Exploring the Impact of Enhancing Fairness on Model's Explainability

Omran Ayoub, Mirna Saad, Franca Corradini
(name.surname@supsi.ch)

May 2, 2025

1. Objective

The objective of this project is to investigate the relationship between three aspects of Machine Learning (ML) model development: predictive performance, model's fairness and quality of explanations. You are tasked with developing and evaluating predictive models, applying explainable AI (XAI) techniques, and analyzing the impact of bias mitigation strategies on both performance and interpretability. The goal is to gain practical insights into how enhancing fairness can influence model behavior and change the explanations extracted using XAI techniques, if any. This project encourages critical thinking on responsible AI development and transparent decision-making.

At the date of submission, you are required to present their findings in the form of the presentation. Additionally, you are also required to submit a brief report outlining your work, and to provide the source code.

Prior to your submission, you have the possibility to check your progress with the instructor of the teaching assistant. Please reach out via email in case you would like to organize such a session.

2. Project Instructions

2.1 Group Formation

You must work in groups of 3. You can find an Excel sheet to select your group and choose the dataset you want to work on. Link to the sheet.

2.2 Important dates:

- Group selection: 09/05/2025
- Check with Instructore and TAs Period: 20/05-30/05
- Final submission and Presentation: 03/06/2025 (Final report, Final powerpoint, and code)

3. Task Description

The general tasks of each project are as follows:

- **Preprocess and Prepare Data.** Perform necessary preprocessing steps to clean the data (handle missing values, remove duplicates, encode categorical variables...).
- **Analyze Data.** Perform an exploratory data analysis to gain a deep understanding of the data at hand. Make sure to include the necessary data visualization.
- **Analyze Data Focusing on Protected Attribute.** Perform a specific data analysis involving the protected attribute. Try to extract insights from this analysis, if any.
- **Develop Machine Learning Models and Evaluate their Performance.** Build and train ML models following best practices. Ensure using the ML model with the best performance in the following steps. Also make sure to use the adequate predictive performance metrics in your comparison.
- **Explain the ML model(s) using an XAI technique.** Explain the ML model using an XAI technique of your choice, either with feature importance or counterfactual explanations, or both. Extract insights from the explanations.
- **Enhance the Fairness of the ML Model.** Identify the protected attribute and select the bias quantification metrics to use. Then, apply bias mitigation techniques to enhance the fairness of the ML model.
- **Explain the ML Model Developed with Enhanced Fairness.** Explain the ML model using the same techniques used to explain the model in an earlier step. Extract insights from the explanations.
- **Compare Performance of ML Models.** Compare the two ML models (with and without fairness constraints) in terms of predictive performance and fairness. Make sure to use the suitable metrics for your comparison.
- **Analyze Change in Behavior Qualitatively.** Highlight the impact of fairness on model explainability on selective local explanations and, if possible, global explanations.
- **Quantify Change in Behavior Across Explanations.** Conduct a research to identify metrics used to quantify change in explanations (i.e., to quantify change across two vectors). Analyze the change in explanations of a number of local explanations extracted from the models developed with and without fairness in mind.

4. Dataset

We provide datasets used frequently in fairness-aware learning:

Adult Dataset

- **Prediction Task**: Predicts whether a person's income exceeds $50K based on demographic characteristics.
- **Protected Attributes**: gender, race
- **Data Link**: `https://raw.githubusercontent.com/columbia/fairtest/master/data/adult/adult.csv`
- **Features**: All features have been treated as categorical, except for 'capital-gain' and 'capital-loss' which are numeric, and 'education-num' and 'hours-per-week' which are treated as ordinal.

German Credit Dataset

- **Prediction Task**: Predicts whether it is risky to grant credit to a person.
- **Protected Attributes**: gender
- **Data Link**: `https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data`
- **Features**: The features 'duration', 'credit', 'rate', 'residence', 'age', 'cards', and 'liables' are treated as numerical, while the remaining features are regarded as categorical.

COMPAS Dataset

- **Prediction Task**: Predicts if an individual is rearrested within two years after the first arrest. The latter predicts if an individual is rearrested for a violent crime within two years.
- **Protected Attributes**: race, gender
- **Data Link**: `https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis`
- **Features**: We treat the features 'juv_fel_count', 'juv_misd_count', 'juv_other_count' as numerical and the rest as categorical.

SSL

- **Protected Attributes**: race
- **Data Link**: `https://raw.githubusercontent.com/samuel-yeom/fliptest/master/exact-ot/chicago-ssl-clean.csv`
- **Features**: we treat all features as numerical (apart from the protected race feature)

5. Guidelines for Report, Presentation and Code

5.1 Report Requirements

Every group is required to write a report that clearly describes the main ideas and methodologies employed in their work. The report should detail the approach taken, key findings, and the results of the experiments conducted.

5.2 Presentation Requirements

Each group will asked to present their work in a maximum of 20-minutes presentation.

5.3 Code Requirements

The code should be clear, well-documented, and implement all tasks described in Section 3 of this document. Additional features or extensions to the tasks are encouraged and will be considered a valuable contribution to the project. Ensure that the code is reproducible and easy to understand.