Addressing Issues Related To Trustworthy, Responsible AI, Hallucination-Mitigated Models

Principal Investigators:

Amar Gupta and Peter Szolovits

Students and Collaborators:

Ariba Khan, Unyime Usua, Ishanvi Kommula, Rashmi Nagpal, Mihir Borkar, Pamina Lässing (EY), Rasoul Shahsavarifar (EY), Andrew Lagworthy (BT), Ali Payani (CISCO)

Addressing Issues Related To Trustworthy, Responsible AI: A Multi-Objective Framework for Balancing Fairness and Accuracy in Debiasing Machine Learning Models

With Ariba Khan, Mihir Borkar and Rashmi Nagpal

Problem Statement:
The primary problem addressed is the challenge of balancing accuracy and fairness in machine learning algorithms, especially in critical domains like banking and healthcare. It is difficult to satisfy multiple fairness metrics in machine learning, where efforts to achieve group fairness lead to conflicts with individual fairness metric, as similar individuals with different protected attributes (gender, age group) receive biased outcomes.

Impact: By expanding the scope of fairness, we facilitated the creation of robust machine learning framework which enabled responsible and equitable data-driven decisions.

Nagpal R, Khan A, Borkar M, Gupta A. A Multi-Objective Framework for Balancing Fairness and Accuracy in Debiasing Machine Learning Models. Machine Learning and Knowledge Extraction. 6(3):2130-2148. https://doi.org/10.3390/make6030105

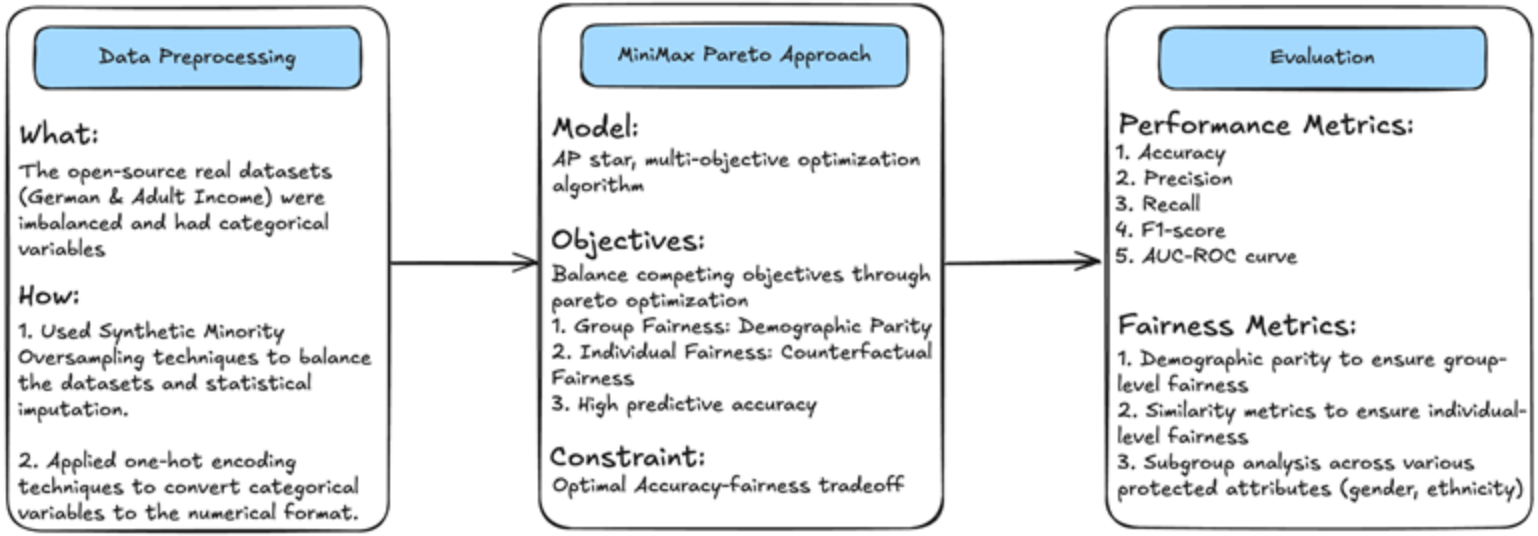
Addressing Issues Related To Trustworthy, Responsible AI: Optimizing Fairness and Accuracy: a Pareto Optimal Approach for Decision-Making

Rasoul Shahsavarifar, Rashmi Nagpal, Peter Szolovits, Amar Gupta

Problem Statement:

The primary problem addressed is the challenge of balancing accuracy and fairness in machine learning algorithms, especially in critical domains like banking, healthcare. It is difficult to satisfy multiple fairness metrics in machine learning, where efforts to achieve group fairness lead to conflicts with individual fairness metric, as similar individuals with different protected attributes (gender, age group) receive biased outcomes.

Solution and Impact: By incorporating individual and group fairness metrics, we developed robust MiniMax Pareto Optimal Framework which enabled responsible and equitable data-driven decisions.



Nagpal, R., Shahsavarifar, R., Goyal, V. *et al.*: Optimizing fairness and accuracy: a pareto optimal approach for decision-making. Springer, *AI and Ethics*