Seminari del Departament d'Estadística i Investigació Operativa

Universitat Politècnica de Catalunya

1 de diciembre de 2025

Título: From Black to White Boxes: Interpretable Regression with the TRUST Algorithm

Autor: Albert Dorador (Whitebox Lab)

Web: https://sites.google.com/view/albertdorador/homeresearch

Abstract: In regression problems, leading machine learning models are often "black boxes", offering strong predictive performance but little interpretability. In this workshop, we will review the regression problem and the main algorithmic approaches, classifying them into black- and white-box methods. We will then present the TRUST algorithm, a new interpretable regression framework recently accepted at PRICAI 2025, and demonstrate its application to forecasting medical insurance charges. In this example, a simple two-leaf TRUST tree outperforms fully tuned Random Forest and XGBoost models, showing that transparency and state-of-the-art accuracy can go hand in hand. The algorithm is implemented in the Python package *trust-free*, freely distributed in binary form for easy adoption in modern data science workflows.

Sobre el Autor: Albert Dorador holds a PhD in Statistics from the University of Wisconsin–Madison and previously worked at the European Central Bank on financial risk management and machine learning applications. His research focuses on trustworthy and interpretable machine learning. He is the creator of the TRUST algorithm and the *trust-free* Python package, and is currently advancing this work through his AI start-up Whitebox Lab. For more information, please visit https://adc-trust-ai.github.io.