Facultad de Economía

IA más allá del poder predictivo: equidad, interpretabilidad, privacidad e incertidumbre

Curso de Verano · Julio 6–17 de 2026

Descripción del curso

Este curso brinda a los estudiantes las bases teóricas y prácticas del desarrollo responsable de sistemas basados en inteligencia artificial como insumo para la toma de decisiones. Se organiza la discusión en torno a cuatro atributos fundamentales que estos sistemas pueden (y deben) tener más allá de su capacidad predictiva: justicia algorítmica, interpretabilidad, cuantificación de la incertidumbre, y privacidad. Para abordarlos, se tomará una aproximación transversal cubriendo el ciclo de desarrollo de sistemas basados en datos, detallando cómo se manifiestan estas propiedades en modelos predictivos, causales y generativos. Las herramientas adquiridas en este curso permiten a los estudiantes involucrarse de manera informada en la construcción y el monitoreo de sistemas basados en inteligencia artificial, comunicarse con equipos técnicos y no técnicos sobre sus riesgos y alcances, y tomar decisiones mejor fundamentadas en entornos académicos y laborales.

Al finalizar, se espera que comprendan los fundamentos de la IA responsable, puedan leer y discutir críticamente las nuevas metodologías propuestas en la creciente literatura sobre justicia algorítmica, interpretabilidad, cuantificación de la incertidumbre y privacidad, y estén en capacidad de analizar de forma holística el diseño, la implementación y la evaluación de sistemas de IA en distintos sectores. El curso combina un componente teórico sobre las propiedades estadísticas y computacionales de las metodologías estudiadas con un componente práctico de programación en Python, en el que los estudiantes implementan y evalúan estos métodos.

Profesores

- **Mateo Dulce** - mateo.d@nyu.edu
- **Melissa Robles -** mv.robles@uniandes.edu.co

Objetivos pedagógicos

- Comprender los fundamentos teóricos y prácticos de la IA responsable (RAI) y sus dimensiones: fairness, interpretabilidad, incertidumbre y privacidad.
- Analizar los sistemas algorítmicos para la toma de decisiones como sistemas socio-técnicos, discutiendo sus potenciales riesgos con argumentos técnicos, éticos y legales relevantes en cada contexto.
- Realizar auditorías de equidad algorítmica con métricas cuantitativas en distintas etapas del ciclo de vida de un sistema basado en datos, e implementar técnicas de

mitigación de sesgos teniendo en cuenta las tensiones entre las distintas métricas y la capacidad predictiva de los modelos.
- Integrar métodos de interpretabilidad (SHAP, LIME) y evaluar su utilidad para explicar decisiones de modelos predictivos.
- Implementar métodos de cuantificación de incertidumbre como parte integral de un modelo predictivo y comprender sus garantías estadísticas.
- Aplicar técnicas de privacidad y protección de datos, y entender sus implicaciones en el desarrollo de sistemas de IA.
- Utilizar modelos predictivos de aprendizaje de máquinas para estimación de efectos causales.
- Discutir críticamente el estado del arte en inteligencia artificial responsable en modelos de lenguaje (LLMs).
- Aplicar los fundamentos, métodos y marcos de RAI para diagnosticar y mejorar sistemas de IA en sus propios contextos profesionales.

Metodología

Cada sesión de tres horas está estructurada en dos momentos complementarios:

- La primera parte es teórica: se presentan los conceptos, fundamentos matemáticos y estadísticos, y propiedades computacionales de las técnicas de inteligencia artificial responsable. Cada clase es motivada y se ilustra con casos de uso en contextos reales de sistemas algorítmicos para la toma de decisiones.
- La segunda parte es práctica: los estudiantes implementan en Python las técnicas estudiadas, explorando su funcionamiento sobre conjuntos de datos y contextos reales, y analizando cómo se comportan bajo distintas condiciones y restricciones.

El curso incluye un proyecto final grupal en el que los equipos aplican y extienden las técnicas estudiadas a un problema de su elección, conectándolas con la literatura académica en RAI. Los proyectos pueden ir desde la replicación y extensión de un artículo académico hasta la aplicación de las técnicas del curso a un conjunto de datos propio. El proyecto tiene dos entregables.

- Una presentación preliminar en clase (julio 17), en la que los equipos reciben retroalimentación del avance del proyecto.
- Un documento escrito y *notebook* de Python en los que se describe el problema abordado, las técnicas aplicadas y los resultados obtenidos.

Evaluación

Componente	Descripción	Peso
Ejercicios y talleres en clase	8 talleres cortos con ejercicios teóricos y prácticos (diarios)	50%
Presentación	Presentación de avance del proyecto	20%
Documento final	Análisis de una metodología de RAI del estado del arte y su aplicación a un caso de uso.	30%

Plan de temas

Fecha	Tema	Contenido	Referencias principales
Semana 1 - 6 al 10 de julio			
Lunes julio 06	Fundamentos de IA y riesgos	- Introducción al curso - Conceptos fundamentales de IA. - Tipos de sistemas. - Riesgos de IA: taxonomía ISO 42005. - System maps y ciclo de vida de IA.	[1]
Martes julio 07	Introducción a la justicia algorítmica. Fairness en post-processing.	- ¿Qué es la justicia algorítmica? - Marcos éticos para estudiar sistemas de IA para la toma de decisiones. - Auditoría de justicia algorítmica. - Definiciones formales: demographic parity, equal opportunity, calibration. - Tensiones entre métricas.	[2]
Miércoles julio 08	Fairness en otros momentos del ciclo de vida de la IA.	- Técnicas de mitigación de sesgos algorítmicos: pre-, in- y post-processing - Comparación y casos de uso.	[2]
Jueves julio 09	Interpretabilidad: SHAP Values	- ¿Por qué interpretabilidad? - Modelos de caja negra vs. caja blanca. - Fundamentos de SHAP (Shapley values).	[3]
Viernes julio 10	Interpretabilidad: Lime. Transparencia	- LIME: Local Interpretable Model-agnostic Explanations. - Diferencias con SHAP. - Transparencia en sistemas de IA.	[3]
Semana 2 - 13 al 17 de julio			
Lunes julio 13	Cuantificación de incertidumbre y predicción conformal	- Incertidumbre en modelos predictivos. - Predicción conformal para cuantificación de incertidumbre - Incertidumbre como transparencia y otros usos en IA responsable	[4]
Martes julio 14	Privacidad y protección de datos	- Privacidad diferencial. - Aprendizaje de datos preservando la privacidad.	[5]

Miércoles julio 15	IA para inferencia causal	- Predicción vs. causalidad - Modelos predictivos para estimar efectos causales - Estimadores doble robustos de ML y sus propiedades - Otras nociones de causalidad en IA responsable.	[6, 2 ,3]
Jueves julio 16	IA generativa y modelos de lenguaje responsable.	- Introducción a modelos generativos - Detección de sesgos en LLMs: red teaming y benchmarks. - El problema de *Alignment*	[7,8]
Viernes julio 17	Presentaciones finales	- Temas complementarios en uso de IA para la toma de decisiones - Presentaciones de proyectos finales.	

Bibliografía y recursos

[1] ISO/IEC 42005:2025. *Artificial intelligence — AI system impact assessment*. International Organization for Standardization.

[2] Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press. https://fairmlbook.org

[3] Molnar, C. (2025). *Interpretable machine learning: A guide for making black box models explainable* (3rd ed.). https://christophm.github.io/interpretable-ml-book

[4] Angelopoulos, A. N., & Bates, S. (2023). *Conformal prediction: A gentle introduction.* Foundations and Trends in Machine Learning, 16(4), 494-591.

[5] Dwork, C., & Roth, A. (2014). *The algorithmic foundations of differential privacy.* Foundations and trends® in theoretical computer science, 9(3-4), 211-487.

[6] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). *Double/debiased machine learning for treatment and structural parameters.* The Econometrics Journal, Volume 21, Issue 1, 1 February 2018, Pages C1–C68, https://doi.org/10.1111/ectj.12097

[7] Robles, M., Bernal, C., Raigoso, D., & Dulce Rubio, M. (2025). SESGO: Spanish evaluation of stereotypical generative outputs. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 8*(3), 2214–2226. https://doi.org/10.1609/aies.v8i3.36707

[8] UNESCO. (2025). *Red teaming artificial intelligence for social good: The playbook*. https://unesdoc.unesco.org/ark:/48223/pf0000394338