Technologies for Information Systems project on Data Ethics Report

Measure fairness in healtcare-related database using the **Aequitas** toolkit

Simone Callegarin

April 2023

POLITECNICO
MILANO 1863

Reference professors: Letizia Tanca, Chiara Criscuolo and Tommaso Dolci

Prefaction

The practical project on Data Ethics consisted in the creation and presentation of a python notebook.
The assignment required the use of the **Aequitas toolkit** to measure fairness in an healthcare-related dataset that in my case concerns **diabetes**.
The content of the **notebook** can be divided into four main **phases**:

1. Load the dataset
2. Set the inputs
3. Basic preprocessing and prediction algorithm
4. Fairness results with Aequitas

1 Load the dataset

The dataset provided in `Kaggle`[1] is an already preprocessed version of the original dataset available in `OpenML`[2], named *"Diabetes 130-Hospitals"*, collected by *Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios*, and *John N. Clore* in 2014.
The notebook provides a comprehensive description and a preliminary analysis of the dataset, involving libraries such as `Seaborn` and `MatPlotLib` for graphs plotting.
The dataset counts 763 tuples with 9 attributes related to the entities' physical parameters and the *Outcome* value that labels individuals as diabetic.

2 Set the inputs

In order to audit the dataset it's necessary to define some inputs first:

- the attributes to audit (i.e. sensitive attributes)
- the reference group for each attribute (majority groups)
- the fairness metric(s) that we care about (in this case `FNR, TPR`)
- the disparity tolerance ($\tau = 1.25$)

In our analysis the sensitive attribute was the *AgeCategory* and as presented by the authors of Aequitas[5], the interested fairness metrics were the `FNR` and the `TPR` due to the fact that the interventions made by our predictive model are assistive and intervening with a big part of the population[14].

The *false negative rate* (FNR) is the fraction of false negatives of a group within the labeled positives of the group:

$$FNR_g = \frac{FN_g}{LP_g} = Pr(\hat{Y} = 0 | Y = 1, A = a_i)$$

The *true positive rate* (TPR) is the fraction of true positive of a group within the labeled positives of the group:

$$TPR_g = \frac{TP_g}{LP_g} = Pr(\hat{Y} = 1 | Y = 1, A = a_i)$$

Speaking in terms of parity, the *true positive rate parity* is also called equality of opportunity and assert that the protected and unprotected groups should have equal true positive rates (i.e. same rate of positive outcome, assuming the people in the groups qualify for the outcome).
The *disparity tolerance treshold* controls the range of disparity values that can be considered fair.
Our formulation involves the *"80% rule"* represented by $\tau = 0.8$.
This notion of parity requires all biases to be within the range defined by τ:

$$\tau \leq DisparityMeasure_{group_i} \leq \frac{1}{\tau}$$

$$0.8 \leq DisparityMeasure_{group_i} \leq \frac{1}{0.8} = 1.25$$

3 Basic preprocessing and prediction algorithm

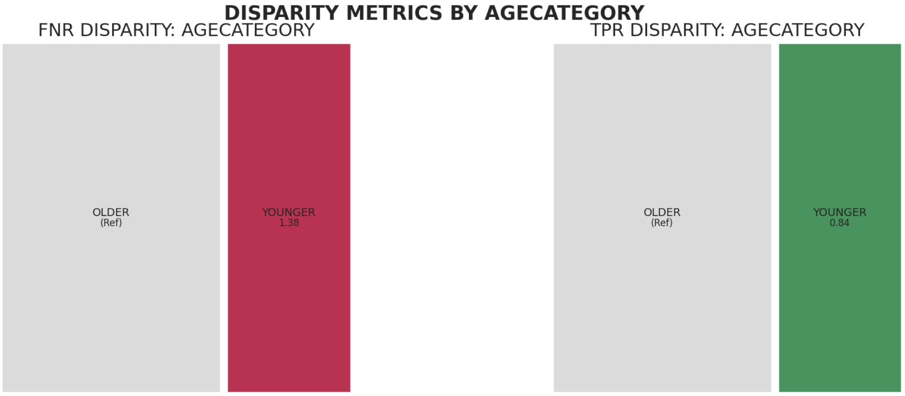
The predictive algorithm used is the `Random Forest Classifier`[3] that is imported from `sklearn` library that also offers tools to handle Data training and test splits.
Data obtained from the predictive algorithm undergoes a preprocessing phase that will lead to the creation of a dataframe, containing entities from the test set and the related algorithm's predictions, that is directly uploaded into the GitHub repository[11].

4 Fairness results with Aequitas

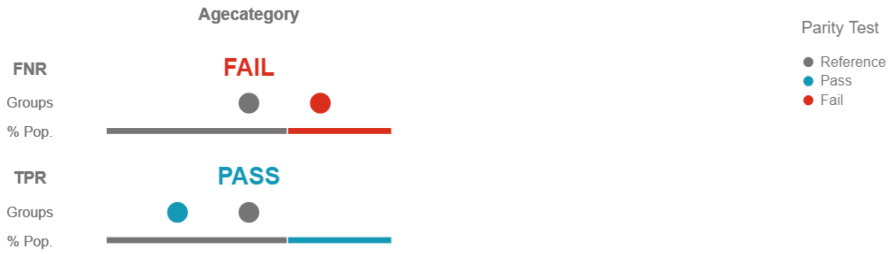
Applying Aequitas[4] programmatically is a three step process represented by this three python classes:

1. `Group`
2. `Bias`
3. `Fairness`

The analysis consist in measuring bias and fairness between the true *outcome* that is the true binary label of a given entity and the *score* assigned to each entity by the predictor[3].
The **results** of the audit highlighted that the model tends to indicate that *people actually with diabetes are more likely to be labeled negative at a younger age.*



This is determined by the fact that the *False Negative Rate* (`FNR`) of the `YOUNGER` group exceed the disparity tolerance treshold, highlighting an unfairness in our model.
On the contrary the *True Positive Rate* (`TPR`) shows that the `YOUNGER` group does not exceed the disparity tolerance treshold, revealing that the predictive model is unbiased over this measure.



For a group to pass the parity test its disparity to the reference group cannot exceed the fairness threshold (1.25).
An attribute passes the parity test for a given metric if all its groups pass the test.

In conclusion we can state that the model didn't pass the **`FNR parity test`**, while it passed the **`TPR parity test`** instead.
The predictive model[3] is then biased because it tends to *label diabetic people as negative at a younger age*

References

[1] Kaggle - diabetes dataset

[2] OpenML - Diabetes-130-Hospitals dataset

[3] ScikitLearn - Random Forest Classifier

[4] Aequitas

[5] Aequitas - paper

[6] Aequitas - textbook

[7] Aequitas - tools guide

[8] Aequitas - tutorial python notebook

[9] Aequitas - tutorial

[10] PyPi

[11] GitHub

[12] StackOverflow

[13] Pandas

[14] Fairness in ML

[15] Confusion metrics

[16] Python projects packaging

[17] Hex calculator

[18] Tkinter colors

[19] SCdataframeFormatter