Instituto Politécnico Nacional

Secretaría Académica
Dirección de Educación Virtual

Secretaría de Investigación y Posgrado
Dirección de Posgrado

SIP-30

Formato para registro de Unidades de aprendizaje 2021

I.- Datos de identificación de la unidad de aprendizaje

Unidad académica:	Centro de Investigación en Computación (CIC)									
Programa académico:	Doctorado en Ciencias de la Computación									
	X	Doctorado						Orientación profesional		
		Maestría					X	Orientado a la investigación		
		Especialidad						Con la industria		
								Especialidad médica		
	Sesión de colegio donde se propuso:		Ordinaria 7, 2023				Fecha de propuesta:		26 de julio de 2023	
Nombre de unidad de aprendizaje:	**INTELIGENCIA ARTIFICIAL EXPLICABLE**									
	Clave de la unidad de aprendizaje:		**23B8374**				Créditos:	5	*REP 2017*	
	Semanas del semestre	18		Horas a la semana:	4			Horas totales:	72	
Tipo de unidad de aprendizaje:	Obligatoria:			Optativa:	x		Observaciones:			
	Semestre:									
	Teórica (%):	100		Práctica (%)				Teórico-prácticas (%):		
Área del conocimiento:	Ingeniería y Ciencias Fisicomatemáticas		Ciencias Sociales y Administrativas			Ciencias Medico Biológicas			Interdisciplinario	X
Modalidad no escolarizada:	No escolarizada		Nombre de la Plataforma:							
	Mixta		Presencial (%):			En plataforma (%):				
Horas establecidas en el programa de estudios:	Presenciales (si procede) (horas x semana)					En plataforma (horas x semana):				

Formato para registro de Unidades de aprendizaje 2021

I. Aprendizajes que el estudiante deberá demostrar al finalizar

Conocimientos	Habilidades y destrezas	Actitudes y valores
1. Comprender los conceptos fundamentales de la explicabilidad en la inteligencia artificial, incluyendo la diferencia entre transparencia y explicabilidad. 2. Conocer los diferentes métodos y técnicas de explicabilidad, incluyendo la visualización de datos, la explicación basada en reglas, la explicación basada en instancias y la explicación basada en modelos. 3. Comprender los beneficios y limitaciones de cada método y técnica de explicabilidad, y saber cuándo y cómo aplicarlos en diferentes contextos. 4. Aplicar los métodos y técnicas de explicabilidad para diseñar sistemas de inteligencia artificial que sean transparentes, explicables y comprensibles para el sector público y privado. 5. Evaluar críticamente los sistemas de inteligencia artificial desarrollados y determinar si cumplen con los requisitos de explicabilidad y comprensibilidad.	1. Conocimiento teórico y práctico de la inteligencia artificial: el estudiante debería tener un conocimiento sólido de los conceptos básicos de la IA, incluyendo las técnicas de aprendizaje automático y de minería de datos, y saber cómo aplicarlos en problemas concretos. 2. Capacidad para interpretar y explicar modelos de IA: el estudiante debería ser capaz de interpretar los resultados de los modelos de IA y explicarlos de manera clara y comprensible a personas que no tienen conocimientos técnicos en IA. 3. Conciencia de las implicaciones éticas y sociales de la IA: el estudiante debería tener una comprensión de las implicaciones éticas y sociales de la IA, incluyendo temas como la privacidad, la discriminación y el sesgo algorítmico. 4. Habilidad para diseñar y ejecutar experimentos con IA: el estudiante debería ser capaz de diseñar y ejecutar experimentos con modelos de IA para probar diferentes hipótesis y evaluar la calidad de los resultados. 5. Destreza en el uso de herramientas y lenguajes de programación para la IA: el estudiante debería ser capaz de utilizar herramientas y lenguajes de programación populares para la IA,	1. Curiosidad y mente abierta: un estudiante debería tener una mente abierta y estar dispuesto a aprender cosas nuevas y a explorar diferentes perspectivas en el campo de la IA. 2. Responsabilidad: un estudiante debería ser responsable y tener en cuenta las posibles implicaciones éticas y sociales de la IA al tomar decisiones y diseñar modelos. 3. Transparencia: un estudiante debería ser transparente en sus procesos y resultados, y estar dispuesto a compartir información y explicar sus modelos a personas que no tienen conocimientos técnicos en IA. 4. Empatía: un estudiante debería tener empatía por las personas que podrían verse afectadas por los modelos de IA y tener en cuenta sus necesidades y preocupaciones. 5. Integridad: un estudiante debería ser honesto y ético en todas las decisiones que tome y en la forma en que utilice los modelos de IA. 6. Colaboración: un estudiante debería ser capaz de trabajar en equipo y colaborar con personas de diferentes disciplinas para lograr objetivos comunes en el campo de la IA.

Formato para registro de Unidades de aprendizaje 2021

	como Python, TensorFlow y Keras. 6. Capacidad para colaborar en proyectos de IA en equipo: el estudiante debería tener habilidades interpersonales y de comunicación que le permitan colaborar eficazmente en proyectos de IA con otros profesionales de diferentes disciplinas.	

Resolución que aborda la propuesta con su enfoque disciplinar

Se plantea cubrir los conceptos de interpretabilidad y explicabilidad en los modelos de IA, y cómo se pueden aplicar técnicas como la visualización y la descomposición de características para explicar los resultados de los modelos de IA a personas no técnicas. Se aborda el sesgo y la discriminación en los modelos de IA, y cómo se pueden aplicar técnicas como el muestreo justo y la corrección de sesgos para mitigar estos problemas.
Esta unidad de aprendizaje revisa a su vez las implicaciones éticas y sociales de la IA, y cómo se pueden aplicar principios éticos como la transparencia, la responsabilidad y la privacidad para garantizar que los modelos de IA se utilicen de manera responsable y ética. Por último, pero no menos importante, esta unidad busca explorar y generar aplicaciones específicas de la IA en diferentes campos, como la atención médica, el marketing y la seguridad cibernética, y cómo se pueden aplicar los conceptos y técnicas aprendidos en la unidad en estos contextos.

II. Proximidad formativa

Áreas multi, inter y transdisciplinarias	Líneas de Generación y Aplicación de Conocimiento	Sectores sociales
La interpretabilidad y explicabilidad en los modelos de IA pueden ser de interés para profesionales en campos multidisciplinares como: • Informática, • Análisis de Decisiones y estadística • Ciencia de datos • Psicología • Sociología • Filosofía	Inteligencia Artificial y Cómputo Científico	1. Investigacion de soluciones para problemas de media y alta complejidad, ubicados en el contexto local y global (pentahélice y Objetivos de Desarrollo Sostenible de la ONU). 2. Aplicaciones en diversos sectores productivos a nivel nacional e internacional, donde el uso de los modelos de IA sea el adecuado.

Formato para registro de Unidades de aprendizaje 2021

• Derecho y Ética. Cada uno de estos campos puede tener una perspectiva única y valiosa para abordar la interpretabilidad y explicabilidad en los modelos de IA. Por ejemplo, los expertos en ética y derecho pueden proporcionar una perspectiva crítica sobre cómo los modelos de IA pueden afectar a los derechos y libertades individuales y colectivas. En áreas interdisciplinares, la interpretabilidad y explicabilidad en los modelos de IA pueden ser objeto de estudio y desarrollo de proyectos en colaboración entre diferentes disciplinas, como la informática, la estadística y la sociología, para abordar problemas específicos. Por ejemplo, en el campo de la salud, la interpretabilidad y explicabilidad en los modelos de IA pueden ser importantes para que los médicos comprendan las decisiones de diagnóstico y tratamiento tomadas por los modelos de IA y confíen en ellas. En áreas transdisciplinares, la interpretabilidad y explicabilidad en los modelos de IA pueden ser parte de una perspectiva más amplia y compleja para abordar problemas sociales complejos, como la desigualdad, la sostenibilidad ambiental y el cambio social. La interpretabilidad y explicabilidad en los modelos de IA pueden ser importantes para comprender cómo los modelos de IA pueden ser utilizados en diferentes contextos y para diferentes propósitos.		

Formato para registro de Unidades de aprendizaje 2021

Estrategia de asociación: es recomendable que esta unidad de aprendizaje se complemente, ya sea en paralelo o consecuente, con las unidades de Datos Masivos y Minería de Datos, Internet de las cosas y Visualización de Datos. Por otra parte, es importante el uso de los recursos gratuitos brindados por parte de plataforma comerciales como AWS (Amazon Web Services), Google o IBM Cloud Solutions a instituciones educativas para el desarrollo práctico de los conocimientos adquiridos en la unidad de aprendizaje.

Promover estancias de investigación en otras instituciones educativas nacionales e internacionales con excelencia en temas relacionados al cómputo en la nube. Así como estancias profesionales en empresas e instancias de gobierno que implemente soluciones relacionadas al cómputo en la nube. Instar al estudiante a realizar actividades como la presentación de trabajos de investigación en diversos foros y congresos, actividades de difusión del conocimiento (videos, radio o podcasts), la publicación de artículos en revistas JCR y la generación de capital intelectual.

Formato para registro de Unidades de aprendizaje 2021

III Metodología de enseñanza – aprendizaje

Descripción
1.

Evidencias como proceso de aprendizaje	Evidencias integradoras (resultados que contribuyen al curriculum)	Ponderación
1.		

IV. Descripción de la participación esperada en el estudiante

Receptiva	Resolutiva	Autónoma	Estratégica

Formato para registro de Unidades de aprendizaje 2021

Contenido temático

1. Introducción a la inteligencia artificial explicativa e interpretable	2 horas
1.1. Definiciones y conceptos básicos	
1.2. Importancia de la explicabilidad e interpretabilidad en los modelos de IA	
1.3. Aplicaciones y usos de la IA explicativa e interpretable	
2. Métodos de interpretación de modelos de aprendizaje automático	12 horas
2.1. Análisis de importancia de características	
2.2. Métodos basados en reglas	
2.3. Visualización de modelos	
2.4. Métodos de análisis de sensibilidad	
2.5. Análisis de atribución	
3. Métodos de explicación de predicciones de modelos de aprendizaje automático	12 horas
3.1. Métodos basados en casos	
3.2. Métodos basados en contraste	
3.3. Métodos basados en preguntas y respuestas	
3.4. Métodos basados en narración	
3.5. Métodos basados en explicaciones interactivas	
4. Evaluación de la explicabilidad e interpretabilidad de modelos de aprendizaje automático	12 horas
4.1. Evaluación de la calidad de las explicaciones	
4.2. Evaluación de la utilidad de las explicaciones	
4.3. Evaluación de la fiabilidad de las explicaciones	
5. Herramientas y marcos de trabajo para la explicabilidad e interpretabilidad de modelos de aprendizaje automático	12 horas
5.1. Marcos de trabajo para la explicabilidad e interpretabilidad	
5.2. Herramientas de software para la interpretación y explicación de modelos de IA	
5.3. Bibliotecas de software y paquetes para la interpretación de modelos de IA	
5.4. Herramientas de visualización de datos para la IA explicativa e interpretable	
6. Aplicaciones de la inteligencia artificial explicativa e interpretable	10 horas
6.1. Seguridad y privacidad	
6.2. Detección de sesgos y discriminación en los modelos de IA	
6.3. Aplicaciones en la toma de decisiones clínicas	

Formato para registro de Unidades de aprendizaje 2021

6.4. Control de calidad y diagnóstico en la atención médica		
6.5. Aplicaciones en la predicción del riesgo crediticio		
6.6. Análisis forense y detección de fraude		
6.7. Aplicaciones en la justicia penal		
7. Ética y responsabilidad social en la inteligencia artificial	12 horas	
7.1. Consideraciones éticas en el diseño y uso de modelos de IA		
7.2. Responsabilidad social en el uso de la IA		
7.3. Riesgos y desafíos de la interpretación y explicación de modelos		
7.4. Regulaciones y marcos legales en la IA		

Formato para registro de Unidades de aprendizaje 2021

V. Secuencia programática

Formato para registro de Unidades de aprendizaje 2021

VI. Habilitadores tecnológicos

	Disposiciones	Especificaciones / descripción de efectos
	Conectividad	
	Habilidades digitales	
	Interoperabilidad	
	Datos abiertos	
	Big Data	
	Machine Learning	
	Simulación	
	Realidad aumentada	
	Otro…	

Conferencias magistrales	Notas complementarias
1. AIUK 2022 WORKSHOP - *ExplAIN: AI explainability in practice*. Disponible en https://www.youtube.com/watch?v=eHKi3gvoCcI&ab_channel=TheAlanTuringInstitute	
2. Fairness, part 1 - Moritz Hardt - MLSS 2020, Tübingen (2020). Disponible en https://www.youtube.com/watch?v=Igq_S_7IfOU&ab_channel=virtualmlss2020	

Formato para registro de Unidades de aprendizaje 2021

VII. Referencias

Documentales / electrónicas

1. Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. (2019) A Survey of Methods for Explaining Black Box Models. ACM Comput. Surv. 51, 5, Article 93, 42 pages. https://doi.org/10.1145/3236009.

2. Tim Miller, Robert Hoffman, Ofra Amir, Andreas Holzinger (2022) Special issue on Explainable Artificial Intelligence (XAI). Artificial Intelligence, Volume 307, 103705, ISSN 0004-3702, https://doi.org/10.1016/j.artint.2022.103705.

3. Explaining decisions made with AI. Co-badged Guidance. Information Commissioner's Office and The Alan Turing Institute. (Accedida en 2023-04-23) Disponible en https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-ai/

4. Wang, X., Zhang, Y. & Zhu, R. (2022). A brief review on algorithmic fairness. MSE 1, 7. https://doi.org/10.1007/s44176-022-00006-z.

5. Solon Barocas, Moritz Hardt, Arvind Narayanan (2019) Fairness and Machine Learning: Limitations and Opportunities. fairmlbook.org http://www.fairmlbook.org

6. The Royal Society (2019) Explainable AI: the basics Policy briefing The Royal Society ISBN 978-1-78252-433-5. www.royalsociety.org/ai-interpretability

7. Two Sigma (2023) Interpretability Methods in Machine Learning: A Brief Survey. (accedida en 2023-04-23) Disponible en https://www.twosigma.com/articles/interpretability-methods-in-machine-learning-a-brief-survey/

8. Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, Klaus-Robert Müller (2019) Explainable AI: Interpreting, Explaining and Visualizing Deep Learning (Lecture Notes in Computer Science, 11700) ISBN: 978-3030289539

9. Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. Inf. Fusion 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

10. Dang Minh, H. Xiang Wang, Y. Fen Li, and Tan N. Nguyen (2022) Explainable artificial intelligence: a comprehensive review. Artif. Intell. Rev. 55, 5, 3503–3568. https://doi.org/10.1007/s10462-021-10088-y.

11. D.Gunning (2017) Explainable Artificial Intelligence, Defense Advanced Research Projects Agency (DARPA). Disponible en https://www.darpa.mil/program/explainable-artificial-intelligence

12. Timo Speith (2022) A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 2239–2250. https://doi.org/10.1145/3531146.3534639.

13. Iván García-Magariño, Rajarajan Muttukrishnan, Jaime Lloret (2019) Human-centric AI for trustworthy IoT systems with explainable multilayer perceptrons. Special Section on Data Mining for Internet of Things. IEEE Access https://doi.org/10.1109/ACCESS.2019.2937521.

14. Wojciech Samek, Thomas Wiegand, Klaus-Robert Müller (2017) Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. arXiv 1708.08296v1 [cs.AI]. https://doi.org/10.48550/arXiv.1708.08296.

15. Harry Surden, Artificial Intelligence and Law: An Overview, 35 GA. ST. U. L. REV. 1305 (2019), available at https://scholar.law.colorado.edu/faculty-articles/1234

16. Owens, Emer, Barry, Sheehan, Martin Mullins, Martin, Cunneen, Juliane Ressel, and German Castignani (2022) Explainable Artificial Intelligence (XAI) in Insurance. Risks 10: 230. https://doi.org/10.3390/risks10120230.

17. Markus A. F., Kors J. A., Rijnbeek P. R. (2021) The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. Journal of Biomedical Informatics, Volume 113, 103655, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2020.103655.

18. Kate Crawford et al. (2019) AI Know Report. AI Know Institute. Disponible en https://ainowinstitute.org

Formato para registro de Unidades de aprendizaje 2021

VIII. Créditos y responsivas

	Responsabilidad	Nombre completo	Clave de nombramiento /No. de empleado
	Coordinador (Autor)	Amadeo José Argüelles Cruz	14976-EJ-20/6
	Participante (Coautor)	Dr. Oscar Camacho Nieto	15403-EH-22/6
	Participante (Coautor)	Dr. Antonio Alarcón Paredes	15782-EA-22
	Participante (Coautor)	Dra. Yenny Villuendas Rey	14160-EG-19/6
	Participante (Coautor)	Dr. Cornelio Yáñez Márquez	15344-EC-22
	Asesor didáctico / Diseñador Instruccional		
	Tecnólogo educativo / Comunicólogo		
	Corrector de estilo		
	Programador multimedia / Diseñador gráfico		
Otro…			

Instituto Politécnico Nacional

Secretaría Académica
Dirección de Educación Virtual

Secretaría de Investigación y Posgrado
Dirección de Posgrado

SIP-30

Formato para registro de Unidades de aprendizaje 2021

VERIFICACIÓN GENERAL DE LA PLANEACIÓN DIDÁCTICA	REVISIÓN DE LA PLANEACIÓN DIDÁCTICA (VIABILIDAD)
Por la División de Operación y Promoción al Posgrado de la SIP Nombre_____ ____ FIRMA_____ _____	Por la Subdirección de Diseño y Desarrollo de la DEV Nombre_____ ____ FIRMA_____ _____

VERIFICACIÓN PARA SU PUESTA EN OPERACIÓN	REVISIÓN TÉCNICO-PEDAGÓGICA PARA LA MODALIDAD
Por la Dirección de Posgrado Nombre_____ ____ FIRMA_____ _____ SELLO DE VALIDACIÓN	Por la Dirección para la Educación Virtual Nombre_____ ____ FIRMA_____ _____