Aprendizaje federado con modelos de reglas descriptivas supervisadas para extraer conocimiento explicable y confiable

María Asunción Padilla Rascón[1] mprascon@ujaen.es, Cristóbal J. Carmona del Jesus, and Ángel Miguel García Vico

Universidad de Jaén, Departamento de Informática, Jaén, España
Fecha de inicio: Enero 2025
Fecha estimada de defensa: Enero 2029
Programa FPU: FPU23/02620

Resumen El Aprendizaje Federado ha emergido como un paradigma de especial interés para entrenar modelos de aprendizaje automático en entornos distribuidos sin compartir datos sensibles. Sin embargo, su adopción en sectores críticos como la sanidad, la justicia o la administración pública exige dos propiedades que la literatura actual no aborda de forma conjunta: la explicabilidad intrínseca de los modelos resultantes y la seguridad criptográfica de las comunicaciones entre nodos. Esta tesis doctoral propone el diseño de arquitecturas avanzadas de aprendizaje federado que integren modelos de reglas descriptivas supervisadas, mediante computación flexible, junto con mecanismos criptográficos, con el fin de construir sistemas de inteligencia artificial confiables, explicables y seguros en entornos distribuidos, de datos complejos y volátiles. La propuesta aborda desde el diseño de nuevos modelos de reglas para flujos de datos continuos hasta arquitecturas FL centralizadas y descentralizadas con garantías de privacidad y auditabilidad. Los resultados obtenidos hasta la fecha evidencian la viabilidad de este enfoque y muestran mejoras en interpretabilidad, simplicidad de los modelos y preservación de la privacidad, validadas en escenarios reales de ámbitos médico e industrial.

Keywords: Aprendizaje Federado · Reglas Descriptivas Supervisadas · IA Confiable · Flujos de Datos · Seguridad Criptográfica

1. Contexto y motivación

El auge del *big data* y la proliferación de dispositivos conectados (*IoT, wearables, smartphones*) ha transformado radicalmente la generación y el consumo de información, dando lugar a entornos donde los datos fluyen de manera continua, distribuida y heterogénea [11]. La cuarta revolución industrial (Industria 4.0), caracterizada por avances en robótica, computación en la nube e inteligencia artificial (IA), ha acelerado esta tendencia hasta el punto de que expertos e instituciones como la Unión Europea ya anticipan una Industria 5.0, centrada en la transformación hacia entornos inteligentes basados en el Internet de las Cosas y la computación cognitiva. Estos desarrollos generan enormes volúmenes de datos

que, procesados adecuadamente, pueden convertirse en conocimiento valioso para organizaciones públicas y privadas [17]. En particular, los flujos continuos de datos presentan características que los hacen especialmente desafiantes: su volumen, dinamismo y volatilidad exigen sistemas capaces de analizarlos en tiempo real, con una velocidad de llegada que a menudo supera la capacidad de almacenamiento y procesamiento disponible [11].

En este contexto, los modelos tradicionales de aprendizaje automático, basados en arquitecturas centralizadas, presentan limitaciones críticas: riesgos de privacidad derivados de la transferencia de datos sensibles, elevados costes de comunicación y dificultades de escalabilidad en entornos reales [6]. El Aprendizaje Federado (FL) surge como respuesta a estos desafíos, permitiendo entrenar modelos de forma descentralizada sin compartir datos en crudo, procesando la información directamente en el dispositivo y compartiendo únicamente actualizaciones del modelo [9]. Sin embargo, como muestran trabajos recientes, el FL sigue enfrentando retos persistentes en heterogeneidad de datos y dispositivos, eficiencia de comunicación y seguridad frente a ataques [21], que deben abordarse para garantizar su adopción efectiva en entornos reales.

No obstante, la mayoría de las propuestas de Aprendizaje Federado (FL) continúan apoyándose en modelos de aprendizaje profundo y enfoques de caja negra, trasladando el problema desde la privacidad hacia una falta de transparencia que genera serios problemas de confianza, responsabilidad y auditabilidad. Esta limitación resulta especialmente crítica en sectores como la salud, la justicia o la administración pública, donde las decisiones no solo deben ser comprensibles y justificables, sino que se exige una propiedad adicional que la literatura actual no aborda de forma conjunta y satisfactoria: la explicabilidad intrínseca de los modelos resultantes combinada con seguridad criptográfica en las comunicaciones [2]. En este contexto, diversos autores han señalado las limitaciones de los enfoques de explicabilidad post-hoc. En particular, Rudin [19] defiende que, en aplicaciones de alto riesgo, resulta preferible utilizar modelos intrínsecamente interpretables en lugar de intentar explicar posteriormente el comportamiento de modelos opacos. Esta reflexión cuestiona la dependencia de técnicas ampliamente utilizadas como SHAP o LIME, cuyas explicaciones constituyen aproximaciones al comportamiento del modelo y no el propio razonamiento seguido durante la toma de decisiones [20]. Frente a esta limitación, los modelos basados en reglas descriptivas ofrecen una alternativa intrínsecamente interpretable, donde la lógica del sistema es comprensible por diseño, sin necesidad de ninguna capa explicativa adicional.

Esta necesidad se ve reforzada por el marco normativo europeo, especialmente el *AI Act* [4] (vigente desde febrero de 2025), que impone requisitos explícitos de transparencia, trazabilidad y auditabilidad a los sistemas de IA de alto riesgo, consolidando la explicabilidad intrínseca no como una opción deseable, sino como una exigencia regulatoria en ámbitos como la sanidad, la justicia o los servicios públicos esenciales [1]. En España, este marco se ha complementado con el proyecto de Ley Orgánica para el buen uso y la gobernanza de la inteligencia artificial, aprobado por el Consejo de Ministros el 26 de mayo de 2026, que

adapta el ordenamiento nacional al Reglamento europeo e incorpora un régimen sancionador propio [10].

2. Hipótesis de partida, objetivos y avances

La hipótesis central de esta investigación sostiene que la combinación del FL con modelos de reglas descriptivas supervisadas junto con tecnologías de seguridad criptográfica avanzadas, puede resolver los principales desafíos actuales de los sistemas de IA distribuidos en entornos de datos complejos y volátiles: la falta de transparencia, los riesgos de privacidad y la escalabilidad en entornos heterogéneos. Esta hipótesis se fundamenta en tres premisas clave.

En primer lugar, el FL permite entrenar modelos de forma descentralizada sin compartir datos en crudo, preservando la confidencialidad de la información local [9]. En segundo lugar, los modelos de reglas descriptivas supervisadas ofrecen una explicabilidad intrínseca frente a los enfoques tradicionales de caja negra. Al generar conocimiento mediante reglas lingüísticas legibles por humanos, eliminan las limitaciones de transparencia del estado del arte; una característica crítica para garantizar sistemas éticamente aceptables y legalmente auditables [5]. En tercer lugar, la seguridad de las comunicaciones entre nodos puede reforzarse mediante técnicas complementarias: el cifrado homomórfico (HE)[22], la privacidad diferencial (DP) [3], y la computación multipartita segura (SMC) [7].

Las premisas anteriores sustentan la hipótesis planteada y permiten identificar una oportunidad de investigación todavía insuficientemente explorada en la literatura actual. En este contexto, la originalidad de esta propuesta reside en que, si bien existen trabajos que combinan FL con explicabilidad [2] o FL con mecanismos de seguridad de forma aislada [18], , no se han encontrado aproximaciones que integren simultáneamente modelos de reglas descriptivas supervisadas, sincronización de etiquetas lingüísticas en entornos non-IID y mecanismos criptográficos adaptados a distintos niveles de confianza en el servidor federado.

La arquitectura general propuesta se detalla en la Figura 1. Cada cliente construye localmente modelos de reglas descriptivas supervisadas y comparte únicamente información agregada mediante mecanismos criptográficos. El resultado es un modelo global interpretable que preserva privacidad y auditabilidad.

El objetivo principal de esta tesis es diseñar arquitecturas avanzadas de FL que integren modelos de reglas descriptivas supervisadas capaces de generar conocimiento altamente explicable e interpretable, junto con mecanismos de seguridad criptográfica, con el fin de construir sistemas de IA confiables y seguros en entornos distribuidos de datos complejos y volátiles.

Este objetivo se desglosa en tres subobjetivos específicos, cuyo grado de avance se detalla a continuación:

- **O1. Análisis, diseño y desarrollo de nuevos modelos de reglas descriptivas en entornos de flujos de datos continuos y discontinuos.** Este objetivo está actualmente completado. Se ha propuesto OFEFS, un modelo adaptativo que opera en entornos online mediante ventana deslizante,

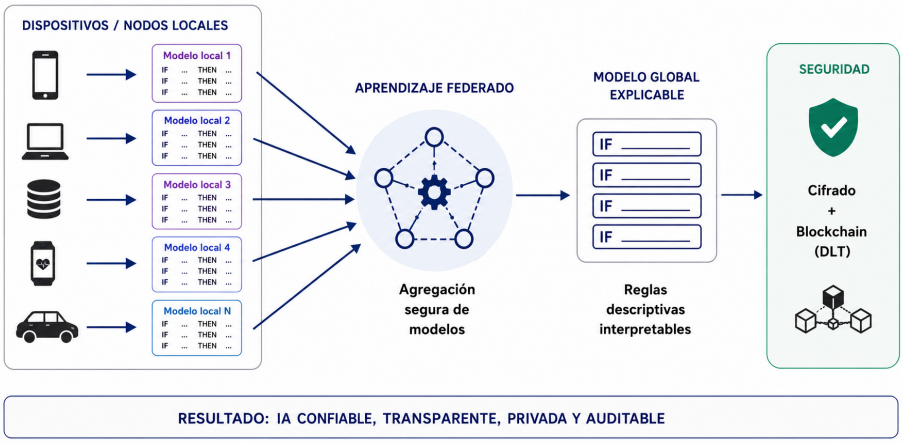


Figura 1. Arquitectura general propuesta. Framework propuesto: dispositivos locales con modelos de reglas descriptivas que se agregan de forma segura mediante FL para construir un modelo global explicable y auditable.

validado sobre la detección de la enfermedad rara NPC1 [16] (actualmente en revisión en el International Journal of Computational Intelligence Systems (Q2)).

- **O2. Análisis y diseño de arquitecturas de FL centralizado y confiable para sistemas de reglas descriptivas.** Se han desarrollado cuatro contribuciones en progresión: TEFeS-SDR con cifrado RSA [13], mecanismo de sincronización de etiquetas lingüísticas fuzzy para entornos non-IID [15], extensión con cifrado homomórfico [14], y sustitución de RSA por SMC con validación en detección de fugas de agua [12].
- **O3. Análisis y diseño de arquitecturas de FL descentralizado, confiable y distribuido.** Constituye el trabajo futuro planificado para el tercer y cuarto año, extendiendo los mecanismos desarrollados en O2 a escenarios totalmente descentralizados mediante blockchain y DLT para garantizar la integridad y soberanía de los datos.

3. Metodología y Plan de Trabajo

La metodología seguida en esta tesis doctoral responde a un enfoque iterativo e incremental estructurado en cinco fases cíclicas: Análisis → Diseño → Implementación → Evaluación → Mejora continua. Este ciclo se repite a lo largo de cada objetivo específico, permitiendo incorporar el conocimiento adquirido en cada iteración para refinar tanto los modelos como las arquitecturas propuestas.

La fase de análisis comprende la revisión de literatura científica en bases de datos especializadas (Scopus, Web of Science, Google Scholar), con el objetivo de identificar el estado del arte, detectar vacíos de investigación y fundamentar las decisiones de diseño. La fase de diseño abarca la formulación de hipótesis,

la propuesta de nuevos algoritmos de reglas descriptivas y el diseño de las arquitecturas federadas. La fase de implementación se materializa en prototipos funcionales desarrollados en Python y Scala. La fase de evaluación contrasta las hipótesis iniciales mediante experimentos sobre conjuntos de datos tanto sintéticos como reales, midiendo métricas de calidad del modelo, explicabilidad e integridad de la seguridad. Finalmente, la fase de mejora continua permite reformular las hipótesis y ajustar los algoritmos en función de los resultados obtenidos, cerrando el ciclo iterativo.

Este enfoque facilita la incorporación progresiva de los resultados obtenidos en cada objetivo, garantizando coherencia metodológica entre las distintas contribuciones de la tesis.

El desglose de tareas, su estado actual y las publicaciones asociadas a cada una se recogen en la Tabla 1, que refleja el avance del proyecto en el segundo año de doctorado.

El proyecto se estructura en torno a tres hitos principales que marcan puntos de control del avance de la investigación y coinciden con la consolidación de los resultados más relevantes de cada etapa.

- **Hito 1 (H1):** Este hito establece las bases conceptuales y metodológicas de la investigación. Incluye la revisión del estado del arte, la identificación de las principales limitaciones de los enfoques existentes y el desarrollo de nuevos modelos de reglas descriptivas adaptados a entornos de datos continuos y discontinuos, priorizando la explicabilidad y la capacidad de integración en entornos federados. Este hito se considera completado mediante la consecución del objetivo O1.
- **Hito 2 (H2):** En este segundo hito se abordará el diseño, implementación y evaluación de arquitecturas de FL que integren los modelos explicables desarrollados previamente. Esta fase se dividirá en dos bloques principales: en primer lugar, se construirá una arquitectura federada centralizada, en la que un nodo coordinador gestiona la federación y se asegura la coherencia y seguridad del proceso de entrenamiento. Posteriormente, se diseñará y desarrollará una arquitectura descentralizada, basada en tecnologías como *blockchain* y DLT. Se generarán prototipos funcionales de ambas arquitecturas, que serán comparados en términos de eficiencia, escalabilidad, seguridad y explicabilidad.
- **Hito 3 (H3):** El último hito del proyecto estará dedicado a la validación de los sistemas desarrollados. Para ello, se evaluarán los sistemas con datos reales, con el objetivo de evaluar su comportamiento en entornos realistas. Se recogerán resultados sobre el rendimiento de los sistemas, su explicabilidad, la privacidad y la robustez. Finalmente, se elaborará la documentación final del proyecto y se realizará la defensa de la tesis doctoral ante el tribunal.

La planificación temporal de este proyecto está descrita en la siguiente Figura 2 donde se indican los cuatro años que debe durar el proyecto dividido por trimestres. Por cada tarea detallada anteriormente se marca en qué trimestres se llevaría a cabo. En el caso de las tareas que se desarrollan a lo largo de todo

Tabla 1. Plan de trabajo doctoral y estado actual

Tarea	Descripción	Años	Obj.	Estado	Publ.
T1: Control y seguimiento del proyecto	Redacción y actualización del Plan de Investigación y Documento de Actividades. Reuniones periódicas con directores. Preparación de publicaciones y ponencias.	1–4	T	●	–
T2: Revisión del estado del arte	Análisis sistemático de literatura sobre reglas descriptivas supervisadas, FL, XAI y métodos de seguridad (HE, DP, SMC) en IEEE Xplore, Scopus y Web of Science. Elaboración de informe técnico de mapeo del área.	1	T	●	–
T3: Diseño de modelos de reglas descriptivas para flujos de datos	Desarrollo del modelo OFEFS, un algoritmo adaptativo para reglas descriptivas supervisadas en entornos online mediante ventana deslizante. Validado sobre detección de la enfermedad rara NPC1.	1–2	O1	✓	[16]
T4.1: Arquitectura FL centralizada con mecanismos de seguridad progresivos	Diseño e implementación de una arquitectura federada centralizada con modelos de reglas descriptivas jerárquicas. Incorporación progresiva de RSA, sincronización fuzzy con HE y SMC. Validación en detección de fugas de agua.	2–3	O2	✓	[12,13, 14,15]
T4.2: Arquitectura FL descentralizada con DLT	Diseño e implementación de una arquitectura federada descentralizada basada en blockchain y DLT, integrando los modelos desarrollados en T3 y T4.1. Evaluación de escalabilidad, seguridad y explicabilidad.	3–4	O3	❍	–
T5: Validación en casos reales y evaluación XAI	Aplicación de los sistemas desarrollados a escenarios reales de alto impacto. Evaluación mediante métricas de interpretabilidad y análisis comparativos. Elaboración de documentación final y defensa doctoral.	4	O2, O3	●	[12,16]

✓ Completada ● En progreso ❍ Trabajo futuro T = Tarea transversal

el proyecto se ha marcado que se realiza en todos los trimestres. Por otro lado, los hitos están destacados en el cronograma con una línea vertical, indicando que se realizaría el entregable al final de ese trimestre. El proyecto se encuentra actualmente en el segundo año de desarrollo, con las tareas transversales T3 y T4.1 prácticamente completadas, la tarea T5 parcialmente completada porque faltaría aplicarlo en arquitecturas descentralizadas y la tarea T4.2 en proceso. El cronograma se ha diseñado teniendo en cuenta que la tesis se realiza a tiempo completo.

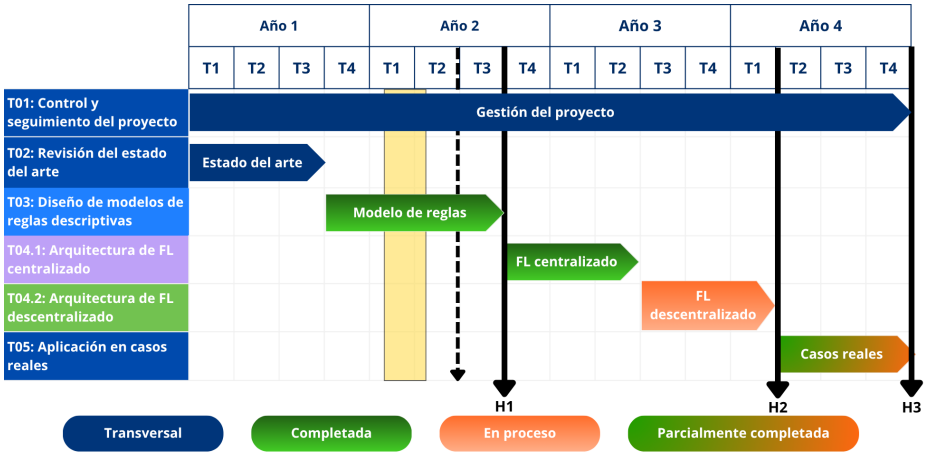


Figura 2. Cronograma de trabajo y planificación temporal del proyecto de tesis. Se ha diseñado teniendo en cuenta que la tesis se realiza a tiempo completo. El diagrama de Gantt detalla la secuenciación de las fases y señala la temporalización de los tres hitos principales. La línea discontinua señala el momento temporal actual y el sombreado amarillo la estancia internacional realizada.

4. Originalidad y aportación

La presente tesis doctoral aborda una combinación de técnicas que, tomadas individualmente, cuentan con antecedentes en la literatura. Sin embargo, la aportación original y diferenciada de este trabajo no reside en cada uno de sus componentes por separado, sino en su integración dentro de una arquitectura unificada capaz de combinar aprendizaje federado, explicabilidad intrínseca y mecanismos avanzados de preservación de la privacidad. A continuación se detallan los aspectos que definen la novedad de esta propuesta frente al estado del arte, respaldados por los resultados empíricos obtenidos.

En primer lugar, el uso de modelos de reglas descriptivas supervisadas en FL aporta explicabilidad intrínseca, a diferencia de los enfoques post-hoc predominantes en la literatura, como SHAP, LIME u otros métodos basados en

perturbaciones o gradientes, que intentan aproximar el comportamiento de modelos opacos a posteriori [21]. Esta característica permite que el modelo sea interpretable por diseño sin depender de aproximaciones externas que pueden ser inconsistentes, infieles al comportamiento real del sistema o incluso susceptibles de ser manipuladas [19]. Además, los modelos de reglas difusas basados en descubrimiento de subgrupos expresan el conocimiento mediante reglas del tipo *SI condición ENTONCES consecuente*, facilitando su comprensión y validación por expertos y usuarios finales.

En segundo lugar, la incorporación de un mecanismo de sincronización de etiquetas lingüísticas fuzzy entre clientes federados constituye una aportación novedosa para abordar la heterogeneidad semántica en entornos non-IID [15,14], un problema habitualmente tratado desde la perspectiva estadística o de agregación de parámetros [8] pero nunca desde la consistencia del espacio lingüístico compartido en sistemas de reglas difusas federados. La solución propuesta garantiza que todos los clientes evolucionen sus modelos utilizando el mismo conjunto de etiquetas lingüísticas, mejorando significativamente la coherencia semántica y la calidad del modelo global.

En tercer lugar, esta tesis propone la adaptación de mecanismos de seguridad ampliamente utilizados en FL para modelos de caja negra al contexto de los modelos de reglas descriptivas supervisadas, adaptando RSA [13], HE [14] y SMC [12]. Estas propuestas se diseñan en función de la confianza depositada en el servidor federado y se evalúan considerando el compromiso entre coste computacional, privacidad y calidad del modelo. En particular, la incorporación de SMC garantiza que ni siquiera el servidor pueda acceder al contenido de las reglas durante la agregación, elevando el nivel de privacidad al máximo posible dentro de una arquitectura centralizada.

La evolución empírica de estas contribuciones queda reflejada en la Tabla 2, que muestra la mejora progresiva del framework a lo largo de las publicaciones realizadas:

Tabla 2. Evolución de resultados del framework federado propuesto

Sistema	#Reglas	#Vars	CONF	WRAccN	FPR	TPR
TEFeS-SDR [13]	101.5	81.5	0.62	0.62	0.27	0.85
SMC-FL [12]	20.91	4.3	0.69	0.53	0.13	0.16
Sync. fuzzy + HE [14,15]	7.33	3.56	0.90	0.64	0.05	0.37

La reducción drástica en el número de reglas (de 101.5 a 7.33) y de variables (de 81.5 a 3.56), junto con la mejora en confianza (de 0.62 a 0.90) y la reducción de falsos positivos (de 0.27 a 0.05), evidencia que la sincronización de etiquetas lingüísticas no solo mejora la seguridad semántica del modelo federado, sino que produce modelos significativamente más simples, precisos e interpretables.

Más allá de los benchmarks sintéticos habituales en la literatura, las soluciones desarrolladas han sido evaluadas en dos escenarios reales de gran complejidad y relevancia social.

El primero aborda la detección de fugas de agua en redes de distribución [12], un entorno industrial distribuido donde los datos de cada nodo son sensibles y no pueden centralizarse por razones operativas y de privacidad. En este escenario, el framework federado con SMC demostró su capacidad para extraer reglas interpretables y directamente accionables por los operadores del sistema, preservando la integridad de la información local.

El segundo se centra en la monitorización de la enfermedad rara NPC1 en un entorno online [16], un escenario especialmente exigente por la combinación de datos extremadamente escasos, clases desbalanceadas y la necesidad de procesar información en tiempo real; condiciones bajo las cuales la mayoría de los métodos estándar presentan serias dificultades. En este contexto, el modelo OFEFS demostró su capacidad para extraer conocimiento interpretable con un número reducido de variables, generando reglas clínicas comprensibles que apoyan directamente la toma de decisiones médicas sin comprometer la privacidad de los pacientes.

En ambos casos, los modelos propuestos demostraron su capacidad para extraer conocimiento interpretable y útil bajo restricciones reales, tal y como se recoge en la Tabla 3.

Tabla 3. Resultados obtenidos en los casos de uso reales

Sistema	**Caso de uso**	**#Reglas**	**#Vars**	**CONF**	**WRAccN**	**TPR**
SMC-FL [12]	Fugas de agua	21.41	6.48	0.859	0.616	0.585
OFEFS [16]	NPC1	35	1.89	0.882	0.654	0.792

5. Relevancia

La relevancia de la presente investigación se sitúa en la intersección de tres áreas de creciente interés dentro de la comunidad de IA: el FL, los modelos de reglas descriptivas supervisadas y la seguridad criptográfica en sistemas distribuidos. Aunque cada una de estas líneas cuenta con una trayectoria consolidada, su integración dentro de un mismo framework sigue siendo limitada en la literatura actual, donde normalmente se abordan de forma independiente [18]. En particular, el proyecto apuesta por modelos con explicabilidad intrínseca basados en reglas difusas, frente a los enfoques post-hoc predominantes en gran parte de la literatura. De este modo, la interpretabilidad no se introduce como una capa adicional sobre un modelo opaco, sino como una propiedad incorporada desde el diseño del sistema. Esta aproximación resulta especialmente relevante en escenarios donde la transparencia y la trazabilidad de las decisiones son requisitos importantes tanto desde el punto de vista técnico como regulatorio. Además, a diferencia de otros paradigmas de aprendizaje automático orientados principalmente a entornos de *big data* o flujos de datos masivos, los modelos de reglas descriptivas supervisadas han demostrado su eficacia en problemas de alta complejidad con conjuntos de datos limitados y baja dimensionalidad, como es

el caso de enfermedades raras o escenarios industriales con información escasa, donde la capacidad de extraer conocimiento interpretable y confiable resulta aún más crítica.

En el ámbito de la investigación, el impacto en el conocimiento y los resultados obtenidos hasta el momento muestran que la integración propuesta es viable tanto a nivel teórico como experimental. La evolución de los trabajos desarrollados durante la tesis refleja una mejora progresiva de los modelos. El sistema TEFeS-SDR [13] estableció las bases del framework utilizando cifrado RSA. Posteriormente, la incorporación de mecanismos de sincronización de etiquetas lingüísticas fuzzy y técnicas de cifrado homomórfico [14,15] permitió mejorar los resultados. Estos avances han dado lugar a publicaciones en revistas y congresos internacionales relacionados con IA y sistemas inteligentes distribuidos, entre ellos Results in Engineering (Q1), IDEAL 2025 e IEEE CAI 2026, que avalan la solidez de las contribuciones realizadas.

Un hito especialmente relevante en este desarrollo científico ha sido la estancia internacional realizada en De Montfort University (Leicester, Reino Unido) entre febrero y abril de 2026. Lejos de ser una visita puntual, esta estancia supuso la colaboración con equipos investigadores multidisciplinares, integrando conocimiento en biomedicina y sistemas industriales que difícilmente hubiera sido posible en el contexto del grupo de origen. Como resultado de esta colaboración, se han realizado dos contribuciones: el trabajo [12], que introduce SMC en el framework federado y lo valida en un caso real de detección de fugas de agua (actualmente en revisión en ACM Transactions on Intelligent Systems and Technology (Q1)); y el trabajo [16], que propone un algoritmo novedoso para la extracción de reglas descriptivas en entornos de flujo de datos online aplicado a la monitorización de la enfermedad rara NPC1, un problema de alta complejidad caracterizado por datos escasos, sensibles y desequilibrados (en revisión en el International Journal of Computational Intelligence Systems (Q2)).

El desarrollo de esta tesis doctoral también está ligado a la actividad docente en el Departamento de Informática de la Universidad de Jaén. En este contexto, se han impartido las sesiones prácticas de la asignatura *Metaheurísticas* durante el primer cuatrimestre del curso académico 2025/2026. Durante este periodo, el alumnado ha trabajado e implementado diferentes métodos de optimización aplicados a solventar el problema planteado, analizando las ventajas e inconvenientes de cada enfoque y evaluando su adecuación a dicho problema.

La transferencia tecnológica y el impacto social se traducen directamente en soluciones tangibles, especialmente en sectores donde la toma de decisiones debe ser simultáneamente privada, transparente y auditable. En este sentido, la entrada en vigor del *AI Act* europeo en febrero de 2025 ha reforzado la importancia de incorporar mecanismos de transparencia y explicabilidad en aplicaciones de alto riesgo. El uso de modelos basados en reglas descriptivas facilita que tanto expertos como usuarios finales puedan comprender el comportamiento del sistema y analizar las decisiones generadas, aspecto especialmente importante en ámbitos sensibles como la sanidad [1]. La aplicabilidad práctica de las soluciones desarrolladas se ha evaluado en dos casos de uso con impacto social relevante:

- El primero aborda la detección de fugas de agua en redes de distribución [12], donde el aprendizaje federado permite trabajar con información distribuida preservando la privacidad de los datos de cada nodo de la red.
- El segundo caso de uso se centra en la monitorización de la enfermedad rara NPC1 [16], un escenario online caracterizado por datos sensibles, heterogéneos y escasos.

Finalmente, esta investigación mantiene un firme compromiso orientado hacia la difusión y la comunicación abierta de resultados. Entre las actividades de divulgación y transferencia realizadas se incluyen la publicación del código fuente en repositorios públicos de GitHub, promoviendo los principios de la ciencia abierta y facilitando la reproducibilidad de los trabajos. Asimismo, se participa activamente en eventos de divulgación científica orientados a la sociedad, como la Noche Europea de los Investigadores en Jaén, y se efectúa la presentación de los resultados en congresos nacionales e internacionales del área, entre ellos CAEPIA, MAEB, IDEAL e IEEE CAI. Todo ello favorece el intercambio abierto de conocimiento tanto con la comunidad científica como con otros sectores socioeconómicos interesados.

Referencias

1. Comisión Europea: Directrices éticas para una IA fiable. Tech. rep., Oficina de Publicaciones de la Unión Europea, LU (2019), accedido: 12 de junio de 2025
2. Corcuera Bárcena, J.L., Ducange, P., Marcelloni, F., Renda, A.: Increasing trust in ai through privacy preservation and model explainability: Federated learning of fuzzy regression trees. Information Fusion **113**, 102598 (2025). https://doi.org/https://doi.org/10.1016/j.inffus.2024.102598, https://www.sciencedirect.com/science/article/pii/S1566253524003762
3. El Ouadrhiri, A., Abdelhadi, A.: Differential privacy for deep and federated learning: A survey. IEEE access **10**, 22359–22380 (2022)
4. European Parliament and Council of the European Union: Regulation (EU) 2024/1689 (Artificial Intelligence Act). Tech. rep., Official Journal of the European Union (2024), http://data.europa.eu/eli/reg/2024/1689/oj, laying down harmonised rules on artificial intelligence and amending certain Regulations and Directives
5. Gunning, D.: Explainable artificial intelligence (xAI). Technical report, Defense Advanced Research Projects Agency (DARPA) (2017)
6. Kairouz, P., et al.: Advances and open problems in federated learning. arXiv preprint arXiv:1912.04977 (2021). https://doi.org/10.48550/arXiv.1912.04977
7. Li, Y., Zhou, Y., Jolfaei, A., Yu, D., Xu, G., Zheng, X.: Privacy-preserving federated learning framework based on chained secure multiparty computing. IEEE Internet of Things Journal **8**(8), 6178–6186 (2020)
8. Mahmud, F., Mahmud, F., Rahman, R.M.: A survey of federated learning: Advances in architecture, synchronization, and security threats. Computers, Materials and Continua **86**(3) (2026). https://doi.org/https://doi.org/10.32604/cmc.2025.073519, https://www.sciencedirect.com/science/article/pii/S1546221826000524
9. McMahan, H.B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.Y.: Communication-efficient learning of deep networks from decentralized data. In:

Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS) (2016), accedido: 12 de junio de 2025
10. Ministerio de Transformación Digital y de la Función Pública: El gobierno aprueba el proyecto de ley que garantizará una super... [completar título] (5 2026), https://digital.gob.es/comunicacion/notas-prensa/mtdfp/2026/05/el-gobierno-aprueba-el-proyecto-de-ley-que-garantizara-una-super
11. Muthukrishnan, S.: Data streams: Algorithms and applications. Theoretical Computer Science (TCS) **1**(2), 117–236 (sep 2005). https://doi.org/10.1561/0400000002
12. Padilla-Rascón, M.A., Bakar, M., Deka, L., Elizondo, D., García-Vico, A.M., Carmona, C.J.: An interpretable and secure federated framework for supervised descriptive rule learning. ACM Transactions on Intelligent Systems and Technology (2026), en revisión
13. Padilla-Rascón, M.A., García-Vico, A.M., Carmona, C.J.: Trustworthy and explainable federated system for extracting descriptive rules in a data streaming environment. Results in Engineering **25** (2025)
14. Padilla-Rascón, M.A., García-Vico, A.M., Carmona, C.J.: A fuzzy based-rule adaptive federated learning for constructing a trustworthy gpais. In: IEEE Conference on Artificial Intelligence (CAI) 2026 (2026)
15. Padilla-Rascón, M.A., García-Vico, Á.M., Carmona, C.J.: Synchronising fuzzy labels in federated evolutionary systems: A preliminary study. In: Intelligent Data Engineering and Automated Learning – IDEAL 2025. Springer Nature (2026)
16. Padilla-Rascón, M.A., García-Vico, A.M., Grootveld, M., Elizondo, D., Carmona, C.J.: Towards transparent rare disease monitoring: Federated rule extraction for online streaming npc1. International Journal of Computational Intelligence Systems (2026), en revisión
17. Raja Santhi, A., Muthuswamy, P.: Industry 5.0 or industry 4.0S? Introduction to industry 4.0 and a peek into the prospective industry 5.0 technologies. International Journal of Interactive Design and Manufacturing (IJIDeM) **17**(2), 947–979 (apr 2023). https://doi.org/10.1007/s12008-023-01217-8
18. Rodríguez-Barroso, N., Jiménez-López, D., Luzón, M.V., Herrera, F., Martínez-Cámara, E.: Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges. Information Fusion **90**, 148–173 (feb 2023). https://doi.org/10.1016/j.inffus.2022.09.011
19. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x
20. Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., Zhong, C.: Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistics Surveys **16**, 1–85 (2022). https://doi.org/10.1214/21-SS133
21. Tunduny, T., Shibwabo, B.: Explainable AI approaches in federated learning: Systematic review. JMIR AI **5**, e69985 (feb 2026). https://doi.org/10.2196/69985
22. Zhang, L., et al.: Homomorphic encryption-based privacy-preserving federated learning in IoT-enabled healthcare system. IEEE Transactions on Network Science and Engineering **10**, 2864–2880 (2022)