PLAN DE GESTIÓN DE DATOS -KiLearn

PROYECTO	
Código del proyecto:	PID2020-112617GB-C22
Acrónimo del proyecto:	KiLearn
Nombre del proyecto:	**Kinetic equations and Learning control**
Área científica o temática del proyecto	**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. Buscamos contribuir a sus fundamentos matemáticos, prestando especial atención a los desafíos en el área de control de sistemas dinámicos, donde el paradigma clásico basado en modelos debe complementarse con una serie de técnicas basadas en datos. Nuestro objetivo es fomentar el rigor en los fundamentos analíticos y computacionales de las metodologías de control basadas en modelos/datos, construyendo un puente específico y productivo entre la ciencia de datos y las matemáticas aplicadas y generando enfoques y algoritmos híbridos novedosos y de mayor rendimiento. En este contexto, se están logrando grandes avances en el amplio campo de la ingeniería de control, fuertemente influenciados por ML. Esto permite abordar un número creciente de aplicaciones desafiantes, pero los fundamentos teóricos y analíticos aún son escasos en términos de rigor. Los enfoques basados en datos proporcionan una nueva gama de potentes herramientas para diseñar estrategias de control superiores. Del mismo modo, algunos de los conceptos clave y los resultados del control arrojan luz sobre algunos de los principales desafíos en las arquitecturas para redes neuronales, por ejemplo. Pero los pilares teóricos son endebles. El objetivo de este subproyecto es abordar cuestiones analíticas y computacionales clave que no se comprenden bien o quedan sin resolver, pero que aparecen en toda la gama de aplicaciones tecnológicas y de la vida real (aeronáutica, conducción autónoma, gestión de recursos, epidemiología, robótica, biomedicina, internet, etc.) y requieren un tratamiento unificado.
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	Junio 2022
Versión:	V1.0

A. Justificación del Uso de Datos y Principios FAIR

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

Para el desarrollo de esta investigación, se ha optado por utilizar tanto conjuntos de datos sintéticos como bases de datos públicas ampliamente reconocidas. Esta decisión se fundamenta en la necesidad de contar con datos adecuados que permitan evaluar, de manera objetiva y controlada, el desempeño de los métodos propuestos.

El uso de datasets sintéticos (por ejemplo, puntos en el plano generados aleatoriamente), generados específicamente para nuestros estudios, 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.

Complementariamente, se emplean conjuntos de datos públicos y de acceso libre, tales como el **EMNIST** (Extended MNIST), disponible a través del Instituto Nacional de Estándares y Tecnología (NIST), y diversos datasets disponibles en la plataforma Kaggle. 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.

La combinación de datos sintéticos y reales 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, todos los datasets empleados son accesibles (Accessible) públicamente, 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.

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 KiLearn 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 KiLearn 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 la investigación utiliza únicamente datasets 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. Además cabe matizar que las *actividades de investigación no incluyen a niños, pacientes, población vulnerable, el uso de células madre embrionarias, cuestiones de privacidad y protección de datos o investigación en animales y primates,* Ética

D. Cuestiones éticas o jurídicas

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