Responsible AI Special course Fall 2023
Project on Algorithmic Fairness

Aasa Feragen

March 1, 2023

Hand-in: By the end of March 12 on Inside.

For your first project, you will train an algorithm to predict recidivism in a juvenile justice dataset from Catalonia. You will first perform a diagnostic analysis of your algorithm, and next you will try to mitigate whatever bias you may have found.

More precisely:

1. You will be working with a dataset found in the zip file

   `catalan_data_course.zip`,

   where the dataset you should use is found in the file

   `catalan-juvenile-recidivism-subset.csv`,

   and an English explanation of the features is found in the file

   `recidivismJuvenileJustice_variables_EN.pdf`.

   First, load the data, perform a first exploration, and provide a description of the dataset in your report.

2. Split the dataset into training, validation and test sets, and train a machine learning algorithm of your choice to predict the variable Recidivism (year 2015). Consider and explain your choice of dataset splits: You want a sufficiently big test split to stably assess bias. Please describe your data splits and your algorithm of choice, as well as any training details needed to reproduce the model.

Image from Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner (ProPublica): Machine Bias

3. Please design your own criteria for diagnosing the three main definitions of fairness: Independence, Separation and Sufficiency. Use your criteria to perform a diagnostic test of your algorithm with respect to sensitive groups such as race and/or gender. Please describe your designed diagnostic criteria, as well as the results of your test. The rest of the project will be most interesting if you have picked a classifier where you observe unfairness at this point.

4. Try a bias mitigation technique of your choice on the model. It is OK to use something that is implemented by others, and it is OK to implement it yourself

5. Redo your diagnostic analysis with the updated model – is it more fair? Did you introduce new problems? Can you visualize your results, e.g. examples of unfair predictions?

Please summarize what you did in a 2-4 page report. Please make sure to explain *why* you made your different choices and discuss your results. Did you find particular aspects of fairness more appropriate than others? How did the fairness mitigation affect your performance? What are the pros and cons of your updated algorithm?