CIENCIA VIOLETA

I Encuentro Científico sobre
Investigación con Perspectiva de Género

Inteligencia artificial explicable para la detección y análisis de sesgos de género en modelos de aprendizaje automático

Aurora Ramírez Quesada

INTRODUCCIÓN Y OBJETIVOS

CIENCIA VIOLETA
I Encuentro Científico sobre
Investigación con Perspectiva de Género

INTRODUCCIÓN Y OBJETIVOS

Motivación:

- Aprendizaje automático basado en datos (IA)
- Decisiones en base a predicciones son "opacas"
- No están exentas de errores o sesgos
- Gran impacto en el día a día (medicina, justicia, finanzas, ocio, etc.)

Objetivos:

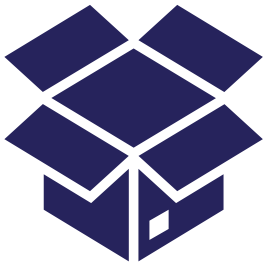
[O1] Recolección de conjuntos de datos con información relevante para el análisis de sesgos
[O2] Análisis de los conjuntos de datos y de las predicciones que los algoritmos de aprendizaje automático realizan sobre ellos
[O3] Aplicación de métodos XAI para estudiar el comportamiento de los modelos predictivos obtenidos

Inteligencia artificial explicable (XAI): ¡abramos la "caja negra"!

¿Qué factores influyen más en las predicciones?

¿Por qué se predijo la opción X y no la opción Y?

¿Qué habría que cambiar para predecir Y en lugar de X?



Proyecto GENIA: Estudio de sesgos de género en modelos de aprendizaje automático mediante inteligencia artificial explicable. *Modalidad UCOImpulsa del plan propio de investigación de la Universidad de Córdoba (Curso 22/23)*
Participantes: Aurora Ramírez (IP), José Raúl Romero, Sebastián Ventura, Amelia Zafra
Grupo KDIS – Dpto. Informática y Análisis Numérico (EPSC) – Instituto DaSCI

KDISLab DaSCI Instituto Andaluz Interuniversitario en Data Science and Computational Intelligence

METODOLOGÍA Y DATOS

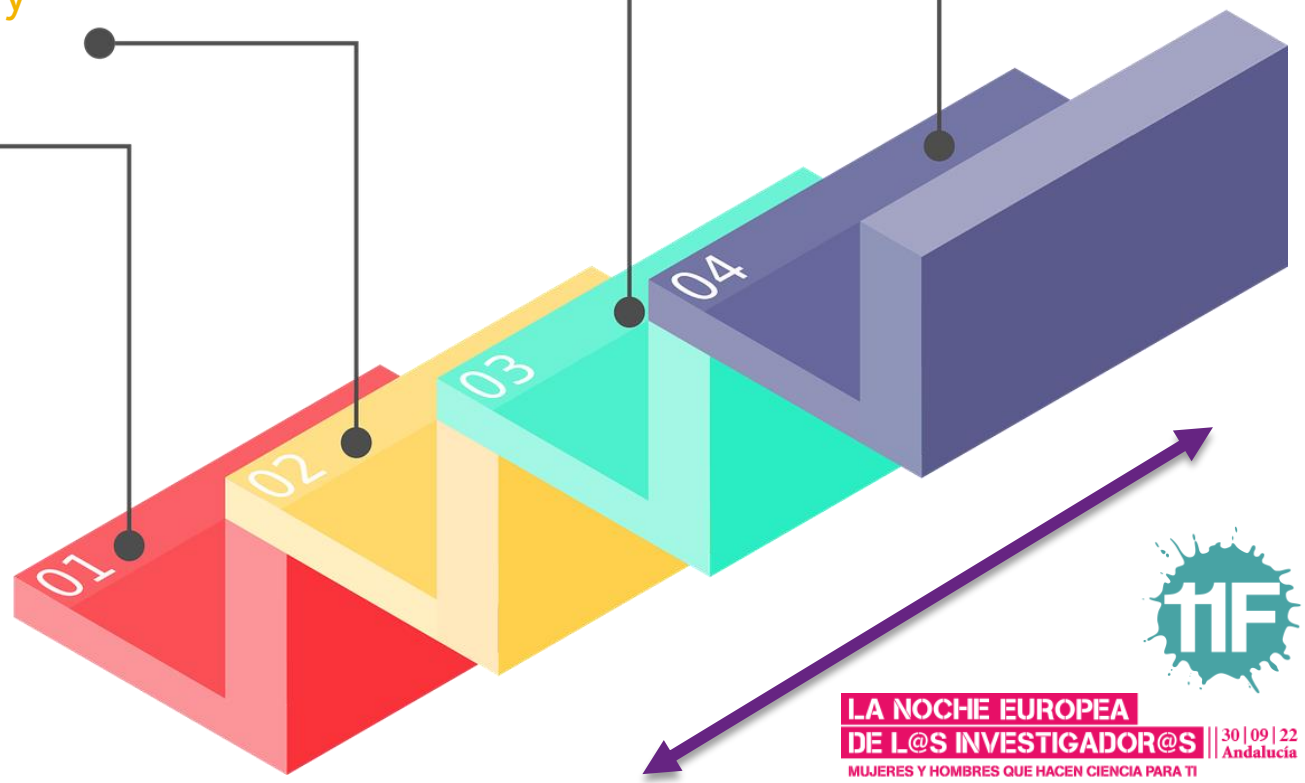
METODOLOGÍA Y DATOS

Generación de explicaciones contrastivas con métodos XAI

Estudio de errores en la predicción con método XAI

Obtención de predicciones globales y disgregadas por género

Recopilación de conjuntos de datos y análisis de sesgos

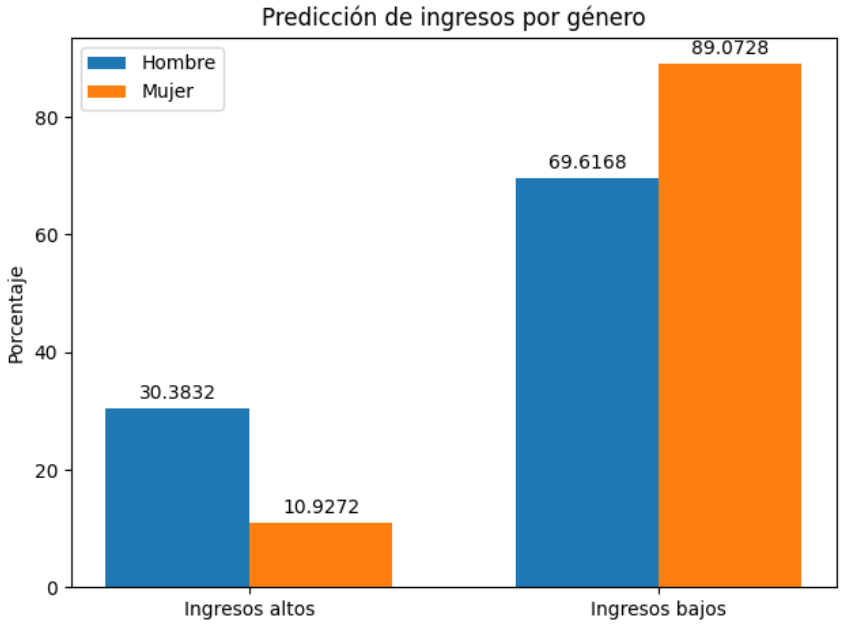


- ✓ Datos de dominio público
- ✓ Librerías de código abierto: sklearn, dalex
- ✓ Métodos XAI propios y de la literatura
- ✓ Repositorio GitHub al finalizar el proyecto

Imagen de Trang Le en Pixabay

RESULTADOS

RESULTADOS



¿Y si la concesión de una hipoteca dependiese de la predicción de ingresos de la persona solicitante?

Precisión global (probabilidad acierto)	79.10%	
Precisión por género	74.12%	90.53%
Factor más influyente	Ocupación laboral	Estado civil
¿Cómo puedo aumentar mis opciones de obtener la hipoteca?	Cambiar ocupación laboral; aumentar capital; aumento ligero de edad y de horas de trabajo/semana	Cambiar de estado civil y ocupación laboral; mayor incremento de edad (x3) y horas de trabajo (x2)

CONCLUSIONES Y FUTURO

CONCLUSIONES Y FUTURO

Conclusiones:

- Los primeros resultados evidencian que los modelos de aprendizaje automático **se comportan de manera diferente** en función de si se considera información de género o no.
- La generación de explicaciones permite detectar **situaciones discriminatorias** y descubrir que existen **diferentes exigencias** según el género.

Trabajo futuro:

- Ampliar el estudio a los distintos **conjuntos de datos recopilados**.
- Extraer conclusiones tanto **generales** como **particulares**.
- Incorporar el conocimiento obtenido a la generación de métodos de aprendizaje automático que **reduzcan o eliminen** la influencia del género.

Contacto:

https://www.uco.es/users/aramirez/ aramirez@uco.es @aurora_rq

¡GRACIAS!