leaders

DEMO

Tool for auditing predictors

www.aileaders-project.eu

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.

CONTENTS

Abstract

Type of OER:	Demo using AEQUITAS in a *Google Colab Environment*
Goal/Purpose:	To provide students a tool for auditing machine learning predictors regarding bias and fairness. By focusing on a fraud detection scenario, students learn how to critically assess algorithmic performance not only in terms of accuracy, but also in terms of equitable treatment across demographic groups.
Expected Learning Outcomes:	By the end of the demo, students will be able to: • Apply Aequitas to audit classification models (decision Tree, Radom Forest, FairGBM) for fairness; • Interpret key fairness metrics such as False Positive Rate (FPR), False Discovery Rate (FDR), and Statistical Parity; • Identify and explain group-level disparities in prediction outcomes; • Reflect on the role of bias auditing and fair ML techniques in high-risk decision-making contexts,
Suggested Methodological Approach	Problem-Based Learning.
Keywords:	Machine Learning, Aequitas, Auditing, Biases, Fairness

Introduction

INTRODUCTION

As machine learning becomes increasingly integrated into decision-making processes—particularly in high-stakes domains such as finance—it is essential to critically assess the ethical implications of model deployment. While predictive models are typically evaluated based on performance metrics such as accuracy, these alone are insufficient to guarantee fairness, especially when the underlying data include sensitive attributes such as gender, ethnicity, or socioeconomic status (Jesus et al., 2024).

This demo-based activity invites students to engage with *Aequitas*, an open-source audit toolkit, through a hands-on exploration of a realistic fraud detection scenario. The objective is twofold: to deepen students' understanding of algorithmic bias, and to familiarise them with practical tools and techniques that support the development of more transparent and accountable machine learning systems.

INTRODUCTION

Aequitas provides a comprehensive framework for evaluating fairness in classification models by examining how predictive outcomes are distributed across demographic groups. It offers a range of fairness metrics—including False Positive Rate (FPR), False Discovery Rate (FDR), and Statistical Parity—which help to uncover disparities in model performance that may not be visible through conventional evaluation methods. Even when trained on imbalanced or biased datasets, models can be audited to assess whether they treat different subgroups equitably.

By engaging directly with *Aequitas* in this applied context, students will gain both technical skills in fairness auditing and a broader ethical awareness of the challenges inherent in deploying machine learning in domains where the consequences of bias can be particularly severe.

Problem presentation

Source: https://huggingface.co/stable-bias

PROBLEM PRESENTATION

The scenario involves auditing a **binary classification model** trained to detect **bank account fraud**.

Fraud detection systems typically suffer from:

- **Severe class imbalance** (fraud cases are rare);
- **High stakes** (misclassification can harm individuals or institutions);
- **Hidden biases** (sensitive features may correlate with label outcomes).

PROBLEM PRESENTATION

The simulation uses the **Bank Account Fraud (BAF) dataset**, a large-scale, privacy-preserving synthetic dataset that replicates real-world patterns of bank fraud.

Key features include:

- Variants simulating sampling bias, temporal drift, and feature imbalance;
- Demographic attributes that enable fairness analysis;
- Strong class imbalance for realism.

These conditions allow for meaningful experimentation with how fairness metrics behave under different model assumptions and data configurations.

Simulation execution

SIMULATION EXECUTION

1. **Access the Simulation Notebook**
   - Go to https://tinyurl.com/4b9u9hun
2. **Run all cells**
   - Execute the notebook fully to load the dataset, train a binary classification model, and generate predictions. To do so, please click *play* at the top-left of each cell.
3. **Perform a Fairness Audit with Aequitas**
   - Use Aequitas to generate a fairness report
   - Focus on metrics like FPR, FDR, and Statistical Parity
   - Identify which groups are treated unfairly in model predictions

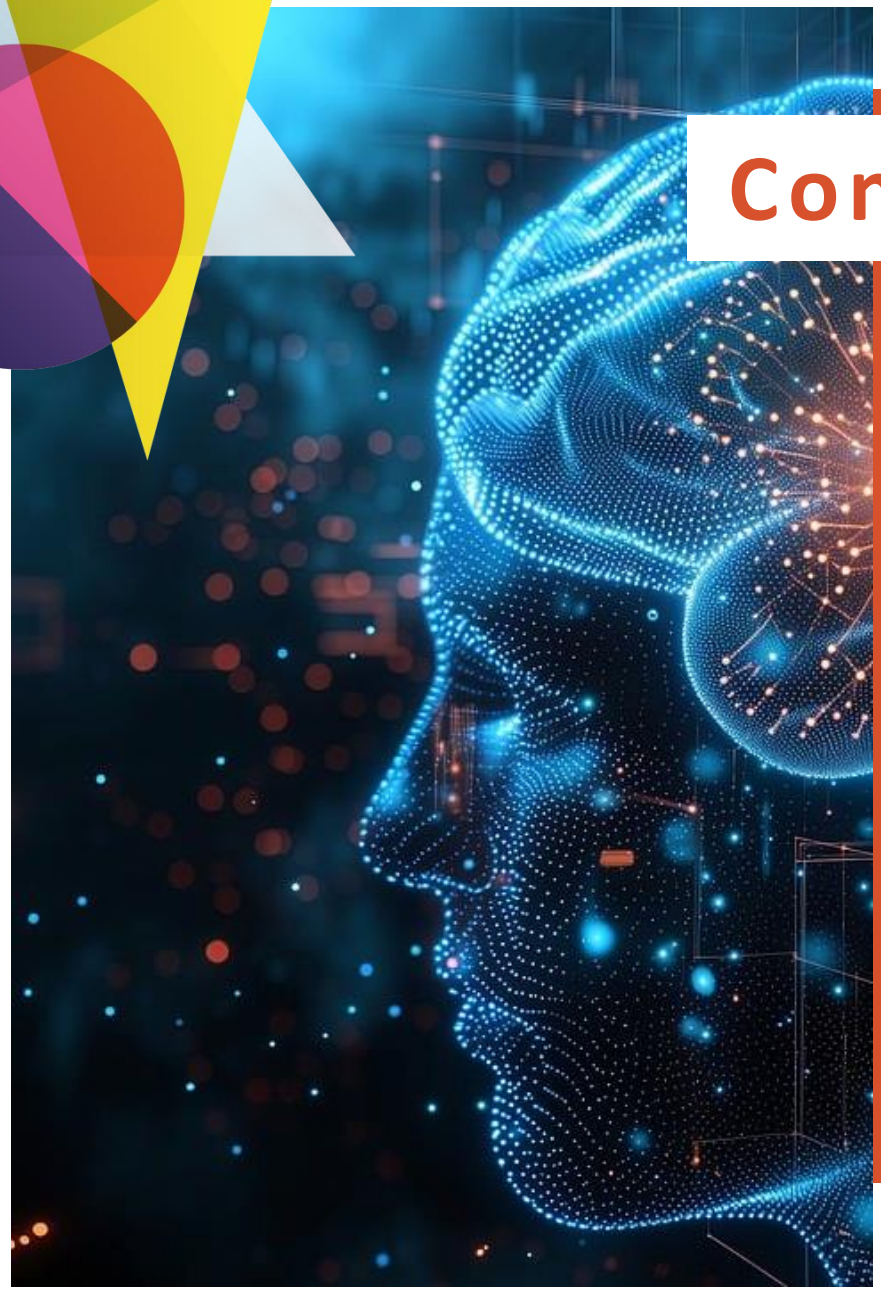
SIMULATION EXECUTION

5. Compare Results and Interpret Metrics

- Review fairness disparities across groups. Compare performance metrics (e.g., accuracy) with fairness metrics to assess trade-offs.

6. Reflect and Discuss

- What patterns of unfairness were observed?
- How might these results affect real individuals?
- How can such tools be incorporated into the ML pipeline to improve ethical outcomes?

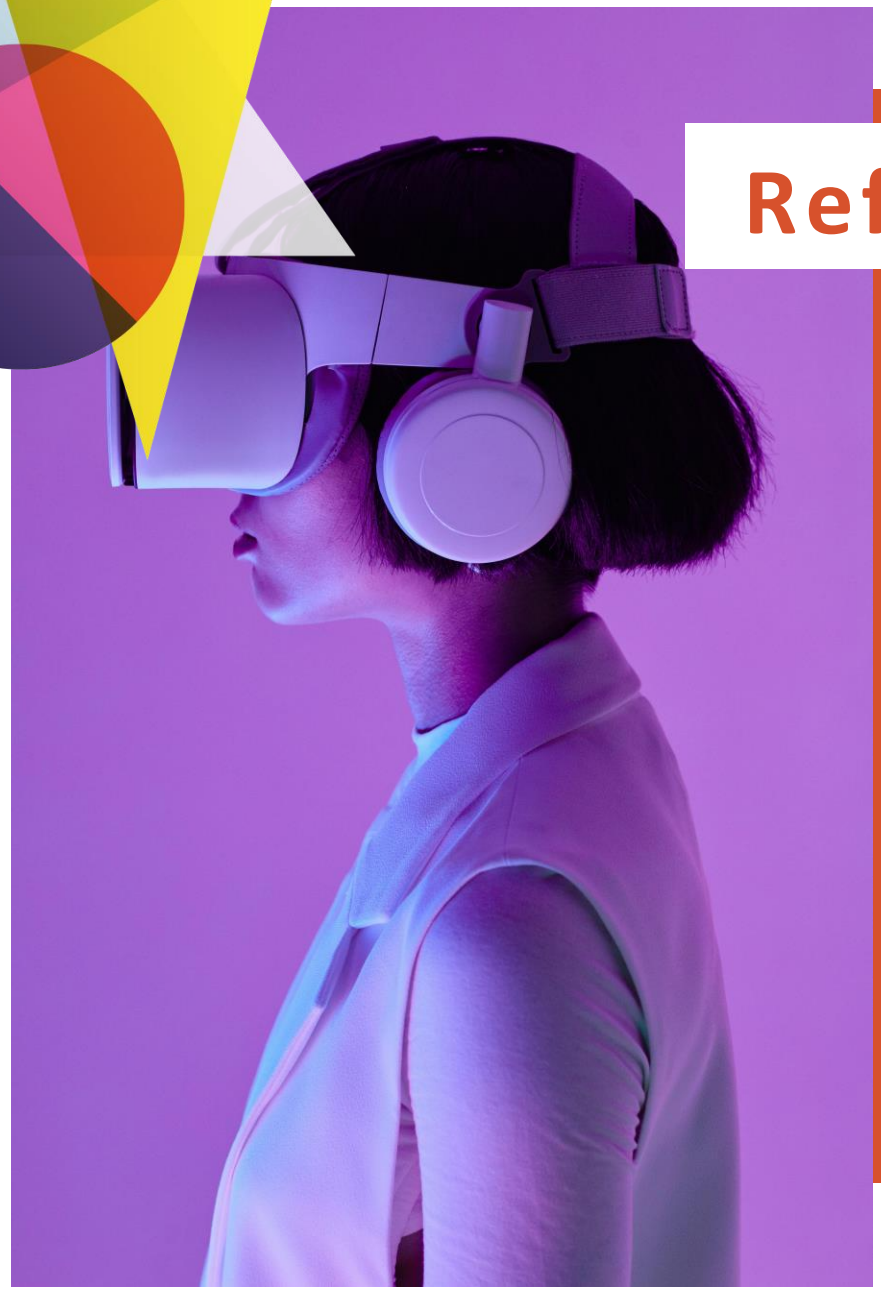


Conclusion

CONCLUSION

This demo shows that even accurate models can lead to **unfair outcomes** when deployed without fairness auditing. Using Aequitas, students uncover how **bias can persist** in binary classification settings and how different demographic groups can experience different rates of misclassification.

By engaging directly with fairness metrics and conducting real audits, students develop technical competencies and ethical awareness. More importantly, they learn that **bias detection is not an add-on**, but an essential part of building **trustworthy AI systems**—particularly in domains like finance where model predictions have serious consequences.



References

REFERENCES

Jesus, S., Saleiro, P., e Silva, I. O., Jorge, B. M., Ribeiro, R. P., Gama, J., ... & Ghani, R. (2024). Aequitas flow: Streamlining fair ml experimentation. *Journal of Machine Learning Research*, *25*(354), 1-7.

Follow our journey

www.aileaders-project.eu

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the author or authors only and do not necessarily reflect those of the European Union or the Foundation for the Development of the Education System. Neither the European Union nor the entity providing the grant can be held responsible for them.