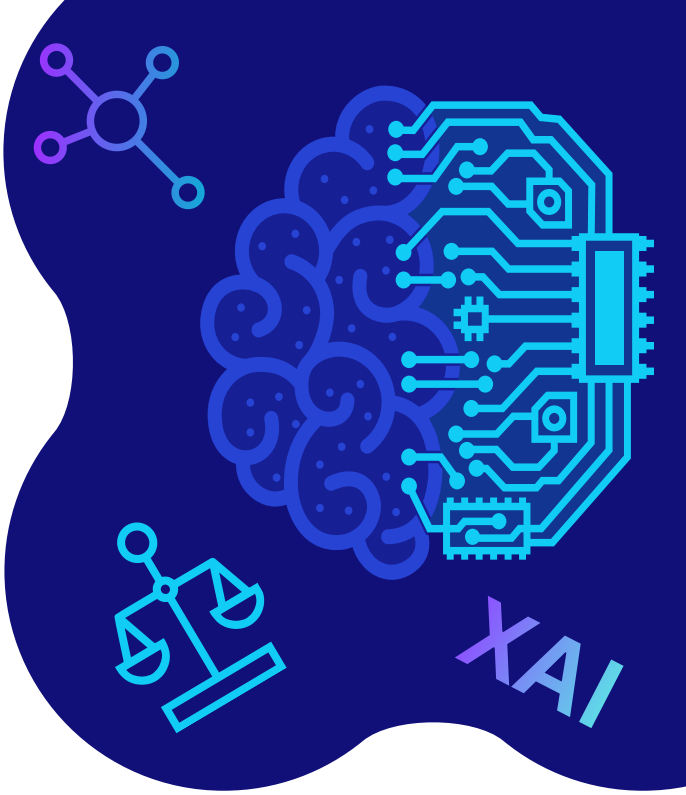
uc3m | Universidad Carlos III de Madrid

Universidad Carlos III de Madrid
DEPARTAMENTO DE ESTADISTICA

WORKSHOP

Towards Equitable and Explainable AI for Responsible Decision-Making



February, 25, 2026
16:00pm - 20:00pm

School of Law and Social Sciences
Costas Goutis room (10.0.23)

Topics covered

- Interpretable machine learning models
- Fairness-aware learning
- Statistical and optimization-based methods
- Applications in high-impact decision-making

Registration

https://shorturl.at/2UiVx

More details at:
https://sbenitez1.github.io/workshop-AI/

Keynote speaker

Dolores Romero Morales
SEIO Medallist 2025

Program

16:00 - Dolores Romero Morales (Copenhagen Business School)
"Some (mORe) steps to enhance the explainability of algorithmic decisions"

16:45 - Elisa Cabana Garcerán del Vall (CUNEF Universidad)
"Interpretable Sign Language Recognition"

17:10 - Pablo Morala Miguélez (Universidad Carlos III de Madrid)
"Benchmarking Feature Interactions in XAI with Controlled Polynomial Ground Truth"

17:35 - Jorge Moral Fernández (Bankinter)
"Mathematical Analysis of Fairness and Bias in Recommender Systems"

18:00 - Coffee Break

18:30 - Antonio Navas Orozco (Universidad de Sevilla)
"Building robust counterfactual explanations for GLMs"

18:55 - Cristina Molero del Río (Universidad de Sevilla)
"Estimating shape-constrained functional regression models using P-splines"

19:20 - Arturo Pérez Peralta (Telefónica & Universidad Carlos III de Madrid)
"New methods in fair AI"

Details of the talks

16:00 - Dolores Romero Morales

"Some (mORe) steps to enhance the explainability of algorithmic decisions"

Abstract: The use of Artificial Intelligence (AI) and Machine Learning (ML) algorithms is ubiquitous in our daily life. The responsible use of AI / ML methodologies requires not only pursuing a high accuracy when representing the data at hand, but also providing explanations on how these algorithms arrive at their decisions. The literature is quite rich on the form these explanations may have, but it has also raised some shortcomings. In this talk, we will navigate through some novel models that enhance the explainability of Supervised and Unsupervised Learning methodologies, addressing some of these limitations.

16:45 - Elisa Cabana Garcerán del Vall

"Interpretable Sign Language Recognition"

Abstract: "Deep learning models have achieved remarkable performance in computer vision, yet their decision-making processes remain largely opaque, limiting interpretability and user trust. Class Activation Mapping (CAM) techniques, including Gradient-weighted CAM, are widely used to visualize model attention and provide insights into model reasoning. In sign language recognition, such visual explanations are particularly valuable, as they can serve as diagnostic tools to identify sources of misclassification and to guide improvements in both model performance and dataset quality. We present the Interactive CAM Visualization Application, a web-based platform that enables intuitive exploration and comparison of multiple CAM variants. Applied to American Sign Language recognition with convolutional neural networks, our framework generates richer and interpretable activation maps without requiring model retraining or architectural modifications."

17:10 - Pablo Morala Miguélez

"Benchmarking Feature Interactions in XAI with Controlled Polynomial Ground Truth"

Abstract: Many popular explainability methods summarize models through single-feature importances, yet real predictions are often driven by feature interactions. Therefore, there exists a need for XAI methods capable of explaining feature interactions. To solve this problem, several SHAP-based extensions have been proposed, but their evaluation is difficult on real datasets because ground truth for interpretable methods is rarely available. In this seminar, a simulation benchmark based on polynomial data-generating processes is presented, where interaction effects are explicitly controlled and therefore known. Using rank-based metrics, we compare interaction recovery across multiple SHAP interaction methods and NN2Poly, an explanation approach based on polynomial representations of neural networks.

17:35 - Jorge Moral Fernández

"Mathematical Analysis of Fairness and Bias in Recommender Systems"

Abstract: This talk focuses on a key challenge in recommender systems: achieving not only high accuracy, but also fairness. Instead of looking at fairness through sensitive user attributes, it explores popularity bias, the tendency of systems to over-recommend already popular items while neglecting niche content.
The results highlight a clear trade-off: models that optimize accuracy often reinforce popularity, while fairer approaches may reduce performance.
The main takeaway is that good recommendations should be not only accurate, but also balanced and fair.

18:30 - Antonio Navas Orozco

“Building robust counterfactual explanations for GLMs”

Abstract: Counterfactual explanations (CEs) are well-established tools of explainable Machine Learning (ML), consisting in finding a minimal perturbation of a given input so that the output through the ML model is above a certain threshold. Lately, there has been an increased interest towards the introduction of robustness in CEs. In this talk, we explore two notions of robustness in counterfactual analysis. The first notion is robustness against data distribution shift: we seek to guarantee that the output remains sufficiently high when the nominal data distribution (the training sample) is replaced by a probability distribution with the same support but different frequencies. The second notion is robustness understood as guaranteeing that the computed CE is valid, not only in the empirical model (the model fed with the training sample), but also in the ground-truth model. We particularize the resulting optimization problems for Generalized Linear Models. For the first notion, we propose a tailored cutting-planes and Gaussian Variable Neighbourhood Search-based numerical algorithm, while for the second, the resulting problem is a Mixed Integer Second Order Cone Program, readily solvable by off-the-shelf solvers such as Gurobi. Lastly, we validate the robust CEs by comparing them to ordinary CEs through numerical experiments in regression, classification and counting regression settings.

18:55 - Cristina Molero del Río

"*Estimating shape-constrained functional regression models using P-splines***"**

Abstract: Functional regression refers to regression models involving functional covariates and/or a functional response. Often, prior knowledge about the relationship between the covariates and the response – arising from biological, medical, or engineering processes - requires that the estimated functional coefficients meet specific shape constraints, such as non negativity, monotonicity or convexity/concavity. However, these requirements are not straightforwardly satisfied when using an unconstrained estimation approach. In this work, we investigate whether the combination of conic optimization and penalized splines (P splines) developed by Navarro-García, Guerrero and Durbán (2023) can outperform recent approaches in the literature for estimating shape-constrained functional regression coefficients. The comparison will be carried out using both simulated and real datasets.

19:20 - Arturo Pérez Peralta

"New methods in fair AI"

Abstract: The staggering performance of Machine Learning models across a wide range of tasks and application domains has lead to their widespread adoption in many areas of society. However, the evidence suggests they can amplify harmful biases against certain demographic groups, thus putting into question their deployment in critical decision-making contexts and creating the need for algorithmic fairness in this field. This presentation presents a compendium of our last few works tackling this question, starting with a comparison of the state of the art for fairness in Natural Language Processing, followed by the study of novel hybrid methods aiming at mitigating intersectional bias, and ending with the introduction of new processors inspired by topology and geometry.