1. Datos Generales de la asignatura

Nombre de la asignatura:	Inteligencia artificial explicable
Clave de la asignatura:	IAD-2416
SATCA[1]:	2-3-5
Carrera:	Ingeniería en Inteligencia Artificial

2. Presentación

Caracterización de la asignatura

La asignatura "Inteligencia Artificial Explicable" se destaca dentro del plan de estudios de la carrera de "Ingeniería en Inteligencia Artificial" por su enfoque en la transparencia, ética y comprensibilidad de los sistemas de IA. A través de cuatro temas principales, prepara a los estudiantes para diseñar, analizar e implementar sistemas de IA que los usuarios finales puedan entender y en los cuales puedan confiar.

Aportación al perfil de egreso:
Esta asignatura equipa a los estudiantes con habilidades críticas para desarrollar y evaluar sistemas de IA en términos de transparencia y ética, fundamentales para el ejercicio responsable de la profesión. Los egresados podrán liderar iniciativas para el desarrollo de tecnologías de IA más accesibles y justificables, acordes con las normativas y expectativas sociales actuales.

Importancia de la asignatura:
La Inteligencia Artificial Explicable (XAI, por *eXplainable Artificial Intelligence*) es clave en una era donde la IA tiene un impacto significativo en múltiples sectores. La capacidad de explicar y justificar las decisiones automatizadas es crucial para la integración efectiva de estas tecnologías en la sociedad, asegurando que sean justas, seguras y aceptadas.

Contenido de la asignatura:
Análisis de errores: Fundamentos de los errores en los modelos de IA, tipos de errores, y métricas de rendimiento.
Métodos de Interpretabilidad global: Técnicas como Importancia de las Características (Feature Importance), SHAP y ALE para obtener una comprensión global del comportamiento de los modelos.
Métodos de interpretabilidad Local: Herramientas como LIME, contraejemplos y razonamiento basado en casos (CBR) para explicaciones detalladas de decisiones individuales.
Extracción de reglas del modelo e Interacción Humano-IA: Métodos para extraer reglas de modelos complejos y diseñar interfaces efectivas para la interacción entre humanos y sistemas de IA.

[1] Sistema de Asignación y Transferencia de Créditos Académicos

Relación con otras asignaturas:

- Introducción a la Inteligencia Artificial: Esta asignatura proporciona las bases teóricas y prácticas iniciales sobre las que se construye "Inteligencia artificial explicable", permitiendo a los estudiantes comprender cómo los principios básicos de IA se aplican en la creación de sistemas explicables y éticos.
- Modelos de Aprendizaje Automático y Aprendizaje Profundo: Las técnicas de aprendizaje automático y profundo son esenciales para desarrollar modelos de IA. "Inteligencia artificial explicable" se enfoca en cómo estos modelos pueden ser interpretados y explicados, proporcionando herramientas para mejorar la transparencia y comprensibilidad, esenciales en la práctica profesional.
- Big Data y sus Aplicaciones y Analítica de Datos: Estas asignaturas se centran en el manejo y análisis de grandes volúmenes de datos. La habilidad para explicar los resultados obtenidos de estas técnicas es crucial, lo cual se aborda específicamente en la asignatura de explicabilidad, haciendo énfasis en la interpretación y justificación de los hallazgos derivados del análisis de grandes datasets.
- Visión Computacional y Análisis Computacional del Lenguaje: Estos campos de especialización en IA se benefician enormemente de las técnicas de explicabilidad para entender mejor los modelos complejos y sus decisiones. La asignatura ayuda a los estudiantes a aplicar conceptos de explicabilidad en áreas específicas, mejorando la accesibilidad y eficacia de las aplicaciones de IA.
- Taller de ética: Ambas asignaturas se complementan al enfocarse en la importancia de la ética en la IA. "Inteligencia Artificial Explicable" profundiza en cómo la transparencia y la comprensión de los sistemas de IA son fundamentales para adherirse a principios éticos y legales, una extensión práctica de los debates teóricos vistos en el Taller de Ética.

Desarrollo de sistemas inteligentes: Esta asignatura se relaciona directamente con la implementación práctica de sistemas de IA. "Inteligencia artificial explicable" proporciona las herramientas para asegurar que estos sistemas no solo sean efectivos sino también accesibles y comprensibles para los usuarios finales, fomentando su adopción y confianza.

Intención didáctica

- Abordaje de los contenidos:
- Los contenidos de la asignatura se tratan a través de una metodología híbrida, integrando clases teóricas para establecer bases conceptuales sólidas y actividades prácticas como laboratorios, estudios de caso y proyectos para aplicar estos conceptos en contextos reales. Esto permite a los estudiantes no solo aprender teoría, sino también practicar y experimentar con sus aplicaciones.
- Enfoque de tratamiento:
- El enfoque está centrado en la aplicabilidad y relevancia de la explicabilidad en IA, enfatizando la importancia de construir sistemas que sean tanto técnicamente eficaces como éticamente responsables. Se promueve un entendimiento profundo de cómo y por qué ciertas decisiones de IA se toman, para garantizar que los futuros ingenieros puedan desarrollar y justificar tecnologías de IA de manera transparente.
- Extensión y profundidad de los contenidos:

- Cada tema se aborda con una profundidad que permite a los estudiantes no solo comprender los conceptos fundamentales sino también aplicar técnicas avanzadas de explicabilidad. Se espera que los estudiantes logren una comprensión detallada de las técnicas más complejas, como SHAP y LIME, aplicándolas en escenarios variados y multidisciplinarios.

- Actividades estudiantiles destacadas:
- Proyectos colaborativos: fomentan el desarrollo de habilidades de trabajo en equipo y liderazgo.
- Presentaciones y defensas de casos: mejoran las habilidades de comunicación y argumentación.
- Análisis crítico de estudios de caso: desarrollan el pensamiento crítico y la capacidad de aplicar teoría a la práctica.
- Simulaciones en laboratorio: refuerzan la aplicación práctica de técnicas y teorías aprendidas.
- Competencias genéricas desarrolladas:
- La asignatura contribuye al desarrollo de competencias genéricas tales como:
- Pensamiento crítico y solución de problemas: a través del análisis y corrección de errores en modelos de IA.
- Comunicación efectiva: al explicar las decisiones de IA a audiencias no técnicas.
- Trabajo en equipo y colaboración: mediante proyectos grupales y actividades colaborativas.
- Responsabilidad ética y profesional: entendiendo y aplicando principios éticos en la creación de tecnología.
- Papel del docente:
- El docente debe actuar como facilitador y guía, proporcionando recursos y apoyo mientras fomenta un ambiente de aprendizaje activo y participativo. Deberá integrar desafíos contemporáneos y estudios de caso reales para contextualizar la teoría, además de evaluar el progreso a través de un enfoque basado en competencias, asegurándose de que los estudiantes no solo adquieran conocimiento, sino que también desarrollen habilidades prácticas y éticas relevantes. La retroalimentación continua y el fomento de un entorno de crítica constructiva son esenciales para el desarrollo integral de los estudiantes.

3. Participantes en el diseño y seguimiento curricular del programa

Lugar y fecha de elaboración o revisión	Participantes	Observaciones
Tecnológico Nacional de México del 4 al 06 de marzo del 2024.	Representantes de los Institutos Tecnológicos de: Celaya, Chihuahua, Iztapalapa III, La Paz, Matehuala, Mérida, Minatitlán, Querétaro, Saltillo, Tijuana. Instituto Tecnológico Superior de Teziutlán. Tecnológico de Estudios Superiores de Ixtapaluca.	Propuesta sintética de la carrera de Ingeniería en Inteligencia Artificial.
Tecnológico Nacional de México del 22 al 26 de abril del 2024	Representantes de los Institutos Tecnológicos de: Celaya, Chihuahua, Iztapalapa III, La Paz, Matehuala, Mérida, Minatitlán, Querétaro, Saltillo, Tijuana. Instituto Tecnológico Superior de Teziutlán, Tecnológico de Estudios Superiores de Ixtapaluca.	Diseño y/o desarrollo curricular de la carrera de Ingeniería en Inteligencia Artificial.
Tecnológico Nacional de México del 27 al 31 de mayo del 2024.	Representantes de los Institutos Tecnológicos de: Celaya, La Paz, Matehuala, Mérida, Minatitlán.	Consolidación curricular de la carrera de Ingeniería en Inteligencia Artificial.

4. Competencia(s) a desarrollar

Competencia(s) específica(s) de la asignatura
Analiza y aplica métodos de inteligencia artificial explicable para diseñar, implementar y evaluar sistemas de IA que sean transparentes y éticos, proporcionando decisiones que los usuarios puedan comprender y justificar, asegurando la fiabilidad y equidad de estos sistemas en una variedad de aplicaciones, enfocándose en cumplir con los estándares éticos y normativos vigentes.

5. Competencias previas

- Entiende los principios básicos y las técnicas fundamentales de la inteligencia artificial, lo que le permite abordar problemas complejos y desarrollar soluciones innovadoras en este campo.
- Posee una sólida base en matemáticas, especialmente en álgebra lineal, cálculo y optimización, necesaria para modelar y resolver problemas relacionados con la inteligencia artificial.
- Tiene conocimientos intermedios de estadística y probabilidad, fundamentales para entender y aplicar modelos predictivos y de clasificación en IA.
- Está familiarizado con conceptos y herramientas de programación avanzada, incluyendo el dominio de lenguajes y el uso de bibliotecas especializadas de IA, esencial para implementar y experimentar con algoritmos de inteligencia artificial.
- Tiene experiencia en el diseño y análisis de algoritmos, así como en estructuras de datos, lo que facilita la construcción de soluciones de IA eficientes y optimizadas.
- Posee la capacidad para trabajar eficazmente en equipo, fomentando la colaboración y el intercambio de conocimientos en proyectos tecnológicos, crucial para el éxito en entornos multidisciplinarios.
- Cuenta con habilidades de comunicación efectiva, tanto verbal como escrita, permitiéndole presentar ideas técnicas de manera comprensible a audiencias variadas, una habilidad esencial para la defensa y explicación de proyectos de IA.
- Identifica y evalúa las responsabilidades legales de los actores involucrados en el desarrollo y uso de sistemas de IA, e integra principios éticos y normativas de protección de datos desde el diseño hasta la implementación de soluciones basadas en IA.
- Reflexiona sobre el impacto ético y social de la inteligencia artificial, relaciona la ética con el desarrollo científico y tecnológico para evaluar sus implicaciones sociales, y asume una responsabilidad profesional en su aplicación.
- Fundamenta la práctica ética en la toma de decisiones estratégicas y operativas, aplicándola a la resolución de problemas en instituciones y organizaciones, enfatizando en la integridad y la justicia.

6. Temario

No.	Temas	Subtemas
1	Análisis de errores.	1.1. Fundamentos de la IA explicable. 1.1.1. Principios de transparencia en IA. 1.1.2. Ética y responsabilidad en algoritmos. 1.2. Tipos de errores. 1.2.1. 1.2.1 Errores de sesgo. 1.2.2. Errores de varianza. 1.2.3. Errores de predicción. 1.3. Métricas de rendimiento del modelo. 1.3.1. Precisión y exactitud. 1.3.2. Sensibilidad y valor F1. 1.3.3. Curvas ROC y AUC

2	Métodos de interpretabilidad global.	2.1. Importancia de la característica (Feature Importance). 2.1.1. Métodos basados en modelos. 2.1.2. Métodos basados en permutaciones. 2.2. Explicaciones aditivas de Shapley (SHAP). 2.2.1. Fundamentos teóricos de SHAP. 2.2.2. Aplicaciones prácticas de SHAP. 2.3. Efectos locales acumulados (ALE). 2.3.1. Introducción a ALE. 2.3.2. Comparación de ALE con otros métodos de interpretación.
3	Métodos de interpretabilidad local.	3.1. Explicaciones agnósticas del modelo local interpretable (LIME). 3.1.1. Principios de LIME. 3.1.2. Implementación de LIME en diferentes tipos de modelos. 3.2. Contraejemplos. 3.2.1. Creación de contraejemplos. 3.2.2. Uso de contraejemplos para mejorar modelos. 3.3. Razonamiento basado en casos (CBR). 3.3.1. Metodología de CBR. 3.3.2. Aplicaciones de CBR en IA.
4	Extracción de reglas del modelo e interacción humano-IA.	4.1. Extracción de reglas del modelo. 4.1.1. Técnicas de extracción de reglas. 4.1.2. Evaluación de la efectividad de las reglas extraídas. 4.2. Interacción humano-IA. 4.2.1. Diseño de interfaces de usuario para IA. 4.2.2. Estudios de caso en interacción humano-IA.

7. Actividades de aprendizaje de los temas

1. Análisis de los errores	
Competencias	Actividades de aprendizaje
Específica(s): ● Analiza y clasifica los tipos de errores en modelos de IA para evaluar y mejorar su rendimiento. *Genéricas:* ● Capacidad de pensamiento analítico. ● Capacidad de atención al detalle. ● Capacidad de aprendizaje autónomo.	Ejercicio de diagnóstico de errores: Los estudiantes realizarán un análisis detallado de errores utilizando métricas de rendimiento específicas. Esta actividad incluirá la búsqueda y selección de información relevante, análisis crítico de resultados, y la propuesta de estrategias de mejora basadas en observaciones y datos recopilados. Además, se fomentará la comunicación efectiva mediante la presentación de los hallazgos en forma oral y escrita, promoviendo la colaboración y el intercambio de ideas en grupos.
2. Métodos de interpretabilidad global	
Competencias	Actividades de aprendizaje
Específica(s): ● Aplica técnicas de interpretabilidad global como Importancia de las Características (Feature Importance), SHAP y ALE para explicar las decisiones de modelos de IA a nivel global. *Genéricas:* ● Capacidad de pensamiento crítico. ● Capacidad de comunicación efectiva. ● Capacidad de trabajo en equipo.	Taller de interpretabilidad global: Implementación en grupos de métodos de interpretabilidad en modelos preentrenados. Esta actividad promoverá el uso de tecnologías de la información, análisis crítico de las técnicas aplicadas, y la organización y planeación del trabajo grupal. Los estudiantes presentarán sus hallazgos, propiciando el uso adecuado de terminología técnica y fomentando el debate y la reflexión crítica entre los participantes.
3. Métodos de interpretabilidad local	
Competencias	Actividades de aprendizaje
Específica(s): ● Utiliza técnicas de interpretabilidad local como LIME, contraejemplos y CBR para explicar decisiones individuales de modelos de IA. *Genéricas:* ● Capacidad de resolución de problemas. ● Capacidad de innovación. ● Capacidad de pensamiento crítico.	Laboratorio de interpretabilidad local: Aplicación de herramientas como LIME en casos de estudio específicos para explicar predicciones de modelos. Los estudiantes llevarán a cabo observaciones detalladas, identificarán y manejarán variables relevantes, y plantearán hipótesis sobre la influencia de características individuales en las decisiones del modelo. Esta actividad también incentivará la investigación y aplicación de conocimientos en un contexto práctico y relevante para su futuro desempeño profesional.

4.- Extracción de reglas del modelo e interacción humano-IA	
Competencias	Actividades de aprendizaje
Específica(s): ● Extrae reglas comprensibles de modelos de IA y diseñar interfaces para la interacción efectiva humano-IA. *Genéricas:* ● Capacidad de diseño y creatividad. ● Habilidades de comunicación interpersonal. ● Capacidad de trabajo colaborativo.	Proyecto de diseño de interfaz: Desarrollo de interfaces de usuario que permitan a los no técnicos interactuar con sistemas de IA. Esta actividad integra el diseño creativo, la planificación de proyectos y la implementación técnica. Los estudiantes trabajarán en equipo, aplicando conceptos de accesibilidad y usabilidad, y realizarán pruebas y ajustes basados en feedback del grupo, fomentando un enfoque sostenible y responsable en el diseño tecnológico.

8. Práctica(s)

- Práctica de diagnóstico y corrección de errores en modelos de IA
  - Objetivo: Aplicar técnicas de análisis y corrección de errores para mejorar la transparencia y el rendimiento de modelos de IA.
  - Competencias Específicas: Analizar y clasificar errores en modelos de IA.
  - Competencias genéricas: Pensamiento crítico, resolución de problemas, y comunicación efectiva.
  - Actividades: Los estudiantes realizarán ejercicios prácticos para identificar, analizar y proponer soluciones a errores en conjuntos de datos reales, seguido de la presentación de sus resultados y metodologías en un informe técnico.

- Proyecto de desarrollo de un modelo de IA explicable
  - Objetivo: Diseñar, implementar y evaluar un modelo de inteligencia artificial con un enfoque en la transparencia y la interpretabilidad.
  - Competencias específicas: Aplicar métodos de interpretabilidad global y local para justificar decisiones tomadas por sistemas de IA.
  - Competencias genéricas: Innovación, trabajo en equipo, y habilidades de diseño y creatividad.
  - Actividades: En grupos, los alumnos desarrollarán un proyecto donde construirán modelos de IA explicables, documentando cada etapa del proceso y finalizando con una demostración y defensa del proyecto ante la clase.

- Talleres de resolución de problemas con casos prácticos
  - Objetivo: Utilizar conocimientos adquiridos para resolver problemas complejos en escenarios del mundo real mediante la aplicación de técnicas de IA explicables.
  - Competencias específicas: Utilizar técnicas de IA explicables en casos prácticos reales.
  - Competencias genéricas: Capacidad de análisis y síntesis, y aplicación de conocimientos en la práctica.
  - Actividades: Los estudiantes participarán en sesiones de taller donde trabajarán en equipos para resolver casos prácticos, utilizando estrategias de IA explicables y presentando sus soluciones y razonamientos en un formato interactivo.

- Investigación y análisis crítico de avances en IA explicable
  - Objetivo: Profundizar en el entendimiento de los desarrollos recientes y los desafíos éticos de la IA explicable.
  - Competencias específicas: Investigar y analizar críticamente la literatura científica y tecnológica.
  - Competencias genéricas: Capacidad de aprendizaje autónomo y comunicación efectiva.
  - Actividades: Los alumnos llevarán a cabo una investigación sobre un tema actual en IA explicable, escribirán un ensayo crítico y lo presentarán a la clase, fomentando el debate y la reflexión sobre temas éticos y tecnológicos.

9. Proyecto de asignatura

El objetivo del proyecto que planteé el docente que imparta esta asignatura, es demostrar el desarrollo y alcance de la(s) competencia(s) de la asignatura, considerando las siguientes fases:

Fundamentación: marco referencial (teórico, conceptual, contextual, legal) en el cual se fundamenta el proyecto de acuerdo con un diagnóstico realizado, mismo que permite a los estudiantes lograr la comprensión de la realidad o situación objeto de estudio para definir un proceso de intervención o hacer el diseño de un modelo.

Planeación: con base en el diagnóstico en esta fase se realiza el diseño del proyecto por parte de los estudiantes con asesoría del docente; implica planificar un proceso: de intervención empresarial, social o comunitario, el diseño de un modelo, entre otros, según el tipo de proyecto, las actividades a realizar los recursos requeridos y el cronograma de trabajo.

Ejecución: consiste en el desarrollo de la planeación del proyecto realizada por parte de los estudiantes con asesoría del docente, es decir en la intervención (social, empresarial), o construcción del modelo propuesto según el tipo de proyecto, es la fase de mayor duración que implica el desempeño de las competencias genéricas y específicas a desarrollar.

Evaluación: es la fase final que aplica un juicio de valor en el contexto laboral-profesión, social e investigativo, ésta se debe realizar a través del reconocimiento de logros y aspectos a mejorar se estará promoviendo el concepto de "evaluación para la mejora continua", la metacognición, el desarrollo del pensamiento crítico y reflexivo en los estudiantes.

10. Evaluación por competencias

Son las técnicas, instrumentos y herramientas sugeridas para constatar los desempeños académicos de las actividades de aprendizaje.

- Evaluación de competencias específicas:
  - Metodología: La evaluación de las competencias específicas se centrará en la capacidad del estudiante para aplicar y utilizar adecuadamente las técnicas y herramientas de IA explicable aprendidas en cada tema. Esta evaluación se llevará a cabo mediante:
    - Proyectos prácticos: Evaluación del diseño, implementación y evaluación de modelos de IA, centrados en la transparencia y la interpretabilidad.
    - Resolución de problemas y casos prácticos: Evaluación basada en la aplicación de conocimientos a situaciones reales, demostrando la habilidad para aplicar técnicas de interpretación tanto globales como locales.
    - Presentaciones de informes y resultados: Evaluación de la claridad, profundidad y precisión en la comunicación de los hallazgos y procesos utilizados.
- Evaluación de competencias genéricas:
  - Metodología: La evaluación de las competencias genéricas se realizará observando cómo los estudiantes aplican habilidades transversales en el contexto de la inteligencia artificial:
    - Participación en discusiones: Evaluación de la capacidad para contribuir con ideas relevantes, argumentar posiciones y participar activamente en debates relacionados con la ética y la aplicación de la IA.
    - Colaboración en proyectos grupales: Evaluación del trabajo en equipo, destacando la coordinación, el liderazgo, la resolución de conflictos y la contribución equitativa.
    - Principios éticos en soluciones: Evaluación del grado en que los estudiantes incorporan consideraciones éticas en sus proyectos y decisiones técnicas.
- Instrumentos de evaluación:
  - Exámenes escritos: Para evaluar el conocimiento teórico y la comprensión de los principios fundamentales de la IA explicable.
  - Evaluación continua de proyectos: Incluyendo revisiones de avances, evaluaciones finales y defensas de proyectos.
  - Presentaciones orales: Para evaluar la capacidad de comunicar efectivamente los hallazgos, justificar decisiones técnicas y defender puntos de vista éticos.

- Evaluación de informes: Para medir la habilidad de documentar procesos y resultados de manera clara y profesional.

- Retroalimentación y mejora continua:
  - Retroalimentación constante: Se proporcionará a los estudiantes después de cada evaluación para facilitar su comprensión de áreas de fortaleza y oportunidades de mejora.
  - Reflexión dirigida: Los estudiantes realizarán autoevaluaciones y reflexionarán sobre sus aprendizajes y el desarrollo de competencias a lo largo del curso.

11. Fuentes de Información

1. Bhargava, A. (2016). Grokking Algorithms: An Illustrated Guide for Programmers and Other Curious People. Manning Publications.
2. Cestero, V., & Caballero, E. (2023). Inteligencia Artificial: Fundamentos matemáticos, algorítmicos y metodológicos (Spanish Edition).
3. Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). Introduction to Algorithms. MIT Press.
4. Ebel, F. (2019). Algoritmia - Técnicas fundamentales de programación: Ejemplos en Python (numerosos ejercicios corregidos). BTS, DUT informática recursos informáticos. ediciones ENI.
5. Goodrich, M. T., & Tamassia, R. (2014). Data Structures and Algorithms in Java. John Wiley & Sons.
6. Hemant Jain. (2021). Estructuras de datos y algoritmos simplificados.
7. Lafore R. (2021). Estructura de datos y algoritmos en java.
8. Nadal, M. (2019). Estructuras de datos y algoritmos: Guía ilustrada para programadores [Pasta blanda]. Anaya multimedia. ISBN-13: 978-8441545199
9. Skiena, S. S. (2008). The Algorithm Design Manual. Springer Science+Business Media, LLC.
10. Sedgewick, R., & Wayne, K. (2011). Algorithms. Addison-Wesley.