Ciencia de Datos Responsable

Syllabus

Equipo Docente

Gabriela Arriagada

Profesora asistente del Instituto de Éticas Aplicadas (IEA) y del Instituto de Ingeniería Matemática y Computacional (IMC) de la Pontificia Universidad Católica de Chile. Es investigadora joven en el Centro Nacional de Inteligencia Artificial (CENIA), trabajando en la línea de investigación 5: IA centrada en el ser humano. Además, es directora de Latinoamérica para la World Ethical Data Foundation (WEDF). Es Licenciada en Filosofía de la UC, Master of Science en Filosofía de la Universidad de Edimburgo, Escocia, y doctor © en la Universidad de Leeds, Inglaterra.

Ha sido profesora invitada en la Universidad Técnica de Eindhoven y la Universidad de Tilburg, en Holanda, y en el Centre for Doctoral Training in Artificial Intelligence for Medical Diagnosis and Care (CDT), de la Universidad de Leeds.

Descripción del curso

En este curso, los y las estudiantes aprenderán principios y valores éticos, así como las herramientas y metodologías que fundamentan el uso responsable de los datos en distintos contextos. Las lecturas, clases y tutoriales tienen como objetivo que los y las estudiantes comprendan tanto la crítica como las posibles soluciones a problemas relacionados con

transparencia, interpretabilidad, explicabilidad, sesgos, entre otros. Asimismo, se instruye en las responsabilidades asociadas a la ética profesional en el manejo de datos.

Resultados de Aprendizaje

- Analizar las dificultades éticas que puede presentar un proceso de análisis de datos, identificando los componentes y técnicas que se pueden utilizar para mitigarlas.
- Aplicar herramientas matemáticas y computacionales que garanticen un uso responsable de los datos.
- Distinguir principios y valores éticos en la toma de decisiones basadas en datos.
- Evaluar la importancia de la privacidad y protección de datos, especialmente en relación con la recopilación y uso de datos sensibles.
- Implementar estrategias para minimizar los sesgos durante la recolección, análisis e interpretación de datos.
- Comunicar resultados de manera accesible e identificar las limitaciones del análisis de datos.

Estructura del curso

El curso está estructurado de la siguiente forma:

1. Introducción al mundo de la ética

- Ética de la inteligencia artificial y ética de datos
- Alfabetización ética crítica: teorías éticas y principios
- Directrices para una ciencia de datos responsable
- Herramientas de apoyo para seguimiento

2. Sesgos, justicia y discriminación

- Sesgos en inteligencia artificial y ciencia de datos
- Definiciones y tipos de justicia
- Previniendo la discriminación algorítmica

3. Transparencia, interpretabilidad y explicabilidad

- Transparencia ética y técnica
- Explicabilidad
- Interpretabilidad

4. Privacidad y manejo de datos

- Privacidad
- Gobernanza de datos

Evaluaciones

Actividad	Evaluación
Cuestionarios	30% nota final
Tareas	45% nota final
Foro de discusión	15% nota final
Trabajo Final	**10% nota final**

Las evaluaciones se aprueban si se obtiene el 50% de las respuestas correctas.

Plataforma e Información General

- **Duración**: 90 horas de dedicación total (24 directas y 66 indirectas) en 8 semanas.
- **Créditos:** 5 créditos UC
- **Requisitos**: MCD3010
- **Restricciones**: MDS O MAN
- **Conector:** Y
- **Carácter:** Mínimo
- **Tipo:** Cátedra y Laboratorio
- **Calificación:** Estándar
- **Nivel formativo:** Magíster

Política de entregas de evaluaciones calificadas fuera de plazo

En caso de entregar una evaluación calificada, sea esta Tarea o Cuestionario, fuera del plazo informado (fecha límite), se aplicará un descuento progresivo a la nota máxima por entrega tardía. El plazo para entregar evaluaciones o tareas fuera de plazo será de 7 días desde la fecha límite. Luego de los 7 días de plazo adicional, el alumno obtendrá una nota de 0% en dicha evaluación.

Si por razones de fuerza mayor, el alumno/a no pudiera rendir la prueba dentro del plazo regular o excepcional, deberá enviar una solicitud al correo de Soporte de su programa, adjuntando respaldos para que su requerimiento sea evaluado por la Unidad Académica (UA). La resolución de esta solicitud quedará a criterio de la UA.

Bibliografía

Mínima:

- Broussard, M. (2018). People problems. Artificial Unintelligence: How Computers Misunderstand the World (pp. 67-85). The MIT Press.
- Saleiro, P., Kuester, B., Stevens, A., Anisfeld, A., Hinkson, L., London, J., & Ghani, R. (2018). Aequitas: A Bias and Fairness Audit Toolkit. ArXiv, abs/1811.05577.
- Varsha, P.S. (2023). How can we manage biases in artificial intelligence systems – A systematic literature review. International Journal of Information Management Data Insights, 3(1), 100165. https://doi.org/10.1016/j.jjimei.2023.100165.
- Mulligan, D. K., Kroll, J. A., Kohli, N., & Wong, R. Y. (2019). This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1–36.
- Leavy, S., O'Sullivan, B., & Siapera, E. (2020). Data, Power and Bias in Artificial Intelligence. ArXiv, abs/2008.07341.
- Diakopoulos, N. (2020). Transparency. En Dubber, M. D., Pasquale, F., & Das, S. (Eds), The Oxford Handbook of Ethics of AI (pp. 197-213). Oxford University Press.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38.

- Kearns, M. & Roth, A. (2019). Algorithmic Privacy: From Anonymity to Noise. En The Science of Socially Aware Algorithm Design (pp. 22-36). Oxford University Press.
- Kroll, J. A. (2018). Data Science Data Governance [AI Ethics]. IEEE Security Privacy, 16(6), 61–70.

Complementaria

- Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI (SSRN Scholarly Paper No. 3518482). Social Science Research Network. https://doi.org/10.2139/ssrn.3518482.
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), Article 9. https://doi.org/10.1038/s42256-019-0088-2.
- Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency.
- Datta, A., Tschantz, M.C., & Datta, A. (2014). Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. ArXiv, abs/1408.6491.
- Zhong, Z. (2018). A Tutorial on Fairness in Machine Learning. Medium.
- Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. Journal of Decision Systems, 29(4), 260-278.
- Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2019). Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.
- Tomova, G., Tennant, P.W.G., Arriagada Bruneau, G.C., Gilthorpe, M.S. (2022). Distinguishing the transparency, explainability, and interpretability of algorithms – American Causal Inference Conference 2022 (ACIC), UC Berkeley, California, USA.
- McDermid, J. A., Jia, Y., Porter, Z., & Habli, I. (2021). Artificial intelligence explainability: the technical and ethical dimensions. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 379(2207), [20200363].
- O'Sullivan, C. (2023). Interpretable vs Explainable Machine Learning. Medium.