# PyThia: A Reporting Tool on Bias Evaluation and Mitigation


## Abstract

In his book 21 lessons from the 21st century. U. N. Harari states, *"Already today, 'truth' is defined by the top results of the Google search."*. So far we may afford the gravity of such a statement, but the recent events in the U.S.A. with police brutality make apparent the fact that authority is increasingly algorithmic and this is very alarming. As society and individuals we need to strive for technology to be governed properly and entrust that it operates in a fair and just way. PyThia, the proposed tool here is based on our thesis that we need technology to control technology and to augment the humans in the loop who are responsible for the task. When it comes to the aspect of Fairness, this can be evaluated to a certain extent using quantitative methods, merely the ones related to bias identification and mitigation. By using metrics such as Disparate Impact, one may assess whether a given dataset and/or model bear certain biases. This may not exclude all forms of bias but it is a first good step towards governing technology, especially Artificial Intelligence/Machine Learning models. PyThia, the proposed tool has the following novelties:

- It monitors through time how the group fairness (based on the Disparate Impact Ratio) evolves depending on the data a given model receives/processes,
- It provides certain bias mitigation approaches given the type of dataset and model,
- It helps its users to gain insights for all sensitive attributes and their reference groups for the Group Fairness as well as the mitigated scores.


## Tool Description

PyThia is a web-based tool whose goal is to identify and mitigate potential biases to both dataset as well as model level. It does so by :

- Monitoring the fairness of a given predictive model over time. It does so by using a set of statistical metrics related to bias identification (e.g. Disparate Impact Ratio),
- Demonstrating the impact of a given bias mitigation technique to the respective bias identification measurements again over time.

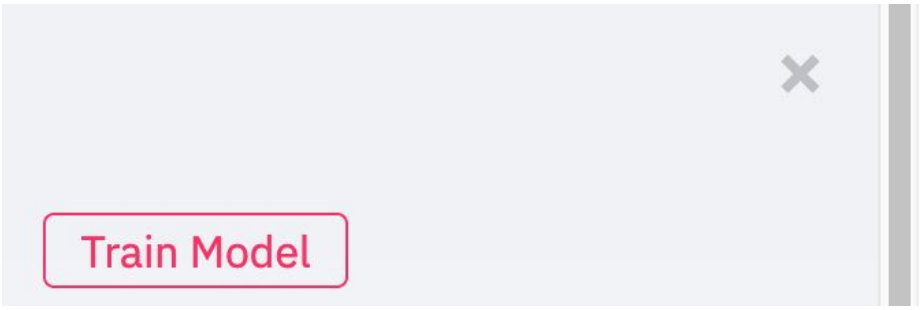
The tool is based on an open and extensible architecture that caters for plugging in more algorithms for either identifying or mitigating bias. The interactive demo of the tool can be accessed at http://md4sg.code4thought.eu and is based on the Bank Marketing dataset as retrieved by Open ML (https://www.openml.org/d/1461).

There, the user needs to perform the following set of actions:

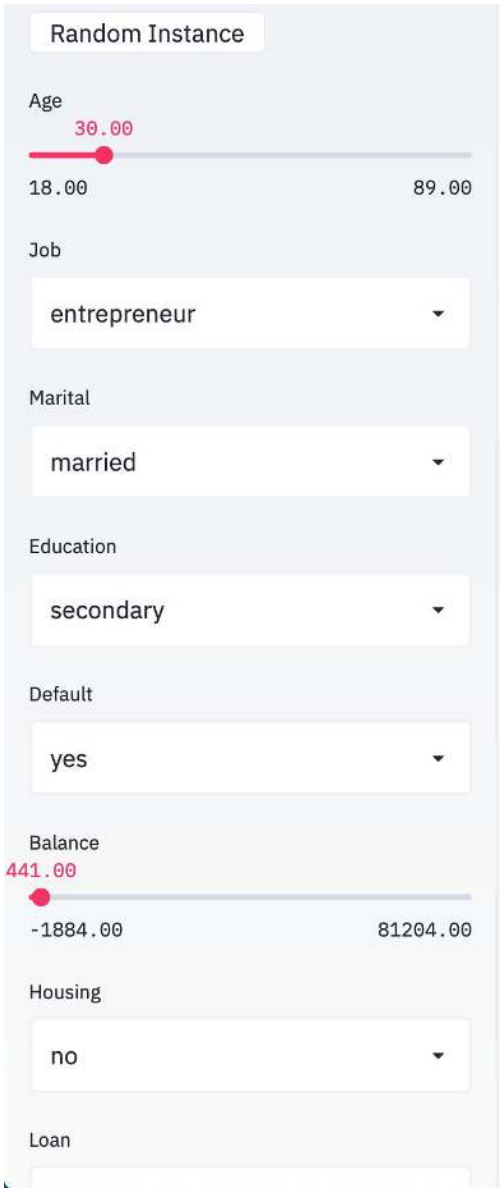
- First, to define the scope of the analyses. More specifically the user needs to define:

- The sensitive attribute(s) against which the bias identification and mitigation analysis will take place,
- Their respective reference group(s); for instance if the user selects to check for bias against the Marital status, then subsequently has three choices to choose from, Single, Married, Divorced.
- The label (i.e. the result/verdict of the model's classification).


- The next step is to train the model (for the purposes of the demo a Random Forest model was used) .

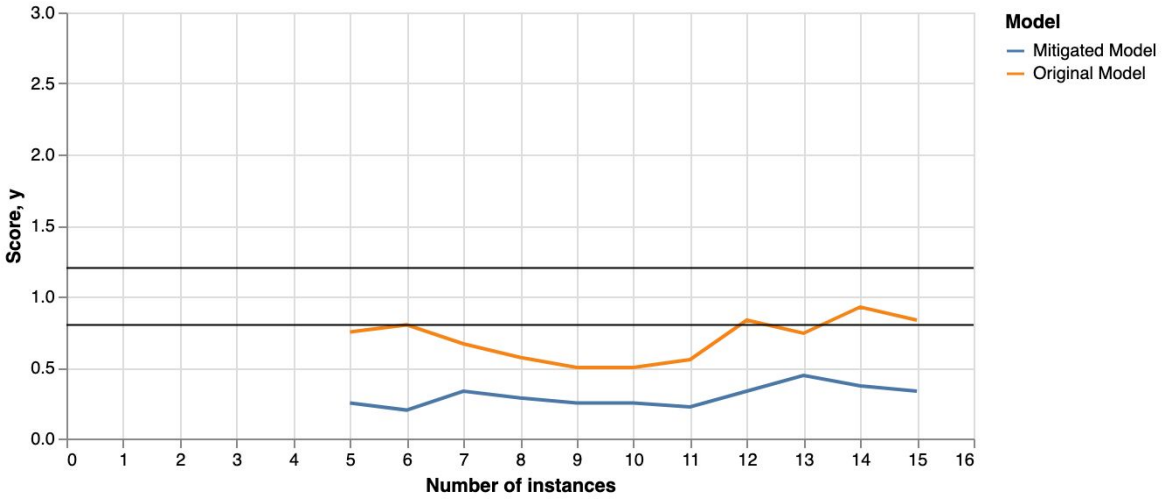



- Then the user has the ability to generate more instances, by pressing the "Random Instance" button and check how the Group Fairness metric evolves for both the original model and the one for which a mitigation technique was applied.




For instance, the following graph indicates how the fairness on the Marital status for both the original and the mitigated model evolves for a set of 15 instances/individuals.

It has to be noted that the user has also the ability to check the fairness  for a combination of sensitive attributes.


Model
Mitigated Model
Original Model
Score, y
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Number of instances