
Universidad
Tecmilenio®

Universidad Tecmilenio

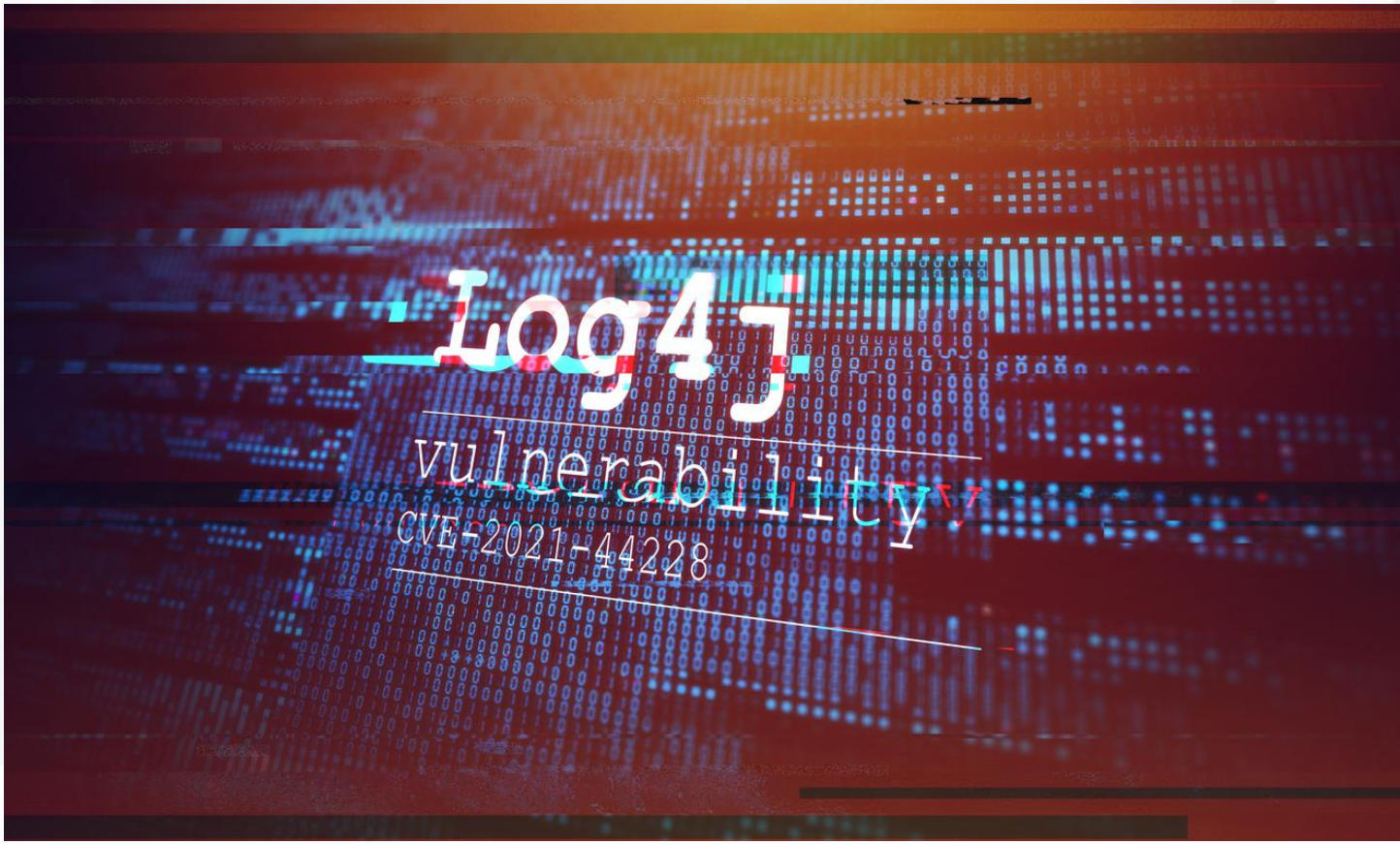
Ética aplicada a la inteligencia artificial

Protección de la privacidad

Introducción

En la medida que las compañías adoptan nuevas tecnologías, se abren nuevas brechas de seguridad, es decir, la inteligencia artificial y el aprendizaje automático no solo aportan grandes beneficios, sino nuevas vulnerabilidades.

En este tema conocerás:
•Las formas para maximizar la privacidad y la seguridad de los datos.
•La definición de privacidad diferencial.



Explicación

Es posible aprovechar las comodidades que ofrece la tecnología y también reducir el riesgo de exposición de los datos mediante acciones que fortalecen la seguridad de la información:

1	Reforzar las contraseñas de aplicaciones y dispositivos más externas, con a menos un número, una mayúscula y un carácter especial, así como utilizar una contraseña única para cada dispositivo o aplicación.
2	Agregar una capa adicional de protección habilitando la autenticación multifactor (MFA) o la autentificación de dos factores (2FA).
3	Revisar el nivel de privacidad seleccionada en cada aplicación o dispositivo.
4	No asumir que la configuración por defecto es la mejor para el usuario. Personalizarlas alternativas ofrecidas por cada aplicación o dispositivo.
5	Conectar con dispositivos de Smart home en una red distinta a la utilizada por las computadoras personales, teléfonos o tabletas.
6	Reducir la cantidad de información personal que se comparte por medio de aplicaciones o dispositivos.

Explicación

La principal premisa de la privacidad diferencial es garantizar que las personas, cuya información personal se encuentra en alguna base de datos, no se vean afectadas de algún modo. Gracias a esta técnica, se pretende entrenar modelos de aprendizaje automático con los datos de los clientes sin que estos aprendan detalles sobre ellos.
Utilizando la privacidad diferencial, una consulta de información en un sistema regresará un conjunto de datos modificado, al que se le agrega un cierto nivel de ruido extraído de forma aleatoria a partir de una distribución generada con los datos originales.

Explicación

La privacidad diferencial tiene cuatro propiedades valiosas:

1. La cuantificación de la pérdida de privacidad.

Tiene mecanismos que miden la pérdida de privacidad.

2. Composición.

Permite diseñar algoritmos diferenciales complejos a partir de elementos más simples.

3. Privacidad grupal.

Permite el control de la pérdida de privacidad de grupos (información de familias, por ejemplo).

4. Cerrada en postprocesamiento.

Es inmune al postprocesamiento. No se puede vulnerar.

Actividad

¿Cuáles son los riesgos de utilizar la tecnología *Clearview AI*? ¿Debería permitirse el uso de este tipo de algoritmos? ¿Qué opinión tienes sobre el uso de datos públicos para este tipo de situaciones? Puedes considerar el siguiente artículo

Cierre

De la misma manera que las organizaciones implementan estrategias de seguridad de la información, cada persona debe ser consciente del riesgo que implica compartir información sensible en aplicaciones y dispositivos, con el fin de realizar las mejores prácticas orientadas a proteger su privacidad.

Universidad Tecmilenio

Ética aplicada a la inteligencia artificial

Modelos transparentes

Introducción

Si bien es cierto que un modelo no está realizando alguna acción ilegal o incorrecta, los datos lo han entrenado para reforzar cierto tipo de prejuicios.

En este tema conocerás:
Causas que producen predicciones sesgadas.
Los conceptos de modelos transparentes, explicables, justos y de caja negra.
Distintos enfoques de la inteligencia artificial explicable usados para crear modelos explicables.

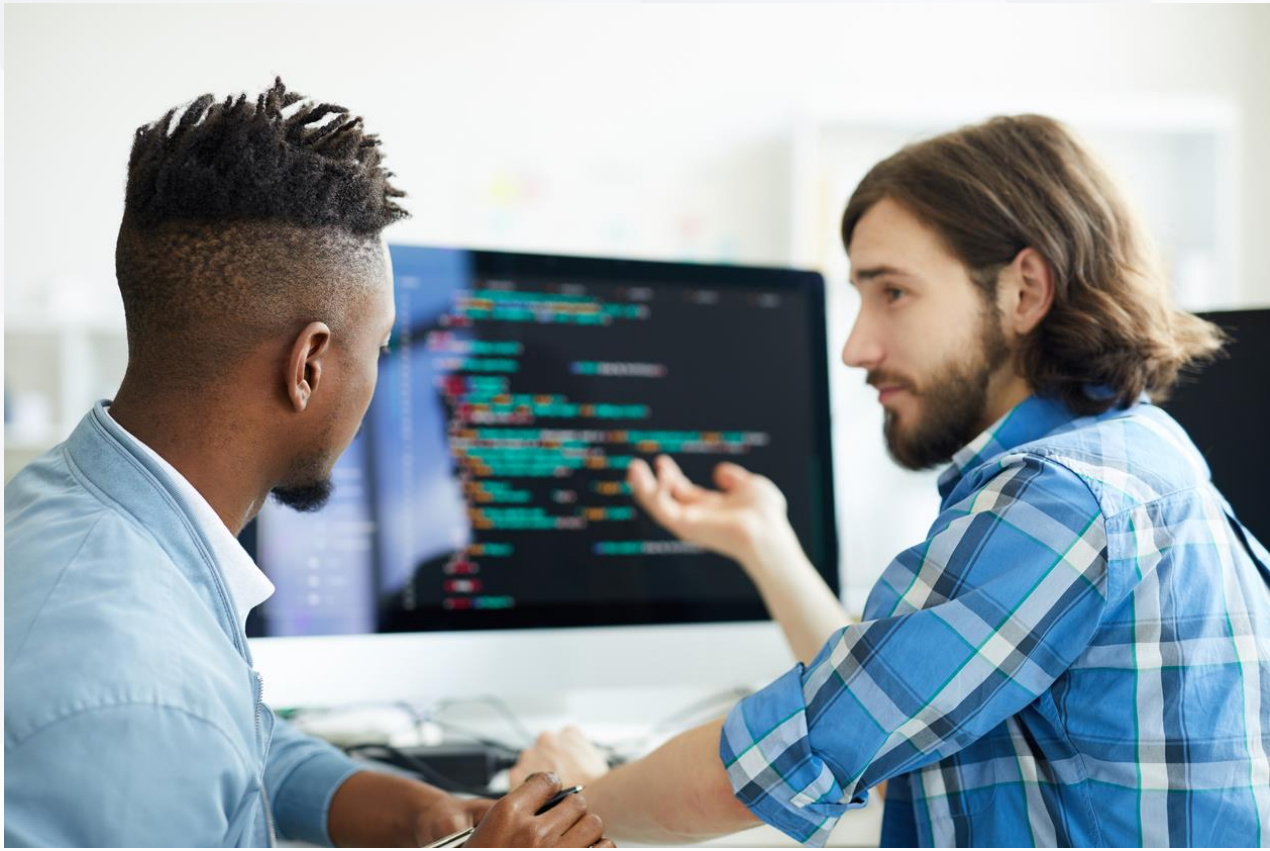
Explicación

Existen múltiples causas que pueden provocar sesgos en las predicciones hechas por los modelos de aprendizaje automático:

1. La idoneidad de los datos para representar a diferentes grupos.	Los patrones poco frecuentes o específicos pueden ser minimizados por el modelo al generalizar, por lo que esos registros menores en cantidad pueden descuidarse injustamente.
2. El sesgo inherente a los datos.	La falta de datos no necesariamente se debe al tamaño del grupo que representa, sino a la metodología de recolección de datos que puede dar ventaja o excluir a ciertos grupos, por ejemplo, recopilación de datos en un solo idioma.
3. La idoneidad del modelo para describir a cada grupo.	La arquitectura del modelo puede describir mejor a un grupo sobre los otros. Por ejemplo, un modelo lineal puede ser adecuado para describir a un grupo específico, pero no a los demás dentro del conjunto de datos.

Explicación

El principio de **explicabilidad** de un algoritmo sostiene que cualquier decisión que se tome debe ser comprensible y accesible para que las personas involucradas en esta decisión puedan refutar los errores observados o los datos incorrectos.
Se explica el fundamento del proceso de la toma de decisiones, revela las fortalezas y debilidades del proceso y da una idea de cómo se puede comportar el algoritmo en el futuro. La explicabilidad se refiere a entender por qué un modelo produce un resultado particular sin necesidad de comprender los aspectos matemáticos del algoritmo.



Explicación

El uso de modelos transparentes está encaminado a clarificar los siguientes aspectos:

- La forma en la que el algoritmo entrenó al modelo para tomar las decisiones.
- El tipo de información que el modelo consideró para su toma de decisiones.
- La forma en la que se ponderaron y trataron los parámetros del algoritmo de entrenamiento.
- El tipo de operaciones realizadas durante el entrenamiento.

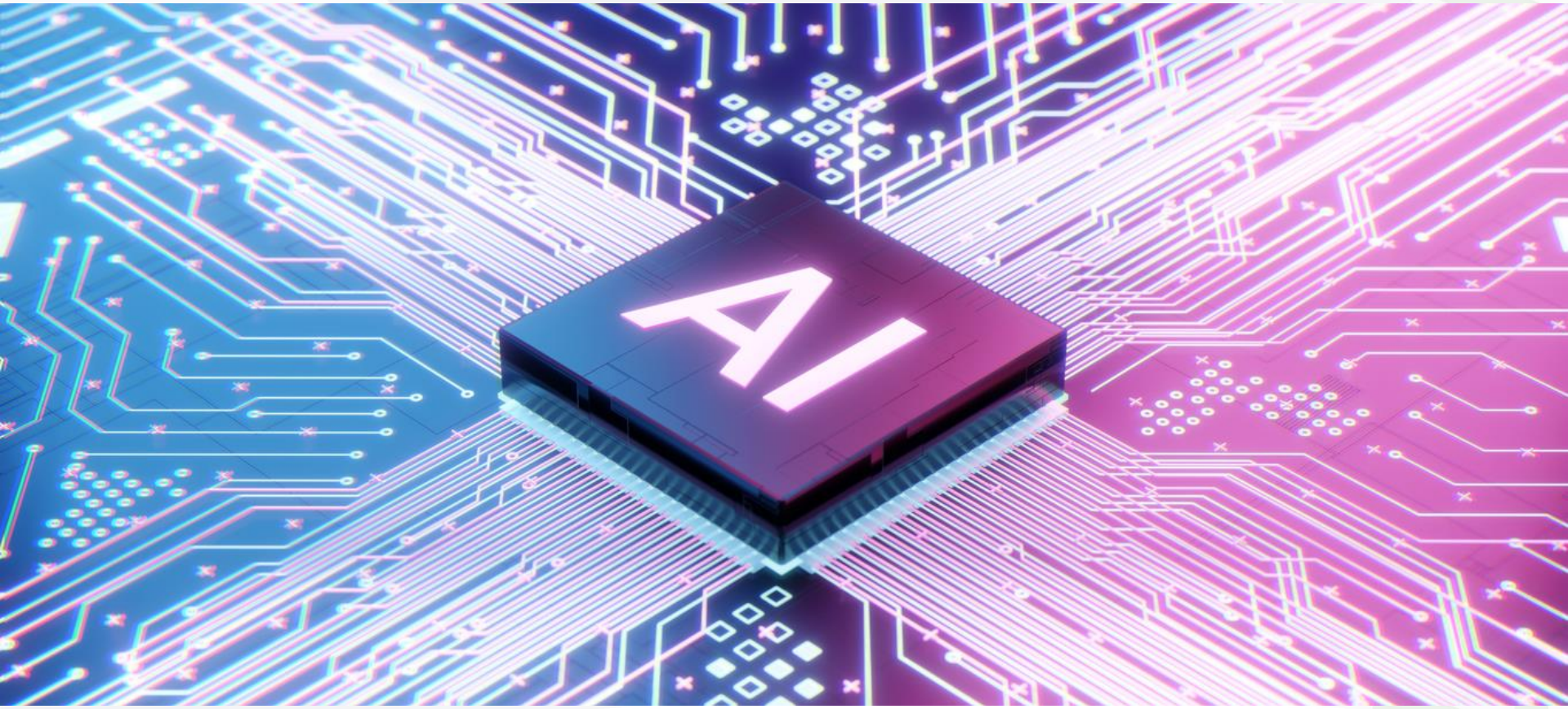
Explicación

La inteligencia artificial explicable (XAI, por sus siglas en inglés) es una nueva rama de la inteligencia artificial que intenta desmitificar los modelos de aprendizaje automático, ofreciendo a los usuarios finales, no solo las predicciones, sino además la evidencia de cómo se llegaron a ellas (Barredo, Díaz, Del Ser, Bennetot, Tabik, Barbado, García, Gil, Molina, Benjamins, Chatila y Herrera, 2020).
En XAI existen diferentes enfoques para generar modelos explicables:

SAHP (Shapley Additive Explanations)	Este enfoque tiene como objetivo explicar la predicción de un modelo, midiendo la contribución de cada atributo de los datos a esa predicción (Merryck y Taly, 2020).
LIME (Local Interpretable Model-Angnostic Explanations)	Por lo general se utiliza con modelos no lineales e intenta linealizar los espacios no lineales, separando el espacio de atributos original en subespacios lineales que puedan utilizarse en modelos lineales explicables (Ribeiro, Singh y Guestrin (2016).
Importancia de atributos con base en árboles	Usa la reducción promedio de las impurezas en un árbol (o en un bosque de árboles) provocada por cada atributo de entrada. Según esta metodología, los atributos que separan los nodos más cerca de la parte superior del árbol tendrían mas peso al crear la predicción.
Gráficos de dependencia parcial (PDP)	Sintetiza la relación entre las variables de entrada y las predicciones. Se utiliza para visualizar la dependencia de la predicción con las diferentes variables del problema.
Gráficos de esperanza individual condicional (ICE)	Se utiliza para tratar de comprender la predicción de cada observación al simular lo que sucederá si los datos de entrada fueran ligeramente distintos.

Actividad

¿Cómo describirías la explicabilidad en los algoritmo computacionales? Considera el siguiente artículo:
Ribeiro, M., Singh, S., y Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. Recuperado de https://arxiv.org/abs/1602.04938



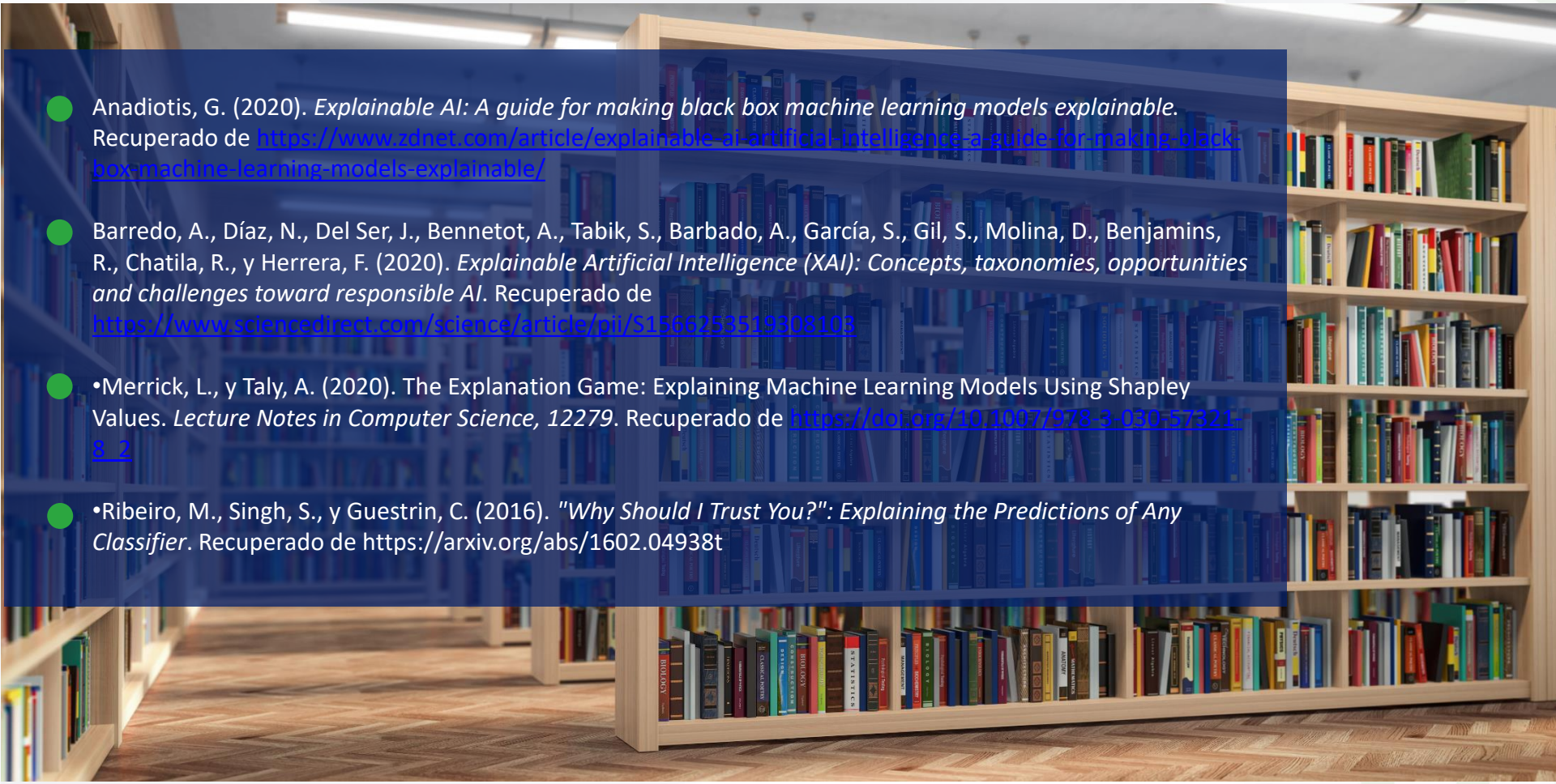
Cierre

El aprendizaje automático ha demostrado múltiples veces un asombroso poder predictivo debido a su capacidad para aprender relaciones complejas a partir de los datos, sin embargo, la aplicación de modelos en áreas como la medicina, las finanzas y la educación es actualmente complicada, debido a las preocupaciones éticas que rodean el uso de algoritmos como herramientas en la toma de decisiones.

La transparencia es un requisito indispensable para generar confianza y adopción de estos modelos. Una forma de lograrlo es utilizando modelos de familias que se consideran explicables, como los lineales, árboles de decisión, conjuntos de reglas y de decisiones, aditivos generalizados y métodos de razonamiento basados en casos que a menudo proporcionan un equilibrio adecuado entre explicabilidad y desempeño (Anadiotis, 2020).

El uso de modelos explicables puede beneficiar a usuarios (¿por qué se rechazó la solicitud de crédito?), entidades reguladoras (evidencia de que los modelos son justos) y desarrolladores (¿qué tan bien funciona el modelo?).

Bibliografía



- Anadiotis, G. (2020). *Explainable AI: A guide for making black box machine learning models explainable.* Recuperado de https://www.zdnet.com/article/explainable-ai-artificial-intelligence-a-guide-for-making-black-box-machine-learning-models-explainable/

- Barredo, A., Díaz, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil, S., Molina, D., Benjamins, R., Chatila, R., y Herrera, F. (2020). *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI*. Recuperado de https://www.sciencedirect.com/science/article/pii/S1566253519308103

- •Merrick, L., y Taly, A. (2020). The Explanation Game: Explaining Machine Learning Models Using Shapley Values. *Lecture Notes in Computer Science, 12279*. Recuperado de https://doi.org/10.1007/978-3-030-57321-8_2

- •Ribeiro, M., Singh, S., y Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. Recuperado de https://arxiv.org/abs/1602.04938t