Mitigando vulnerabilidades del Aprendizaje Federado Mediante Blockchain

Mario García Márquez

Tutores: Francisco Herrera Triguero y Nuria Rodríguez Barroso

UNIVERSIDAD DE GRANADA

Motivación

Aprendizaje Federado

Blockchain

Optimización no lineal

Operador Krum

Blockchain aplicado al Aprendizaje Federado

Krum Federated Chain

Conclusiones y Trabajo Futuro

MOTIVACIÓN

Contexto

DÍA MUNDIAL DEL MEDIO AMBIENTE

Inteligencia artificial al rescate de la naturaleza

ÓSCAR GRANADOS

La nueva tecnología se abre paso en diversos ámbitos para mejorar su eficiencia y reducir la emisión de gases contaminantes

MOTIVACIÓN

Contexto

DÍA MUNDIAL DEL MEDIO AMBIENTE

Inteligencia artificial al rescate de la naturaleza

ÓSCAR GRANADOS

La nueva tecnología se abre paso en diversos ámbitos para mejorar su eficiencia y reducir la emisión de gases contaminantes

EXTRA ENERGÍA

Algoritmos contra el derroche energético

MIGUEL ÁNGEL GARCÍA VEGA

La inteligencia artificial aplicada a las energías limpias tiene la enorme facilidad de encontrar fallos, errores o ineficacias en el sistema energético

MOTIVACIÓN

Contexto

DÍA MUNDIAL DEL MEDIO AMBIENTE

Inteligencia artificial al rescate de la naturaleza

ÓSCAR GRANADOS

La nueva tecnología se abre paso en diversos ámbitos para mejorar su eficiencia y reducir la emisión de gases contaminantes

EXTRA ENERGÍA

Algoritmos contra el derroche energético

MIGUEL ÁNGEL GARCÍA VEGA

La inteligencia artificial aplicada a las energías limpias tiene la enorme facilidad de encontrar sistema energético

INTELIGENCIA ARTIFICIAL

¿Reducirá la inteligencia artificial nuestras capacidades?

ALICIA TRONCOSO

Las personas no debemos solo usar bien la IA, sino que debemos comprender sus limitaciones, sus riesgos y desarrollar habilidades que la complementen

MOTIVACIÓN

Contexto

DÍA MUNDIAL DEL MEDIO AMBIENTE

Inteligencia artificial al rescate de la naturaleza

ÓSCAR GRANADOS

La nueva tecnología se abre paso en diversos ámbitos para mejorar su eficiencia y reducir la emisión de gases cc

ESTUDIO

E Los jóvenes también temen perder su trabajo por la inteligencia artificial

MANME GUERRA

Los nacidos a partir de 1995 consideran que podría influir en sus decisiones laborales y optarían por roles menos vulnerables a la automatización

MIGUEL ÁNGEL GARCÍA VEGA

La inteligencia artificial aplicada a las energías limpias tiene la enorme facilidad de encontrar fallos, errores o ineficacias en el sistema energético

INTELIGENCIA ARTIFICIAL

¿Reducirá la inteligencia artificial nuestras capacidades?

ALICIA TRONCOSO

Las personas no debemos solo usar bien la IA, sino que debemos comprender sus limitaciones, sus riesgos y desarrollar habilidades que la complementen

Contexto

DÍA MUNDIAL DEL MEDIO AMBIENTE

Inteligencia artificial al rescate de la naturaleza

ÓSCAR GRANADOS

La nueva tecnología se abre paso en diversos ámbitos para mejorar su eficiencia y reducir la

INTELIGENCIA ARTIFICIAL

E Europa apuesta por una inteligencia artificial que no alucine, industrial, fiable y menos costosa

RAÚL LIMÓN

Las empresas de la UE desconfían de los grandes modelos de lenguaje para aplicaciones conversacionales y el desarrollo de agentes

La inteligencia artificial aplicada a las energías limpias tiene la enorme facilidad de encontrar fallos, errores o ineficacias en el sistema energético

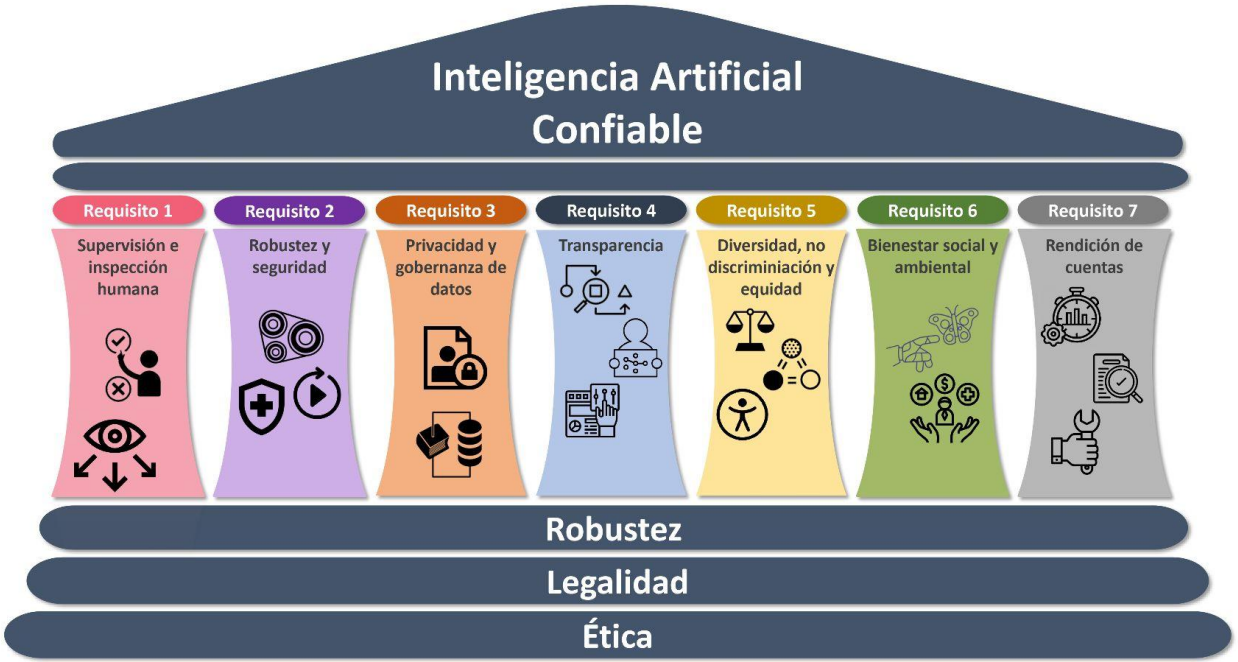
INTELIGENCIA ARTIFICIAL

¿Reducirá la inteligencia artificial nuestras capacidades?

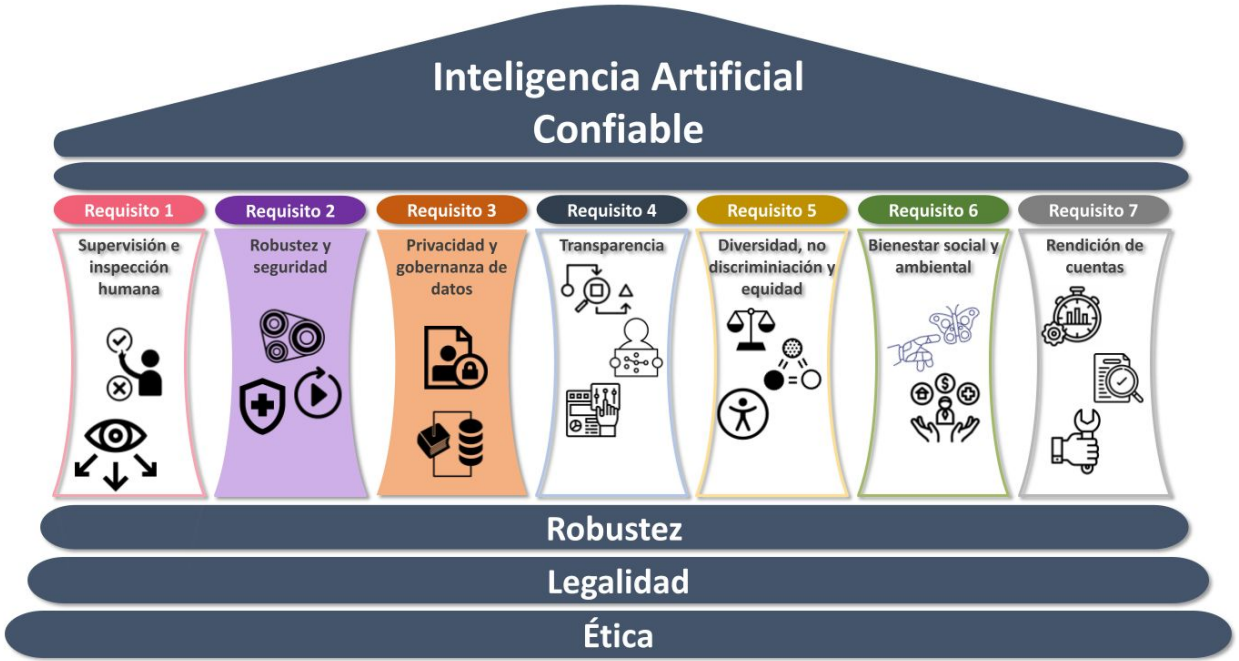
ALICIA TRONCOSO

Las personas no debemos solo usar bien la IA, sino que debemos comprender sus limitaciones, sus riesgos y desarrollar habilidades que la complementen

IA Confiable



IA Confiable



Motivación

Aprendizaje Federado

Blockchain

Optimización no lineal

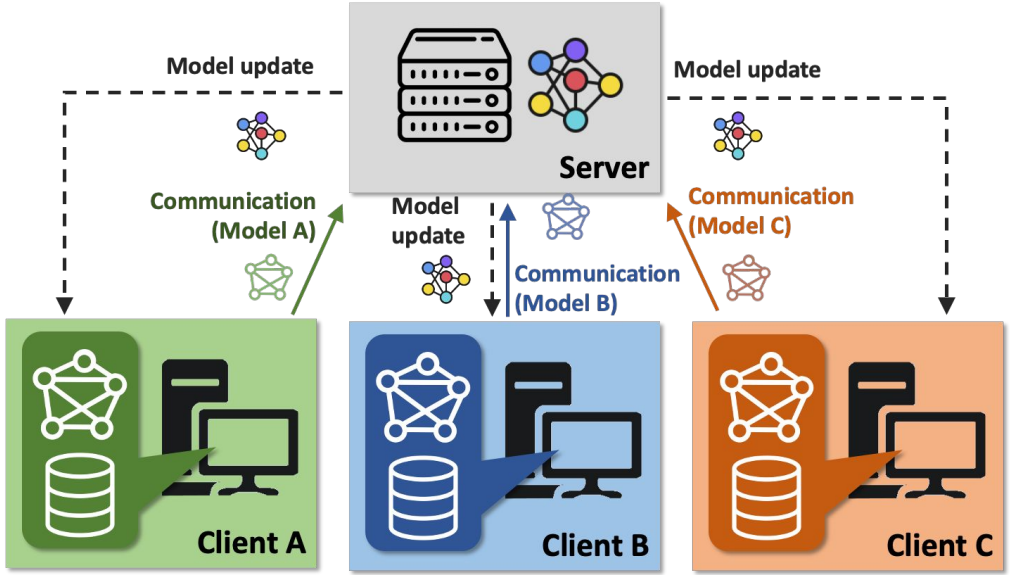
Operador Krum

Blockchain aplicado al Aprendizaje Federado

Krum Federated Chain

Conclusiones y Trabajo Futuro

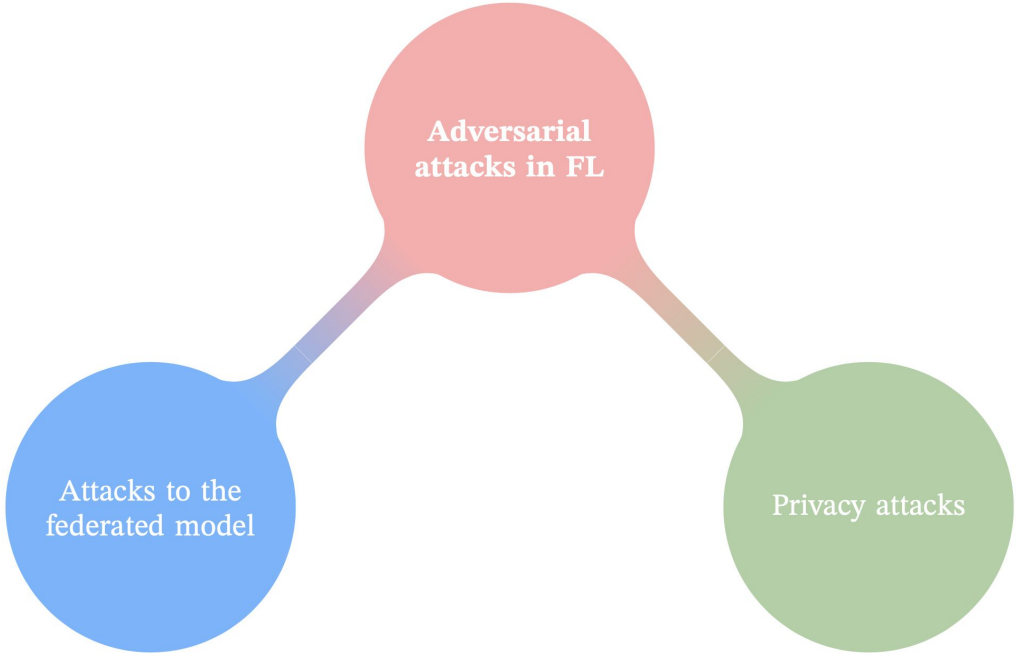
Aprendizaje Federado: Enfoque ML en el que un modelo se entrena a través de **dispositivos descentralizados**, permitiendo que los datos permanezcan en los dispositivos locales mientras el modelo global se actualiza de forma **colaborativa**.



¡Los datos no abandonan los dispositivos!

- Costes de comunicación.
- Robustez.
- **Privacidad.**

Ataques adversarios



- **Ataques al modelo federado,** buscan modificar su comportamiento.

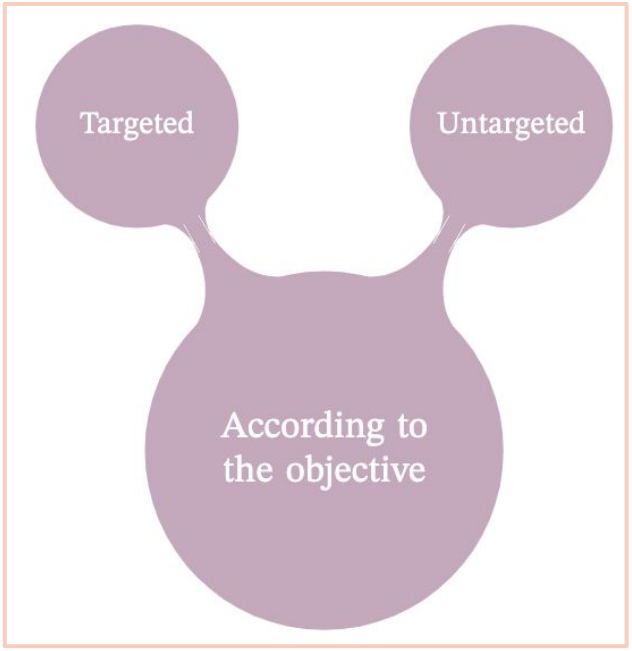
- **Ataques a la privacidad,** su objetivo es inferir información sensible mediante los datos en el entrenamiento.

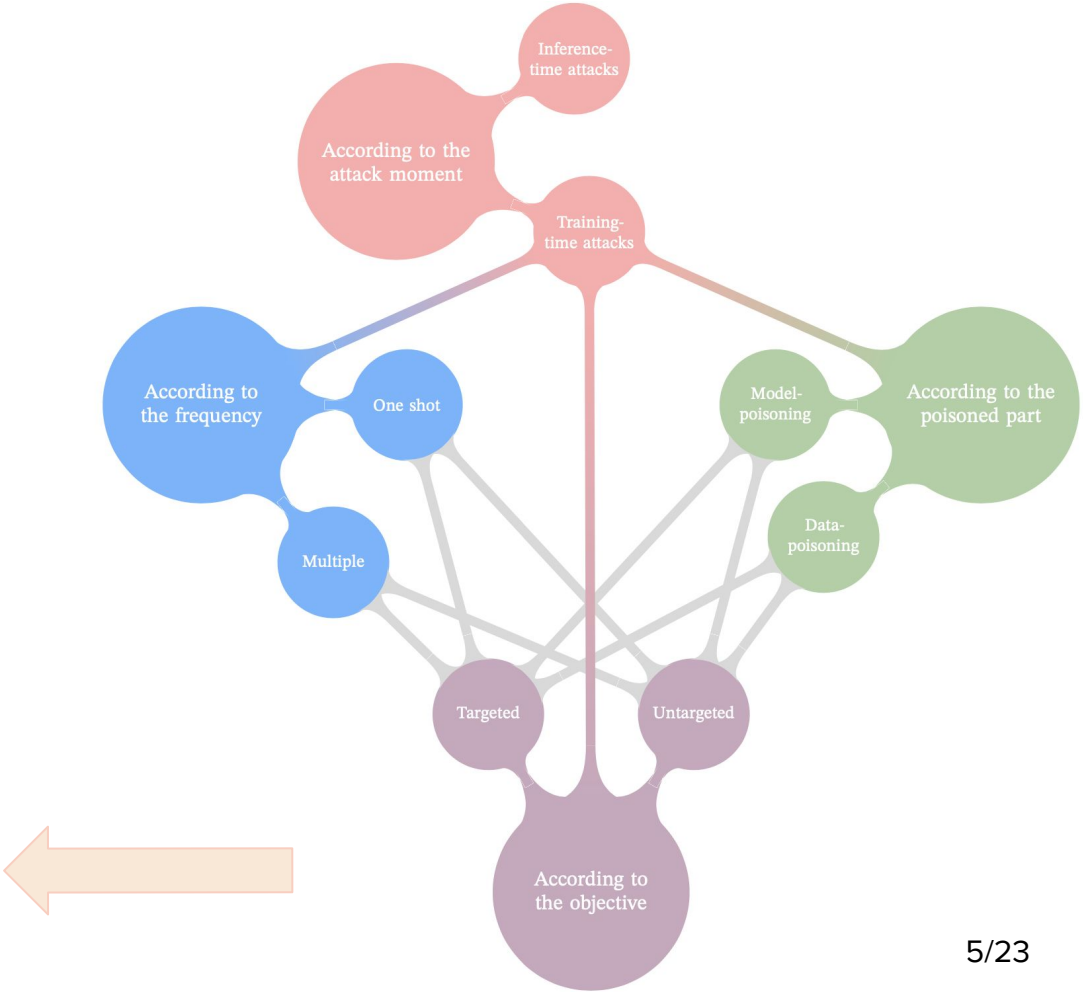
Ataques adversarios al modelo

Nos centraremos en los **ataques adversarios al modelo**.

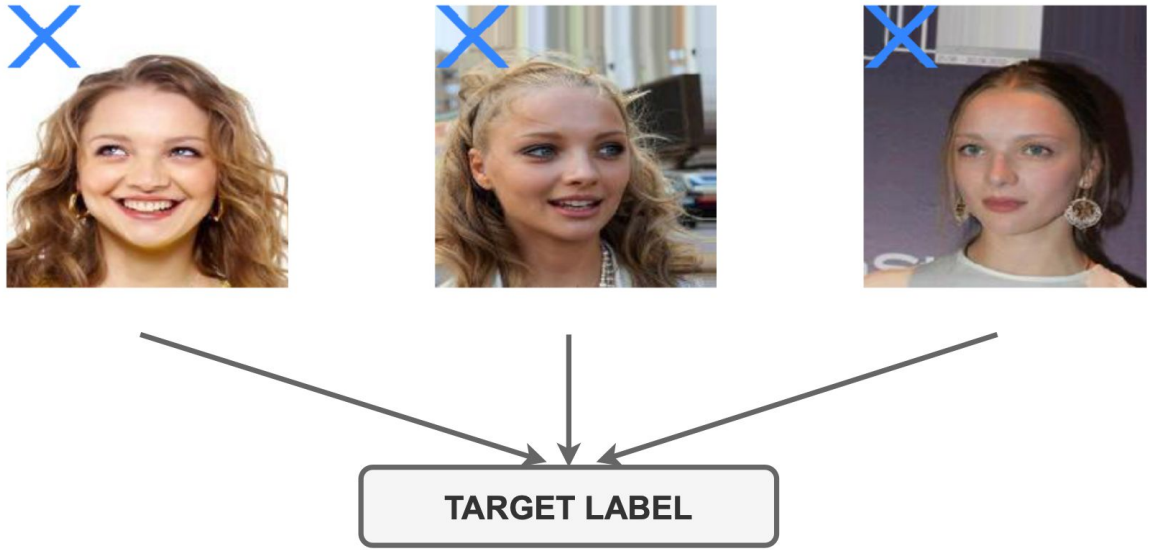
APRENDIZAJE FEDERADO

Ataques adversarios al modelo

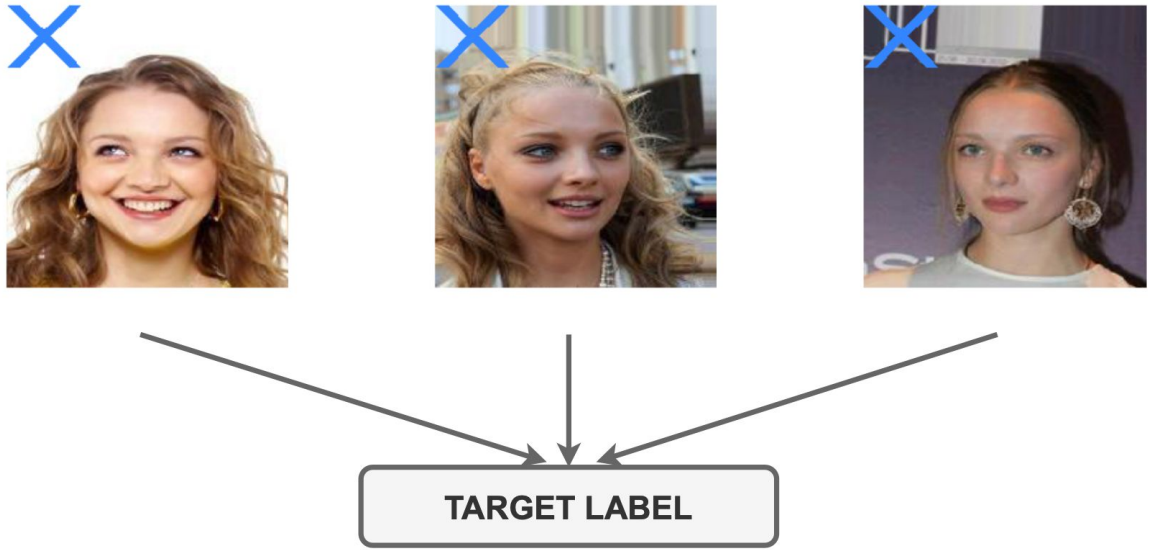




Dirigidos/*Backdoor*: El objetivo es el de inyectar una tarea secundaria en el modelo.



Dirigidos/*Backdoor*: El objetivo es el de inyectar una tarea secundaria en el modelo.



No Dirigidos/Bizantinos: El objetivo es únicamente el de degradar el rendimiento del modelo.

Tecnología para **registrar transacciones** y procesamiento, que protege de pérdida o alteración de la información, y **sin una entidad central** en la que confiar.

Para realizar una **comunicación fiable** entre los nodos y mantener el **estado correcto** del sistema, se usan los mecanismos de consenso. Los más populares son:

- ***Proof of Work***.
- ***Proof of Stake***.

Entrenar un modelo que aproxime una función. Intentamos **minimizar** el error. El problema a tratar es el siguiente:

$$\min_{x \in \mathcal{D}} f(x).$$

Donde

$$f : \mathcal{D} \subset \mathbb{R}^n \to \mathbb{R}$$

y

$$f \in C^1(\mathcal{D}).$$

Uno de los resultados más importantes de la teoría de optimización es la conocida **condición de primer orden**. Si un elemento es un mínimo, entonces el gradiente de la función en ese punto es 0.

$$\nabla f(x^*) = 0$$

Por lo tantos muchos métodos se dedican a **encontrar aquellos puntos en los que el gradiente se anula**.

Un método para esto es el **descenso por el gradiente**. Da una sucesión de puntos para que el gradiente converja a 0.

$$x^{k+1} = x^k - \gamma_k \nabla f(x^k)$$

Se necesita que el gradiente sea Lipschitz:

$$||\nabla f(x) - \nabla f(y)|| \leq L||x - y||$$

y que la función esté acotada inferiormente:

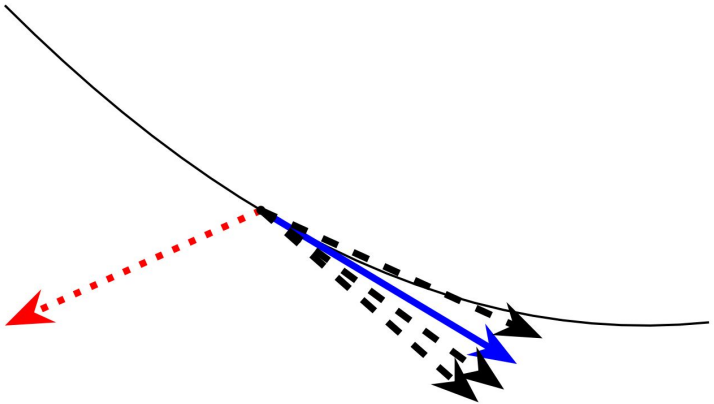
$$f(x) \geq f^* > -\infty.$$

Es inviable en caso de que la dimensión sea muy alta. Se usa el **descenso estocástico por el gradiente**. En lugar de usar el gradiente se usa un **estimador insesgado** de este.

$$\hat{g} = \frac{1}{m} \nabla_\theta \sum_i L(f(x^{(i)}; \theta), y^{(i)})$$

El problema de agregar vectores

Obtener una “buena” **estimación de un gradiente** dadas una serie de estimaciones previas. Entre estas estimaciones pueden estar presentes ***outliers*.**



Cualquier combinación lineal de estos vectores **no es un mecanismo robusto ante *outliers***. Se puede dar un *outlier* tal que el resultado sea un vector arbitrario.

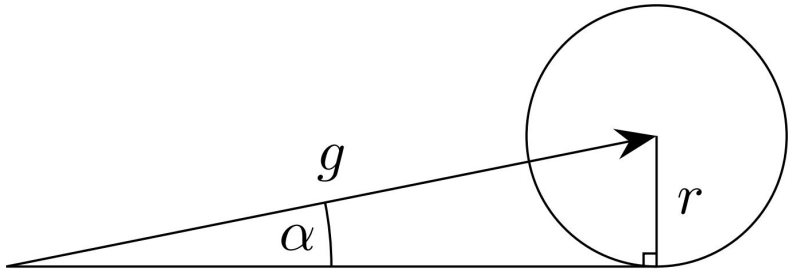
$$F_{lin}(V_1, \ldots, V_n) = \sum_{i=1}^{n} \lambda_i \cdot V_i$$

$$V_n = \frac{1}{\lambda_n} \cdot U - \sum_{i=1}^{n-1} \frac{\lambda_i}{\lambda_n} V_i$$

Definición 40. Sea $0 \leq \alpha < \pi/2$ cualquier ángulo, y cualquier entero $0 \leq f \leq n$. Sean $V_1, \ldots, V_n$ vectores aleatorios i.i.d. en $\mathbb{R}^d$, $V_i \sim G$, con $E[G] = g$. Sean $B_1, \ldots, B_f$ vectores aleatorios en $\mathbb{R}^d$, posiblemente dependientes de los vectores V_i. Diremos que una regla de agregación F es (α, f)-Resistente bizantina si, para cualesquiera $1 \leq j_1 < \ldots < j_f \leq n$, el vector

$$F = F(V_1, \ldots, \underbrace{B_1}_{j_1}, \ldots, \underbrace{B_f}_{j_f}, \ldots, V_n) \tag{86}$$

cumple que $\langle E[F], g\rangle \geq (1 - \sin\alpha) \cdot ||g||^2 > 0$ y que para $r \in \{2, 3, 4\}$, $E[||F||^r]$ está superiormente acotada por una combinación linear de los términos $E[||G||^1], \ldots, E[||G||^{r_{n-1}}]$ con $r_1 + \ldots + r_{n-1} = r$.



Se introduce el operador de Krum. A cada vector se le asigna una puntuación:

$$s(i) = \sum_{i \to j} ||V_i - V_j||^2.$$

El resultado la regla de agregación es aquel vector que minimize la puntuación.

$$KR(V_1, \ldots, V_n) = V_{i_*}$$

El operador de Krum es (α,f)-resistente en condiciones bastantes generales.

Ventajas

Algunos de los motivos para **combinar** las tecnologías blockchain con el aprendizaje federado han sido:

- Mayor escalabilidad.
- No depende de ninguna entidad central.
- Heterogeneidad de los sistemas.
- Falta de incentivos.

Algunos de los motivos para **combinar** las tecnologías blockchain con el aprendizaje federado han sido:

- Mayor escalabilidad.
- No depende de ninguna entidad central.
- Heterogeneidad de los sistemas.
- Falta de incentivos.

Para ello se crea ***Proof of Federated Learning (PoFL)*.**

- *Pooled-mining*.
- Basado en rendimiento.
- Eficiencia energética.

PoFL puede ser un **mecanismo de defensa**.

- *Pooled-mining*.
- Basado en rendimiento.
- Los atacantes reducen el rendimiento.
- Existe un minero sin atacar.

Propuesta: Krum Federated Chain

PoFL puede ser un **mecanismo de defensa**.

- *Pooled-mining*.
- Basado en rendimiento.
- Los atacantes reducen el rendimiento.
- Existe un minero sin atacar.

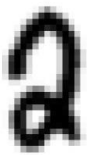
Proponemos **Krum Federated Chain (KFC).**

- PoFL.
- Krum.
- Los atacantes son outliers.

Comparamos varias arquitecturas ante varios ataques.

Bizantino, *label flipping*.

Etiqueta: ~~9~~ 2



Etiqueta: ~~2~~ 7

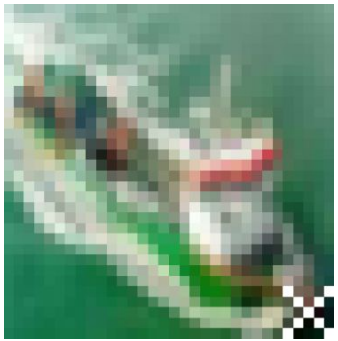
Backdoor, basado en patrones.

Comparamos varias arquitecturas ante varios ataques.

Bizantino, *label flipping*.

- Medimos la precisión.
- Más es mejor.

Backdoor, basado en patrones.



Comparamos varias arquitecturas ante varios ataques.

Bizantino, *label flipping*.

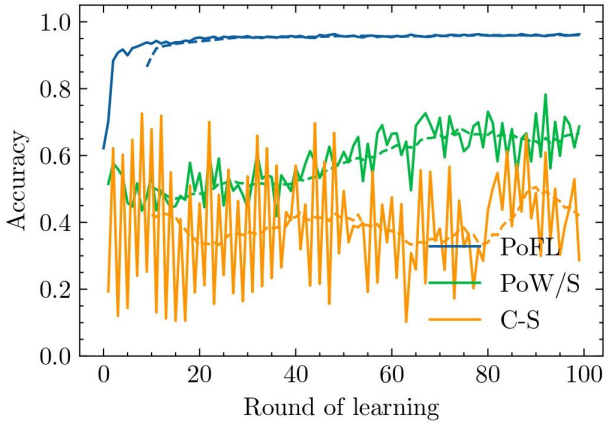
- Medimos la precisión.
- Más es mejor.

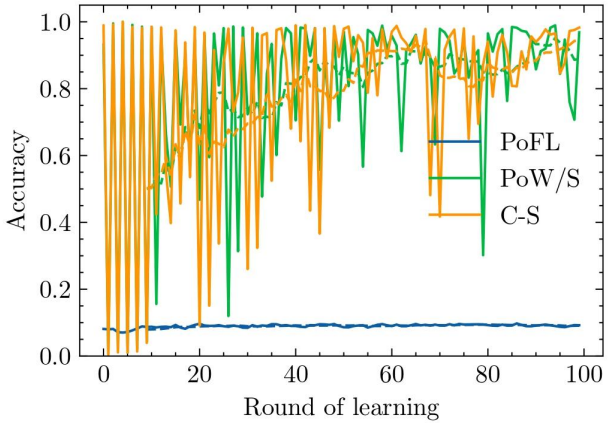
Backdoor, basado en patrones.

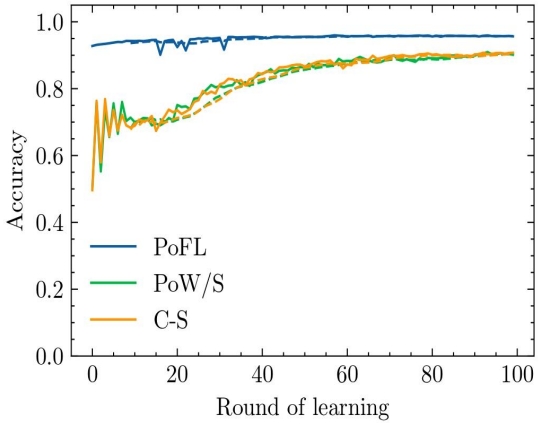
- Medimos la precisión.
- Dos tareas.
  - Original.
  - Inyectada.
- Más es mejor en la original.
- Menos es mejor en la inyectada.

Resultados PoFL

PoFL demuestra ser **válido** bajo hipótesis.

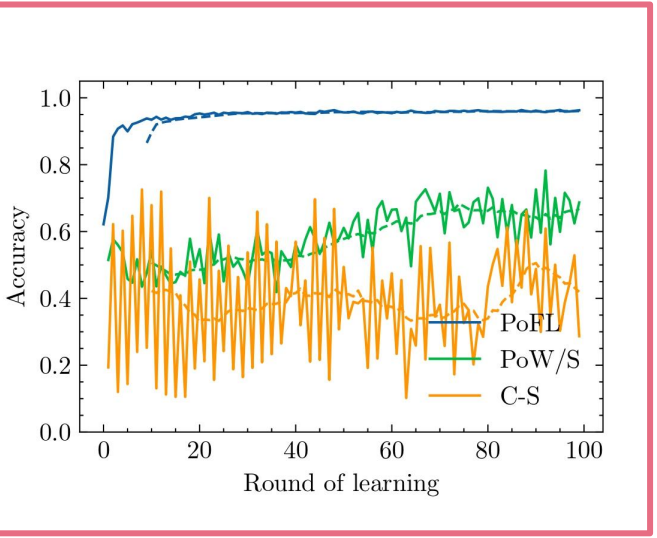


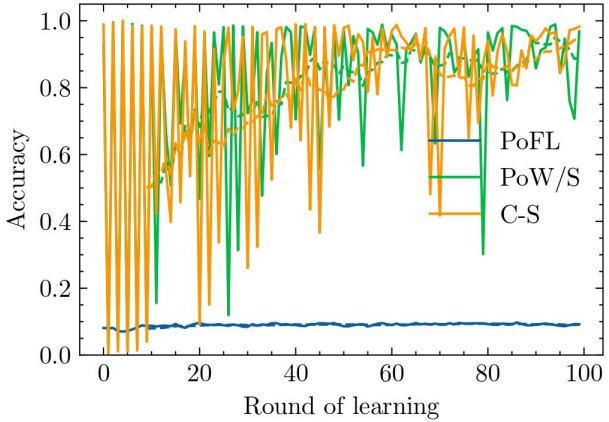


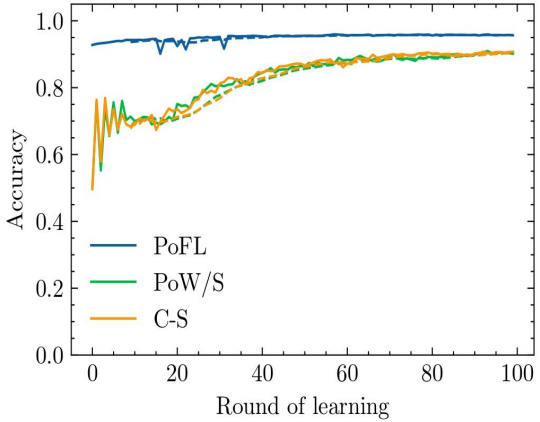


Resultados PoFL

PoFL demuestra ser **válido** bajo hipótesis.



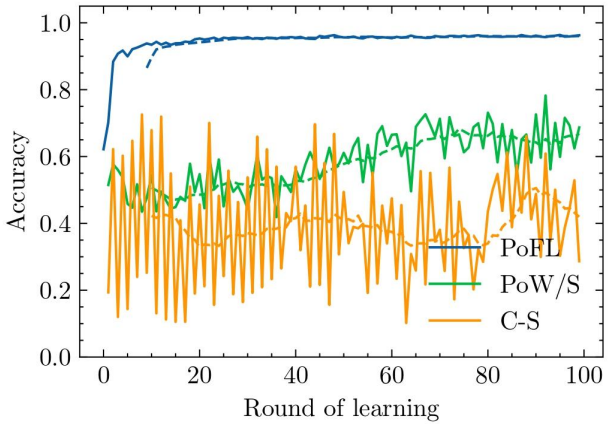


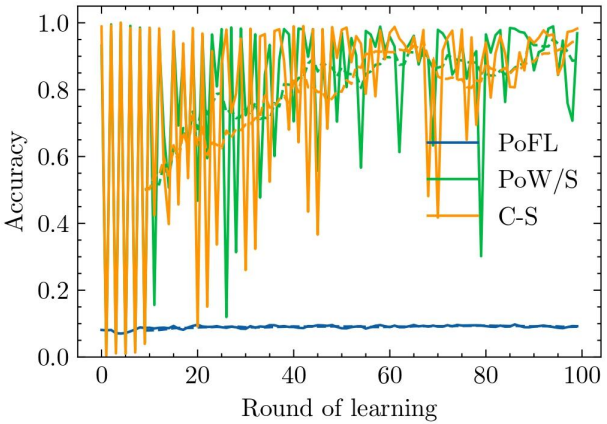


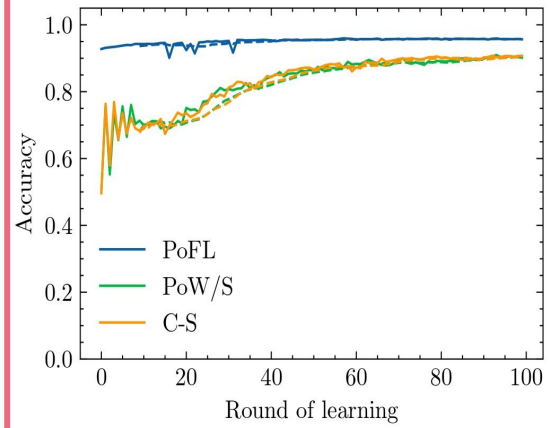
Resultados en ataque **bizantinos**.

Resultados PoFL

PoFL demuestra ser **válido** bajo hipótesis.



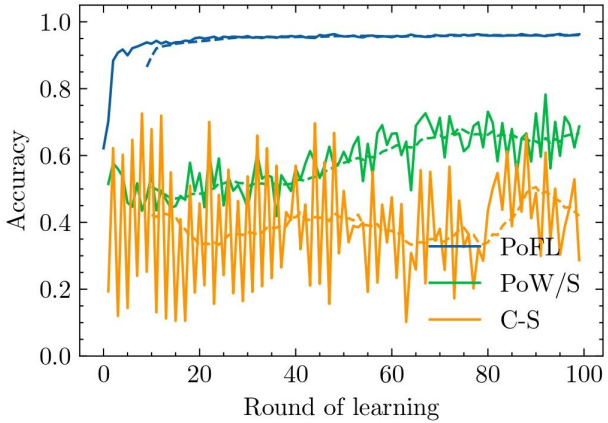


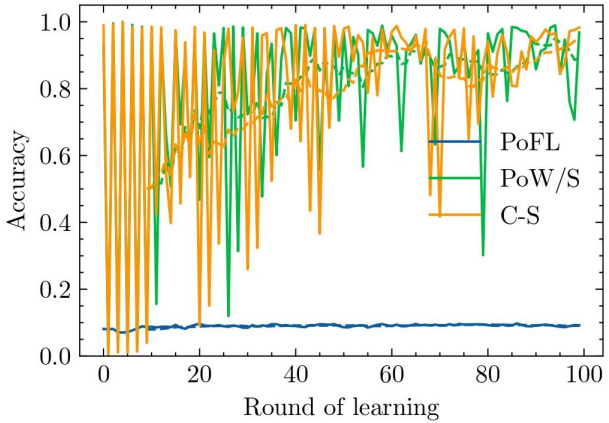


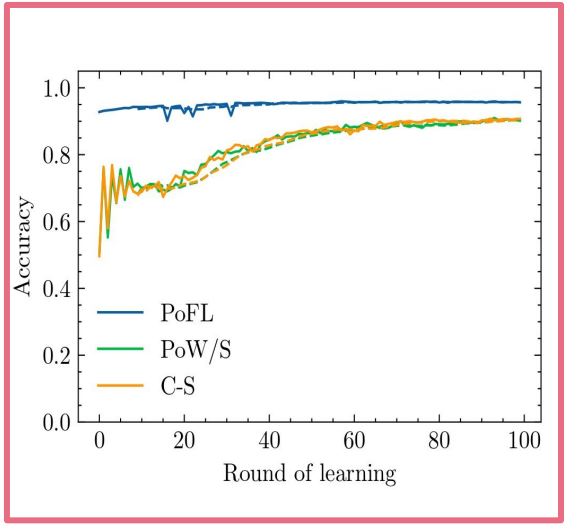
Resultados en **tarea inyectada** *backdoor*.

Resultados PoFL

PoFL demuestra ser **válido** bajo hipótesis.



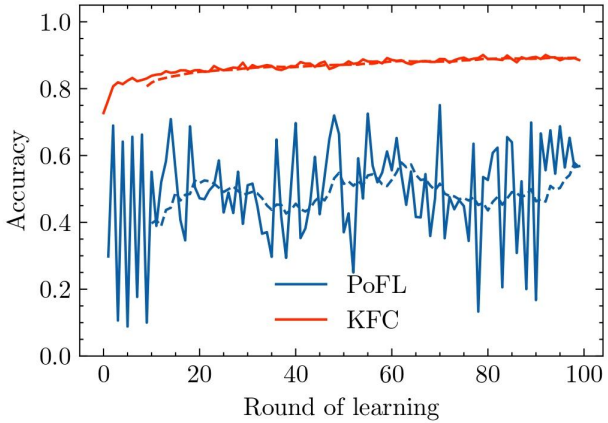


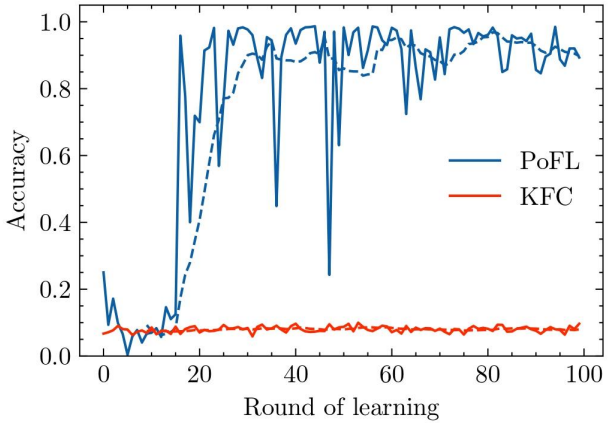


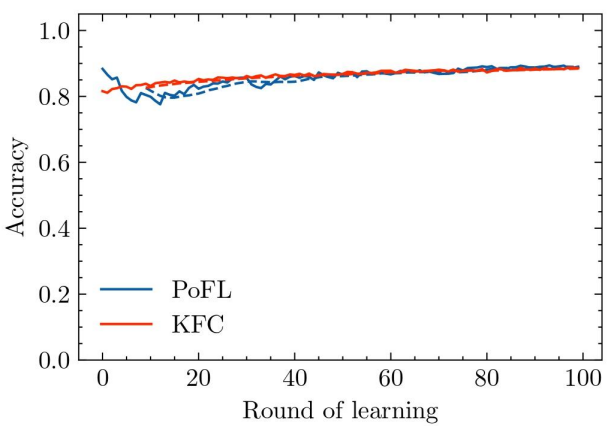
Resultados en la **tarea original** durante *backdoor*.

Resultados KFC

Nuestra propuesta **mejora** a PoFL.

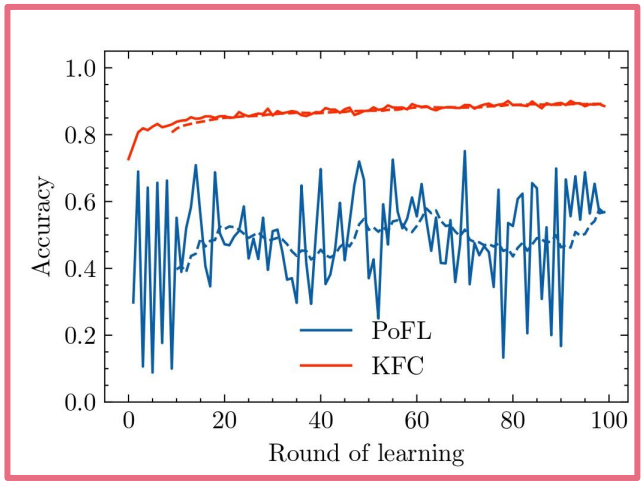


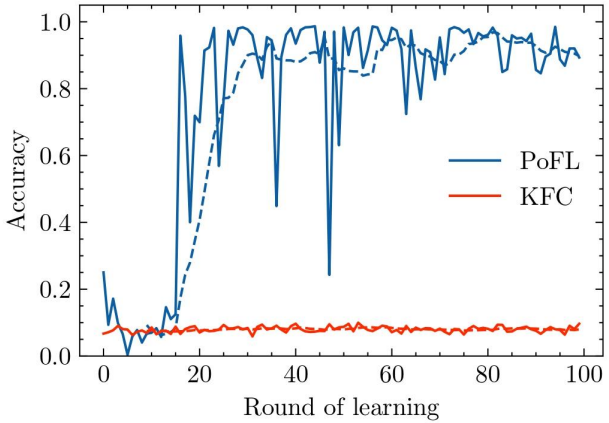


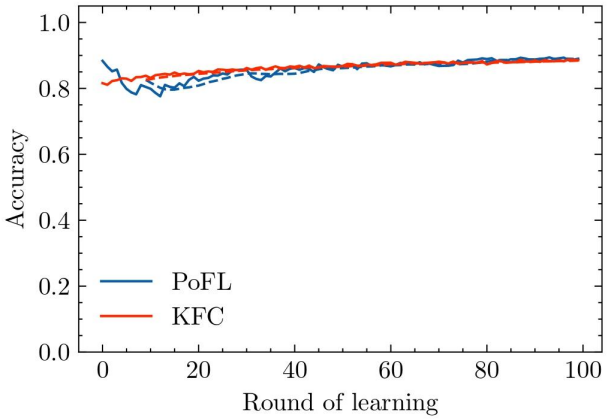


Resultados KFC

Nuestra propuesta **mejora** a PoFL.



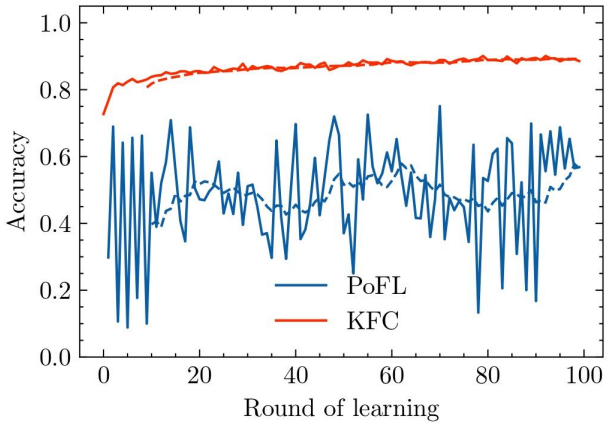


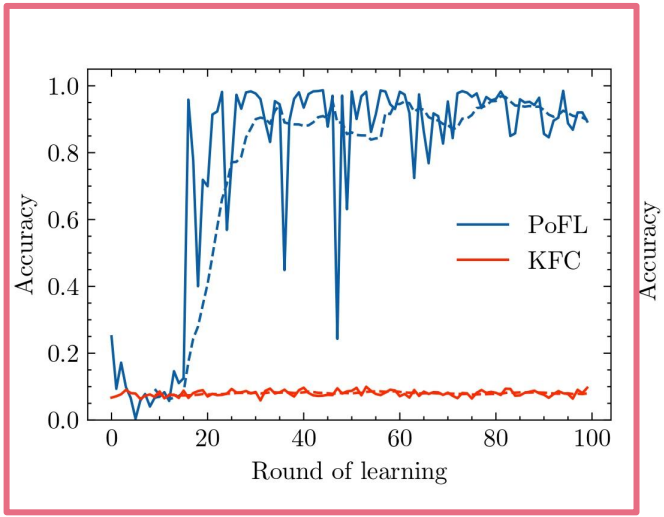


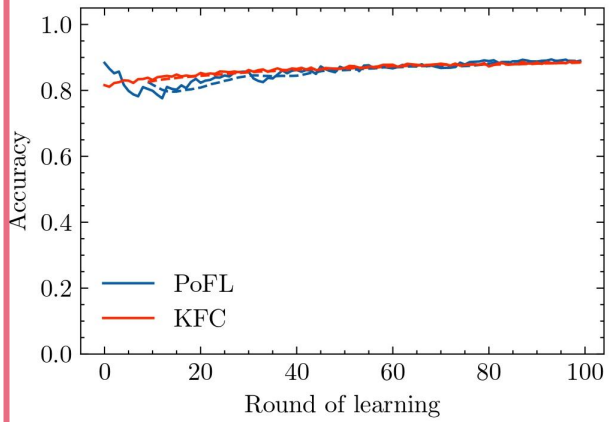
Resultados en ataque **bizantinos**.

Resultados KFC

Nuestra propuesta **mejora** a PoFL.



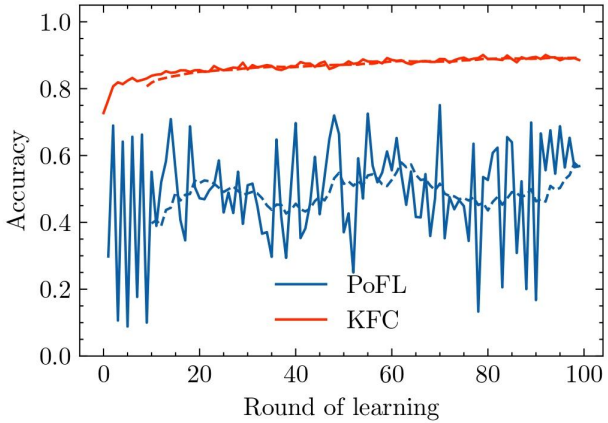


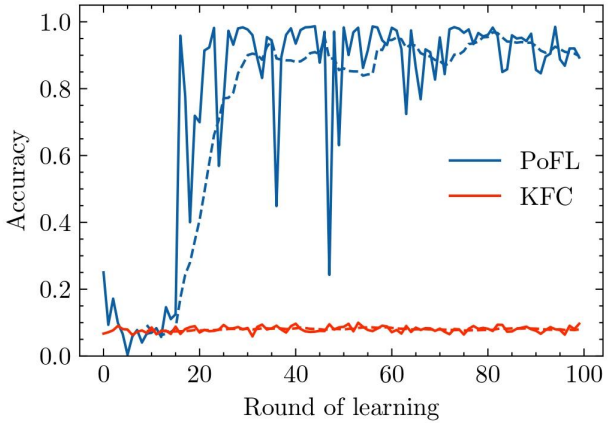


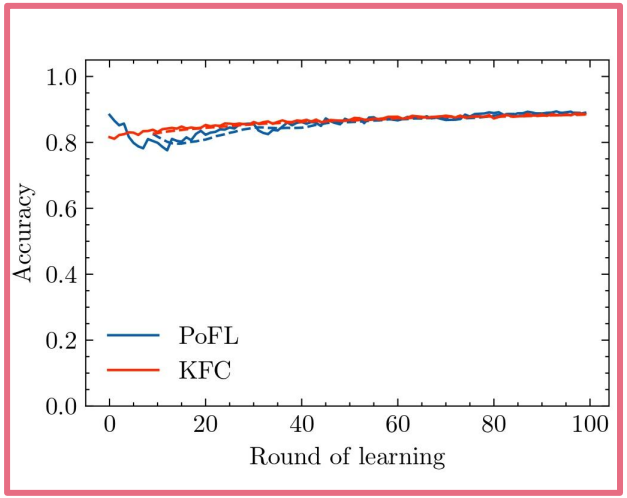
Resultados en **tarea inyectada** *backdoor*.

Resultados KFC

Nuestra propuesta **mejora** a PoFL.







Resultados en la **tarea original** durante *backdoor*.

Motivación

Aprendizaje Federado

Blockchain

Optimización no lineal

Operador Krum

Blockchain aplicado al Aprendizaje Federado

Krum Federated Chain

Conclusiones y Trabajo Futuro

- Se ha estudiado la combinación de **Blockchain** y **Aprendizaje Federado.**
- Se ha analizado y estudiado los **fundamentos matemáticos** de los mecanismos de defensa en el Aprendizaje Federado.
- Se ha analizado *Proof of Federated Learning* como **mecanismo de defensa**.
- Se ha propuesto KFC, un **mecanismo novedoso de defensa** mejorando a PoFL.
- Se ha contribuido a la **plataforma FLEX** de DaSCI con un módulo de *blockchain*.
- Se han presentado los resultados obtenidos a **congreso ECAI**.

Trabajo Futuro

- Estudiar **mecanismos alternativos** a Krum.
- Probar más ataques, como aquellos a la **privacidad**.
- Ataques en **tiempo de inferencia**.
- Mejor **integración** entre blockchain y aprendizaje federado.

Muchas Gracias por su atención :)

Mario García Márquez

Tutores: Francisco Herrera Triguero y Nuria Rodríguez Barroso

UNIVERSIDAD DE GRANADA