Juvenile Justice: Examining Bias in Facility Security Levels and Resource Availability

Nandini Nag and Jane Andrews
Thomas Lord Department of Computer Science, University of Southern California, Los Angeles, California 90089, USA

(Dated: May 8, 2025)

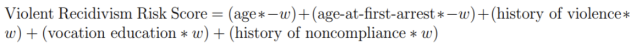
Using data and new technology in the criminal justice system is not a novel idea, as seen by Northepointe's Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) model. While much research has been done on the validity and accuracy of this recidivism model, less attention has been given to how early interactions with the juvenile justice system influence these outcomes. In this study, we inspect whether racial biases exist in juvenile correctional facilities by analyzing whether facilities with majority Black populations are subject to higher security levels as opposed to their White counterparts. Utilizing a variety of fairness tools – exploratory data analysis, counterfactual predictive analysis, NLP analysis, fairness evaluation metrics, etc. – we continue to broaden the discussion on systemic bias in the criminal justice system.

Keywords: Machine Learning, Natural Language Processing, Fairness, Bias, Juvenile Justice, COMPAS, Counterfactual Fairness

I. INTRODUCTION

According to the U.S. Department of Justice, many justice agencies across the country have "adopted data-driven decision making to supervise, manage, and treat justice-involved populations" (Bureau of Justice Assistance, 2025). In many cases, these decision-making processes, which include risk assessments and AI predictive policing, are utilized to assess an individual's risk of re-offending. But do these tools paint an unbiased picture or are their algorithms inherently flawed? How do other parts of the criminal justice system interplay with the COMPAS algorithm, and how do these interactions impact offenders' lives? After finding a stark racial disparity between youth offenders in the Broward County data set (Appendix A, Section 2), we became curious about potential racial bias in the juvenile criminal justice system, especially since juvenile incarceration may impact one's COMPAS scores for the rest of their life.

Northpointe's COMPAS algorithm is proprietary, so details of how scores are calculated with COMPAS are not public. However, Northpointe has disclosed the basic equation of their algorithm used to calculate their Violent Recidivism Risk Score (Northpointe, 2015):



$$\text{Violent Recidivism Risk Score} = (\text{age}*-w)+(\text{age-at-first-arrest}*-w)+(\text{history of violence}*w)+(\text{vocation education}*w)+(\text{history of noncompliance}*w)$$

FIG. 1. Northepointe's COMPAS algorithm formula.

The negative weights of the first two elements mean that the younger offenders currently are and were at their first arrest, the higher their scores will be. The final three elements all have positive weights, which indicates the more extensive an offender's history of violence, noncompliance, and "educational and vocational problems" is, the higher their violent recidivism risk score will be.

Our study aims to investigate racial bias in juvenile detention facilities by analyzing whether facilities with a majority Black juvenile population are subject to higher security levels compared to their White counterparts. We will pay special attention to potential causal links between the racial makeup of a facility and whether it has features that would influence whether an offender recidivates. We consider security level, as studies show that higher security levels are correlated with higher rates of recidivism, regardless of inmate history (Peck, 2016 and Gaes, 2009). Information about these potential links will provide valuable insight into whether biases exist at levels of the criminal justice system that may have lifelong impacts on the minors that go through the system. Another gap in the literature we identified and intend to fill is a more technical analysis of the questions on the COMPAS questionnaire, which will be done using NLP-based tools.

II. RELATED WORK

Since the release of ProPublica's seminal work, "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks.", debate around and investigation into the potential for bias in algorithmic risk assessment has flourished. The ProPublica study posited that Northpointe's COMPAS tool produced uneven error rates between Black and White criminal offenders, by showing a tendency to incorrectly label Black offenders as a high recidivism risk and White offenders as a low recidivism risk (Angwin et al., 2016). Northpointe, the creator of the proprietary COMPAS software, hit back soon after with their own analysis, arguing that their algorithm was not biased, due to the fact that it was well calibrated and predicted recidivism risk equally accurately for White and Black offenders (Dieterich, 2016). The publication of the initial study and Northpointe's response ignited a fierce debate among academics and professionals in the fields of machine learning, criminology, and law.

III. METHODS

To examine bias in juvenile facility security arrangements, we will employ a combination of fairness-aware machine learning algorithms and natural language processing techniques. Our approach builds on predictive models, while also incorporating more advanced techniques, ensuring a more rigorous evaluation of bias mitigation strategies.

A. Data

The main dataset used in our analysis is the publicly available Census of Public and Private Juvenile Detention, Correctional, and Shelter Facilities, 1986-1987: [United States]. These data include the population, demographics, and characteristics of every facility holding youths in custody in the United States in early 1987. Each row in the dataset represents one facility. (For more details, see Appendix A, Section 1.) This dataset comes with a codebook [19] listing all variable names and their meanings, please follow along to understand all variable names used from now on.

The vintage of this dataset is appropriate for our purposes, as youth offenders in custody in 1987 would currently be middle-aged adults, and still potentially interacting with the criminal justice system.

B. Exploratory Data Analysis

To begin our research, we first performed simple Exploratory Data Analysis on the data. We found that Black male youths were overrepresented in juvenile delinquent custody facilities (making up 40.6% of the population of all facilities, despite African-Americans only making up approximately 12% of the United States population at the time) (U.S. Department of the Census, 1987). Figures 2 and 3 similarly show potential bias, in this case, how facilities where Black male youth residents outnumbered while male youth residents skew higher-security than facilities where the opposite is true.

C. Data Pre-Processing

After sourcing our juvenile detention facilities data, we created a pandas DataFrame and filtered on numeric columns to use in our analysis. Additionally, we created our sensitive attribute, majority race, by comparing the non-Hispanic White and Black proportions of population in each facility and assigning the majority group as the label. This variable was encoded binarily, with 0 being White-majority and 1 being Black-majority.

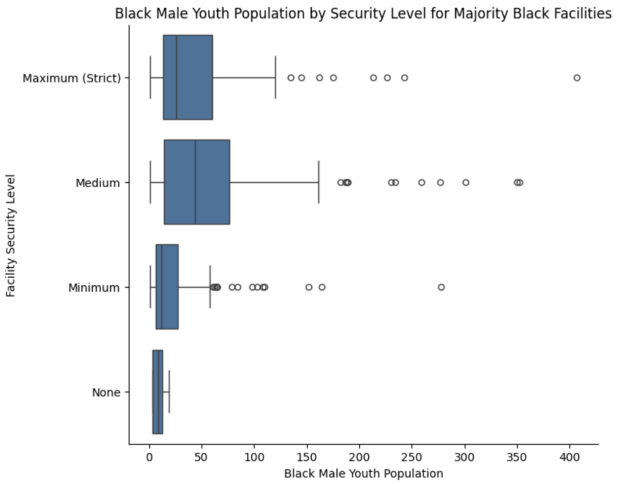


FIG. 2. A categorical boxplot showing the Black male youth population distribution in facilities where Black male youths outnumber White male youths, by security level.

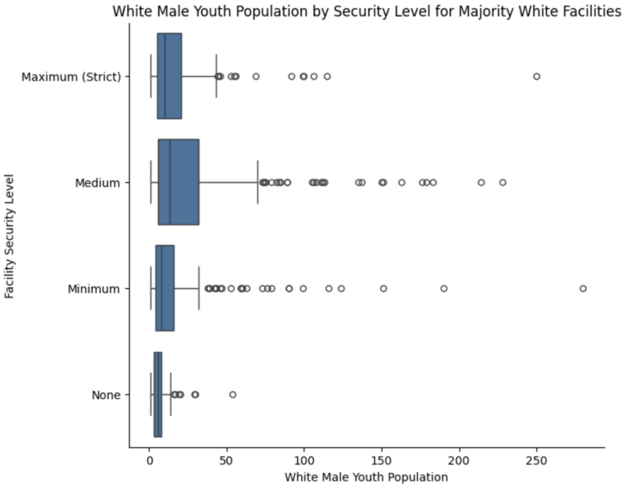


FIG. 3. A categorical boxplot showing the White juvenile male population distribution in facilities where White male youths outnumber Black male youths, by security level.

D. Correlation Analysis

Our dataset contained over 600 variables, which presented a challenge in selecting the most relevant features for our analysis. In order to combat this, we calculated Pearson correlation coefficients between each feature and the target variable, the categorical variable V40 (security arrangements). We decided to only select those with a score above |0.20|, which resulted in 25 features, which you can see in the Results section. On these 25 features, we ran two-component Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) in order to assess clustering and class separation colored by both security level and majority race.

E. Training and Counterfactual Fairness

We split our full dataset with a 70/30 train test split and stratified on V40, our target. We then employed two different models: 1) Random Forest Classifier (RF): An ensemble learning method that builds multiple decision trees and averages their predictions to reduce overfitting and improve generalization. 2) Ordinal Logistic Regression: a logistic regression model used to predict ordinal, or ordered, variables.

To proceed with counterfactual fairness, we defined a function to flip the majority race variable, while keeping all other attributes unchanged. We then predicted on both the original and counterfactual inputs, computing counterfactual accuracy and prediction difference, the latter which measures the proportion of instances where the model's predictions differ between the original and counterfactual datasets. A significant difference would suggest that the model is sensitive to the majority race, potentially indicating bias (Huang et al., 2022).

We also performed a focused analysis looking specifically at security levels 1 (maximum) versus 3 (minimum), as well as a model with upweighted Black-majority facilities. Details of these methods and their results can be found in Appendix B, Sections 2 and 3.

F. Interaction Features

To further augment our analysis, we created feature x race interaction terms for each of the 25 selected features. Utilizing only an Ordinal Logistic Regression model, we fit this model on the entire dataset, including our new interaction features. Once again, we utilized the counterfactual framework on this enhanced model. We extracted the top 10 interaction terms and plotted the feature value against the predicted probability of Maximum Security by racial group for the three most meaningful features. These visuals and their interpretations are in the Results section.

G. BERT NLP Analysis

The last step in our methodology is an NLP analysis of survey language bias. We preprocessed the text of a sample risk assessment survey uploaded by ProPublica in their COMPAS study (Bowers, 2018). After tokenization, lowercasing, binarizing, and stop word removal, we used a Sentence-Transformer (BERT) model to embed each survey question and statement. We then performed K-means clustering to group similar questions together and inspected the clusters for racial connotations. We also performed Zero-shot classification to assign "bias" or "neutral" labels and compute per question bias scores.

IV. RESULTS

Our comprehensive methodology produced several key insights. In the following section, we summarize our findings and their possible implications.

After performing our initial Pearson correlation analysis, we found that majority race illustrates a negligible correlation with our target, V40 (security arrangements); it was not even in the top 25 features. Despite this, we still selected features with a correlation coefficient above |0.20|, which resulted in the 25 features seen in Figure 4.

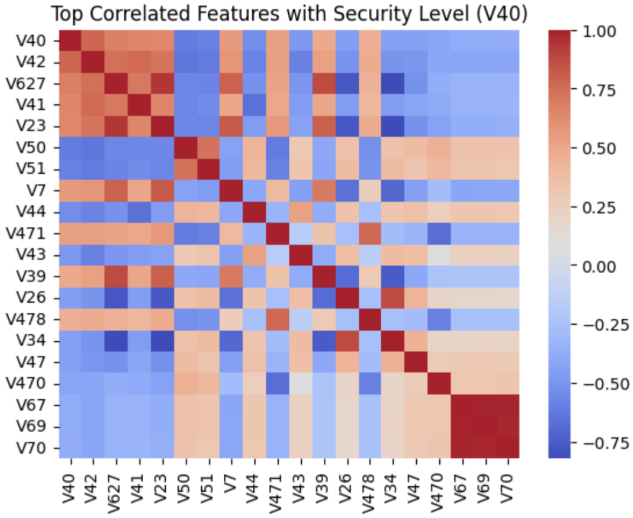


FIG. 4. Correlation heat map showing top 25 features.

On these features, we ran PCA and LDA colored by security level as well as majority race. Figure 5 shows the result of running PCA colored by security level. The rest of the three graphs can be found in Appendix B, Section 1. As one can see in PCA, there is partial separation between Maximum (1) and Minimum (3) security levels on PC1, but no clustering by race. For the LDA results, there are clearer groupings for security levels 1 versus 3, but still no separation by majority race.

The results of our correlation analysis suggest that our sensitive attribute (majority race) is not correlated with security levels, indicating no bias. However, we still wanted to complete our counterfactual analysis to fulfill our research pipeline, which is described next.

After performing counterfactual fairness on the full dataset – all four security levels – we found the results in Table 1. A prediction difference of 0 tells us that flipping majority race has no effect.

Model	Accuracy	CF Accuracy	Pred. Diff.
RF	0.708709	0.708709	0.0
Ordinal Reg	0.117117	0.117117	0.0

TABLE I. Results of CF analysis on full dataset.

However, after adding the feature x race interaction

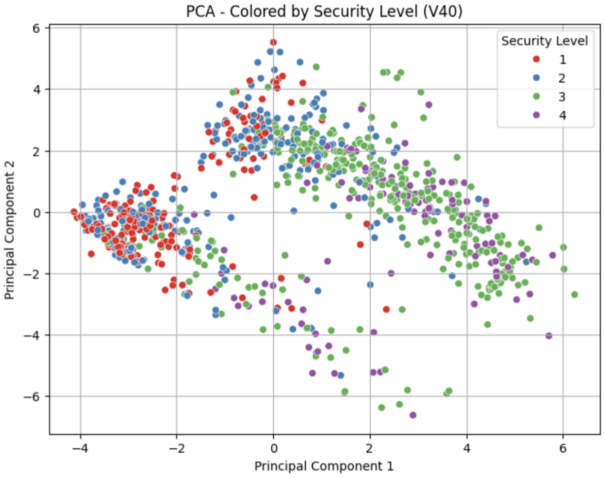


FIG. 5. PCA results by security level.

terms, our Ordinal Logistic Regression model yields a prediction difference of 0.159, meaning that 15.9% of our test set predictions change when we counterfactually switched the facility's majority race (Table 3). While the accuracy is not the highest, this still suggests that our model is now sensitive to race and interaction-driven bias may exist.

Model	Accuracy	CF Accuracy	Pred. Diff.
Ordinal Reg w Interactions	0.582583	0.576577	0.159159

TABLE II. Results of CF analysis on data with interaction terms.

The following are the top 10 features with influence on our model:

Variable	Coefficient
1/2	2.708309
V627	2.207219
V42	1.489815
3/4	1.439601
V23	1.230478
2/3	1.140132
V23 x race	0.973708
V35 x race	0.795748
V7 x race	0.708801
V39 x race	0.673286

TABLE III. Top 10 features.

Only looking at the interaction variables, V7 x race appears among the top ranked, however, we chose not to further analyze this variable because it represents Agency ID, which is merely an administrative code and does not have any meaning for security levels. In contrast, V23 (facility type) × race, V35 (reason for custody is commitment) × race, and V39 (reason for custody) × race represent concrete facility attributes, thus making them more relevant to our research. We plotted probability curves by race for these three interaction terms and explain the results below.

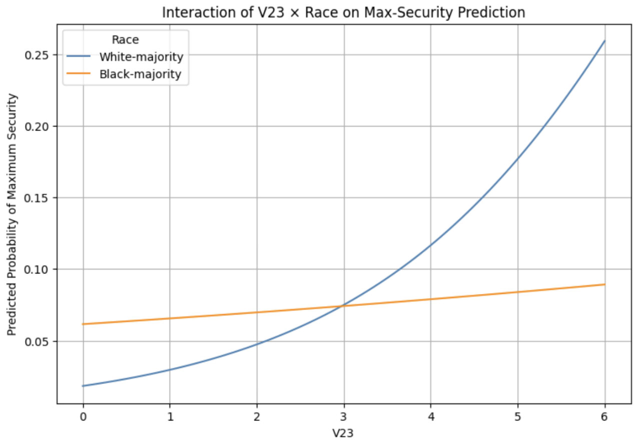


FIG. 6. Probability curve of V23 x race.

V23 corresponds to facility type. As seen in the codebook [19], the values for this variable are 0 (Detention Center), 1 (Shelter), 2 (Reception or diagnostic center), 3 (Training school), 5 (Ranch, forestry camp, or farm), 6 (Halfway house or group home). There is no value for 4. In Figure 6, we see two interesting details; the first is that for detention centers, shelters, and diagnostic centers, Black-majority facilities will have a higher probability of being maximum security than White-majority facilities. Second, we found it troubling that White-majority facilities of types 5 and 6 had much higher probabilities of maximum security than their Black counterparts. This may open conversations on the racial demographics of those facility types and who gets to go to those in the first place.

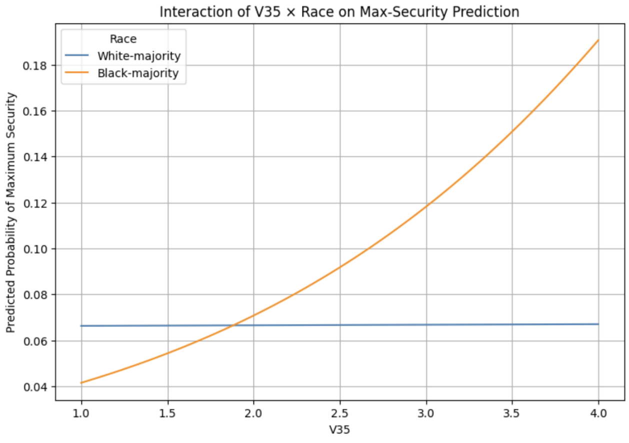


FIG. 7. Probability curve of V35 x race.

V35 is a variable representing the reason for custody is commitment/placement for treatment, informing us for what reason the facility usually holds juveniles. Since only a value of 3 represents commitment/placement for treatment being the reason for custody, we only look at this value on the x-axis in Figure 7. It is evident that when V35 is 3, Black-majority facilities have a higher probability of maximum security than White-majority facilities.

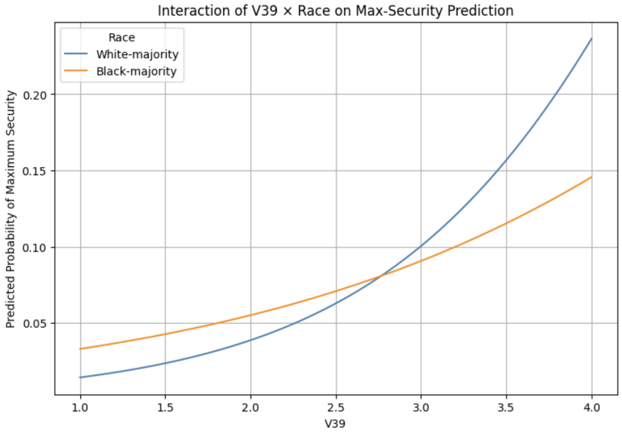


FIG. 8. Probability curve of V39 x race.

Lastly, V39 represents reason for custody. The possible values for this variable are 1 (Diagnosis and/or classification), 2 (Detention pending adjudication, commitment), 3 (Commitment/placement for treatment), 4 (Probation or aftercare), or 5 (Voluntary admission). Figure 8 reveals that for diagnostic and detention cases, Black-majority facilities have a higher probability of being maximum security. This is interesting, as these two stages in the juvenile carceral process involve the initial assessment of a juvenile offender, determining appropriate interventions or treatments, and holding an individual in a secure facility.

We finally supplemented our methodology with a BERT NLP analysis. After performing Zero-Shot Classification, we identified several survey items with a bias score greater than 0.60, predominantly disadvantaging and weaponizing terms that could be used as proxy for more overt race-related questions, as seen in Table 4. Results of our K-Means clustering are in Appendix B, Section 4.

Question/Statement	Bias Score
Is it easy to get drugs in your neighborhood?	0.611785
How hard is it for you to find a job ABOVE min. wage compared to others?	0.605769
Are there gangs in your neighborhood?	0.678274
How many of your friends/acquaintances are gang members?	0.823683

TABLE IV. Examples of language with bias higher than 0.60.

V. CONCLUSIONS

As predictive algorithms are increasingly used in decision making, investigating the potential bias baked into these tools is only becoming more salient. This is especially true for juvenile offenders, as their early criminal history and experiences while incarcerated as minors have the power to influence how algorithms used within the criminal justice system will predict the level of risk they pose to society.

Juvenile defenders' success in rehabilitation is influenced in part by their treatment while in correctional facilities, with research indicating that holding other factors constant, youths held in higher security level facilities have higher rates of recidivism. Our results on the connection between majority race and facility security level were mixed; while statistical analysis such as PCA, LDA, and Pearson Correlation Analysis did not support our hypothesis that race had significant explanatory power over facility security levels, looking at variable interaction revealed instances where facilities' Black majorities are correlated with heightened security levels. These results, combined with the racial bias identified in the COMPAS questionnaire by BERT NLP analysis, suggest worrying opportunities for racial bias to impact offenders' experiences in the justice system.

Limitations of our research include the lack of intersectional bias analysis, the relative simplicity our BERT model, and the short time scale that our data span. The scope of this paper is relatively narrow; only Black and White male youth offenders are included in our analysis, and the data on these youths represents a singular snapshot in time. Researchers interested in furthering this study may want to complete similar analysis over a longer time period and include more genders and races in their analysis, which will allow their results to be more generalizable, complete, and relevant to a wider group of offenders.

Additionaly, while our results did show some bias, with more time, we recommend researchers explore a variety of fairness-aware algorithms aside from counterfactual analysis, as we believe these may provide more concrete and accurate results.

DATA AVAILABILITY

Data is available at `https://github.com/nnag00/DSCI-531-Final-Project/tree/main`

CODE AVAILABILITY

Code is available at `https://github.com/nnag00/DSCI-531-Final-Project/tree/main`.

ACKNOWLEDGMENTS

We would like to thank our peer reviewers and TAs, whose valuable comments shaped our final paper.

Appendix A: COMPAS EDA

The impetus of this project was a troubling finding from our exploratory analysis of the COMPAS data mentioned briefly earlier in the paper. This first appendix provides detail on this EDA and the relevant findings.

1. Data

We performed some exploratory analysis on the publicly available compas.db database published by ProPublica along with their initial COMPAS analysis, which includes detailed information on individual offenders, as well as their COMPAS scores. (For more details, see Appendix A.) These data include COMPAS scores, (sourced from the Broward County Sheriff's Office through a public records request by ProPublica), and criminal history and incarcerate records of those receiving COMPAS scores, (collected from the publicly accessible Broward County Clerk's Office and Florida Department of Corrections website (ProPublica, 2025)

2. Exploratory Analysis Results

The following histograms illustrate the disproportionate racial makeup of youth offenders, i.e. 84% of youth offenders in the dataset are Black while Broward County is only about 8-10% Black.

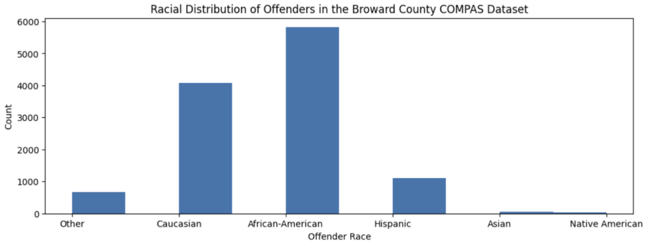


FIG. 9. A histogram showing the racial demographics of offenders in the Broward County COMPAS dataset.

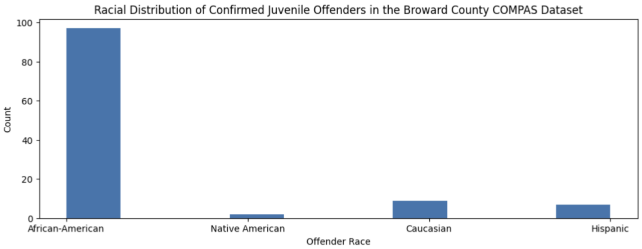


FIG. 10. A histogram showing the racial demographics of confirmed youth offenders in the Broward County COMPAS dataset.

Appendix B: Supplemental Analysis and Results

This second appendix details some of our methods and results that were not the most informative, but we still wanted to share to illuminate our research pipeline.

1. PCA and LDA

The following figures are the remaining results of our PCA and LDA graph analysis.

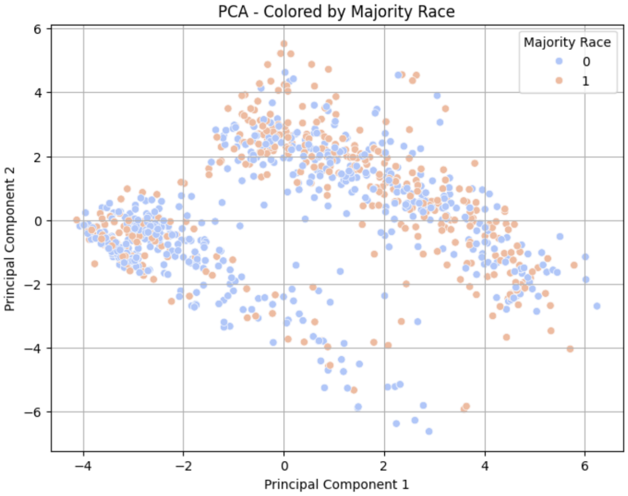


FIG. 11. PCA results by majority race.

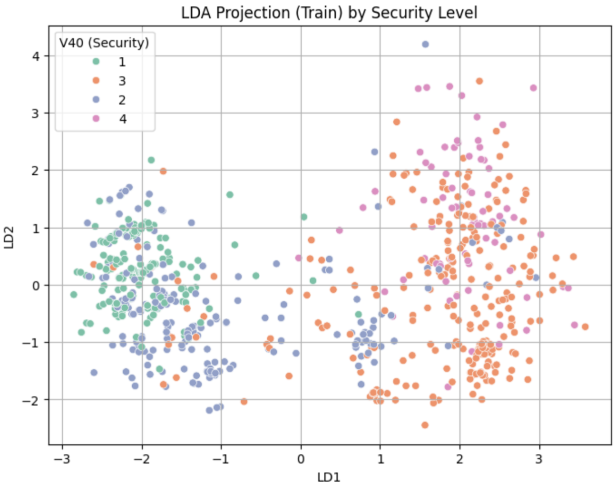


FIG. 12. LDA results by security level.

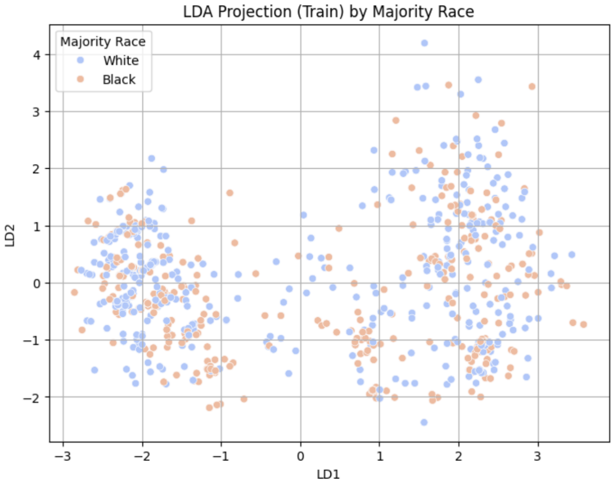


FIG. 13. LDA results by majority race.

2. Focused Analysis: Security Level 1 vs. 3

Based on our correlation analysis, specifically our PCA and LDA exploration, we decided to create a binary subset of our data, where V40 was either 1 (Maximum security) or 3 (Minimum security) and re-coded these as 0/1. We repeated our previous counterfactual step on this subset and prediction differences were recomputed to assess whether a focused analysis would have different results.

In this focused analysis, the prediction difference was once again 0 as seen in Table 5. This shows again that flipping majority race has no effect, even when looking at data that have more separation and clustering.

Model	Accuracy	CF Accuracy	Pred. Diff.
RF	0.971	0.971	0.0
Ordinal Reg	0.947	0.947	0.0

TABLE V. Results of CF analysis on subset.

3. Upweighting Black-majority Facilities

In order to test whether majority race was being ignored by our models or truly not significant, we calculated weights that were inversely proportional to racial group frequencies, thereby giving more weight to the underrepresented group, which in this case is the Black-majority facilities. Both RF and Logistic Regression were re-trained with these weights and evaluated again using the same counterfactual framework. However, even when up-weighting Black-majority facilities, no predictions flip in RF or Log Reg (Table 6).

Model	Accuracy	CF Accuracy	Pred. Diff.
RF Weighted	0.967	0.967	0.0
Ordinal Reg Weighted	0.947	0.947	0.0

TABLE VI. Results of CF analysis on weighted data.

4. BERT Analysis

Based off the graph in Figure 14, the clustering reveals around five different cluster groups, proving that our UMAP cloud can be partitioned into five different clusters or groupings of semantically similar sentences.

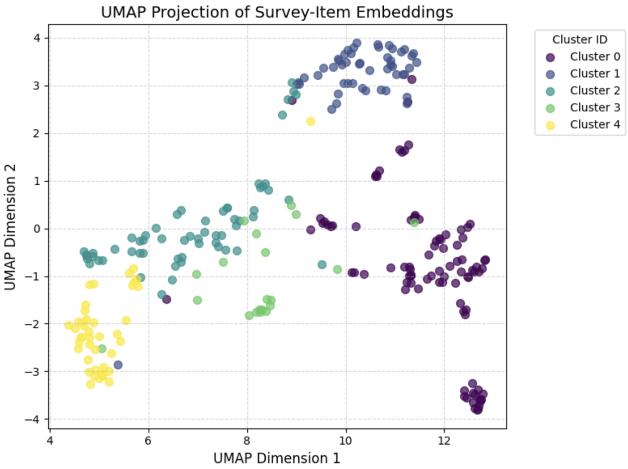


FIG. 14. K-Means clustering results.

[1] Angwin, J., Larson, J., Kirchner, L., and Mattu, S. (2016). Machine bias. Retrieved from `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

[2] Bao, M., Zhou, A., Zottola, S., Brubach, B., Desmarais, S., Horowitz, A., Lum, K., and Venkatasubramanian, S. (2021). It's compaslicated: The messy relationship between RAI datasets and algorithmic fairness benchmarks. *CoRR*, abs/2106.05498.

[3] Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44.

[4] Blomberg, T., Bales, W., Mann, K., Meldrum, R., and Nedelec, J. (2010). Validation of the COMPAS risk assessment classification instrument. *Technical Report.*

[5] Bureau of Justice Assistance (2025). What is risk assessment: PSRAC. Available at `https://bja.ojp.gov/program/psrac/basics/what-is-risk-assessment`.

[6] Bureau of Justice Statistics (2025). Recidivism and reentry. Available at `https://bjs.ojp.gov/topics/recidivism-and-reentry`.

[7] Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

[8] Carbo, B. and Patricia, M. (2021). *Machine Learning Applications in the United States Criminal Justice System: A Critical Content Analysis of the COMPAS Recidivism Risk Assessment.* Thesis, Charles University, Faculty of Social Sciences, Prague, CZ.

[9] Caton, S. and Haas, C. (2024). Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(7).

[10] Chen, M. and Shapiro, J. (2007). Do harsher prison conditions reduce recidivism? a discontinuity-based approach. *American Law and Economics Review*, 9:1–29.

[11] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv.*

[12] Chukwudi, W. (2024). Mitigating intersectional bias in machine learning a novel approach to fairness in automated decision-making.

[13] Dieterich, W., Mendoza, C., and Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 7(4):1–36.

[14] Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4:eaao5580.

[15] Engel, C., Linhardt, L., and Schubert, M. (2024). Code is law: How compas affects the way the judiciary handles the risk of recidivism. *Artificial Intelligence and Law.*

[16] Florida-Health (2024). Population dashboard. Retrieved from `https://www.flhealthcharts.gov/ChartsReports/rdPage.aspx?rdReport=PopAtlas.PopulationAtlasDASHBOARD`.

[17] Gaes, G. and Camp, S. (2009). Unintended consequences: experimental evidence for the criminogenic effect of prison security level placement on post-release recidivism. *Journal of Experimental Criminology*, 5:139–162.

[18] Huang, J., Galal, G., Etemadi, M., and Vaidyanathan, M. (2022). Evaluation and mitigation of racial bias in clinical machine learning models: Scoping review. *JMIR Medical Informatics*, 10(5):e36388.

[19] Inter-University Consortium For Political and Social Research (2025). ICPSR 8973: Census of Public and Private Juvenile Detention, Correctional,and Shelter Facilities, 1986-1987: [United States] Codebook. Accessed: April 2, 2025. Available at: `https://drive.google.com/file/d/1lHMqqTbZzHZRKAiaIIQ2mdHLCIT5TxVZ/view?usp=sharing`.

[20] Islam, M. S. and Zhang, L. (2024). A review on bert: Language understanding for different types of nlp task. *Preprints.*

[21] Johndrow, J. E. and Lum, K. (2019). An algorithm for removing sensitive information: Application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1):189 – 220.

[22] Lagioia, F., Rovatti, R., and Sartor, G. (2023). Algorithmic fairness through group parities? the case of compassapmoc. *AI & Society*, 38:459–478.

[23] Larson, J., Angwin, J., Kirchner, L., and Mattu, S. (2016). How we analyzed the COMPAS recidivism algorithm. *ProPublica.*

[24] May, C., Wang, A., Bordia, S., Bowman, S. R., and Rudinger, R. (2019). On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

[25] Northpointe (2015). Practicioner's guide to compas score. Retrieved from `https://s3.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf`.

[26] Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Peixoto, R. M., Guimarães, G. A. S., Cruz, G. O. R., Araujo, M. M., Santos, L. L., Cruz, M. A. S., Oliveira, E. L. S., Winkler, I., and Nascimento, E. G. S. (2023a). Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, 7(1).

[27] Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Peixoto, R. M., Guimarães, G. A. S., Cruz, G. O. R., Araujo, M. M., Santos, L. L., Cruz, M. A. S., Oliveira, E. L. S., Winkler, I., and Nascimento, E. G. S. (2023b). Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, 7(1):15.

[28] Peck, J. and Jennings, W. (2016). A critical examination of being black in the juvenile justice system. *Law and Human Behavior*, 40:219–232.

[29] ProPublica (2025). Data store archive. Retrieved from `https://projects.propublica.org/datastore/#compas-recidivism-risk-score-data-and-analysis`.

[30] Rudin, C., Wang, C., and Coker, B. (2020). The Age of Secrecy and Unfairness in Recidivism Prediction. *Harvard Data Science Review*, 2(1). https://hdsr.mitpress.mit.edu/pub/7z10o269.

[31] Saadani, T. (2022). 6 natural language processing models you should know. Retrieved from https://medium.com/ubiai-nlp/5-natural-language-processing-models-you-should-know-836958b03da3

[32] Turner Lee, N. (2018). Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society*, 16(3):252–260.

[33] U.S. Bureau of the Census (1987). Population profile of the united states: 1984–85. Technical Report P-23, No. 150, U.S. Government Printing Office, Washington, D.C.

[34] USDoJ (2011). Census of public and private juvenile detention, correctional, and shelter facilities, 1986-1987: [united states]. Retrieved from https://www.icpsr.umich.edu/web/NACJD/studies/8973#.

[35] Wang, C., Han, B., Patel, B., et al. (2023). In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *Journal of Quantitative Criminology*, 39:519–581.