# Explicabilidad de modelos de aprendizaje automático utilizados en la industria aeroespacial

Francisco Javier Cantero Zorita

XAI: Inteligencia Artificial eXplicable

31 de octubre de 2025

IA3 Cátedra de Inteligencia Artificial en Aeronáutica y Aeroespacio
Universidad de Alcalá
indra
Financiado por la Unión Europea NextGenerationEU
MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
Plan de Recuperación, Transformación y Resiliencia

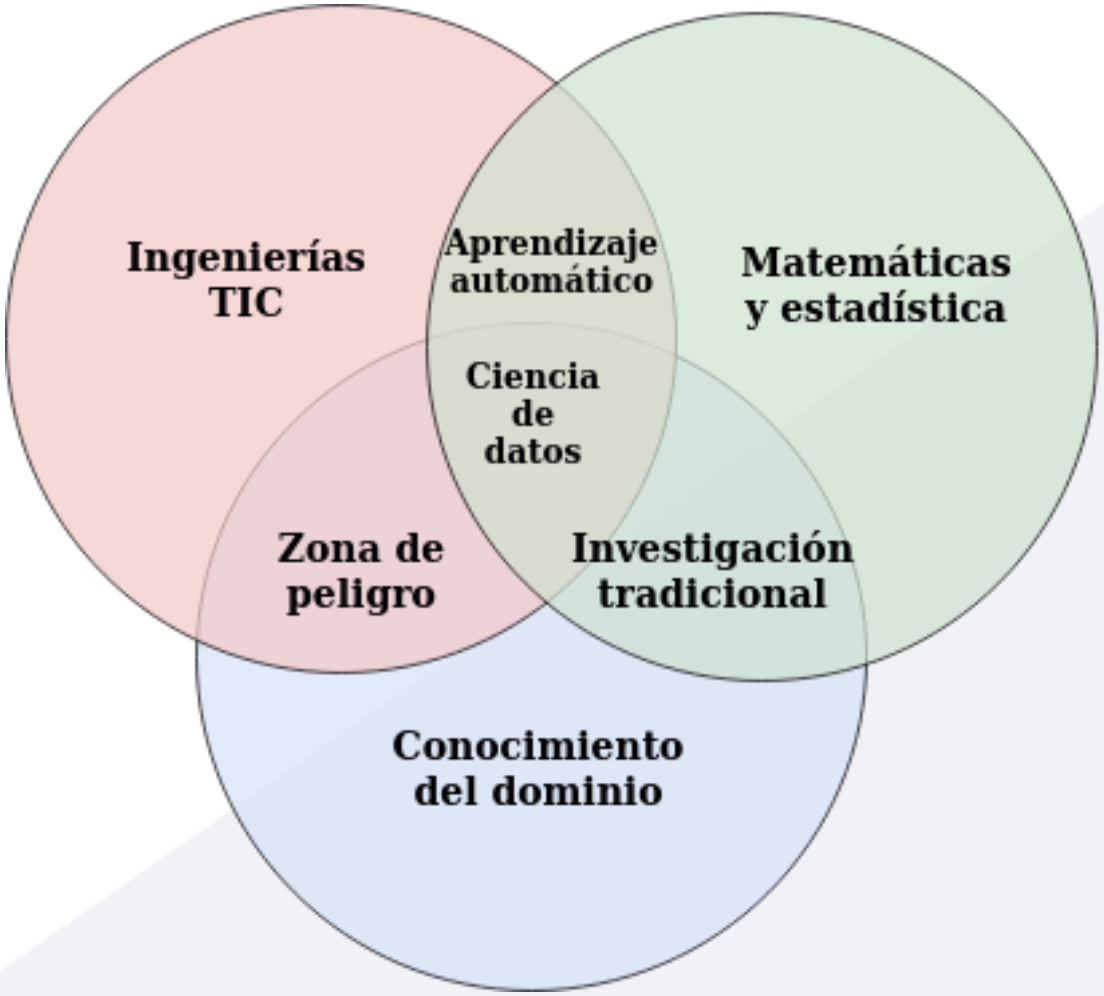
# Índice

1. Fundamentos de la ciencia de datos
2. XAI: Inteligencia Artificial eXplicable
3. Paquete de explicabilidad de la cátedra IA3
4. Conclusiones

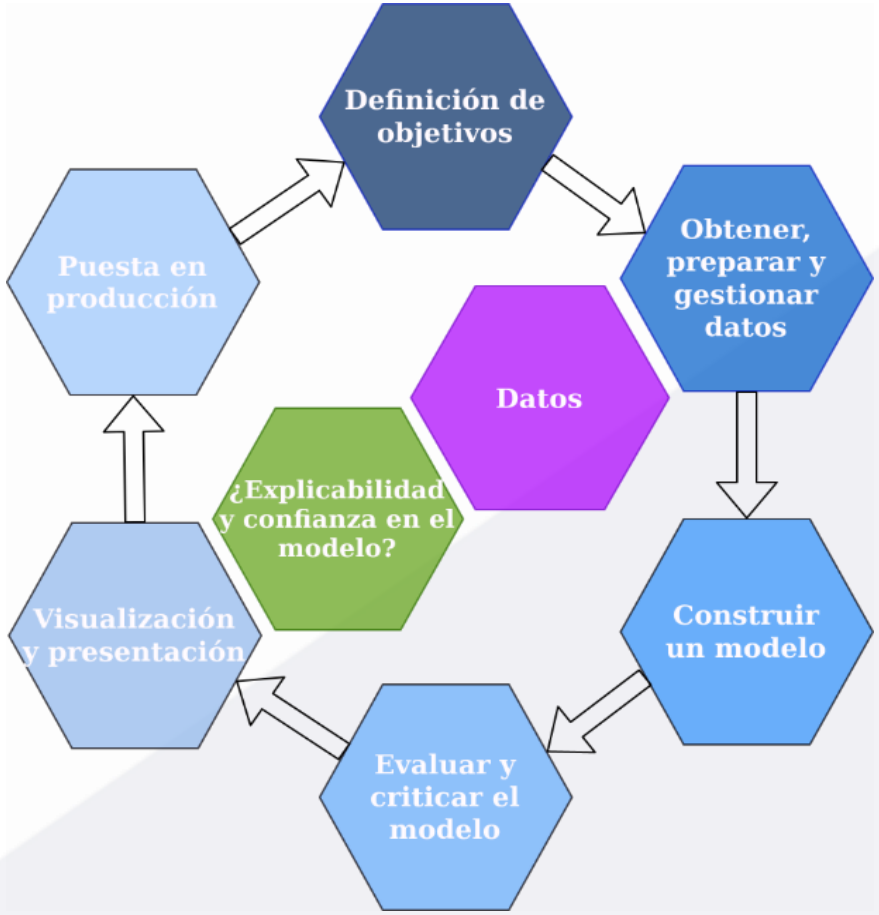
# Fundamentos de la ciencia de datos

IA3 Cátedra de Inteligencia Artificial en Aeronáutica y Aeroespacio
Universidad de Alcalá
indra
Financiado por la Unión Europea NextGenerationEU
MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
Plan de Recuperación, Transformación y Resiliencia

# Fundamentos de la ciencia de datos



# Etapas de un proyecto de ciencia de datos



# Motivación

¿Utilizarías ciegamente un modelo de Aprendizaje Automático para tomar una decisión de la que dependiese tu vida? Considera estas situaciones:

- No tienes acceso al modelo.
- Tienes acceso al modelo, pero no puedes entender sus predicciones.
- Tienes acceso al modelo y puedes "entender" sus predicciones.
- ¿Y si hubieseis diseñado vosotros el modelo?
- ¿Y si vuestro modelo lo utilizasen otros?

GDPR (Reglamento General de Protección de Datos), Articulo 22:

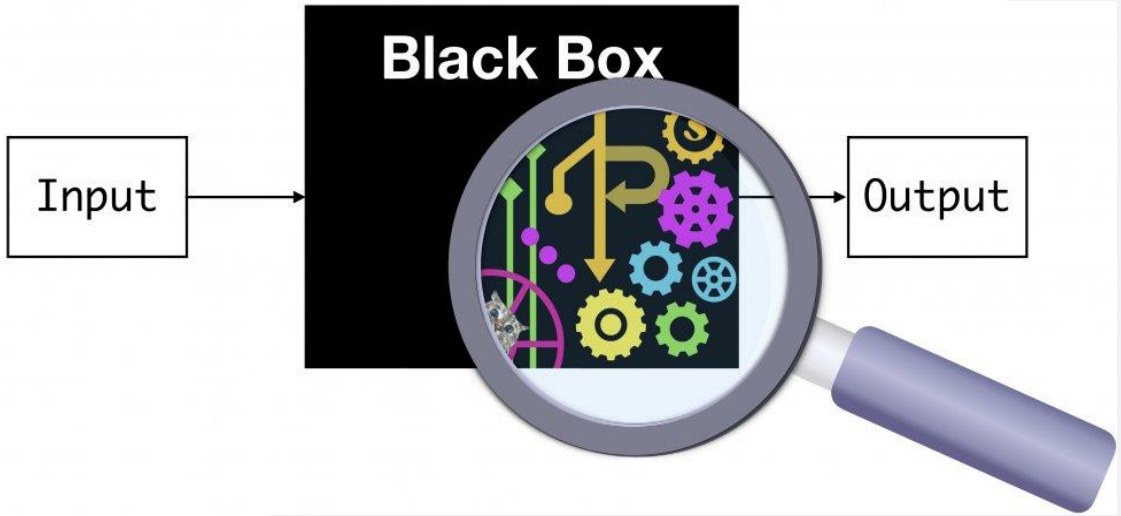
*"Los individuos tienen el derecho a demandar una explicación de como un sistema automático ha tomado una decisión que les afecta."*

# XAI: Inteligencia Artificial eXplicable

IA3 Cátedra de Inteligencia Artificial en Aeronáutica y Aeroespacio
Universidad de Alcalá
indra
Financiado por la Unión Europea NextGenerationEU
MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
Plan de Recuperación, Transformación y Resiliencia

# Percepción de modelos de Inteligencia Artificial

Desde fuera, se considera a los modelos de IA como *"cajas negras"* a la hora de tomar sus decisiones...




... Ante esa problemática, la **IA eXplicable (XAI)** aparece como solución frente a este problema.

# XAI al rescate: Definición y objetivos

La **IA eXplicable o eXplanaible AI (XAI)** surge en el año 2016 tras el programa XAI lanzado por la Agencia de Proyectos de Investigación Avanzado de Defensa (DARPA - Defense Advanced Research Projects Agency) del Departamento de Defensa de los Estados Unidos. Esta puede definirse como:

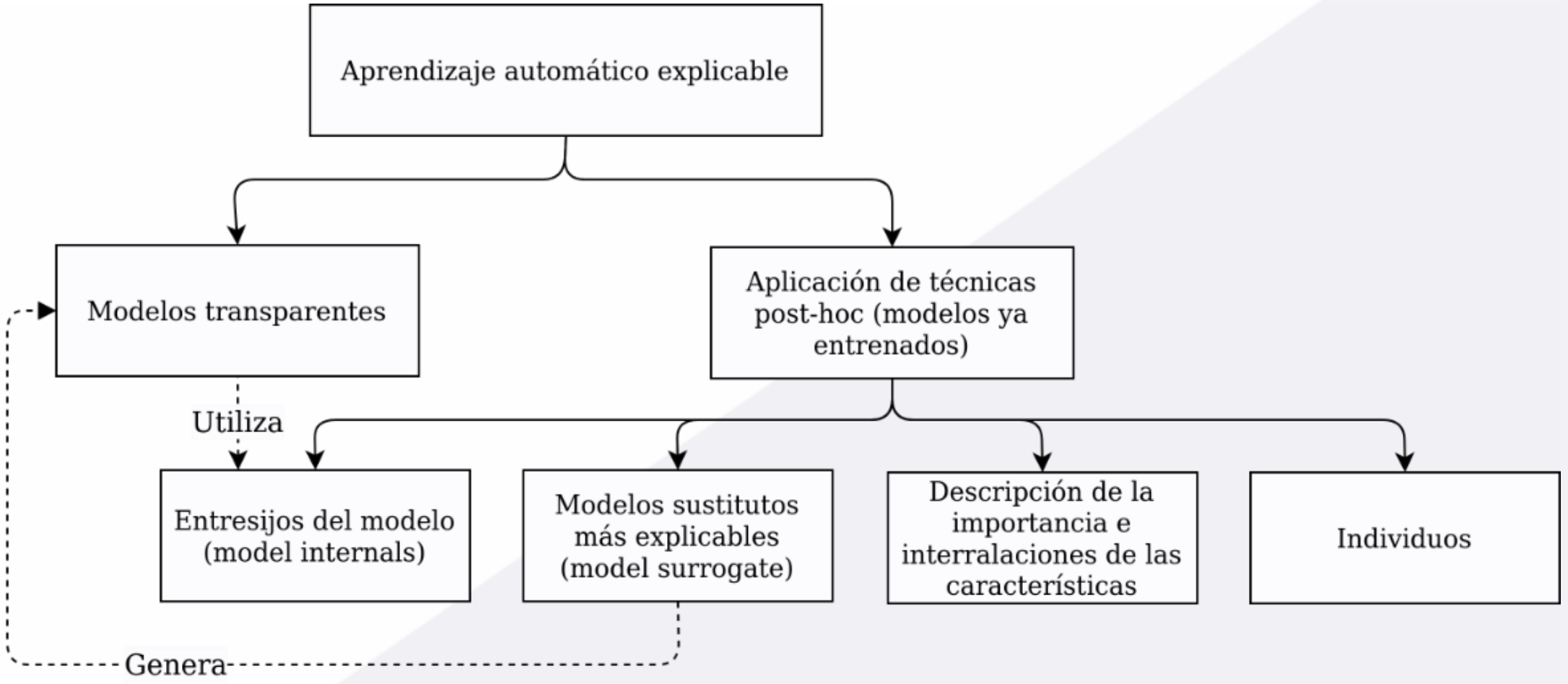
***Gunning en 2017***

*"Un conjunto de técnicas de aprendizaje automático que permitirá a los usuarios humanos comprender, confiar adecuadamente y gestionar de manera eficaz la generación emergente de socios con inteligencia artificial"*

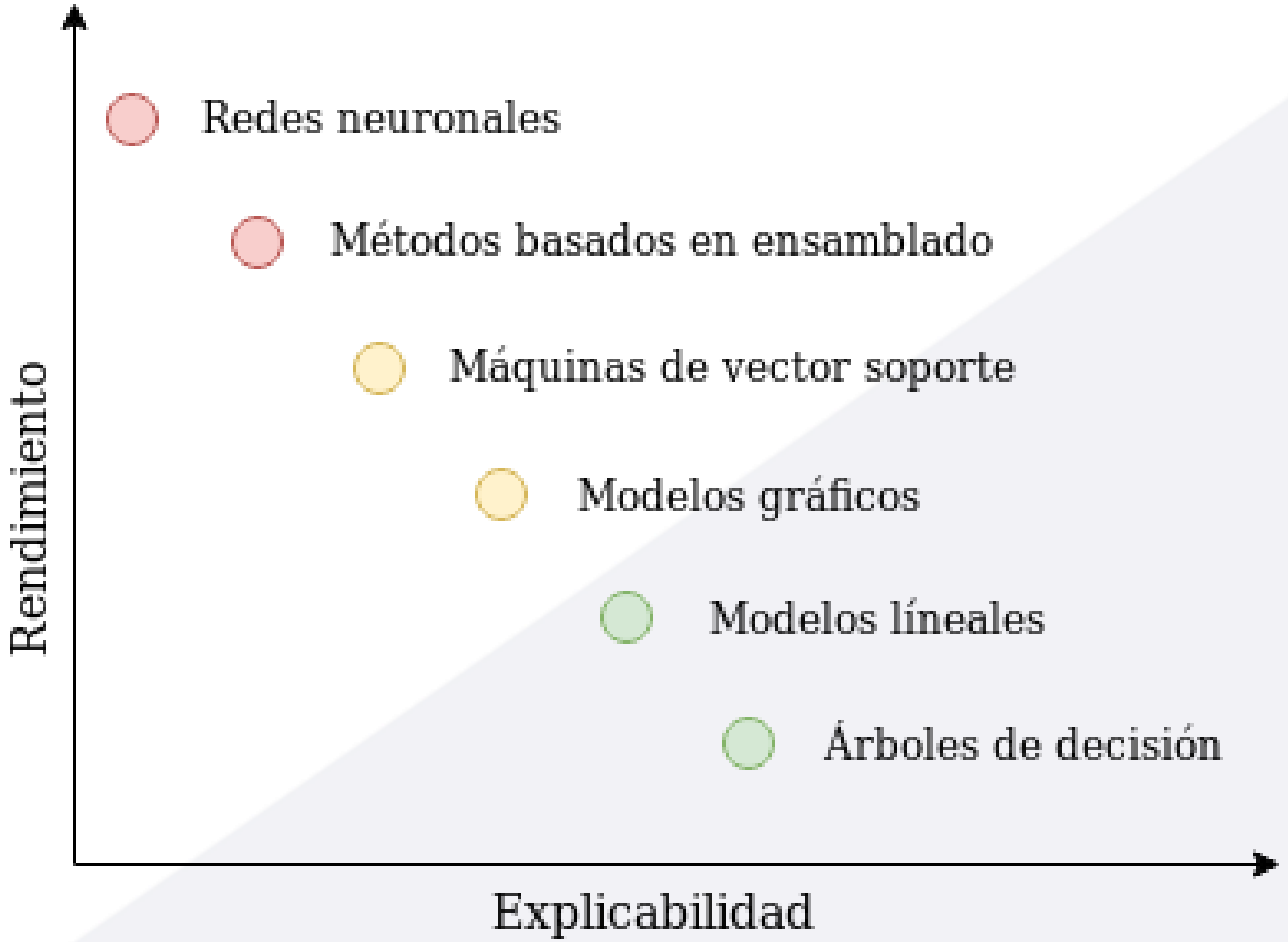
La **IA eXplicable** tiene como **objetivos principales**:

- Crear un conjunto de técnicas que permitan a los seres humanos entender los modelos de Inteligencia Artificial y confiar en ellos.
- Agrupar los diferentes términos que pueden verse relacionados en cuanto al entendimiento del funcionamiento de los modelos y sistemas de Inteligencia Artificial.
- Hacer que las explicaciones generadas para poder entender los modelos de IA sean realizadas en función del tipo de perfil de usuario que las reciba.

# Taxonomía de la XAI



# Rendimiento vs Explicabilidad de modelos




IA3 Cátedra de Inteligencia Artificial en Aeronáutica y Aeroespacio
Universidad de Alcalá
indra
Financiado por la Unión Europea NextGenerationEU
MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
Plan de Recuperación, Transformación y Resiliencia

# Paquete de explicabilidad de la cátedra IA3

IA3
Cátedra de Inteligencia Artificial en Aeronáutica y Aeroespacio

Universidad de Alcalá

indra

Financiado por la Unión Europea
NextGenerationEU

MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA

Plan de Recuperación, Transformación y Resiliencia

# Contexto (I)

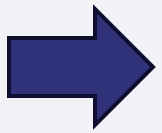
- Las **observaciones contrafácticas** de un individuo (denominadas contrafácticos) **son observaciones hipotéticas similares a dicho individuo para las que un determinado modelo proporciona valores de predicción diferentes.**

- Los contrafácticos **permiten obtener las diferencias entre el caso real y el contrafáctico** revelan los factores que mueven la decisión del modelo y cuánto habría que cambiar para lograr un resultado deseado.

- **La interpretación** de los contrafácticos, por tanto, **depende del funcionamiento del modelo en el contexto en que se esté utilizando**, y no de su proceso de construcción.

***Observación original***

A una persona que cobra **2000€** y no tiene deudas **NO** se le ha concedido un crédito



***Observación contrafáctica***

Si la persona cobrase **2500€** y no tiene deudas **SI** se le ha concedido un crédito

# Contexto (II): Ejemplos reales

## ***Explicabilidad de modelos***

Se quiere desarrollar un satélite y se utiliza un sistema basado en aprendizaje automático para determinar el coste de su diseño, fabricación y lanzamiento. El sistema indica que el coste de diseño y fabricación es de 900 millones de euros, sin embargo, el presupuesto disponible es de 850 millones de euros. El sistema propone las siguientes alternativas para reducir el coste total:

- Reducir el tamaño del satélite.
- Que el satélite no emita señales (no es posible).
- Subcontratar parte de la fabricación.

## ***Reducción de sesgos***

Determinar si el modelo está utilizando correctamente la información de entrada para reducir los sesgos:

- El sistema de Amazon se enseñó a sí mismo que los candidatos masculinos eran preferibles. Penalizaba los currículums que incluían la palabra "mujeres", como en "capitana de club de ajedrez femenino". Y bajó la calificación de los graduados de dos universidades para mujeres, según personas familiarizadas con el tema. No especificaron los nombres de las escuelas. (*Fuente:* https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-airecruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G)

# Trabajo realizado

**El grupo de trabajo** englobado en el PT4 **se centra en la investigación, desarrollo y divulgación de técnicas** de IA eXplicable en los sectores de la aeronáutica y el aeroespacio. Dentro de los logros más relevantes conseguidos se encuentran:

- Participación y desarrollo de seminarios relacionados con la divulgación de la inteligencia artificial y el uso de la explicabilidad en los sectores de la aeronáutica y el aeroespacio.

- Publicación de estado del arte científico sobre el rol de la Explicabilidad de la IA en el entorno de la aeronáutica y el aeroespacio. Publicación: *"The Role of XAI in Transforming Aeronautics and Aerospace Systems"* (https://arxiv.org/abs/2412.17440)

- Investigación y desarrollo de nuevas técnicas de explicabilidad para modelos de regresión, basadas en el usos de ejemplos contrafácticos. El trabajo ha sido aceptado y está pendiente de publicación en el congreso IDEAL 2025 (https://ideal2025.ujaen.es/)

# FUCO: Contrafácticos difusos

**FUCO** es un marco de trabajo que permite la **identificación de observaciones contrafácticas en los modelos de regresión** siguiendo el siguiente proceso:

1. **Definición de umbrales** teniendo en cuenta las restricciones de los actores que interfieren en el modelo.
2. **Identificación de los conjuntos de respuesta** mediante la división del espacio de respuesta.
3. **Evaluación de las observaciones** de cada uno de los conjuntos de respuesta para obtener las más representativas de cada uno de estos.
4. **Generación de explicaciones** adaptadas a las necesidades de los actores del modelo para un mayor entendimiento.

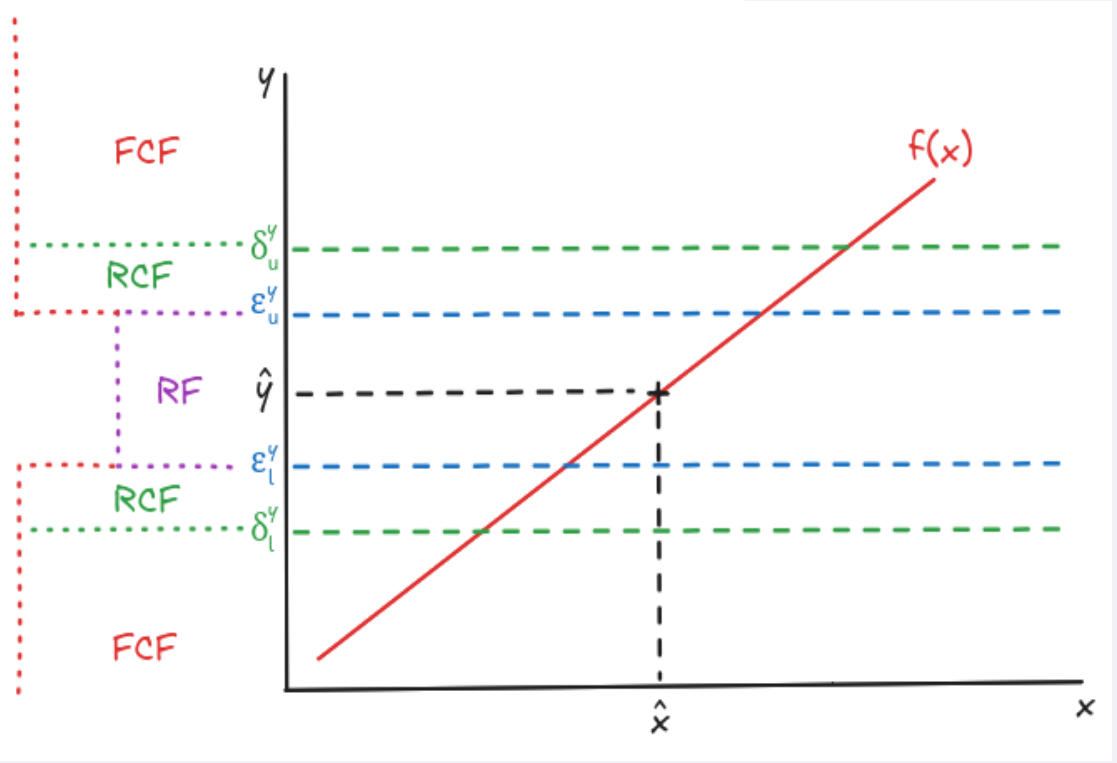



**Figura 1.** *Representación visual de FUCO de la división del espacio de respuesta para la identificación de contrafácticos*

# Conclusiones

# Conclusiones

1. Las **técnicas XAI facilitan la interacción entre humanos y modelos**, permitiendo contrastar las predicciones con el conocimiento del dominio.

2. El **objetivo final de las técnicas XAI es**, no solamente interpretar los modelos, sino también **generar la confianza en los usuarios**.

3. **Para cada familia de modelos y dominio se debe elegir la técnica de explicabilidad más adecuada** que permita obtener el máximo tanto en explicabilidad como en rendimiento.