

Inteligencia Artificial basada en aprendizaje automático explicable

Javier M. Moguerza

Sesión Inter-Academias
Inteligencia artificial: El valor de los datos (2^a sesión)
Madrid, 19 de febrero de 2020



1 Fundamentos de la ciencia de datos

2 Aprendizaje automático explicable

- Rendimiento y explicabilidad de los modelos
- Taxonomía de técnicas de aprendizaje automático explicable
- Técnicas de aprendizaje automático explicables basadas en individuos

3 Contrafácticos

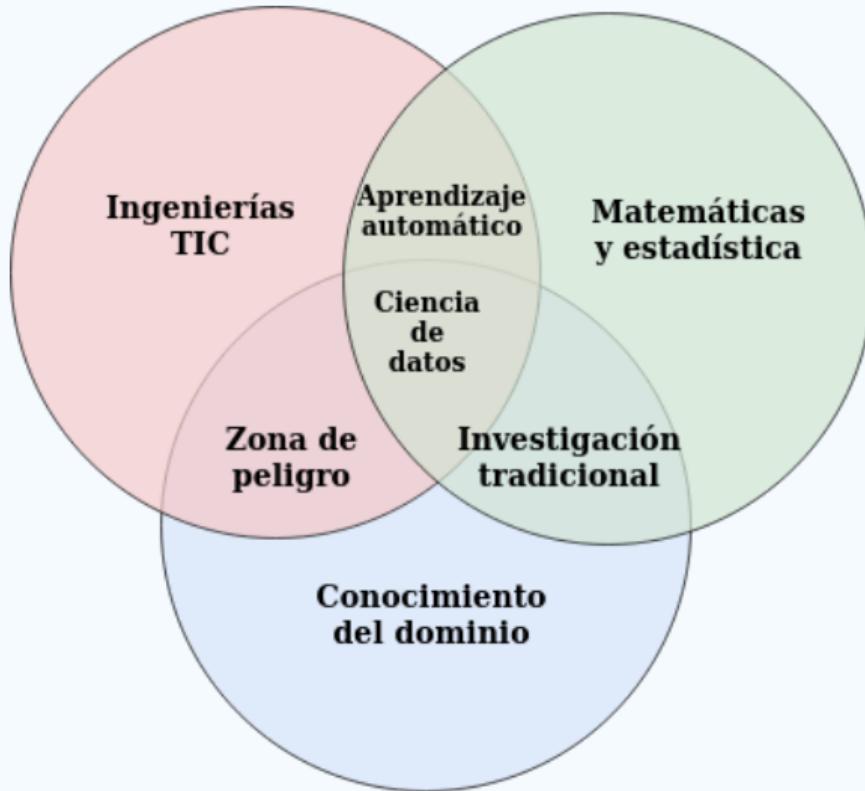
- Definición de contrafáctico
- Ejemplos de contrafácticos

4 Conjuntos de contrafácticos

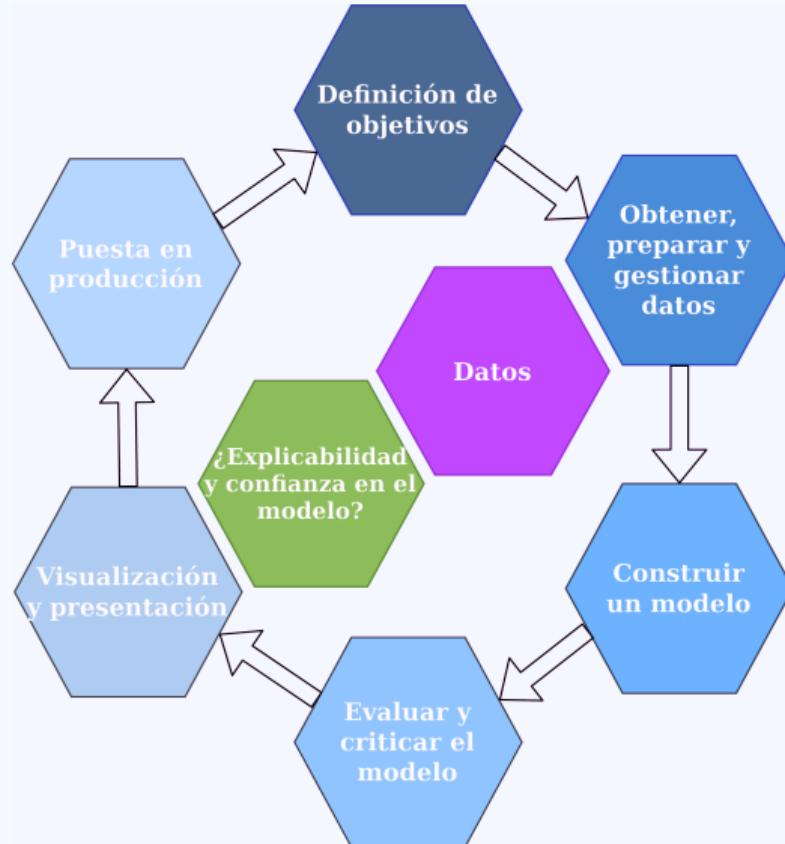
- Definición de conjuntos de contrafácticos
- Ejemplos
- RF-OCSE

5 Conclusiones

Fundamentos de la ciencia de datos



Etapas de un proyecto de ciencia de datos



Aprendizaje automático explicable

Conocemos el “Qué” (predicción) pero no conocemos el “Como” (razonamiento).

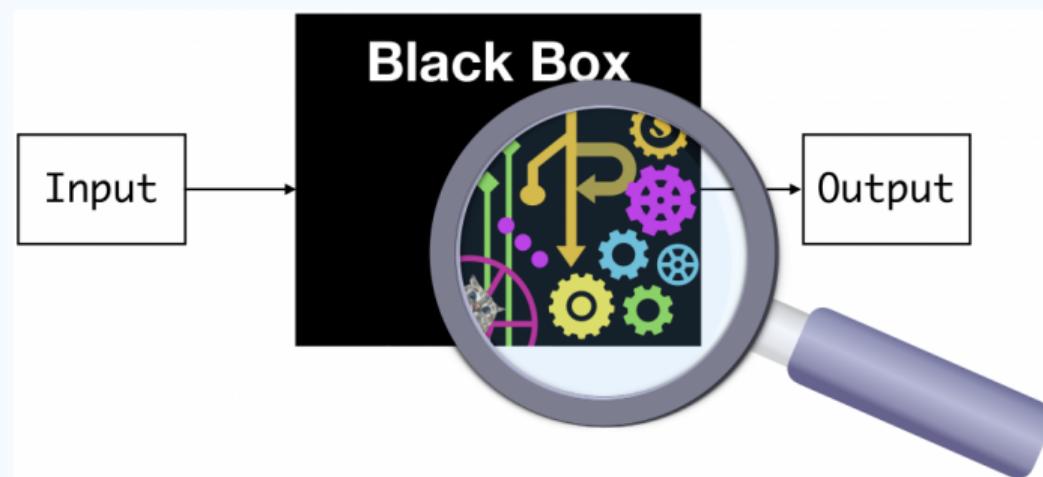


Figure: Representación de un modelo de caja negra.

Fuente: <https://towardsdatascience.com/black-boxes-and-their-intrusion-620aa3c4c56b>

Conocemos el “Qué” y podemos extraer un razonamiento de “Como” se ha hecho.

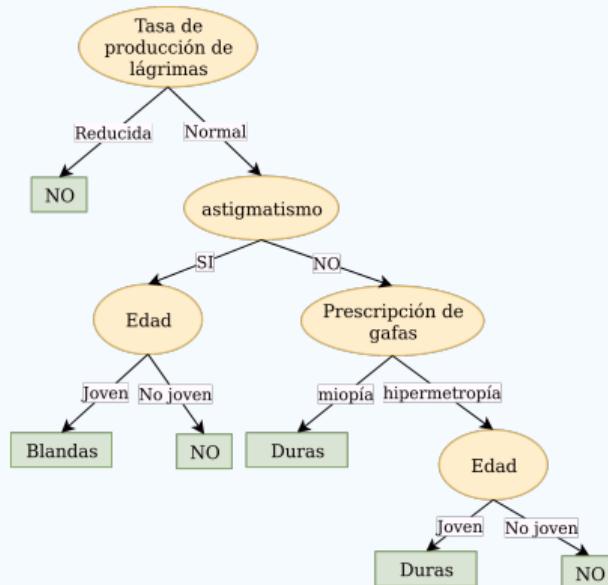
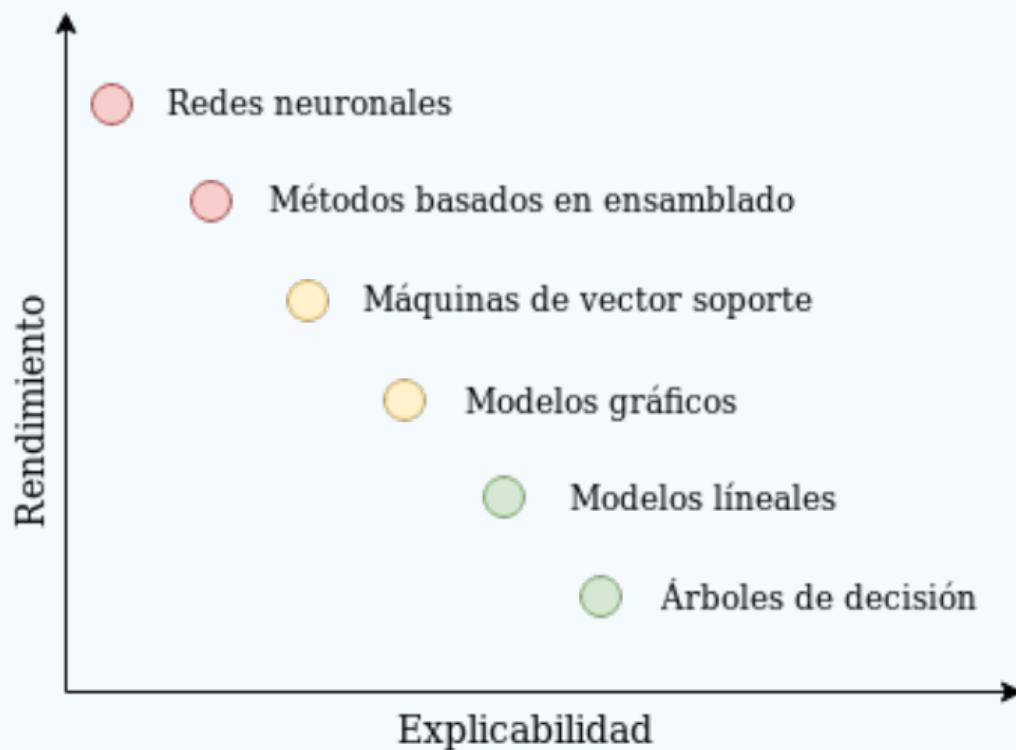


Figure: Árbol de decisión para determinar el tipo de lentes de contacto.



Introducción de aprendizaje automático explicable

El aprendizaje automático explicable es un área del aprendizaje automático que se centra en conseguir que los modelos y sus predicciones sean interpretables para personas no necesariamente expertas en el ámbito.



Figure: Fuente: <https://all-free-download.com/>

El aprendizaje automático explicable tiene muchas aplicaciones, siendo algunas:

- Permitir que podamos confiar en las predicciones de los modelos. Esto es de especial importancia en dominios de alto riesgo, donde una mala decisión puede costar vidas (por ejemplo, medicina o coches autónomos) o grandes cantidades de dinero (por ejemplo, banca).
- En los problemas donde no es suficiente una predicción correcta, sino que también es importante obtener conocimiento del proceso (por ejemplo, procesos físicos en los que se necesita entender el porqué/caracterización del proceso).
- Reducir el sesgo en los modelos (por ejemplo, discriminación de minorías en sistemas de valoración crediticia).
- Desarrollar sistemas más robustos a las perturbaciones o situaciones anómalas (por ejemplo, entender las decisiones de un coche autónomo en situaciones de riesgo).

Taxonomía de técnicas de aprendizaje automático explicable

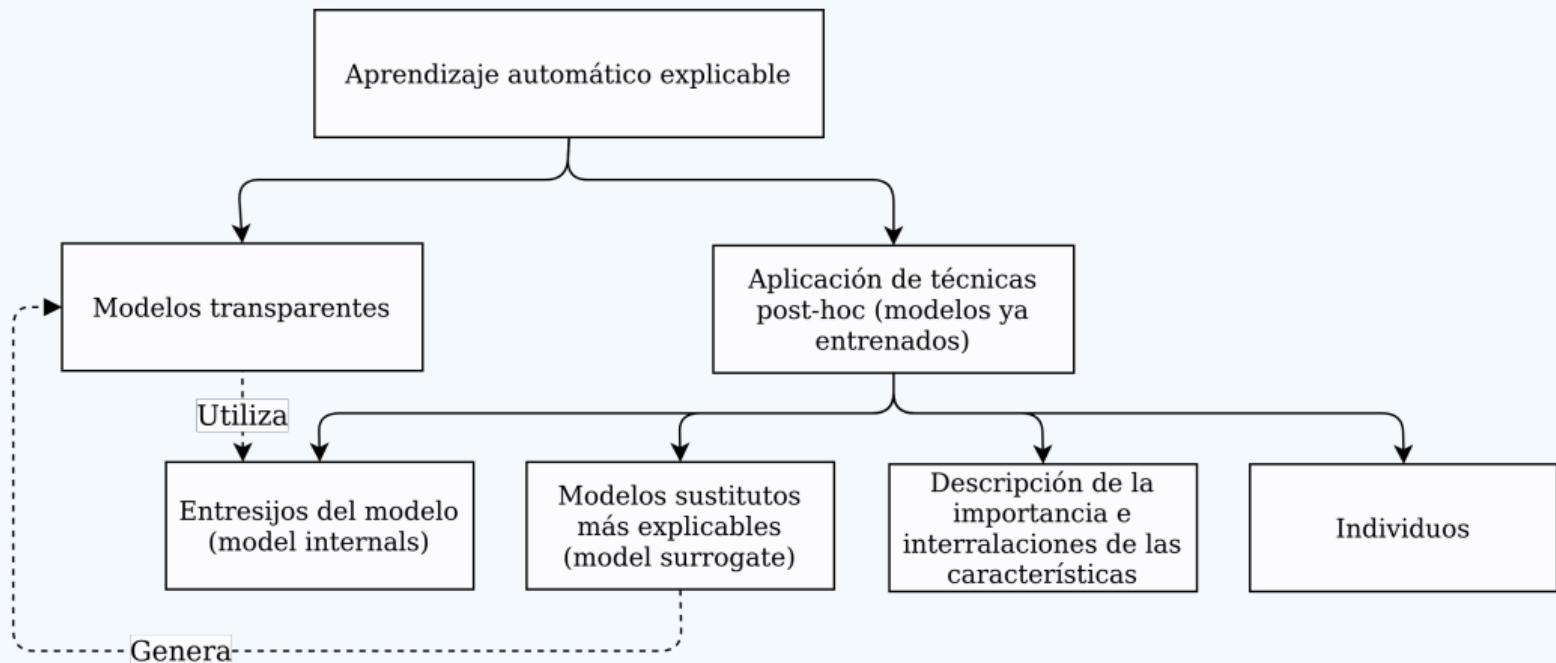


Figure: Inspirado en <https://christophm.github.io/interpretable-ml-book/>.

Los entresijos del modelo (*model internals*) son los coeficientes (por ejemplo, coeficientes en una regresión lineal o pesos en una red neuronal) y estructura interna del modelo (por ejemplo, aristas en una red bayesiana).

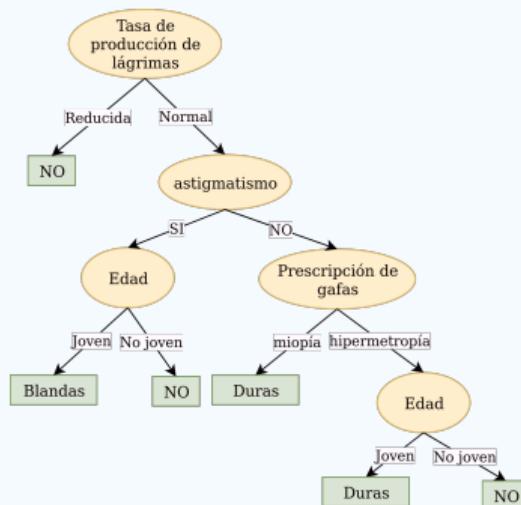


Figure: Árbol de decisión para determinar el tipo de lentes de contacto.

Los modelos sustitutos son aproximaciones de modelos más complejos, locales o globales, utilizando un modelo más explicable (por ejemplo, modelos transparentes).

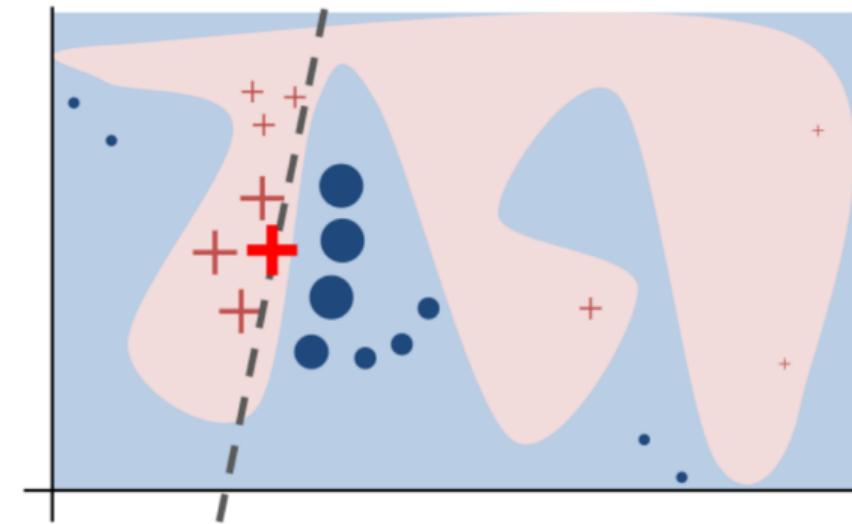


Figure: Ejemplo de modelo sustituto. Fuente: <https://arxiv.org/abs/1602.04938>

Descripción de la importancia e interrelaciones de las características

Las técnicas de descripción de la importancia e interrelaciones de las características, extraen información sobre la utilización de las características por el modelo. Por ejemplo, la importancia de características basada en permutación mide la importancia como el empeoramiento del modelo al permutar los valores de una característica.

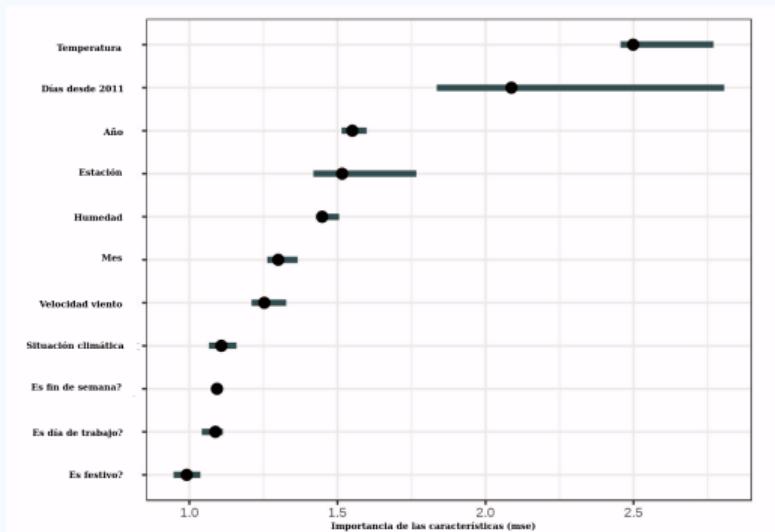


Figure: Fuente: <https://christophm.github.io/interpretable-ml-book/feature-importance.html>

Las técnicas de aprendizaje automático explicables basadas en individuos utilizan observaciones (individuos), reales o sintéticas, para explicar los modelos (explicabilidad global) y/o sus predicciones (explicabilidad local). Algunos ejemplos de estas técnicas son:

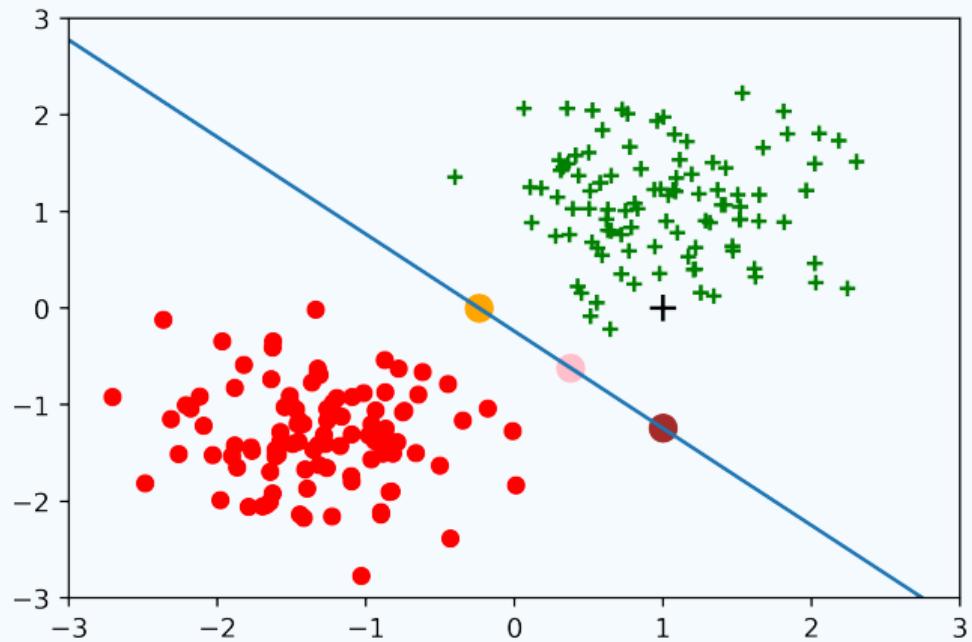
- **Observaciones influyentes:** son observaciones cuya ausencia en la fase de construcción del modelo cambia significativamente su funcionamiento. Su interpretación depende del modelo.
- **Observaciones contrafácticas (contrafácticos):** son observaciones hipotéticas similares a un individuo dado cuya predicción es distinta.

Contrafácticos

- 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 a los que proporciona para el individuo.
- Las diferencias entre la observación hipotética y el individuo permiten explicar el funcionamiento del modelo.
- 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.
- Intuitivamente, los contrafácticos simulan el razonamiento humano.
- Por ejemplo, en un árbol de decisión, los contrafácticos son las observaciones que caen en hojas con clase diferente a la clase predicha para un individuo dado.

Ejemplo simple de contrafácticos

- + clase +
- clase o
- + A explicar (1.00, 0.00)
- contrafáctico (y) (1.00,-1.24)
- contrafáctico (x) (-0.24,0.00)
- contrafáctico (x, y) (0.38,-0.62)



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 currículum 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.¹

¹<https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scaps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>

Conseguir sistemas robustos contra perturbaciones o situaciones anómalas.

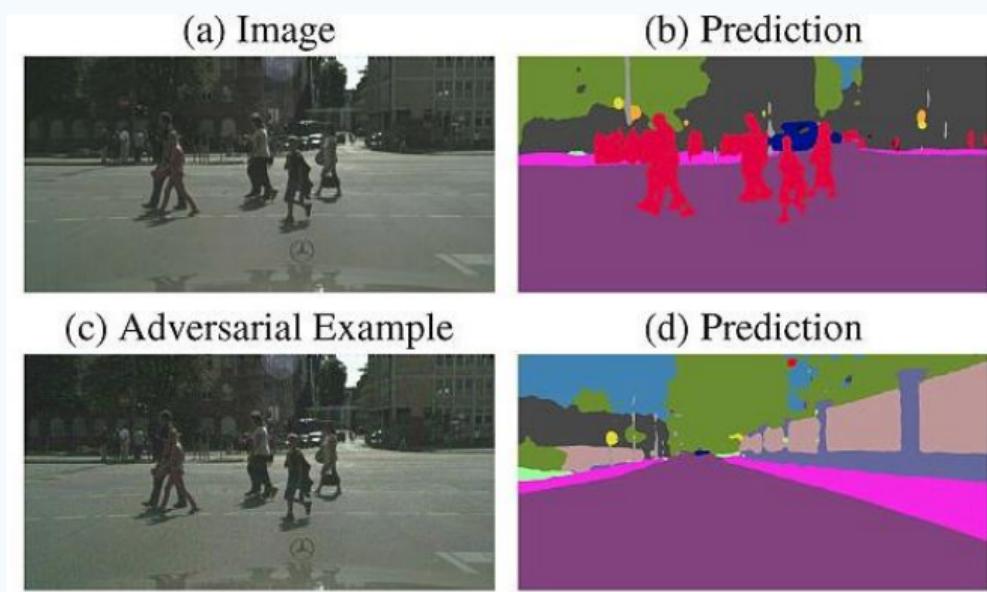
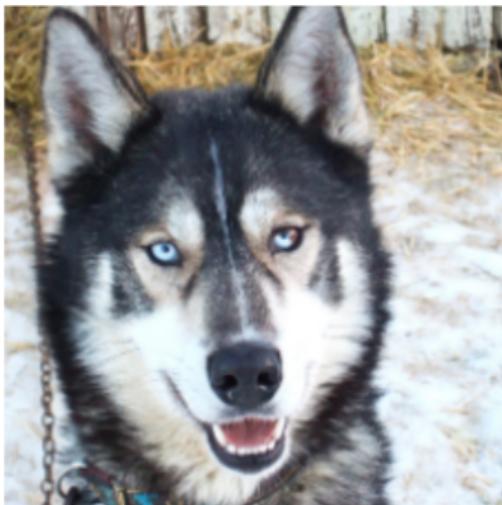
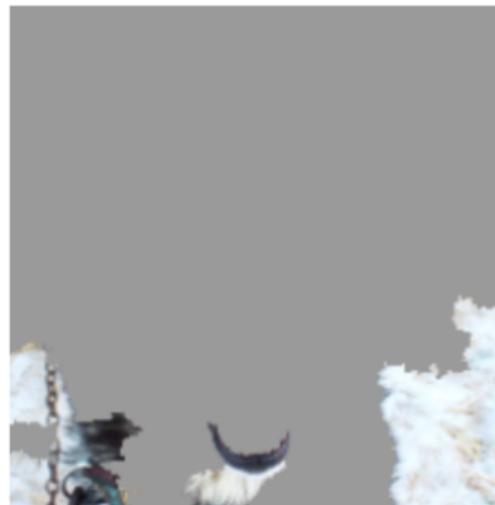


Figure: Ejemplo de ataque adversario. Fuente: <https://arxiv.org/pdf/1704.05712.pdf>

Conseguir sistemas robustos contra perturbaciones o situaciones anómalas.



(a) Husky classified as wolf



(b) Explanation

Figure: Ejemplo de explicabilidad. Fuente: <https://arxiv.org/abs/1602.04938>

Un sistema de puntuación crediticia de un banco ha denegado una propuesta de hipoteca de 150000 euros a 20 años a un usuario con unos ingresos mensuales de 1000 euros. El sistema le propone los siguientes cambios (counterfactuals):

- Aumentar la duración de la hipoteca a 30 años.
- Aumentar los ingresos mensuales a 1210 euros.
- Reducir el importe de la hipoteca a 100000 euros.
- Aumentar la duración de la hipoteca a 25 años y aumentar los ingresos mensuales a 1083 euros.

Un usuario quiere alquilar su apartamento y utiliza un sistema basado en aprendizaje automático para determinar el precio mensual. El sistema le indica que su apartamento se puede alquilar por 850 euros, sin embargo, el usuario quiere alquilarlo por un mínimo de 900 euros. El sistema le propone los siguientes cambios para incrementar el precio del alquiler:

- Comprar una secadora.
- Cambiar la localización de la vivienda (no es posible).
- Renovar el salón.
- Pintar el inmueble.

Conjuntos de contrafácticos

Un conjunto de contrafácticos es una subregión del espacio de características donde el contrafáctico se cumple, y que no contiene no-contrafácticos (es decir, no contiene individuos de la clase factual). Algunas de las características de los conjuntos de contrafácticos son las siguientes:

- La subregión del espacio se puede representar utilizando conjuntos de reglas que define el conjunto de contrafácticos.
- Cuando los cambios se producen sobre características con una alta variabilidad, los conjuntos de contrafácticos son mas fáciles de aplicar que los contrafácticos (por ejemplo, peso o velocidad).
- Por ejemplo, en un árbol de decisión, un conjunto de contrafácticos se puede definir como la subregión del espacio que induce la regla que clasifica un contrafáctico, es decir, la clasificación en hojas del árbol con clase diferente a la predicha para un individuo dado.

Un centro de datos utiliza un sistema basado en aprendizaje automático para determinar si la temperatura de la sala de servidores es adecuada. El sistema indica que una temperatura de 10 grados centígrados es inadecuada, y propone el siguiente cambio:

- Conjunto de contrafácticos: $12.7 \leq \text{Temperatura} \leq 18.33$

Un sistema de puntuación crediticia de un banco ha denegado una propuesta de hipoteca a una familia con 8 tarjetas de crédito porque se excede su capacidad de endeudamiento mensuales. El sistema propone el siguiente cambio:

- Conjunto de contrafácticos: $3 \leq \text{Número de tarjetas} \leq 5$.

Conjunto de contrafácticos en Random Forest

Random Forest Optimal Counterfactual Set Extraction (RF-OCSE) es un método para la extracción de conjuntos contrafácticos con garantía de optimalidad. Se basa en una conversión parcial del Random Forest a un árbol de decisión utilizando una modificación del algoritmo *CART*, y proporciona el conjunto contrafáctico que contiene el contrafáctico más cercano.

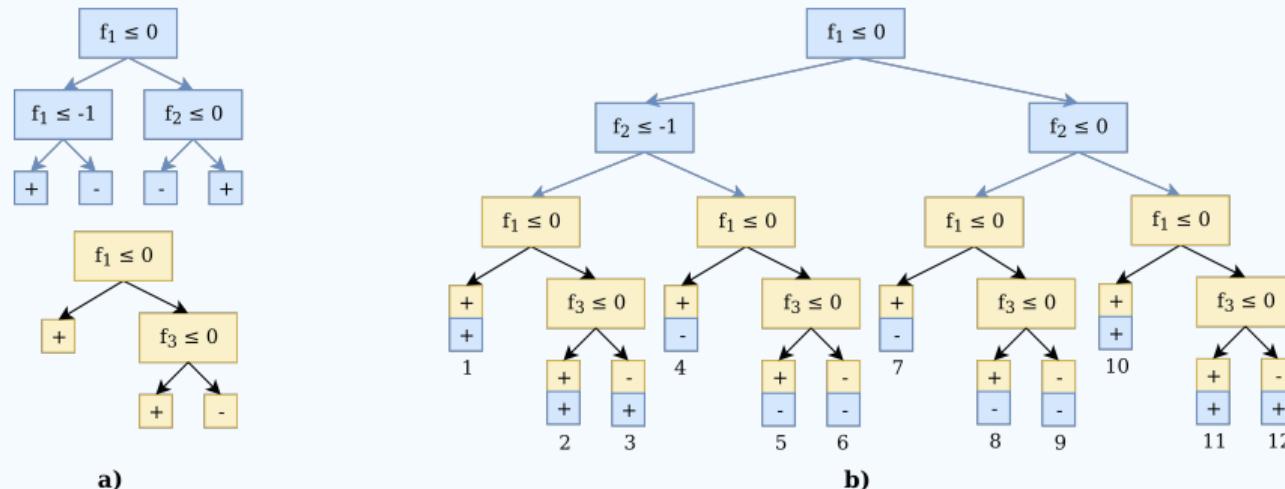


Figure: Intuición del funcionamiento de RF-OCSE

Comparativa de RF-OCSE con *Local Rule-based Explanations (LORE)*:

Método	Metrica	Conjunto de datos		
		adult	compas	credit
LORE	distancia mínima promedio	0.0090	0.0199	0.0092
	% no contrafácticos	0.37	0.40	0.44
	% conjuntos poblados (% cf.)	83.0 (63%)	93.0 (60%)	96.0 (56%)
RF-OCSE	distancia mínima promedio	0.0049	0.0120	0.0004
	% no contrafácticos	0.0	0.0	0.0
	% conjuntos poblados (% cf.)	25.0 (100%)	100.0 (100%)	7.0 (100%)

Conclusiones

- Las técnicas de aprendizaje automático explicable facilitan la interacción entre humanos y modelos, permitiendo contrastar las predicciones con el conocimiento del dominio.
- El objetivo final de las técnicas de aprendizaje automático explicable es, no solamente interpretar los modelos, sino también generar confianza en los usuarios.
- Los contrafácticos son una técnica de explicabilidad que se asemeja al razonamiento humano y susceptible de ser entendida por personas sin conocimientos de aprendizaje automático.
- Para cada familia de modelos y dominio se debe elegir la técnica de explicabilidad más adecuada.
- Como trabajo futuro, se está trabajando en la creación un catálogo de problemas/dominios, modelos y técnicas, de cara a asociar las técnicas de explicabilidad a los problemas y modelos que se adaptan a ellas.
- Además, se tratará de extender este trabajo preliminar a otras técnicas de aprendizaje automático.