PLAN DE GESTIÓN DE DATOS -DyCMaMod

PROYECTO	
Código del proyecto:	PID2023-146872OB-I00
Acrónimo del proyecto:	DyCMaMod
Nombre del proyecto:	**Dynamics and Control for Machine Learning and Modeling**
Área científica o temática del proyecto	**Ciencias Matemáticas**
Objetivos del proyecto	El aprendizaje automático (en inglés Machine Learning - ML) y, más en general, la ciencia de datos están revolucionando las matemáticas aplicadas, lo que lleva a una investigación rica, intensiva e innovadora y a una variedad de nuevas ideas y métodos poderosos. En este marco, el proyecto DyCMaMod desarrolla nuevos fundamentos matemáticos y computacionales para la inteligencia artificial científica, combinando aprendizaje automático, teoría del control, ecuaciones en derivadas parciales (PDEs) y modelado híbrido basado en datos. Su objetivo principal es diseñar modelos sintéticos robustos, interpretables y eficientes capaces de integrarse con modelos mecanicistas inspirados en principios físicos. Entre otros campos, el proyecto busca avances en Neural ODEs, aprendizaje de operadores, modelos generativos de difusión, aprendizaje federado y mecanismos de self-attention. Al mismo tiempo desarrolla nuevas herramientas matemáticas para el análisis y entrenamiento de arquitecturas modernas de inteligencia artificial desde una perspectiva dinámica y de control. Asimismo, el proyecto impulsa nuevas metodologías para la hibridación entre modelos basados en datos y modelos continuos gobernados por PDEs y ecuaciones cinéticas, contribuyendo al desarrollo de modelos híbridos capaces de integrar observaciones experimentales y conocimiento físico de manera coherente y eficiente. Estas investigaciones abren nuevas perspectivas para aplicaciones en campos como la biomedicina y el estudio del envejecimiento saludable, la simulación científica, la dinámica de multitudes, los sistemas industriales complejos, y el desarrollo de gemelos digitales.
Gestión de datos del proyecto	i. El proyecto de investigación no recopila datos sensibles de ninguna naturaleza. Para las simulaciones se utilizan datos sintéticos o de datasets públicamente disponibles. ii. La investigación no contiene datos personales y sensibles. iii. No aplica disponer de un repositorio de datos de la investigación.

PLAN DE GESTIÓN DE DATOS	
Fecha	Diciembre 2024
Versión:	V1.0

A. Justificación del Uso de Datos y Principios FAIR

La investigación realizada en DyCMaMod se basa en explorar los fundamentos matemáticos del *machine learning* y validarlos con simulaciones sobre *datasets benchmark*

Para el desarrollo de esta investigación, se utilizan datos de diferentes orígenes, dependiendo del contexto y del tema de investigación específico:

- **Datos sintéticos**, generados por los investigadores involucrados en el proyecto mediante rutinas de Python dedicadas. El uso de datasets sintéticos (por ejemplo, puntos en el plano generados aleatoriamente) permite diseñar escenarios experimentales bajo condiciones controladas, lo cual facilita la validación inicial de los algoritmos en situaciones específicas. Este tipo de datos es especialmente útil en las etapas preliminares de desarrollo, ya que ofrece flexibilidad en la manipulación de variables y garantiza la repetibilidad de los experimentos.
- Datos descargados de **bases de datos públicas** ampliamente reconocidas tipo Kaggle el dataset **EMNIST** (Extended MNIST), disponible a través del Instituto Nacional de Estándares y Tecnología (NIST) de EE. UU. La inclusión de estos recursos tiene como objetivo validar la aplicabilidad de los métodos desarrollados en contextos más cercanos a problemas reales y facilitar la comparación con trabajos previos, al tratarse de bases de datos ampliamente utilizadas por la comunidad científica.
- **Datos reales**, recolectados gracias a nuestras colaboraciones con diferentes partners tecnológicos, lo que garantiza una mayor fiabilidad, representatividad y validez de los modelos frente al uso exclusivo de datos sintéticos. Estos datos se recolectan respetando los marcos jurídicos pertinentes y con la autorización escrita de todas las personas involucradas.

El uso de diferentes tipos de datos fortalece la rigurosidad metodológica de nuestros estudios, al permitir tanto una evaluación teórica controlada como una validación empírica en escenarios representativos del mundo real. Esta estrategia contribuye a garantizar la solidez de los resultados obtenidos y su relevancia científica.

La elección de los conjuntos de datos utilizados en esta investigación responde tanto a criterios metodológicos como éticos, en línea con los principios FAIR (*Findable, Accessible, Interoperable, Reusable*), los cuales promueven una gestión responsable y sostenible de los datos científicos.

En primer lugar, se han utilizado conjuntos de datos encontrables (Findable), disponibles en repositorios conocidos, garantizando la trazabilidad y la correcta citación de los datos, lo cual es esencial para la transparencia y la validación por parte de la comunidad científica.

Asimismo, los datasets sintéticos y los provenientes de repositorios públicos son accesibles (Accessible) bajo licencias abiertas que permiten su descarga y uso sin restricciones indebidas. Esta característica facilita la replicabilidad de los experimentos realizados y fomenta la colaboración científica. Los datos reales, por razones de privacidad, son accesibles bajo procedimientos de autorización adecuados y con las medidas de seguridad necesarias implementadas.

En cuanto a la interoperabilidad (Interoperable), se ha priorizado el uso de formatos estándar (como CSV o imágenes en formatos comunes), lo que permite integrar los datos fácilmente con herramientas de análisis y software de código abierto, evitando barreras tecnológicas que limitan su reutilización.

Por último, se ha asegurado que los datos sean reutilizables (Reusable), tanto por su clara

documentación como por su disponibilidad en plataformas ampliamente utilizadas. La elección de conjuntos bien estructurados, con metadatos completos y contextos de aplicación definidos, permite que los resultados de esta investigación puedan ser contrastados, ampliados o replicados en estudios futuros.

Esta alineación con los principios FAIR no solo fortalece la calidad científica del presente estudio, sino que también contribuye al avance de una ciencia abierta, ética y responsable.

Las publicaciones generadas en el proyecto DasEL indican específicamente en los apartados relativos a los criterios utilizados en la evaluación el tipo de datasets utilizados.

B. Otros resultados de la investigación

Además de la gestión de los datos, la herramienta de software asociada al dataset, disponible en DyCMaMod Toolbox se gestiona siguiendo los principios FAIR para garantizar su reutilización y accesibilidad. En este espacio compartido se incorpora el código de los experimentos y no los datos base utilizados.

C. **Seguridad de los datos***.

Tal y como se ha mencionado, cuando la investigación utiliza datos reales, estos se recolectan respetando todos los marcos legales pertinentes. Además, su reutilizo por partes no involucradas en la investigación original se permite sólo bajo autorización escrita del detentor de los datos, y únicamente en caso de clara necesidad debidamente motivada.

Estas consideraciones no aplican a los otros tipos de datos mencionados en el apartado A. Dichos datasets están públicamente disponibles por lo que no procede especificar el tratamiento de la recuperación de los datos, así como el almacenamiento seguro y el traspaso de datos sensibles y las medidas adoptadas para la seguridad de los datos.

Por último, cabe matizar que las *actividades de investigación no incluyen a niños, pacientes sensibles, población vulnerable, el uso de células madre embrionarias, o investigación en animales y primates.*

D. Cuestiones éticas o jurídicas

No existen cuestiones éticas o jurídicas que puedan repercutir en la puesta en común de los datos.