Danmarks Tekniske Universitet

DTU

Responsible AI

First assignment: Project on Algorithmic Fairness

Authors

Albert Kjøller Jacobsen - s194253
Carolina Lopez - s212725
Naja Jean Larsen - s184424
Nina Weng - s202997

GitHub repository / code base: `https://github.com/albertkjoller/AlgorithmicFairness`

March 12, 2023

Contents

1 Introduction

This project will study the biases found in an algorithm's prediction, a Neural Network (NN) taking the form of an *encoding-only* Autoencoder (AE) with an associated NN classifying from latent representations. This project investigates whether removing all sensitive attributes $\boldsymbol{S}$, from the data set $\boldsymbol{X}$, provides a basis for training a fairer model. As such, fairness is solely analyzed with respect to *gender*, however, the provided methods easily extends to other issues, such as *race*-related model biases. This hypothesis will be investigated through comparisons to an approach exploiting the full data, $\boldsymbol{X}$, and will be based on discussions of the fairness criteria as well as through analyses of the learned latent space of the AE.

2 Dataset

2.1 Statistical analysis

Initially, an analysis of the given data set was performed in order to find possible biases in the data and the general statistics of the existing attributes. Table 10 found in the Appendix explains the separation carried out considering sensitive attributes from the numerical, ordinal, and categorical ones . Similarly, figures regarding statistics of the raw data set can be found in the Appendix (Section 6). It was found that the percentage of total recidivist subjects corresponds to 34.3% (see Fig. 6). Regarding sex, the data set has a larger fraction of *males*, presenting 82.2% of the data in this group. Furthermore, the *male* group shows higher recidivism (37.3%) compared to what is observed for *females* (20.5%) (see Fig. 7). The imbalance present in the data set regarding the amount of *male* data will also play a critical role in the algorithm's performance.

2.2 Pre-processing of the data

After having diagnosed the respective data types for all attributes, a pre-processing step was carried out in order to appropriately deal with them. Therefore, the ordinal attributes have been grouped into bins and mapped from the original string content to numerical ordered data. An overview of the mappings can be found in Table 5. Furthermore, one-hot encoding has been applied to categorical attributes resulting in a data frame containing only numerical values. Lastly, all data splits were standardized by demeaning and dividing with the standard deviation obtained from the training set split for improving convergence and equal treatment of features.

2.3 Dataset Splitting

Two strategies were used for splitting the data set. First, in order to achieve a stable data set split for a fair comparison between different methods, the data set was split by setting a random seed, which guarantees the same splitting for every run. Second, a large test set (40% of the all data) was kept so that all sensitive attributes are representative. To ensure that, analysis was run and shows that the distribution of sensitive attributes are similar across training, validation and testing set (percentage of female: 17.9%, 18.3%, 17.17% respectively).

3 Algorithm of choice

For conducting the fairness analyses with two different data foundations, a model *with* access to all attributes - denoted as $AE_{\boldsymbol{X}}(\boldsymbol{x})$ - and a model *without* access to sensitive attributes - denoted as $AE_{\boldsymbol{X}\setminus\boldsymbol{S}}(\boldsymbol{x})$ - were trained.

3.1 Model architecture

The network followed the encoder-part of the structure of an autoencoder (AE) mapping input data, $\boldsymbol{X}$, to representations in a latent space, $\boldsymbol{z}$. Through pilot experiments the architecture of the AE was specified as 3 fully connected layers; each of 32 hidden units with dropout of 20% and batch normalization before a linear layer was applied to obtain a latent space of dimensionality 32.

3.2 Training details

Through pilot runs both models were independently optimized wrt. learning rate $\in [10^{-6}, 10^{-3}]$ and batch size $\in \{32, 64, 128\}$, resulting in a learning rate of 10^{-4} and a batch size of 64 for both models. Both models were trained for minimizing the binary cross-entropy loss objective (BCE) for 200 epochs from which model parameters at the epoch with the lowest validation loss were saved. The random seed was set to 42.

4 Results and Discussion

This section presents the results and discussions from the performance of the models, $AE_{\boldsymbol{X}}(\boldsymbol{x})$ and $AE_{\boldsymbol{X}\setminus\boldsymbol{S}}(\boldsymbol{x})$, first after applying the fairness criteria and secondly, through the latent space representations obtained.

4.1 Fairness Criteria and Bias Mitigation

To access how fair the models are related to gender, an evaluation based on three different fairness criteria - namely, *Independence*, *Separation* and *Sufficiency* - will be conducted.

4.1.1 Independence

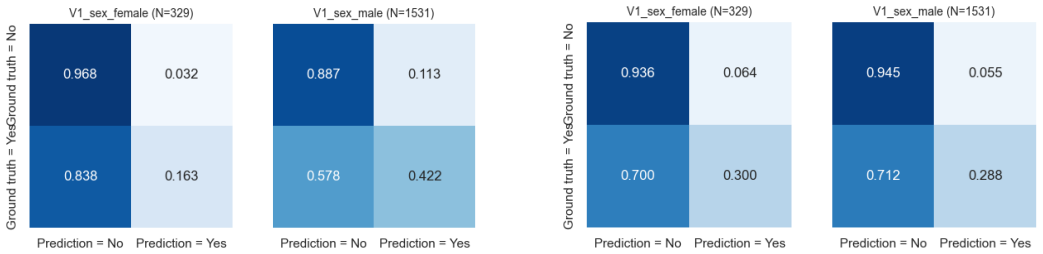
Independence, also known as *equality of outcome*, can be mathematically formulated as follows: $\mathbb{P}\{\hat{Y} = 1 \mid S = male\} = \mathbb{P}\{\hat{Y} = 1 \mid S = female\}$. This means the probability of predicting that a person will re-offend must be independent of the sensitive attribute, in our investigation, *gender*. Table 1 shows that when *gender* information is included, the model is very biased in regards to the independence criteria. *Males* (0.23) are predicted to re-offend more than twice as much as *females* (0.10), whereas when excluding the gender information, they are almost equally predicted to re-offend (*males* 0.23 and *females* 0.22). As a result, based on the Independence criteria, the model that only had access to non-sensitive information is fairer.

	Males	**Females**
$AE_{\boldsymbol{X}}(\boldsymbol{x_{test}})$	0.23	0.1
$AE_{\boldsymbol{X}\setminus\boldsymbol{S}}(\boldsymbol{x_{test}})$	0.23	0.22

Table 1: Predicted recidivism ratio for males and female

4.1.2 Separation

Compared to *Independence*, *Separation* does not only look at the prediction ratio between groups, but also the error rates. For binary classification task, the goal is to obtain an equal true positive rate (TPR) and the false positive rate (FPR) between the sensitive groups, thus equalizing the error rates. Mathematically this is formulated as: $\mathbb{P}\{\hat{Y} = 1 \mid Y = 1, S = male\} = \mathbb{P}\{\hat{Y} = 1 \mid Y = 1, S = female\}$ and $\mathbb{P}\{\hat{Y} = 1 \mid Y = 0, S = male\} = \mathbb{P}\{\hat{Y} = 1 \mid Y = 0, S = female\}$. and can be measured by confusion matrices and ROC curves.



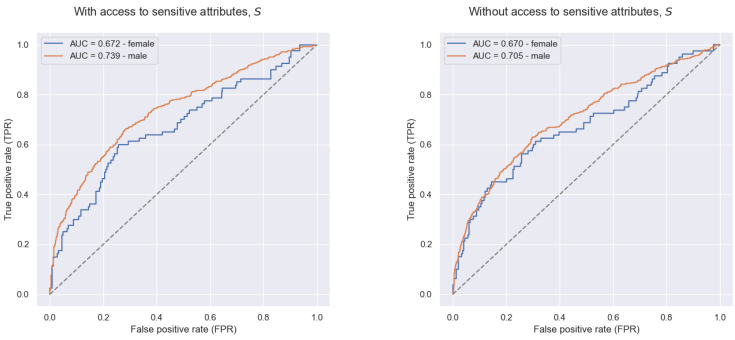
(a) $AE_{\boldsymbol{X}}(\boldsymbol{x_{test}})$ (b) $AE_{\boldsymbol{X}\setminus\boldsymbol{S}}(\boldsymbol{x_{test}})$

Figure 1: Confusion matrices including and excluding sensitive attributes

In Figure 1 an improvement in the bias towards *Females* is seen. In model $AE_{\boldsymbol{X}}(\boldsymbol{x})$, the *Females* group tends to be more often falsely predicted of non-recidivism, compared to *Males*. Not only does model $AE_{\boldsymbol{X}\setminus\boldsymbol{S}}$ reduce this difference, but also illustrates matrices looking much more alike. The ROC curves in Figure 2 provide a similar message, where ROC for both, Female and Male, have less gap in $AE_{\boldsymbol{X}\setminus\boldsymbol{S}}$, especially when FPR is low. In short, from a *Separation* point of view, leaving out sensitive information also seems to make the model fairer with regard to gender. It is worth noticing here the new problems introduced while achieving fairness over gender. As seen in the ROC curve in Figure 2, both female and male obtained a lower AUC score in $AE_{\boldsymbol{X}\setminus\boldsymbol{S}}$ (female: decrease 0.002; male: decrease 0.033). Moreover, the accuracy for the whole test set dropped 0.8% (see Table 4). This fact indicates that there is a trade-off between prediction ability and fairness in this specific case.

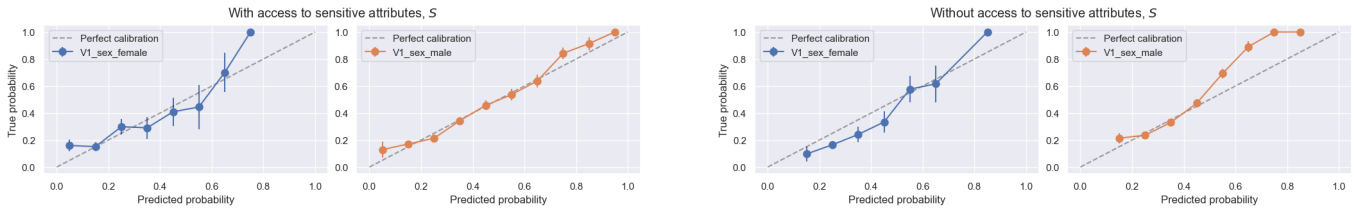
4.1.3 Sufficiency/Calibration

Sufficiency is achieved if the sensitive attributes are statistically independent of the target value Y, given the prediction R. In this case, it could be written as: $\mathbb{P}\{Y = 1 \mid R = r, A = male\} = \mathbb{P}\{Y = 1 \mid R = r, A = female\}$. In this report, calibration by group is used as a proxy for *Separation*. *Calibration* could be satisfied by: $\mathbb{P}\{Y = 1 \mid R = r, A = a\} = r$. Figure 3 shows the probability for *Females* is not calibrated very well for $AE_{\boldsymbol{X}}$, as the outer probabilities are over-represented, while the middle ones are slightly underrepresented. The standard deviation around the middle probabilities for *Females* are also wider. On contrary, the curve for *Males* is much better calibrated. For $AE_{\boldsymbol{X}\setminus\boldsymbol{S}}$, the curve for *Females* is slightly better calibrated, but the calibration for men significantly degraded, with an over-representation at higher probabilities. From a sufficiency/calibration perspective, it is difficult to determine which model is fairer.



(a) ROC curve for $AE_{\boldsymbol{X}}(\boldsymbol{x_{test}})$ (b) ROC curve for $AE_{\boldsymbol{X}\setminus\boldsymbol{S}}(\boldsymbol{x_{test}})$

Figure 2: ROC curve including and excluding sensitive attributes



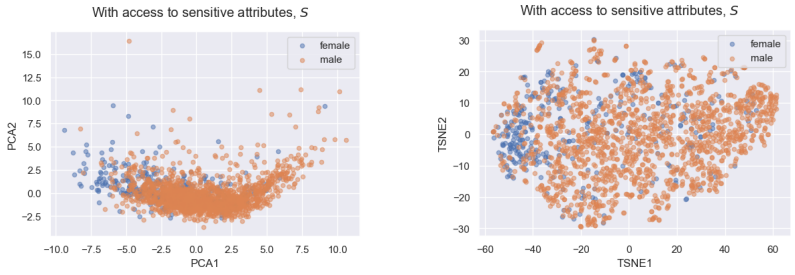
(a) $AE_{\boldsymbol{X}}(\boldsymbol{x_{test}})$ (b) $AE_{\boldsymbol{X}\setminus\boldsymbol{S}}(\boldsymbol{x_{test}})$

Figure 3: Calibration curves including and excluding sensitive attributes for male and female

4.2 Latent Representations

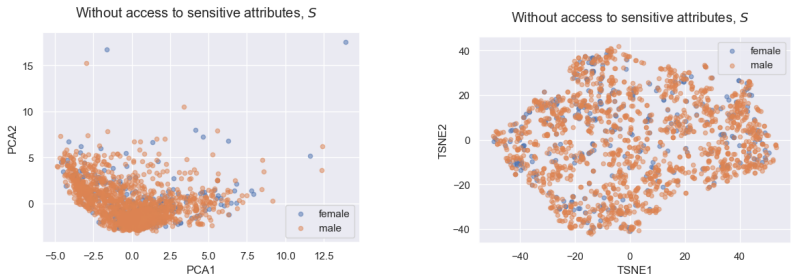
The latent space representation of our model provides another possibility for exploring the effectiveness of our bias mitigation approach. The encoder takes the input data and maps it to a lower-dimensional space, the *latent space*. Both PCA and t-SNE reduced versions of this space are visualized in the graphs below.

Figures 5a and 5b show the distribution in the latent space for $AE_{\boldsymbol{X}\setminus\boldsymbol{S}}(\boldsymbol{x_{test}})$ where it is harder to separate within the sensitive groups, compared to Figures 4a and 4b where the *female* group is skewed to the left of the



(a) PCA in $AE_{\boldsymbol{X}}(\boldsymbol{x_{test}})$ (b) TSNE in $AE_{\boldsymbol{X}}(\boldsymbol{x_{test}})$

Figure 4: Latent space including sensitive attributes for both PCA and TSNE.



(a) PCA in $AE_{\boldsymbol{X}}(\boldsymbol{x_{test}})$ (b) TSNE in $AE_{\boldsymbol{X}}(\boldsymbol{x_{test}})$

Figure 5: Latent space excluding sensitive attributes for both PCA and TSNE.

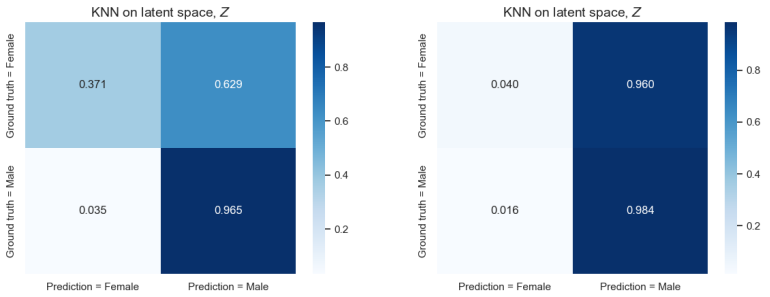
representations, showing an easier separable frontier.

4.2.1 KNN Classifier

With the aim of finding another indicator to confirm the learned bias in $AE_{\boldsymbol{X}}(\boldsymbol{x}_{test})$, a KNN classifier predicting the sensitive attribute has been trained. Figure 3 shows a poor accuracy on the model trained with latent vector of model $AE_{\boldsymbol{X\setminus S}}(\boldsymbol{x_{test}})$, which classifies most samples as *man*, being unable to make a separation between the groups. This reassures the positive outcome of the bias mitigation technique used when it comes to *gender*.

	Accu.	**Accu.(F)**	**Accu.(M)**
Baseline[1]	0.823	-	-
$AE_{\boldsymbol{X}}$	0.860	0.370	0.965
$AE_{\boldsymbol{X\setminus S}}$	0.817	0.039	0.984

Table 2: Accuracy obtained for the KNN Classifier.



(a) $AE_{\boldsymbol{X}}(\boldsymbol{x_{test}})$ (b) $AE_{\boldsymbol{X\setminus S}}(\boldsymbol{x_{test}})$

Table 3: KNN confusion matrix

4.3 Accuracies

Table 4 shows the accuracy compared across the models, including the specific groups. Once again, the baseline model is the majority class predictor result.

	Accuracy	**Accuracy (F)**	**Accuracy (M)**
Baseline$(\boldsymbol{x_{test}})$ [2]	0.659	-	-
$AE_{\boldsymbol{X}}(\boldsymbol{x_{test}})$	0.728	0.772	0.718
$AE_{\boldsymbol{X\setminus S}}(\boldsymbol{x_{test}})$	0.720	0.781	0.707

Table 4: Autoencoder accuracy

5 Conclusion

In conclusion, the work presented shows the bias mitigation strategy followed - removing sensitive attributes - has had a positive impact on the model predictions' reducing the differences between both gender groups, with the tradeoff of also reducing the model accuracy. From the perspective of the *Independence* and *Separation* fairness criteria it was found that the model excluding sensitive attributes was significantly fairer than the model including them, whereas the sufficiency criteria did not show any clear conclusion. It is important to

notice when looking at these results that the data is naturally biased as there's an over-representation of men, so e.g. when analyzing the calibration curve the standard deviation for females will be larger than for males. In addition, this project is only taking gender fairness into consideration, not *Area Of Origin*. For future work, it would be interesting to see the impact of the bias mitigation strategy applied to this sensitive attribute as well. Moreover, the implementation of a new loss function was considered taking the separation criteria into account, such that the TPR and FPR are expressed in it, optimizing for fairness.

6 Appendix

Figure 6: Recidivism rate

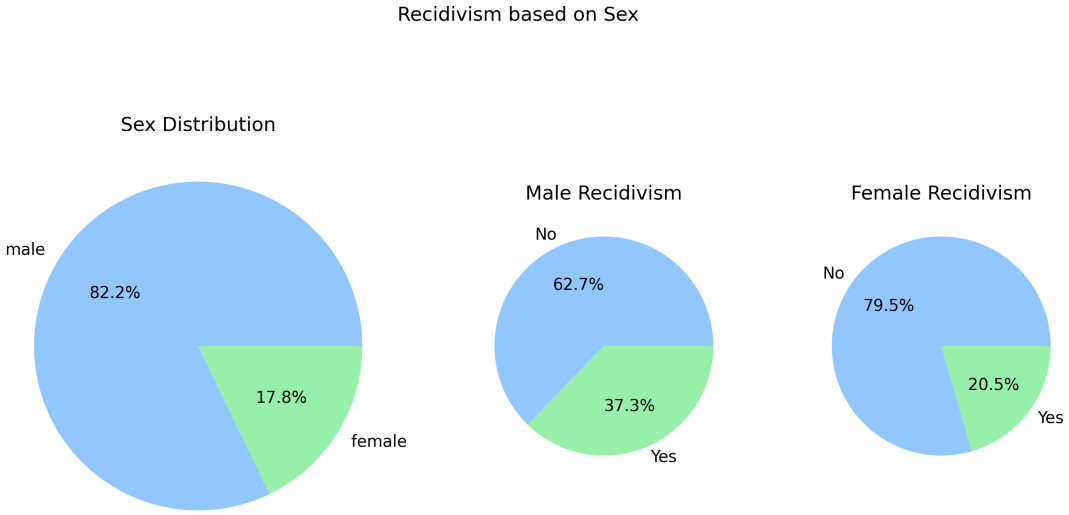


Figure 7: Recidivism based on sex

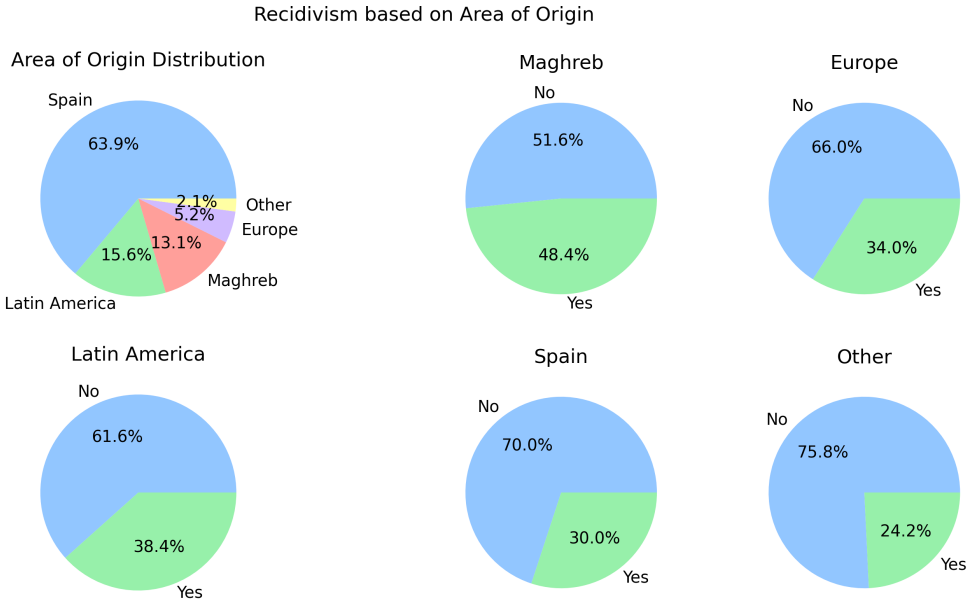


Figure 8: Recidivism based on Area of Origin

V12_n_criminal_record		V13_n_crime_cat		V27_program_duration_cat		V10_date_of_birth_year_grouped		V9_age_at_program_end_grouped	
'0'	0	'1'	1	'<6 months'	0	'(1990, 1993]'	0	'(14, 16]'	0
'1-2'	1	'2'	2	'6 months <1 year'	1	'(1993, 1996]'	1	'(16, 18]'	1
'3-5'	2	'3+'	3	'>1 year'	2	'(1982, 1990]'	2	'(18, 27]'	2
'5+'	3								

Table 5: Conversion of ordinal attributes applied in Data Pre-Processing.

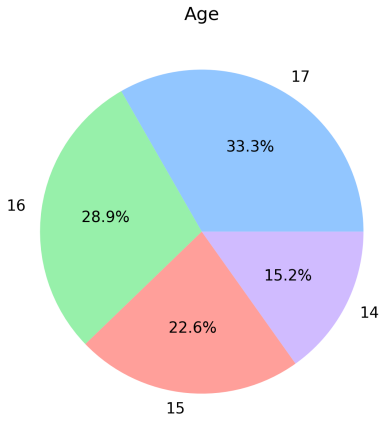


Figure 9: Age distribution across the dataset.

	Sensitive Attributes	Numerical	Ordinal	Categorical
0	V1_sex	V20_n_juvenile_records	V12_n_criminal_record	V6_province
1	V4_area_origin	V28_days_from_crime_to_program	V13_n_crime_cat	V11_criminal_record
2	V8_age	V29_program_duration	V21_n_crime	V15_main_crime_cat
3	V9_age_at_program_end	V22_main_crime_date_year	V27_program_duration_cat	V16_violent_crime
4	V10_date_of_birth_year	V22_main_crime_date_month	None	V17_crime_classification
5	V10_date_of_birth_month	V30_program_start_year	None	V19_committed_crime
6	None	V30_program_start_month	None	V23_territory_of_execution
7	None	V31_program_end_month	None	V24_finished_program
8	None	None	None	V26_finished_measure_grouped

Figure 10: Attributes table separation.