LEPL1109 - Statistics and Data Sciences

HACKATHON 3 - BIAS IN CLUSTERING ALGORITHMS

Group n°18 December 22, 2024

Lastname	Firstname	Noma
Decaluwé	Maxime	50802200
Defrenne	Simon	42242200
Mil-Homens Cavaco	Mathieu	38282200
Peiffer	Thibaut	47352200
Roekens	Raphaël	70732200
Starck	Robin	88952200

Please carefully read the following guidelines:

- Answer in English, with complete sentences and correct grammar. Feel free to use grammar checker tools such as LanguageTools.
- Do not modify questions, and input all answers inside `\begin{answer}...\end{answer}` environments.
- Each question should be followed by an answer.
- You are allowed (and often encouraged) to add figures that support your answers, provided that you explain how they support your claims.
- Clearly cite every source of information (even for pictures!).
- Whenever possible, use the `.pdf` format when you export your images (this usually makes your report look prettier).

Question 1.1: Removing unnecessary features:

Can you already, a priori, detect that some features are useless? If yes, list those (useless) features and explain your choice. If not, then explain why it is better to wait.
Generally speaking, is it a good idea to remove a feature based on *a priori* knowledge, or doesn't it alter the final outcome?

Answer to 1.1: Removing unnecessary features:

We can a priori identify certain features as non-essential. Firstly, irrelevant variables such as "name," "first," "last," "c_case_number," and "r_case_number" do not influence an individual's likelihood to recidivate. Therefore, these can be excluded from the dataset without affecting the model's predictive performance. "c_charge_desc," takes more than 3000 different unique entries. Introducing more than 3000 binary variables is unrealistic. The feature is then also excluded. Additionally, some features may introduce data leakage if retained. Specifically, "decile_score," "is_recid," "is_violent_recid," "decile_score.1," "score_text," "v_decile_score," and "v_score_text" represent predictions of recidivism probabilities generated by external entities or informations that we are supposed to predict. Including these in our model would inadvertently incorporate external predictive information, thereby biasing our results.
Furthermore, variables such as "type_of_assessment" and "v_type_of_assessment," which remain constant across all records, offer no discriminatory value and can be safely removed. Finally, "prior_counts.1" was redundant with "prior_counts" as it is a duplicate. Same went for "age_cat" and "age". This has been checked by observing the correlation matrix.
In general, removing features based on a priori knowledge can enhance model performance by eliminating noise and preventing overfitting. However, it is crucial to ensure that the removal does not discard potentially valuable information.

Question 1.2: Handling missing data:

Given the dataset and the amount / type of missing information, what strategy do you propose to follow regarding missing data (NaNs) ? Justify briefly your choice.

Answer to 1.2: Handling missing data:

When a feature contains a substantial amount of missing data (Null values), our strategy is to remove that feature from the dataset. This decision is based on the fact that features with excessive missing values can hinder the model's ability to learn effectively, leading to decreased performance and reliability. By excluding such features, we ensure that the dataset remains robust and that the model is trained on complete and meaningful information, thereby enhancing its predictive accuracy. When the amount of Null values is low enough, we could fill the missing entries with the mean value of the feature for example. In our case, this situation did not occur as the features with low amount of Null entries have been excluded from the feature selection based on other reasons.

Question 1.3: New features:

What features have you added? If a particular manipulation has been applied, please explain.

Answer to 1.3: New features:

We performed feature engineering to enhance the dataset by creating more informative and relevant variables.
Firstly, we calculated the duration of incarceration (d_time_jail) by subtracting the date of entry from the date of release. Similarly, we computed the duration of custody (d_custody) by calculating the difference between the custody entry and release dates. These duration-based features provide a more meaningful representation of an individual's time in jail and custody compared to the raw dates alone. Secondly, we transformed the screening dates ("compas_screening_date","screening_date","v_screening_date") into the individuals' ages at the time of screening (measured in days) by combining the screening dates with their respective dates of birth ("dob"). This transformation allows us to incorporate age as a continuous variable, offering better insights and improving the predictive performance of our models.
Note however that screening_age, v_screening_age and compass_screening_age turned out to be redudant with the "age" feature with correlation of 1,0.

Question 2.1: (Im)Balanced dataset ?:

Is the dataset imbalanced ? What could be the consequences in terms of fairness i.e. in terms of the model performing equally well across all groups ?

Answer to 2.1: (Im)Balanced dataset ?:

Plotting a pie chart to represent the racial distribution of our dataset reveals a significant imbalance. Approximately 51% of the entries are African-American, while one-third are Caucasian. The remaining races collectively constitute a mere 15%. This disproportionate representation has profound implications for the fairness of our predictive model when incorporating the race feature.
For example, Native Americans account for only 0.2% of our dataset. Such a minimal representation hampers our ability to draw reliable conclusions or make accurate predictions for this group. Consequently, the model may inadvertently perpetuate biases, leading to unfair treatment of underrepresented races. Addressing this imbalance is crucial to ensure that the model performs equitably across all demographic groups.

Question 2.2: Principal Component Analysis:

Do all features have the same importance? If no, which features are less important, and why ? You can use all other graphs from the visualization part to justify your answer.

Answer to 2.2: Principal Component Analysis:

All features don't have the same importance, as we can see with the PCA1, the "days from compas" contains the smallest variance explained. In addition it seems orthogonal to the main direction along which there is variation of the target feature. It it also translated in terms of correlation. "days from compas" feature has the lowest correlation. Same goes for "days b screening arrest". Finally, we notice that the age also seem to be directed in the direction orthogonal to the variation of the target feature. Someone older doesn't seem to have much more chance to recidivate than a younger one.

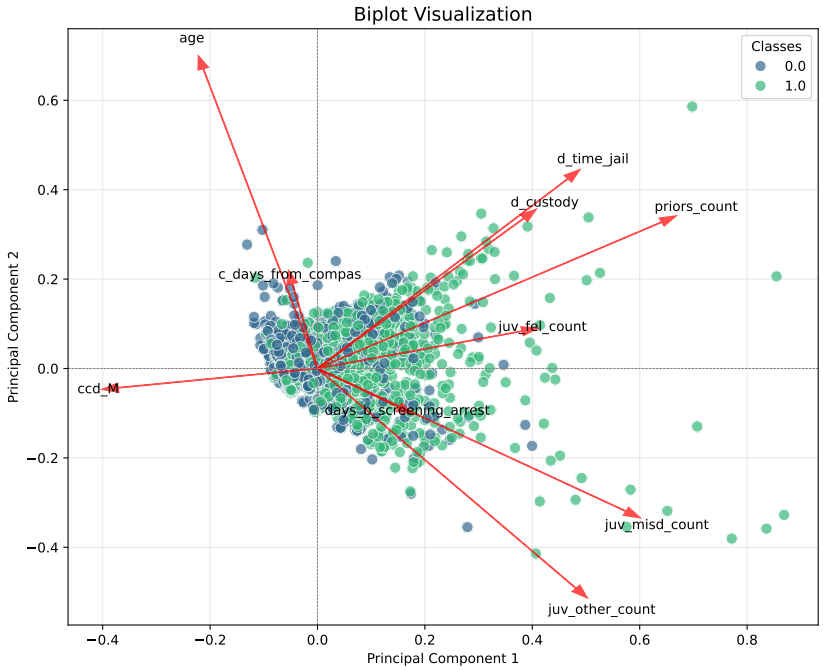

Figure 1: PCA

Question 3.1: Number of clusters:

Accounting for all features, what do you think is the ideal number of clusters ? What will happen if too many or too few clusters are chosen

Answer to 3.1: Number of clusters:

To estimate the ideal number of clusters, we can adopt a qualitative approach by analyzing the results of the PCA. Observing the 2D graph, we notice that the non-recidivists are clustered around a single point, while the recidivists are primarily distributed around two distinct points. Therefore, we believe that 3 clusters can effectively represent the dataset.

Question 3.2: Quality of the clustering:

You considered three different measures for the quality of the clustering: the first one is the silhouette score and is oblivious to the true labels: it is a truly unsupervised metric. The second and third metric use the true label to assess the quality of the clustering. Based on this observation,

1. Comment on the evolution of each metric according to the number of clusters.
2. Comment on what do you now think is the ideal number of clusters.

Answer to 3.2: Quality of the clustering:

1. The mean silhouette score shows a sharp peak of 0.50 at 3 clusters. It then drops back to around 0.20 before rising again to a second maximum of 0.40 at about twenty clusters. Entropy and purity increase and decrease, respectively, in an almost monotonic manner. It is noticeable that the transition from 2 to 3 clusters results in the largest difference in entropy or purity. Each additional increment in the number of clusters has little effect on their values.
2. We believe that the ideal number of clusters is 3. The silhouette score represents this quite well. Although entropy and purity values can still be improved with more clusters, the gains from each additional cluster become significantly smaller after 3.

$$\Delta_{2\to 3\ clust} purity \approx \Delta_{3\to 24\ clust} purity \approx 0.1$$

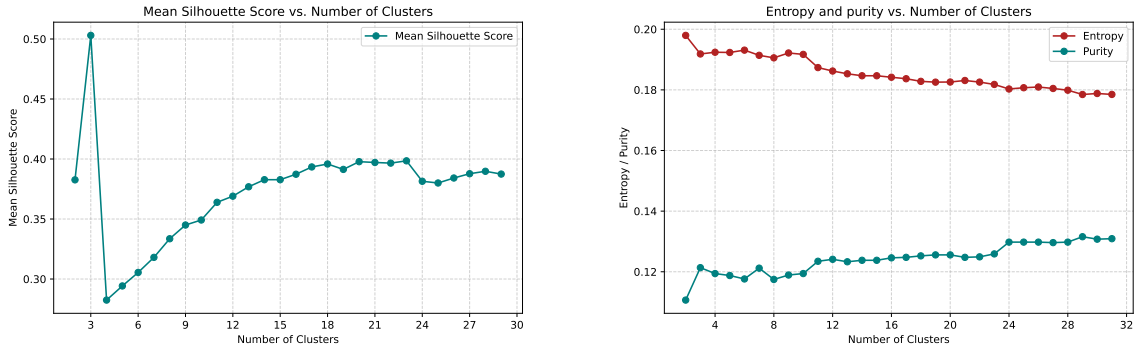

Figure 2: Mean silhouette score and entropy/purity vs. number of clusters

Question 4.1: Fairness of your model:

You considered two different measures for the fairness of your model and checked for various variants of your algorithm (number of clusters) the value of these fairness metrics.
Is your algorithm unfair ? If yes, which ethnic group is penalized by the unfairness of your model ?

Answer to 4.1: Fairness of your model:

Yes, the algorithm is unfair, and male African-Americans are the group penalized by its unfairness. This is evident as the false positive rate (FPR) for male African-Americans is significantly higher than for other groups regardless of the number of clusters, meaning they are more likely to be incorrectly labeled as "high risk", even though the race isn't used in the data set. Additionally, demographic parity metrics show unequal outcomes, with African-Americans consistently being disproportionately impacted.

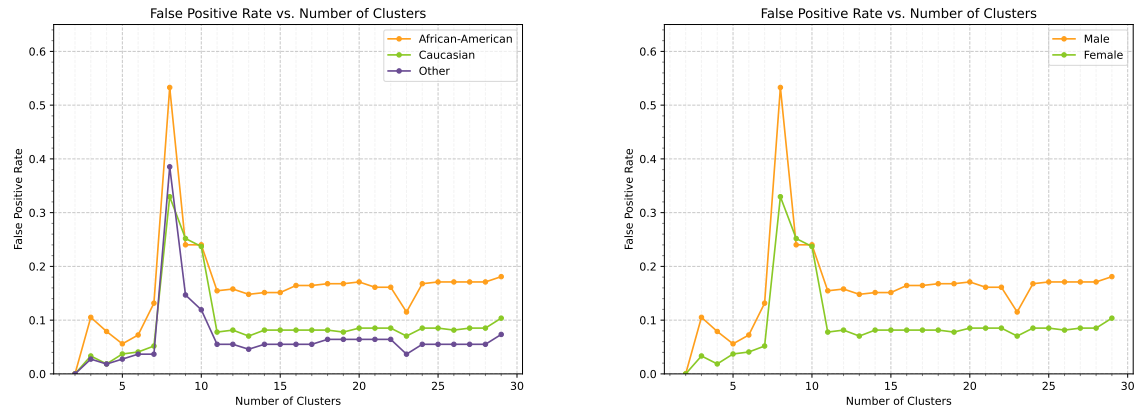

Figure 3: FPR vs. number of clusters

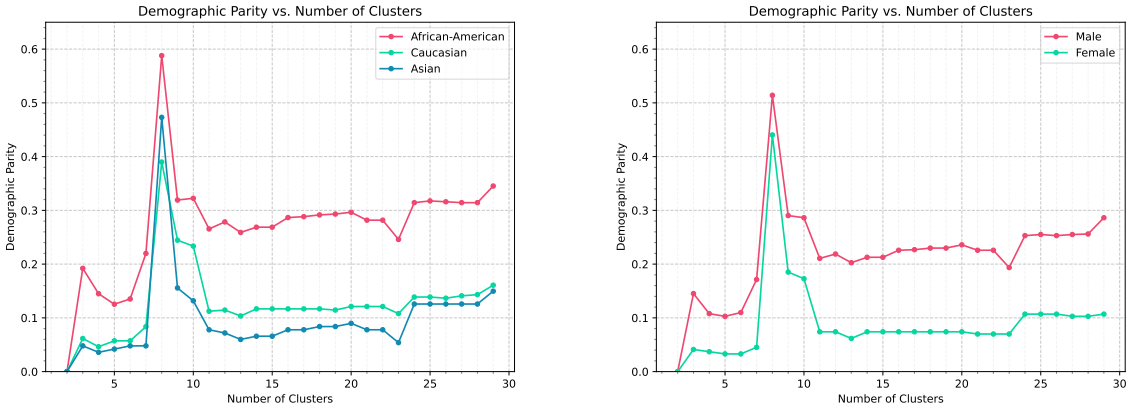

Figure 4: Demographic Parity vs. number of clusters

Question 4.2: Presence of the sensitive features in the dataset [BONUS]:

In Cell 1.5, you removed the sensitive features from your dataset before building your algorithm. Yet, you may have noticed unfairness in your algorithm.

1. Provide reasons why it is not necessarily enough to remove sensitive features from your dataset if you want to have fair predictions.
2. Compute FPR and Demographic Parity for your algorithm when trained on the full dataset. Is the fairness of your classifier worse ?

Answer to 4.2: Presence of the sensitive features in the dataset [BONUS]:

1. A reason we might find to explain why removing sensitive features is not always enough to remove unfairness is that the dataset might have bias introduced through other features. The link between the sensitive feature and the presumed "innocent" features might not be straighforward or be visible through linear correlation but still contain this unfairness. For example, if there are society's inherent injustices resulting in some people from a subgroup to be treated more harshly (in a way similar to more dangerous criminals), they might present caracteristics that look similar to recidivist people and then be categorized by k-means that way (for example the time spent in jail).

2. According to the plot we get when training on the whole dataset, the fairness does not change much. The gap between african-american people and caucasian or others remains about $\Delta \approx 0.1$ for *fpr* rates and about $\Delta \approx 0.15 - 0.20$ for *Demographic parity*.

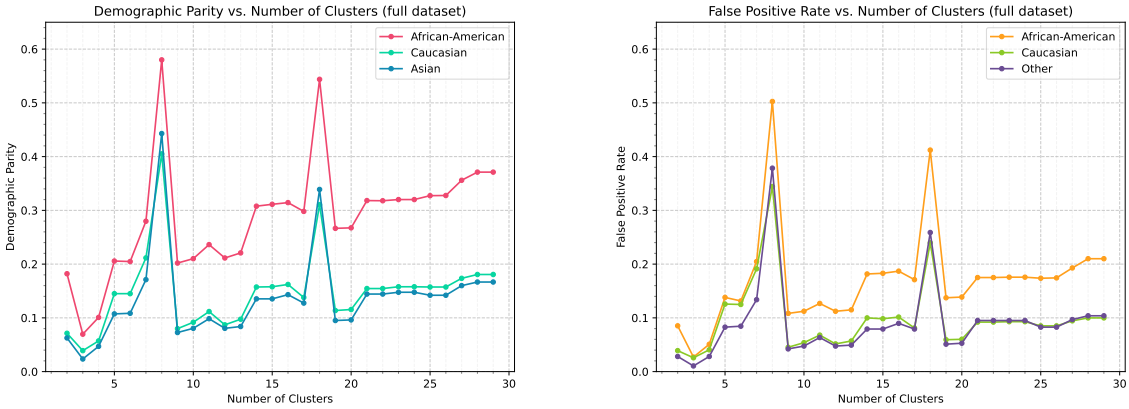

Figure 5: FPR and Demographic Parity for model trained on the full dataset

Question 5.1: Visualization:

Produce a clear, clean figure expressing a result or giving an overall vision of your work for this hackathon. Please feel free to do as you wish. Be original! The clarity, content and description of your figure will be evaluated.

Answer to 5.1: Visualization:

In order to explain the bias within our model, it is useful to investigate whether some of the features are inherently biased and may affect the model's fairness across different demographic groups. The following violin plots showcase features and their distributions across these groups, alongside the distribution of the same feature for the recidivist group. The recidivist group's population has been balanced to ensure that no single demographic group disproportionately influences its overall shape. By doing so, we aim to identify whether the feature's distribution within the recidivist group closely resembles that of any specific community, potentially indicating bias in how this feature correlates with recidivism predictions.
The first two features ('priors_count' and 'd_time_jail') indicate there may be inherent bias within the different demographics, as the recidivist distribution closely resembles the distribution of a particular gender and race (resp. Male and African-American). This explains why our model predicts a bigger share of these groups as recidivist.

Alternatively, when comparing the amount of days each person was kept in custody, we can't conclusively say any demographic is unfairly affected, as while the distributions across groups are not entirely uniform, none resembles the target distribution more than another.

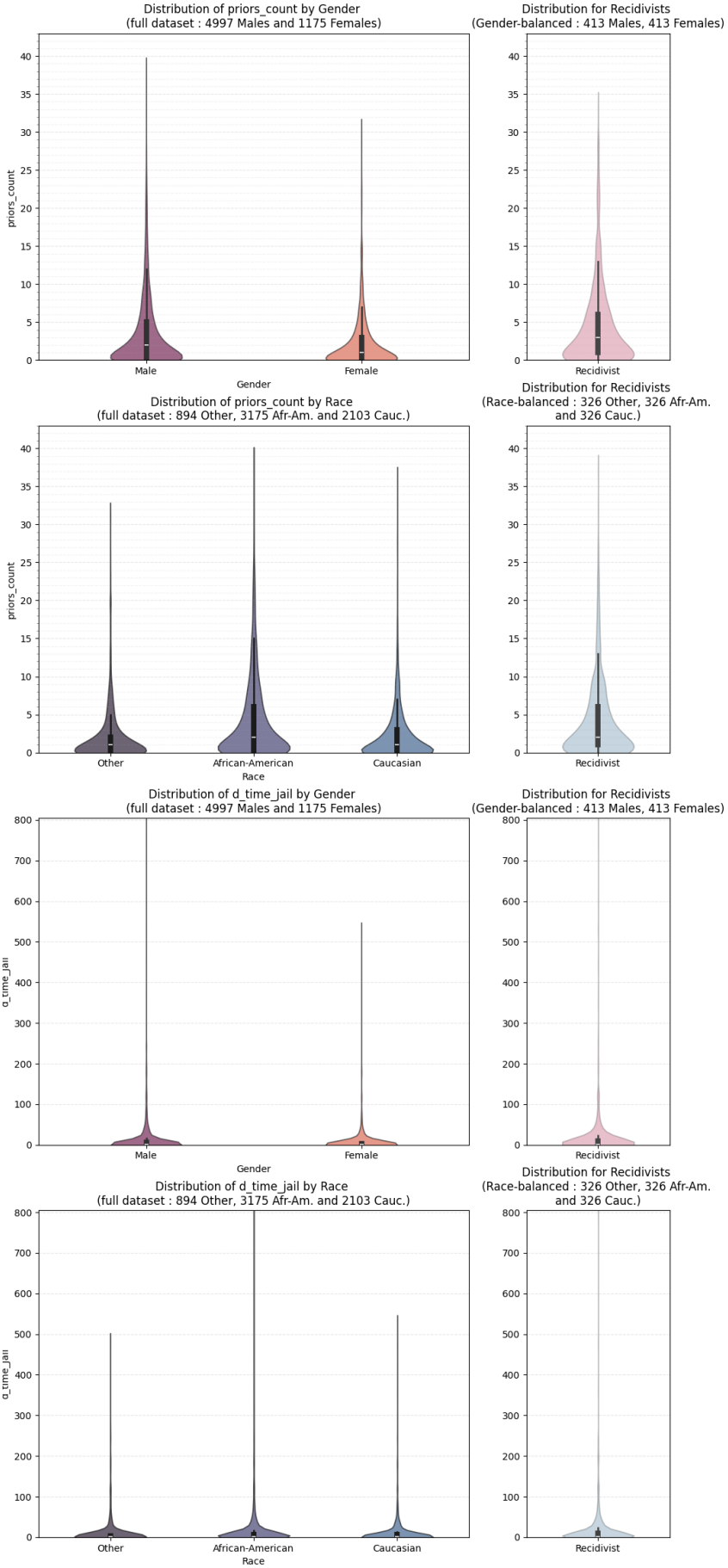

Figure 6: Biased feature distribution across different groups

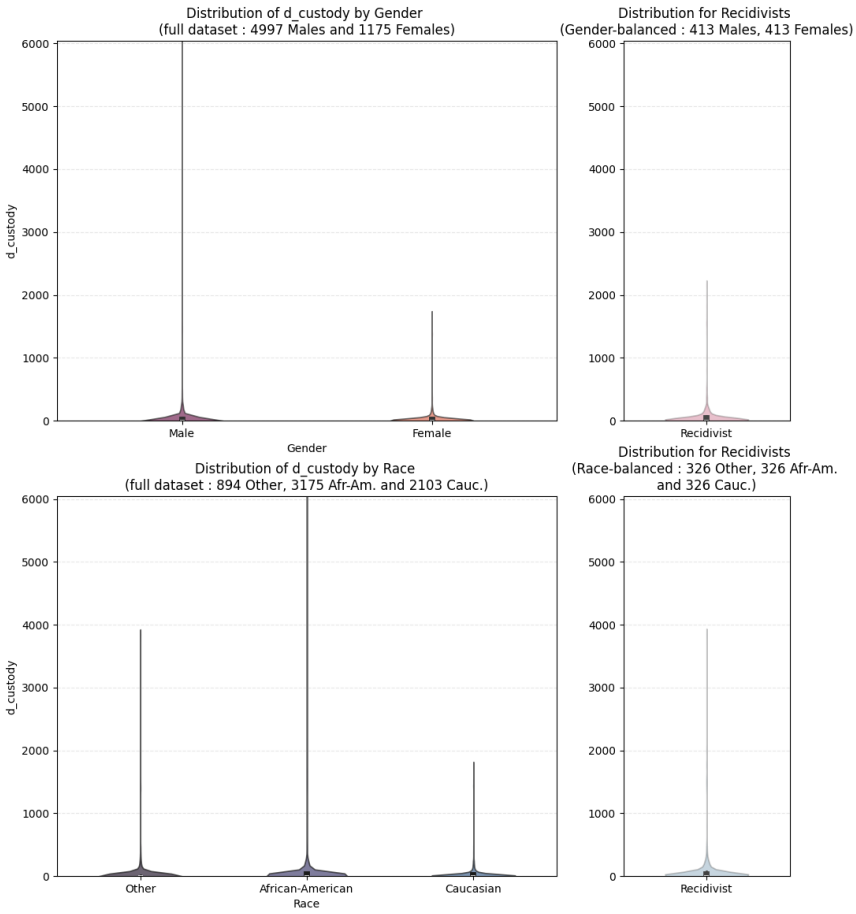

Figure 7: Unbiased feature distribution across different groups