Un estudio comparativo de métodos de MLOps para la protección de datos en aprendizaje federado

NAVARRETE-VILLEGAS Martín martin.navarretevl@uanl.edu.mx

Universidad Autónoma de Nuevo León

Introducción

- **Problema Central:** A pesar de la creciente relevancia del aprendizaje federado, la integración de prácticas de MLOps en estos entornos para fortalecer la seguridad y privacidad de los datos no ha sido suficientemente explorada. Este vacío en la investigación destaca la necesidad de un estudio detallado y comparativo que aborde específicamente esta intersección.

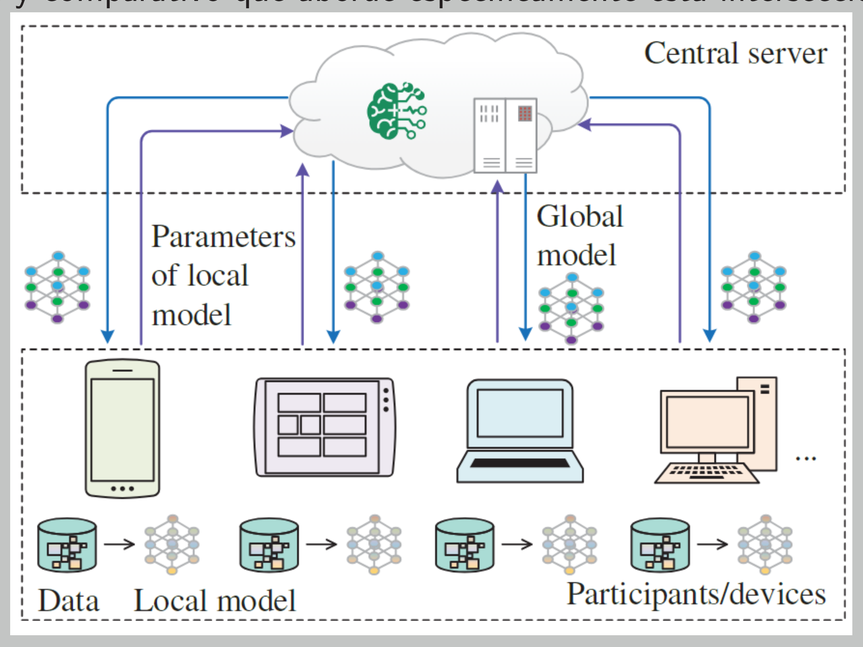

Figura 1:Bin, Wenjie, Yihan, Chen y Yu [1] (2021) Diagrama de arquitectura general de un sistema de aprendizaje federado

- **Objetivos:** Investigar cómo las metodologías de MLOps pueden ser aplicadas efectivamente en entornos de aprendizaje federado para mejorar la protección de datos. Así como proporcionar un análisis comparativo de diferentes métodos de MLOps y evaluar su impacto en la privacidad y seguridad de los datos en el aprendizaje federado.
- **Contribución a la Comunidad Académica y Profesional:** Este estudio aspira a contribuir tanto teórica como prácticamente al campo de la ciencia de datos. Se espera que los insights generados sean de valor para la comunidad académica, así como para los profesionales que operan en el ámbito del aprendizaje automático y la protección de datos.

Metodología

- **Selección de Herramientas y Plataformas**
  - La cantidad de datos usadas en entrenamiento determina la efectividad del sistema [4], es por eso que en este estudio emplea datos de libre acceso proporcionados por Microsoft a través de Kaggle bajo el nombre Medical Imaging y liga: https://github.com/Azure/medical-imaging

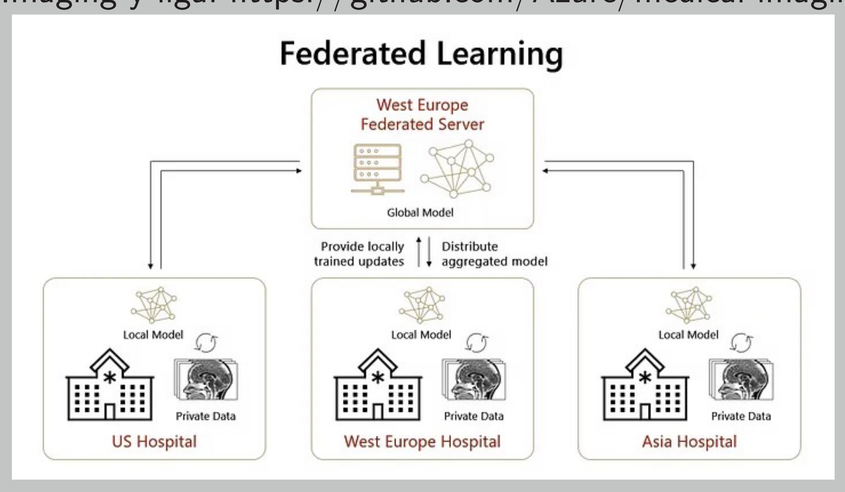

Figura 2:Moreno, C. [2] (2018). Federated Learning with Azure Machine Learning. [Figura 1]. Recuperado de https://caiomsouza.medium.com/federated-learning-with-azure-machine-learning-a774f8df309b

  - Se presentan tablas comparativas detallando las características y capacidades de los cinco tipos de entornos ('sandboxes') analizados: Minimal sandbox, Eyes-on sandbox, Eyes-off sandbox, Confidential VM sandbox, Configurable sandbox
  - Los criterios de selección incluyen precisión, eficiencia, seguridad, y otros atributos relevantes para la protección de datos.
    1. Eficiencia computacional, medida en tiempo de entrenamiento y recursos utilizados.
    2. Seguridad y privacidad de los datos, evaluadas a través de métricas como el nivel de encriptación.
    3. Calidad del modelo, evaluada mediante métricas como la precisión y el F1-score.

Resultados: Implementación y Evaluación de Precisión

- La Figura 3 presenta el flujo de trabajo de nuestro sistema de aprendizaje federado con MLOps integrados.

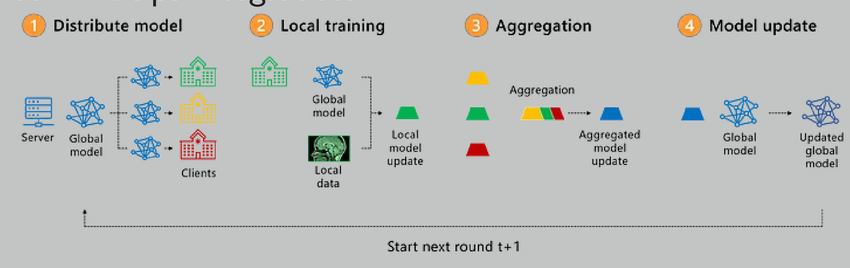

Figura 3:Moreno, C. [2] (2018). Flujo de Aprendizaje Federado. Recuperado de https://caiomsouza.medium.com/federated-learning-with-azure- machine-learning-a774f8df309b

  - Múltiples hospitales entrenan un modelo central, usando un servidor para agregación y orquestación con NVIDIA Flare. Los hospitales están representados por Workspaces en Azure a nivel mundial.
- MLOps en esta investigación automatiza la compilación, ejecución, prueba y despliegue de entrenamientos, datos, pesos y modelos. Se presentan ejemplos de comandos orquestados por el mecanismo de MLOps.

Server Commands

SCOPE	COMMAND	DESCRIPTION
authz	eval_right	check if a user has a right on a site
authz	eval_rule	evaluate a site against a rule
authz	show_config	show authorization config
authz	show_info	show general info of authorization policy
authz	show_rights	show rights configured for authorization
authz	show_rules	show rules configured for authorization
authz	show_sites	show sites configured for authorization
authz	show_users	show users configured for authorization
info	reset_errors	reset errors
info	show_errors	show latest errors
info	show_stats	show current system stats
sys	cat	show content of a file
sys	env	show system environment vars
sys	grep	search for PATTERN in a file.
sys	head	print the first 10 lines of a file
sys	ls	list files in work dir
sys	pwd	print the name of work directory
sys	sys_info	get the system info
sys	tail	print the last 10 lines of a file
training	abort	abort the FL app
training	abort_task	abort the client current task execution
training	check_status	check status of the FL server/client
training	delete_run_number	delete a run
training	deploy_app	deploy FL app to client/server
training	remove_client	remove a FL client
training	restart	restart the FL server/client
training	set_run_number	set the run number
training	set_timeout	set the admin commands timeout
training	shutdown	shutdown the FL server/client
training	start_app	start the FL app

Figura 4:Moreno, C. [2] (2018). Comandos NVFlare. Recuperado de https://caiomsouza.medium.com/federated-learning-with-azure- machine-learning-a774f8df309b

  - La Tabla 1 ilustra la matriz de valores normalizados para una serie de datos en un ambiente de aprendizaje federado. Hay al menos diez nodos que aportan datos al modelo, que será procesado por métodos MLOps.

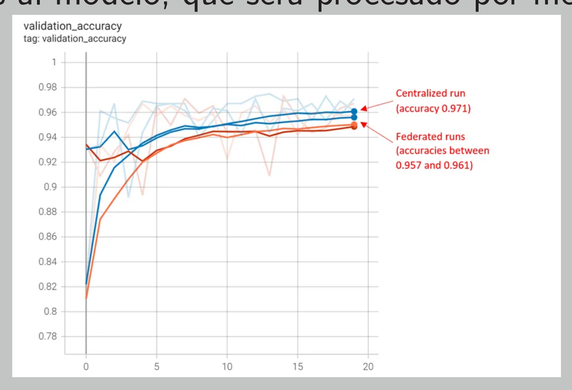

Figura 5:Ghorpade, P. (2020). Medical Image Classification using Azure, Validation Accuracy.

Conclusiones

- La eficiencia computacional emerge como el factor más crucial, impactando directamente la viabilidad del aprendizaje federado en el mundo real.
- Ciertas combinaciones de métricas favorecen consistentemente a un método de MLOps, destacando su óptimo rendimiento en diversos atributos evaluados.

Referencias

1 Bin Yu, Wenjie Mao, Yihan Lv, Chen Zhang y Yu Xie. A survey on federated learning in data mining. Agosto 2021; 3. pp 6

2 Andreas Kopp. Practical Federated Learning with Azure Machine Learning. Agosto 2022; Towards Data Science: pp. 1

3 P. P. Shinde, A review of machine learning and deep learning applications, (2018), 1-6.

4 A. L Heureux, K. Grolinger, H. F. Elyamany and M. A. M. Capretz, Machine learning with big data: Challenges and approaches, IEEE Access, 5 (2017), 7776-7797.