Algorithmic Fairness, Accountability and Ethics
Assignment 1

IT UNIVERSITY OF CPH

-

Deadline 3 March 2025 23:59 (CPH time).
This is a group submission (2-3 person group).

What to hand-in: A report as pdf summarizing the main findings, max. 3 pages, including plots. A jupyter notebook detailing the process. Upload the two files as a single zip file on learnIT.

Where to start: This assignment consists of two tasks. You can find a Jupyter template to get started in the assignment on learnIT.

Dataset: US Census data from https://github.com/zykls/folktables. We use data of individuals from the state California in 2018, as detailed in the template. The template also details which attributes we use as feature vector. More details on the dataset can be found in the accompanying paper at https://arxiv.org/pdf/2108.04884.pdf.

Task 1 – Classifiers and fairness considerations

1. Starting from the template, train two different classifiers on the training data: a white-box model using logistic regression, and a black-box model using a random forest. Describe the feature engineering and scaling steps you took to train the classifiers (e.g. standardization of features is a good idea for the logistic regression model) and summarize the steps in your report.

   For both models, report on the accuracy of the classifier on the test set.

2. For each classifier, measure statistical parity and equalized odds (both in terms of T = 0 and T (see Lecture 2). Plot the results and discuss the differences that you observe.

3. Change the classification pipeline to (approximately) fulfill one of the fairness criteria by post-processing the results (e.g. either statistical parity or equalized odds). How did the intervention influence the different fairness criteria, how did it change the accuracy of the classification?

Task 2 – Explaining models using SHAP

1. Both for the white-box (logistic regression) and the black-box (random forest) classifier, use the SHAP module to explain predictions. Contrast the two models to each other: What are similarities, how do they differ?

2. Given the outcome of your study, which classifier is most suited for the prediction task under accuracy, explainability, and fairness considerations?

Please submit your report (in PDF format) & jupyter notebook on learnIT.

Checklist

To avoid surprises, please make sure that your hand-in covers the following parts:

Overall

- □ Assignment: Concise discussion of the results
- □ Plots: Labels and titles clear and readable?
- □ Code: can run the whole notebook? Modular, concise, and documented code?
- □ Page limit: At most 3 pages. Plots should still be readable!

Task 1

- □ Part 1: Discussion feature engineering and scaling steps
- □ Part 1: Correct implementation of one-hot-encoding (if used for model)
- □ Part 1: Build and train relevant models
- □ Part 2: Code to compute statistical parity and equalized odds
- □ Part 2: Plot(s) comparing the metrics
- □ Part 3: Discussion on accuracy changes and how other measures were affected by the intervention

Task 2

- □ Part 1: Code to load and use SHAP package on models from Task 1 to explain their predictions.
- □ Part 1: One or two SHAP plots comparing the similarities and differences between the models
- □ Part 2: Discussion on which classifier is most suited for the prediction task, factoring in accuracy, explanainability, and fairness