaluación distribuida de cada fuente de datos para evaluar su nivel de calidad. Esto podría hacerse comprobando algunas características de calidad como la integridad, la credibilidad de las fuentes, la actualidad[47], el cumplimiento y otras[48], y comprobando si es posible llevar a cabo un tratamiento efectivo para aumentar el nivel de calidad de los datos. Los datos que no alcanzan un determinado nivel de calidad no son necesarios para el proceso de entrenamiento. Una vez consolidado el conjunto de datos de entrenamiento, se podrían llevar a cabo otras evaluaciones de calidad (como la consistencia entre registros).

En un entorno FL, la comprobación de la calidad de los datos es más difícil, ya que las fuentes de datos no están centralizadas y no se transmiten. Por lo tanto, cada fuente de datos no se puede comparar con las otras fuentes de datos (no hay comprobaciones de calidad de datos entre fuentes), no hay posibilidades de comprobar la calidad de todos los datos de entrenamiento en su conjunto y puede ser difícil comprobar la credibilidad de cada fuente de datos. En FL, necesariamente se deben implementar procedimientos específicos de gestión de la calidad de los datos distribuidos.

Algunos métodos para evaluar y mejorar la calidad de los datos de FL son los siguientes[49].

- **Distribución de datos:** El servidor central (cuando lo hay) puede solicitar información estadística de los conjuntos de datos locales (que se utilizaron para entrenar los modelos locales) para determinar si se debe dar más peso (es decir, relevancia) a los parámetros de los modelos locales con mayor valor. Por ejemplo, el método "Private Set Intersection" (PSI)[50] permite calcular los elementos comunes de diferentes conjuntos de datos sin intercambiar estos conjuntos de datos[51]. Hasta cierto punto, cuanto mayor sea la similitud estadística, mayor será la seguridad que tenga el servidor central de que los datos utilizados para entrenar los modelos son de suficiente calidad.

47 La actualidad, tal como se define en la norma ISO 25012:2008, significa el grado en que los datos tienen atributos que tienen la edad adecuada en un contexto de uso específico
48 Modelo de calidad de datos ISO/IEC 25012
49 A Survey of Federated Evaluation in Federated Learning, Behnaz Soltani et al., 2023
50 Private set intersection: A systematic literature review, Daniel Morales et al., 2023
51 Practical Private Set Intersection Protocols with Linear Computational and Bandwidth Complexity, E. De Cristofaro, 2009

- **Utilidad del modelo:** La calidad de un modelo local puede evaluarse en función de su utilidad, es decir, cuánto afecta el modelo local a la calidad de las predicciones del modelo central de la siguiente iteración (el modelo central resultante del uso de los parámetros de los modelos locales). Para ello, se evalúa el impacto en la calidad del modelo central actual. Se integran los parámetros de un modelo local y se reevalúa el modelo central. Si la calidad de este nuevo modelo central ha mejorado, entonces se puede afirmar que el modelo local se entrenó con datos de calidad.

- **Métricas estadísticas:** Los modelos locales se pueden evaluar en función de métricas estadísticas. Por lo general, esto se hace calculando la distancia entre los parámetros del modelo antes y después de las rondas de entrenamiento; algunos argumentan que si la distancia de los parámetros del modelo es mayor cuando se utilizan datos independientes e idénticamente distribuidos (IID, Identically Distributed Data)[52], entonces el modelo puede considerarse de mejor calidad; otros[53] muestran que para los datos que no son IID puede ser lo contrario.

En cualquier caso, cuando los datos de entrenamiento continuo se recopilan en flujos de datos (es decir, cada nuevo dato se agrega inmediatamente al conjunto de datos de entrenamiento) y se procesan en tiempo real, se deben adoptar otras soluciones específicas para garantizar la calidad de los datos tanto en entornos centralizados como en el FL.

5.2.2 Precisión y sesgo de la salida de ML

Tanto en el entrenamiento de FL como en el de ML en general (sin FL), los desarrolladores deben asegurarse de que el modelo final de ML esté libre de sesgos y permanezca libre de sesgos (éste es un proceso continuo). En el caso del FL, la dificultad proviene de la implementación de un proceso de gestión de la calidad de los datos de entrenamiento distribuidos. Algunas de las técnicas de mitigación disponibles son:

- asegurar que las operaciones de extracción y transformación de datos funcionen correctamente en cada sitio (como sensores o convertidores de formato);

52 La lógica es que si la distancia entre los parámetros del modelo antes y después de una ronda de entrenamiento es mayor, el modelo se ha ajustado de manera más significativa, lo que podría indicar que está aprendiendo efectivamente de los datos. Por el contrario, las pequeñas actualizaciones de parámetros podrían sugerir que el modelo está convergiendo o que los datos podrían no proporcionar nuevas actualizaciones informativas. Sin embargo, esta es una métrica matizada y puede depender del contexto específico. En algunos casos, los cambios de parámetros excesivamente grandes pueden indicar inestabilidad o gradientes ruidosos, lo que podría perjudicar la convergencia del modelo.

53 Federated Learning with Non-IID Data, Yue Zhao et al., 2022

- garantizar la consistencia de los procesos de muestreo y normalización de cada modelo de ML local;

- el seguimiento de la distribución estadística de los datos locales de entrenamiento y el reequilibrio local de su representatividad estadística; Esto debe hacerse hasta que se alcance cierta uniformidad en un conjunto de participantes antes de que pueda comenzar el entrenamiento o actualización del modelo global.

5.2.3 Integridad

En un entorno FL, es necesario asegurarse de que los datos no se modifiquen indebidamente para que el modelo central (o distribuido) resultante sea preciso. En comparación con las arquitecturas que no son FL, el FL tiene diferentes vectores de amenazas[54] que pueden afectar a la integridad de los datos. Esto se debe a que, en las arquitecturas FL, hay múltiples dispositivos que participan en el sistema general y, por lo tanto, hay múltiples modelos y transferencias de datos que defender (los modelos locales y, en su caso, el modelo central).

Una forma de atacar los modelos locales o centrales es realizar un envenenamiento de datos (Data Poisoning)[55,56]. El envenenamiento de datos es un ataque en el que se inyectan datos falsos en el proceso de entrenamiento de cualquier dispositivo para sesgar un sistema de IA en su conjunto y reducir su rendimiento. Por lo general, esto se puede mitigar mediante la detección de valores atípicos (mediante el análisis de las actualizaciones del modelo local recibidas de los dispositivos en busca de anomalías estadísticas)[57,58].

La modificación indebida de las actualizaciones del modelo local (envenenamiento del modelo), en los dispositivos cliente o en tránsito, también tendría un efecto perjudicial en el modelo global en términos de integridad.

En respuesta a los ataques de envenenamiento, los investigadores proponen defensas pasivas y activas. Las defensas pasivas comienzan con el análisis de la agregación de los modelos en el lado del servidor (diseñando estrategias de modelos de agregación relevantes), mejorando así el rendimiento del modelo global. Las defensas activas eliminan el impacto del modelo de envenenamiento en el modelo global al detectar el rendimiento del modelo local y eliminar el modelo envenenado. Actualmente, las defensas activas parecen ser la tendencia más prometedora[59].

54 Varios métodos o vías que utilizan los atacantes para obtener acceso no autorizado a los datos.
55 Robustness and Explainability of Artificial Intelligence, Hamond R. et al., 2020
56 Detection and Prevention Against Poisoning Attacks in Federated Learning, V. Valadi et al., 2022
57 Precision Guided Approach to Mitigate Data Poisoning Attacks in Federated Learning, K. N. Kumar et al., 2024
58 Detection and Prevention Against Poisoning Attacks in Federated Learning, V. Valadi et al., 2022
59 Challenges and Countermeasures of Federated Learning Data Poisoning Attack Situation Prediction, Wu J et al., 2024

La protección de los dispositivos cliente sigue siendo difícil en el FL multidispositivo, ya que, en general, los participantes que poseen esos dispositivos carecen de la experiencia, los recursos y la madurez que las organizaciones pueden poner en la seguridad de su configuración del FL.

5.2.4 Confidencialidad

El FL ofrece una confidencialidad mejorada para el responsable de cada uno de los dispositivos ya que no requiere compartir los datos brutos de entrenamiento con el resto de los dispositivos del ecosistema. Al mismo tiempo, una configuración FL ofrece la oportunidad de atacar los modelos locales a medida que son entrenados por los dispositivos y atacar el eslabón más débil puede poner en riesgo toda la estructura. Los modelos locales de FL deben almacenarse en la ubicación original (dispositivos) y luego transmitirse a la ubicación central[60]. A continuación, podrían ser hackeados en los dispositivos, o podrían ser analizados en tránsito o en destino. Después de recibir el modelo inicial pre-entrenado, los modelos locales comienzan entrenándose en conjuntos de datos locales y son más susceptibles a revelar qué datos (o un subconjunto de estos datos, incluidos los datos potencialmente personales) se usaron para entrenarlos. Esto puede suceder porque los modelos locales pueden conservar características y correlaciones de muestras de datos de entrenamiento que los atacantes podrían usar para reconstruir o extraer registros. Por lo tanto, es posible crear un modelo atacante[61] que intente averiguar los datos de entrenamiento de los modelos locales u observar los cambios del modelo a lo largo del tiempo para ayudar a determinar datos personales usados para el entrenamiento.

Con el fin de protegerse contra este tipo de ataques, se deben implementar salvaguardas en:

- los datos en reposo y los modelos en los dispositivos cliente;
- la comunicación entre los dispositivos cliente y el servidor central (o entre los dispositivos cliente en un enfoque DFL);
- el propio servidor central (cuando lo hay), que contiene los modelos intermedios y finales.

Estas salvaguardas pueden consistir en:

- El uso del **cifrado** en los datos en reposo en los dispositivos cliente para mitigar los ataques que podrían poner en peligro directamente esos dispositivos.

60 en una arquitectura no DFL

61 Leak and Learn: An Attacker's Cookbook to Train Using Leaked Data from Federated Learning, Joshua C. Zhao et al., 2024

- El uso de **Computación Segura Multi-parte (SMPC, Secure Multi-party Computation) o Agregación Inmediata y Segura**[62,63,64] para limitar la exposición de datos. Se trata de métodos criptográficos que permiten a varias partes calcular datos distribuidos sin revelar puntos de datos individuales. Los cálculos se realizan en parámetros cifrados sin revelar nunca los parámetros.
- El uso de los llamados **Enclaves Seguros o Entornos de Ejecución de Confianza (TEE, Trusted Execution Environments)**[65], que permiten que los datos se procesen dentro de una "pieza de hardware segura" y utiliza protecciones criptográficas para habilitar un entorno informático protegido. Se utilizan, por ejemplo, en transacciones de pago seguras.
- El uso de la **Privacidad Diferencial**. Consiste en añadir ruido a los datos para reducir el riesgo de que cualquier persona pueda ser identificada durante la fase de entrenamiento de los modelos locales.

Debido a la complejidad de la configuración del FL y dado que ningún PET es una solución definitiva, los responsables del tratamiento deben considerar la implementación de medidas de seguridad "clásicas" disponibles para proteger los datos como lo harían en cualquier otra operación de tratamiento, con el fin de minimizar los riesgos.

62 Fair and Secure Multi-Party Computation with Cheater Detection, Minhye Seo, 2021
63 https://securecomputation.org/
64 SMPAI: Secure Multi-Party Computation for Federated Learning, Vaikkunth Mugunthan et al., 2019
65 Trusted Execution Environments: Applications and Organizational Challenges, Tim Geppert et al., 2022

6. Conclusión

El FL ofrece un enfoque prometedor para el aprendizaje automático al permitir que múltiples dispositivos entrenen de forma colaborativa un modelo compartido mientras mantienen los datos (incluidos los datos personales cuando corresponda) descentralizados. Este método es particularmente ventajoso para escenarios que involucran el tratamiento de datos personales sensibles o requisitos regulatorios, ya que mitiga los riesgos de privacidad al garantizar que los datos personales originales permanezcan en los dispositivos locales. Al mantener los datos personales descentralizados, el FL se alinea con los principios básicos de protección de datos, como son la minimización de datos, la responsabilidad proactiva y el de seguridad, así como reduce el riesgo de brechas de datos personales a gran escala.

En un entorno no FL, se necesita una evaluación caso por caso para determinar el riesgo de ataques de reidentificación (como ataques de membresía) en los modelos finales, pero en entornos FL, dicha evaluación también debe realizarse en los modelos locales intercambiados.

El FL presenta desafíos que deben abordarse para garantizar una protección efectiva de los datos. Una de las principales preocupaciones es la posibilidad de fuga de datos a través de las actualizaciones de los modelos, donde los atacantes podrían inferir información a partir de gradientes o pesos compartidos entre los dispositivos y los servidores centrales. Este riesgo, junto con los posibles ataques de inferencia de membresía y la dificultad para detectar y mitigar el sesgo o garantizar la integridad de los datos, pone de manifiesto la necesidad de contar con medidas de seguridad sólidas en todo el ecosistema de FL y la combinación de FL con otros PETs.

7. Lecturas recomendadas

- L. Tian, A. Kumar Sahu, A. S. Talwalkar and V. Smith, **Federated Learning: Challenges, Methods, and Future Directions**, IEEE Signal Processing Magazine 37, 2020.

- Q. Li, W. Zeyi, H. Bingsheng, **A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection**, ArXiv abs/1907.09693, 2021.

- P. Kairouz et al, **Advances and Open Problems in Federated Learning, Foundations and Trends in Machine Learning Vol 4 Issue 1**, 2021.

- Nicole Mitchell and Adam Pearce, **How Federated Learning Protects Privacy**, November 2022 – available at https://pair.withgoogle.com/explorables/federated-learning/

Esta publicación es un breve informe elaborado por la Unidad de Tecnología y Privacidad del Supervisor Europeo de Protección de Datos (SEPD) y por la División de Innovación Tecnológica de la Agencia Española de Protección de Datos (AEPD). Su objetivo es ofrecer una descripción objetiva de una tecnología emergente y analizar sus posibles impactos en la privacidad y en la protección de los datos personales. El contenido de esta publicación no implica una posición política del SEPD.

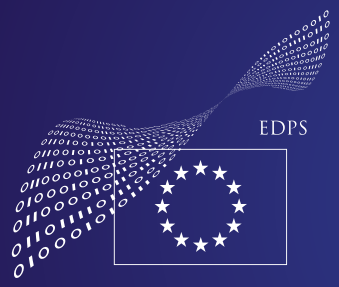
Autores del número: Andy Goldstein, Miguel Peñalba, Luis de Salvador Carrasco

Editores: Luis Velasco, Luis de Salvador Carrasco, Massimo Attoresi y Xabier Lareo

Contacto: techmonitoring@edps.europa.eu

Para suscribirse o darse de baja de las publicaciones de TechDispatch, envíe un correo a: techmonitoring@edps.europa.eu.

La nota informativa sobre protección de datos está disponible en el sitio web del SEPD.



edps.europa.eu

www.aepd.es