FICHA IDENTIFICATIVA

DATOS DE LA ASIGNATURA

Código: 46534
Nombre: Auditoría algoritmos
Ciclo: Máster Universitario Oficial
Créditos ECTS: 4,5
Curso académico: 2025-26

TITULACIONES

Titulación	Centro	Curso	Periodo
2258 - Máster Universitario en Sociedad Digital	Facultat de Ciències Socials	1	Segundo cuatrimeste

MATERIAS

Titulación	Materia	Carácter
2258 - Máster Universitario en Sociedad Digital	Auditoría algoritmos	OBLIGATORIA

COORDINACIÓN

RESUMEN

Esta asignatura explora la creciente influencia y omnipresencia de los algoritmos y la inteligencia artificial en nuestras vidas. En esta era, comprender la estructura, funcionamiento y el impacto de los algoritmos se convierte en un imperativo para evaluar su justicia, transparencia y eficacia. Abordaremos los principios fundamentales de la IA, incluyendo el diseño, implementación y optimización de algoritmos, así como la crítica evaluación de su impacto social, cultural y económico. Se discutirán conceptos clave como el sesgo de los algoritmos, la ética en la IA, la privacidad de los datos y la gobernanza digital, proporcionando a los estudiantes los conocimientos necesarios para realizar auditorías de algoritmos eficientes y responsables. Mediante el análisis de casos prácticos, los estudiantes aprenderán a identificar y mitigar riesgos, asegurando que las tecnologías de IA se apliquen de manera que promuevan la equidad y el bienestar social. Este curso pretende no solo brindar una comprensión técnica profunda de los algoritmos y la IA, sino también fomentar una reflexión crítica sobre sus implicaciones, preparando a los futuros profesionales para contribuir positivamente en la configuración de una sociedad digital más justa e inclusiva

CONOCIMIENTOS PREVIOS

RELACIÓN CON OTRAS ASIGNATURAS DE LA MISMA TITULACIÓN

No se han especificado restricciones de matrícula con otras asignaturas del plan de estudios.

OTROS TIPOS DE REQUISITOS

COMPETENCIAS / RESULTADOS DE APRENDIZAJE

-

Actuar con autonomía en el aprendizaje, tomando decisiones fundamentadas en diferentes contextos, emitiendo juicios en base a la experimentación y el análisis y transfiriendo el conocimiento a nuevas situaciones.

Adquirir y demostrar conocimientos avanzados sobre los principios y aplicaciones de la inteligencia artificial y su influencia en la sociedad digital

Aplicar e integrar los conocimientos teóricos y prácticos adquiridos para analizar casos reales de la economía digital, el trabajo, la educación, la cultura o la gobernanza en la sociedad digital

Comprender y demostrar un conocimiento detallado de las técnicas avanzadas en investigación social aplicadas al estudio de la sociedad digital, incluyendo el uso de big data, análisis de redes sociales, y metodologías digitales

Conocer y comprender, desde el propio ámbito de la titulación, las desigualdades por razón de sexo y género en la sociedad; integrar las diferentes necesidades y preferencias por razón de sexo y de género en el diseño de soluciones y resolución de problemas

Demostrar razonamiento crítico y autocrítico en el ámbito de la titulación, considerando aspectos tales como la ética profesional, los valores morales y las implicaciones sociales de las diferentes actividades realizadas

Diseñar proyectos de investigación en el ámbito de la sociedad digital, utilizando técnicas avanzadas de investigación social

Integrar conocimientos de economía, derecho, comunicación, cultura y sociología para abordar problemas interdisciplinarios en el contexto de la sociedad digital, y ser capaz de transmitir de un modo claro y sin ambigüedades los resultados procedentes de la investigación

Saber evaluar el impacto de las políticas y prácticas digitales, seleccionando la perspectiva teórica adecuada y la metodología precisa para diseñar y presentar propuestas de intervención

Ser capaz de analizar, evaluar e interpretar conjuntos de datos complejos y de gran escala en el contexto de la sociedad digital utilizando herramientas y técnicas avanzadas para extraer conocimientos significativos y tomar decisiones fundamentadas

DESCRIPCIÓN DE CONTENIDOS

Tema 1: Algoritmos y Machine Learning: se introducen los conceptos básicos de algoritmos y machine learning, definiendo la auditoría como proceso sistemático de evaluación y los algoritmos como

secuencias de instrucciones para resolver problemas. Se presenta el machine learning como técnica donde los algoritmos aprenden de datos, ejemplificado con predicciones inmobiliarias usando el dataset Boston Housing. Se examinan y analizan las variables predictivas en este conjunto de datos, realizando el análisis práctico con herramientas como Orange y Google Colab. Se concluye examinando problemas éticos críticos como sesgos algorítmicos, ilustrados con casos reales donde variables discriminatorias o datos sesgados perpetúan prejuicios sociales.

Tema 2: Historia, conceptos y aplicaciones de la IA. Chat GPT: se explora la evolución de la IA desde sus orígenes en los años 40 hasta los modelos actuales como Chat GPT, abarcando los ciclos históricos de auge y declive, los paradigmas de aprendizaje (supervisado, no supervisado y por refuerzo), conceptos técnicos fundamentales y arquitecturas de modelos. Se profundiza en el funcionamiento de los LLMs, su entrenamiento y sus aplicaciones. Se concluye analizando aspectos éticos y regulatorios mediante casos de estudio sobre sesgos algorítmicos, el ciclo de vida de algoritmos según normativas y los tipos de licencias para modelos de IA.

Tema 3: Privacidad de datos y seguridad: se abordan los fundamentos y desafíos actuales en la protección de información personal en entornos digitales, explorando conceptos clave como la k-anonimidad y la privacidad diferencial, e ilustrando sus aplicaciones y limitaciones mediante casos emblemáticos. Se analizan las principales amenazas de seguridad (malware, phishing, ransomware), los mecanismos de protección como la encriptación punto a punto, y las vulnerabilidades físicas y sociales de los sistemas informáticos. El tema también examina los dilemas éticos en la recopilación y uso de datos, las implicaciones de la interconexión digital para la privacidad colectiva.

Tema 4: Equidad algorítmica y sesgos: se examina cómo los algoritmos pueden perpetuar discriminaciones existentes, analizando casos como Word2Vec, y COMPAS, y se exploran diferentes definiciones de equidad algorítmica y la imposibilidad matemática de satisfacerlas todas simultáneamente. Como solución, se presentan herramientas para evaluar y mitigar sesgos durante el propio aprendizaje (la optimización del modelo), introduciendo conceptos como la Frontera de Pareto para visualizar el compromiso entre precisión y equidad. El machine learning no solo presenta riesgos en ámbito de sesgos y equidad, sino también la oportunidad para medirlos y mitigarlos sistemáticamente.

Tema 5: Transparencia y Explicabilidad: se analiza cómo hacer comprensibles los algoritmos de IA mediante dos enfoques: la transparencia, que revela el proceso de creación y funcionamiento del modelo, y la explicabilidad (XAI), que aclara cómo estos toman decisiones específicas. Se estudian explicaciones locales y globales a través de modelos inherentemente interpretables (regresiones, árboles) y métodos post-hoc para sistemas complejos (LIME, SHAP, técnicas de visualización). Aunque estas herramientas son útiles, explicar modelos complejos inevitablemente implica simplificaciones, y es importante entender estas limitaciones.

Tema 06: Seguridad en IA avanzada: se ilustran los riesgos y desafíos de seguridad en sistemas de IA avanzada. Se analizan las capacidades actuales de la IA de propósito general, los riesgos por uso malicioso, mal funcionamiento y problemas sistémicos, junto con estrategias de mitigación. Se estudian fenómenos como la convergencia instrumental y el comportamiento estratégico desalineado en modelos avanzados. Se abordan conceptos como AGI, ASI y singularidad tecnológica, así como los dilemas éticos de la IA autónoma en aplicaciones civiles y militares, destacando la necesidad de auditorías algorítmicas robustas.

VOLUMEN DE TRABAJO (HORAS)

ACTIVIDADES PRESENCIALES

Actividad	Horas
Teoría	30,00
Prácticas en aula	15,00
Total horas	**45,00**

ACTIVIDADES NO PRESENCIALES

Actividad	Horas
Asistencia a otras actividades	0,00
Elaboración de trabajos individuales o en grupo	0,00
Estudio y trabajo autónomo	0,00
Preparación de clases	0,00
Preparación de actividades de evaluación	0,00
Resolución de casos prácticos	0,00
Total horas	**0,00**

METODOLOGÍA DOCENTE

- *Clases magistrales*. Durante estas sesiones, el profesorado expondrá los conceptos fundamentales relacionados con la auditoría de algoritmos. Estas clases están diseñadas para proporcionar a los estudiantes una sólida base teórica que les permita comprender los desafíos y soluciones en el campo de la auditoría de algoritmos.

- *Exposición de artículos por parte de los alumnos.* Una parte crucial del aprendizaje en esta asignatura será la exposición y discusión de artículos científicos y técnicos recientes por parte de los alumnos. Esta actividad está destinada a fomentar el análisis crítico y la capacidad de síntesis de los estudiantes, así como a mantenerlos al día con los últimos avances y debates en el campo de la IA. Cada estudiante, o grupo de estudiantes, seleccionará un artículo de una lista propuesta por el profesorado, lo analizará y lo presentará al resto de la clase, generando un espacio de discusión y reflexión colectiva.

- *Trabajos grupales*. Los trabajos grupales son esenciales para desarrollar habilidades colaborativas y de resolución de problemas en contextos reales. En estos trabajos, los estudiantes aplicarán los conceptos aprendidos para diseñar, implementar o evaluar un proyecto de auditoría de algoritmos. Se incentivará la diversidad de enfoques, incluyendo el análisis de casos reales, la propuesta de soluciones a problemas éticos o técnicos identificados, o el desarrollo de herramientas para mejorar la transparencia y equidad de los sistemas de IA. Estos trabajos culminarán con una presentación ante la clase, donde se discutirán los resultados,

desafíos enfrentados y lecciones aprendidas.

- *Tutorías individuales y grupales*. Para apoyar el proceso de aprendizaje, se ofrecerán sesiones de tutoría, donde los estudiantes podrán resolver dudas específicas sobre el material de estudio, discutir avances en sus proyectos o profundizar en temas de interés particular.

EVALUACIÓN

La evaluación del aprendizaje de los conocimientos y competencias conseguidas por los estudiantes se hará de forma continuada a lo largo del curso, y constará de los siguientes bloques:

¿ 30%: SE1 - Prueba objetiva, consistente en un examen de cuestiones teóricas (evaluación de resultados de aprendizaje RA4, RA5, RA11, CT2)
¿ 60%: SE2 - Evaluación a partir de trabajos individuales y/o grupales a partir de la elaboración de trabajos/memorias y/o exposiciones orales:
  ¿ SE2-1 (10%) Presentación de un artículo: Presentación y análisis crítico de un artículo científico en el ámbito de la Inteligencia Artificial. Actividad NO RECUPERABLE (evaluación de resultados de aprendizaje RA4, RA11, CT2)
  ¿ SE2-2 (50%) Trabajo final: Análisis investigativo sobre los sesgos en un algoritmo de Inteligencia Artificia. Actividad NO RECUPERABLE (evaluación de resultados de aprendizaje RA6, RA8, RA9, RA15, CT4, CT7)
¿ 10%: SE3 - Evaluación continua de cada alumno, en base a participación activa durante las clases y/o la realización de cuestionarios. Actividad NO RECUPERABLE (evaluación de resultados de aprendizaje CT4, RA5, RA15)

Consideración sobre la evaluación: Para superar la asignatura, será necesario obtener una **calificación mínima de 4 sobre 10 en el apartado de evaluación SE1**

BIBLIOGRAFÍA

O. Santos, P. Radanlieve. Beyond the algorithm: AI, Security, Privacy and Ethics. Addison-Wesley, 2024

M. Kearns, A. Roth, Aaron. The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford University Press, 2022

M. Broussard. Artificial Unintelligence: How Computers Misunderstand the World. MIT Press, 2018

B. Franks. 97 Things About Ethics Everyone in Data Science Should Know. OReilly Media, 2020

Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter and Julia Lane. Big Data and Social Science, Capítulo 11, https://textbook.coleridgeinitiative.org/chap-bias.html, 2024

Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.). https://christophm.github.io/interpretable-ml-book

Bengio, Yoshua, et al. "International AI Safety Report." arXiv preprint arXiv:2501.17805 (2025).

Unión Europea. (2024). Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n.° 300/2008, (UE) n.° 167/2013, (UE) n.° 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial) (Texto pertinente a efectos del EEE). Diario Oficial de la Unión Europea, https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:32024R1689