AFRICAN INSTITUTE FOR MATHEMATICAL SCIENCES

QUANTUM LEAP AFRICA

Assignment 3 of Social and Ethics:

Fairness

03 JUIN 2019

SONNA MOMO BELONA MARY
bsonna@aimsammi.org

I- SUMMARY

Table des matières

II- INTRODUCTION AND DESCRIPTION OF THE DATASET

The objectives of the assignment named Fairness in Classification and representation Learning are:

- Identify the sensitives attributes (features that can impact the prediction of the model) of a dataset and build a model classifier which is fair wrt these attributes.
- How we can include the metrics such as MMD to improve the fairness of the model a representation learning

To accomplish these tasks, we are working with Adult dataset that is described in the README file. Basically, we have train dataset and test dataset files. Each of this file contains, the samples x, the labels y and the sensitive attribute 'a'. the sentitive attribute A represents the gender. This attribute is represented in the dataset by two features, sex_Male and sex_female. For a given sample, if the value of sex_male is 1, the value of sex_female will be 0. That is why we can consider that sex male is A and Sex Female is 1-A with A the gender. Our work is divided into two notebooks, the first one is solving the classification and the second one is about the representation learning.

III- CLASSIFICATION NOTEBOOK

In this part, we will sequentially:

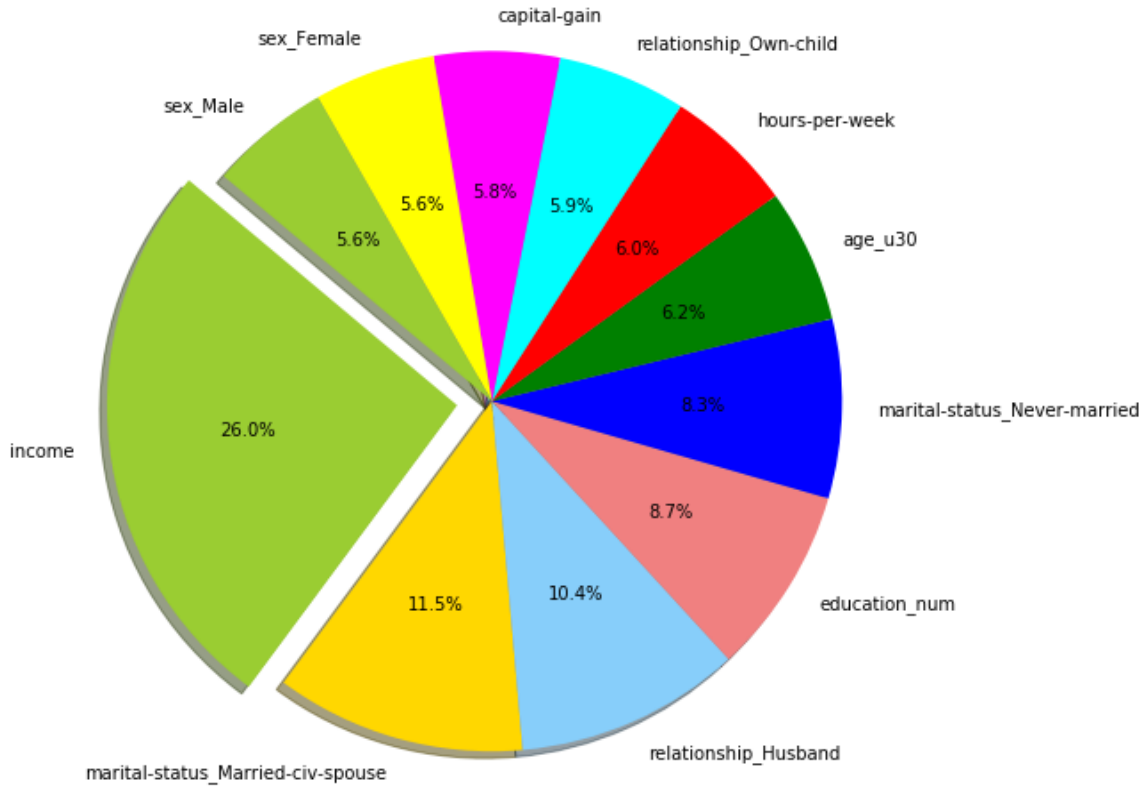
- Find the attributes that are mostly correlated with y ;
- Find the attributes that are mostly corelated with the sensitive attribute A;
- Build the model classifier;
- Train and evaluate the previous model with all the features and report the score;
- Train and evaluate the previous model without the attributes that are most corelated with sensitive attribute A. This latter is to build a classifier which is fair wrt to the gender.
- And other several tasks given in the exercises.

1- Build the Train data

After the extraction of the data to the dictionary file; we have built a Pandas Data frame having as labels the content of the headers.txt and as the content, the concatenation of 'x',' y' and 'a'. 'Y' stands for the income and 'a' the gender. To compute the correlation between the features of x and y or a; we have used the predefine function corr() of pandas since it uses the Pearson algorithm.

2- Attributes mostly corelated with Y

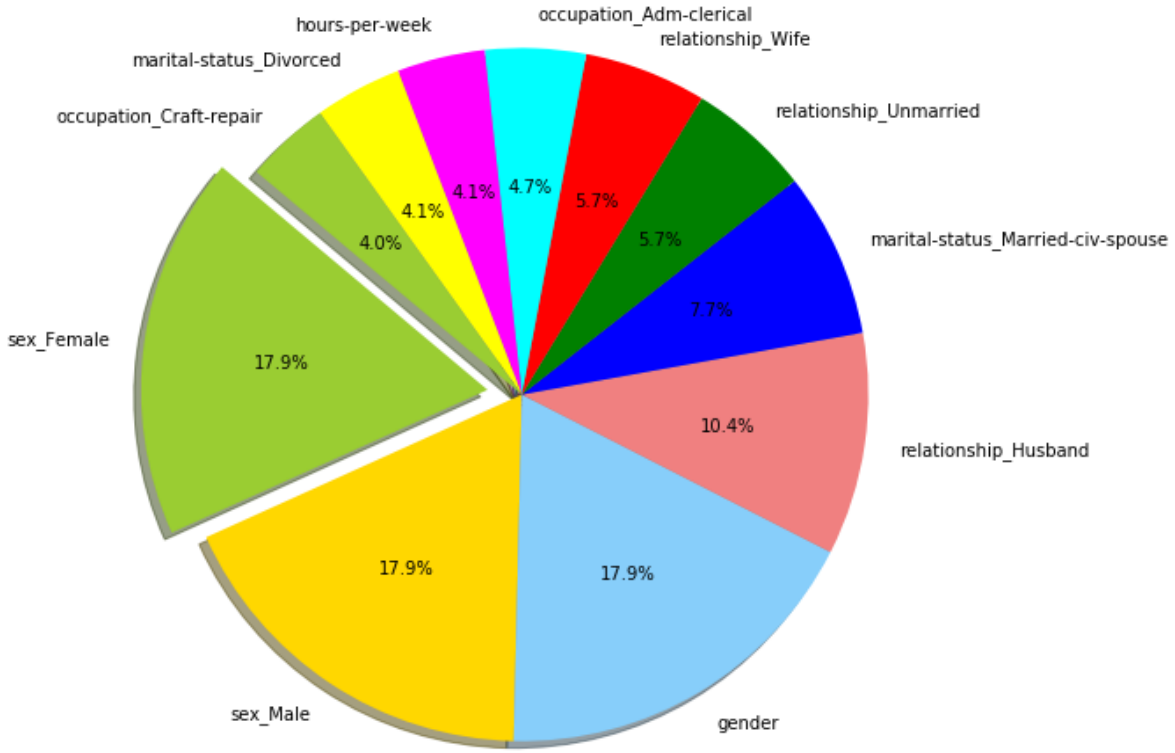
```
income                                1.000000
marital-status_Married-civ-spouse     0.444696
relationship_Husband                  0.401035
education_num                         0.335154
marital-status_Never-married          0.318440
age_u30                               0.238133
hours-per-week                        0.229689
relationship_Own-child                0.228532
capital-gain                          0.223329
sex_Female                            0.215980
sex_Male                              0.215980
Name: income, dtype: float64
```



Comments:

3- Attributes mostly corelated with A

```
sex_Female                          1.000000
sex_Male                            1.000000
gender                              1.000000
relationship_Husband                0.580135
marital-status_Married-civ-spouse   0.431805
relationship_Unmarried              0.321273
relationship_Wife                   0.319311
occupation_Adm-clerical             0.263148
hours-per-week                      0.229309
marital-status_Divorced             0.228621
occupation_Craft-repair             0.223128
Name: gender, dtype: float64
```



Comments: we noticed that the sex_female and the sex_male are strongly correlated to the gender. It makes a lot of sense because these two features are representing the gender.

4- Scores of the model classifier to predict Y with all the features of x on test data

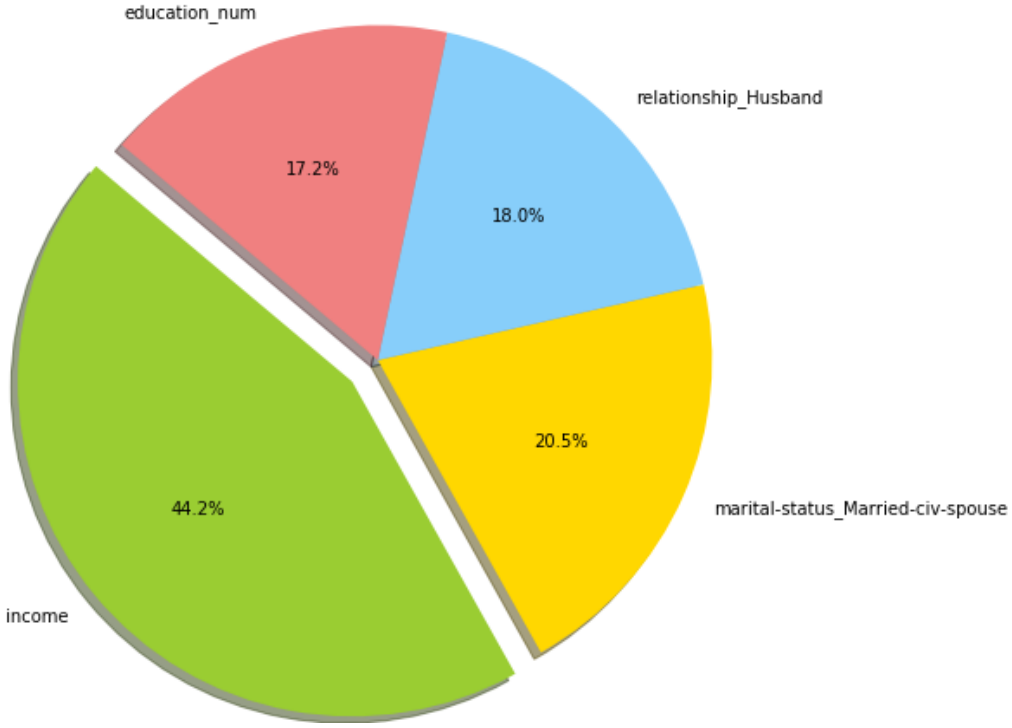
Metrics	values
Accuracy	0.84
Mean Difference	-0.17
Demographic parity	0.17

5- Scores of the model classifier to predict Y without the 10 features mostly corelated to A on test data

Metrics	values
Accuracy	0.83
Mean Difference	-0.12
Demographic parity	0.12

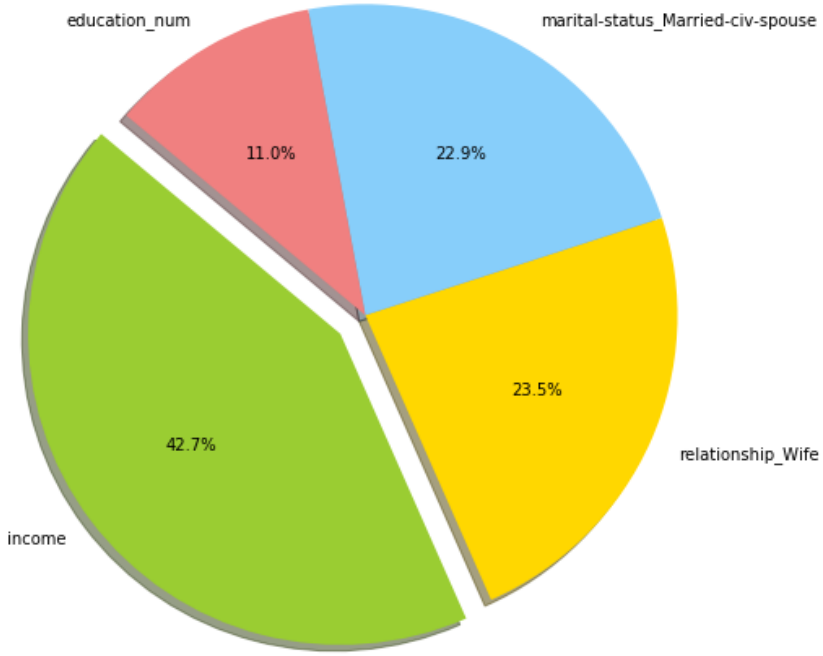
Comments: Regarding the mean differences of the sensitive groups over the predicted value $\hat{Y}$, we can conclude that the group with attribute A has higher value of $\hat{Y}$ on the average.

6- Three features of our data mostly corelated with the predicted value

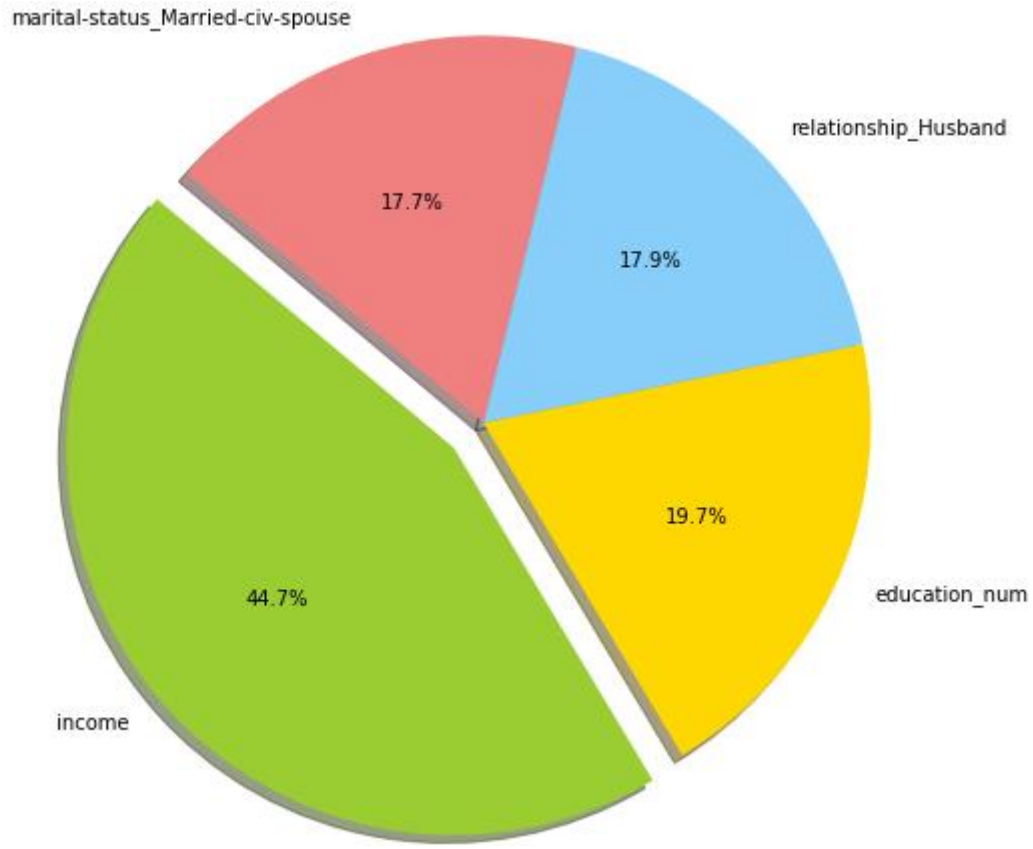


Comments: it makes sense, because these three attributes where also the most three features most corelated to the Y.

7- Three features of our data mostly corelated with the predicted value only looking at A=0



8- Three features of our data mostly corelated with the predicted value only looking at A=1



9- Scores of Model Classifier to predict A without the features sex_female and sex_male in the features of x

Metrics	values
Accuracy	0.81
Reweighted Accuracy	0.77

10-Scores of Model Classifier to predict A without the 10 features mostly corelated with A

Metrics	values
Accuracy	0.72
Reweighted Accuracy	0.65

Comments: we noticed that both the accuracy and the reweighted accuracy are going down when we remove a lot of attributes corelated to the sensitive attribute. This remark only confirms that there is always a trade-off between the fairness and the accuracy.

IV- REPRESENTATION LEARNING

1- Scores of Model classifier g to predict Y

Metrics	values
Accuracy	0.82
Demographic Parity	0.17

2- Scores of Model classifier h to predict A

Metrics	values
Accuracy	0.99
Reweighted Accuracy	0.99

Comments: we noticed that the accuracy and the accuracy of the predictor of A is almost 100%. This result was expected because the processing applied to the data is a kind of normalisation and standardization of the dataset. Hence, we confirm that the normalisation is welcome to improve the scores of a ML model.

3- Scores of Model classifier g to predict Y with MMD

Metrics	values
Accuracy	0.83
Demographic Parity	0.17

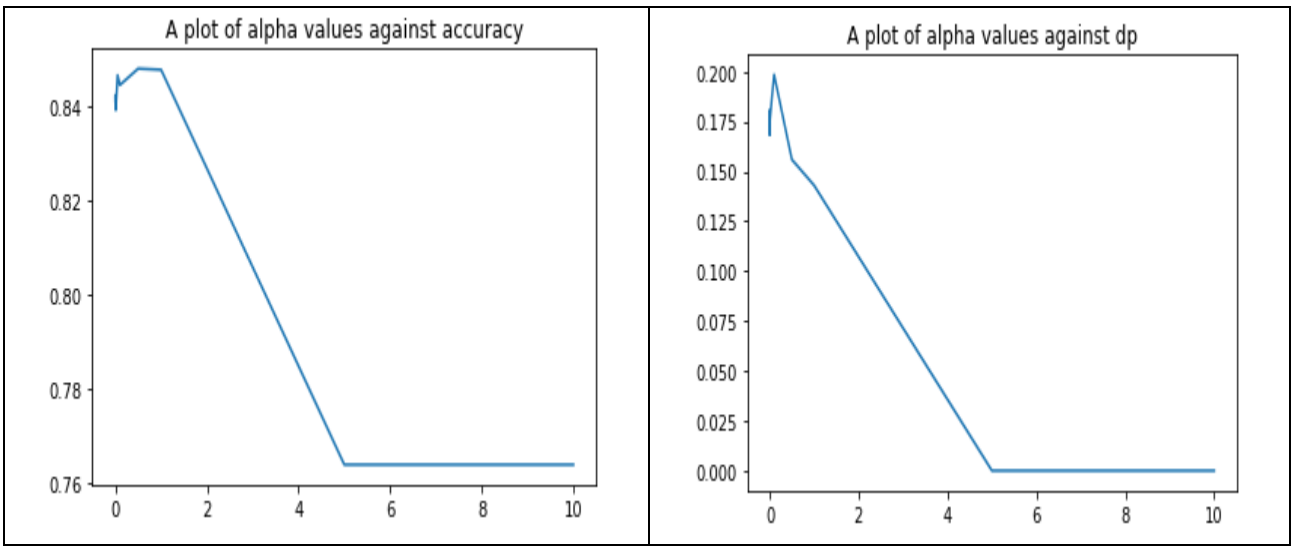
4- Scores of Model classifier h to predict A with MMD

Metrics	values
Accuracy	0.96
Reweighted Accuracy	0.95

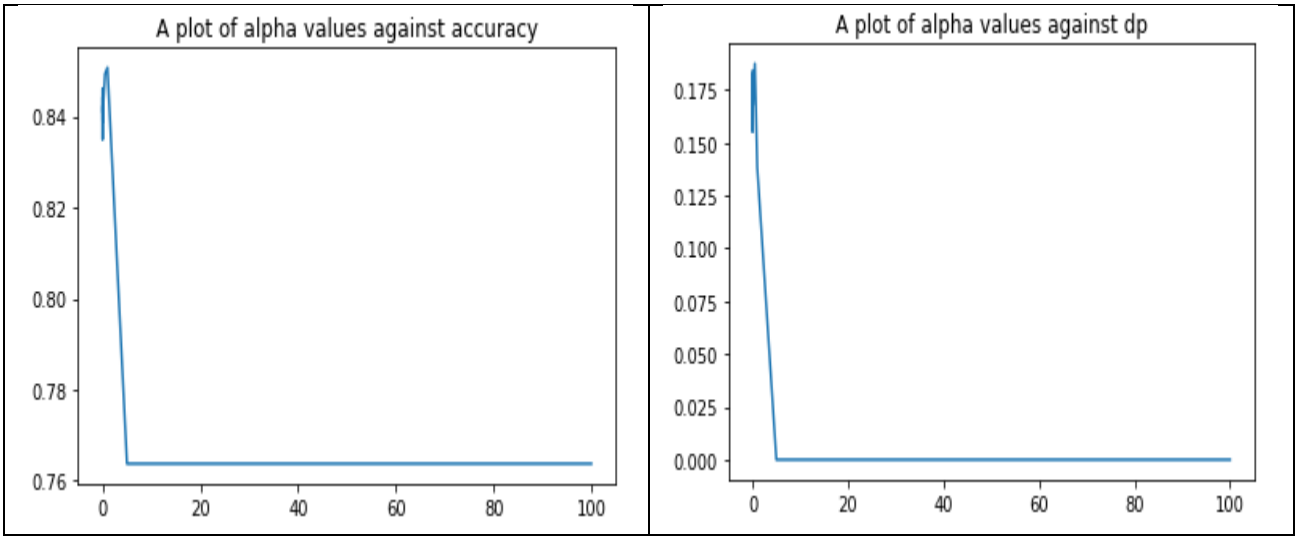
Comments: the model was better to predict A without MMD but was better to predict Y with MMD.

5- Plot the Accuracy against the range of values of alpha

- First range of alpha between 0 and 10



- First range of alpha between 0 and 100



Comments: we noticed that when the value of alpha increases, both the values of accuracy and the demographic privacy decrease. This result was expected because this alpha can be considered as a kind of regulariser. So, we should always find the value of alpha that maximises the accuracy. In this work, the best value of alpha is between 0 and 1.

V- CONCLUSION [1]

- In this work we have learned two methods of bias mitigation for Machine learning namely reweighting and fair representation. Both of these methods are part of pre-processing algorithms. Each of this method has its particularity. For the first method, the idea is to apply

appropriate weights to different tuples in the training dataset to make the training dataset discrimination free with respect to the sensitive attributes. For the fair representation method, the idea is to find a latent representation that encodes the data well while obfuscating information about protected attributes. For this specific task, I think the best way to compare these methods is the accuracy of the model on the test data.

- Another way of removing information about A is Optimized pre-processing. The idea of this latter is to learn a probabilistic transformation that can edit the features and labels in the data with group fairness, individual distortion and data fidelity constraints and objectives.

VI- REFERENCES

[1] https://dzone.com/articles/machine-learning-models-bias-mitigation-strategies