Universidad del Valle de Guatemala

Faculty of Engineering

Responsible AI

Department of Computer Science

Semester II, 2025

Project 01

Identifying and Mitigating Biases in Machine Learning Models

Emilio José Solano Orozco	21212
José Ricardo Méndez González	21289
Sara María Pérez Echeverría	21371
Emily Elvia Melissa Pérez Alarcón	21385
Diego Alberto Leiva Pérez	21752

Guatemala, August 29, 2025

Introduction

Artificial Intelligence (AI) has the potential to improve the decision-making process in many fields. However, these systems also carry significant risks of bias and unfairness, especially when they are applied to sensitive decisions that affect people's lives. A widely studied example is the COMPAS algorithm (Correctional Offender Management Profiling for Alternative Sanctions), a commercial tool used by judges and parole officers in the United States to predict the likelihood that a defendant will re-offend.

Investigations led by ProPublica revealed that COMPAS systematically produced biased outcomes: Black defendants who did not re-offend within two years were nearly twice as likely as White defendants to be misclassified as high risk, while White defendants who did re-offend were often misclassified as low risk. These disparities persisted even after controlling factors such as prior crimes, age, and gender, raising concerns about racial bias embedded in the system.

The controversy around COMPAS highlights a broader challenge: how to ensure that predictive models respect principles of Responsible AI. Fairness, accountability, transparency and ethics are essential for trustworthy AI. This project appears these principles to analyze the ProPublica COMPAS dataset, identify potential sources of bias, train predictive modes, and evaluate their fairness across sensitive groups such as race, gender and age. Beyond measuring performance, the project shows strategies to mitigate bias and discusses the ethical implications of deploying AI in high-stakes contexts like criminal justice.

Exploratory Data Analysis (EDA)

Dataset Overview

This project uses the FairML-preprocessed ProPublica COMPAS dataset (propublica_data_for_fairml.csv). This version was selected instead of the raw ProPublica data because it is cleaner, smaller in feature size and explicitly encodes sensitive attributes such as race, gender and age. These characteristics make it more suitable for fairness analysis, as the dataset highlights variables directly relevant to Responsible AI evaluation.

The dataset contains 6,172 entries and 12 variables, with no missing values. The main features include:

- Target Variable: *Two_yr_Recidivism*, indicating whether a defendant re-offended within two years.
- Demographic attributes: race (African American, Hispanic, Asian, Native American, Other), gender (Female = 1, Male = 0), and age group indicators (Age_Below_TwentyFive, Age_Above_FourtyFive).
- Criminal history features: *Number_of_Priors* (count of prior offenses), *Misdemeanor* (indicator for misdemeanor offenses), and *score_factor* (derived risk score).

The overall two-year recidivism rate is 45.5%. While 54.5% of defendants did not re-offend within the same period. The completeness and structure of this dataset ensures that no imputation or additional preprocessing was required prior to the analysis.

Distributions of Sensitive Attributes

The dataset includes several sensitivities attributes that are relevant for fairness analysis, the 3 more important ones are race, gender and age group. Examining the distributions of these variables is essential in the identification of potential imbalances that may contribute to a biased model outcome.

The dataset is heavily imbalanced in terms of racial representation. African American defendants account for approximately 85.5% of the records, while Hispanics represent 8.2%, individuals labeled as Other represent 5.6%, Asians constitute only 0.5%, and Native Americans just 0.2%. This strong overrepresentation of African Americans raises concerns about disproportionate influence on model training and evaluation (Figure 1).

The gender distribution is also uneven. Male defendants represent approximately 81.0% of the dataset, while female defendants account for 19.0%. This imbalance indicates that any model trained on this dataset will be disproportionately influenced by male outcomes.

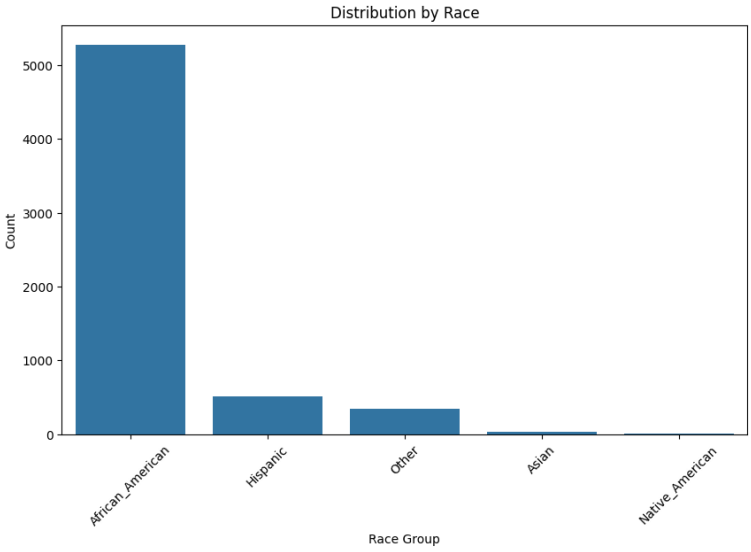

Figure 1: Distribution of defendants by Race

The dataset is somewhat more balanced by age, although disparities remain. Defendants aged 25–45 make up the largest group (57.2%), followed by those below 25 (21.8%) and those above 45 (20.9%). Younger defendants are thus underrepresented compared to the central age group but still account for a significant proportion.

Recidivism outcomes by Group

To evaluate the potential disparities in outcomes, recidivism rates were analyzed across race, gender and age groups. This step is critical for identifying whether predictive patterns differ among sensitive subpopulations.

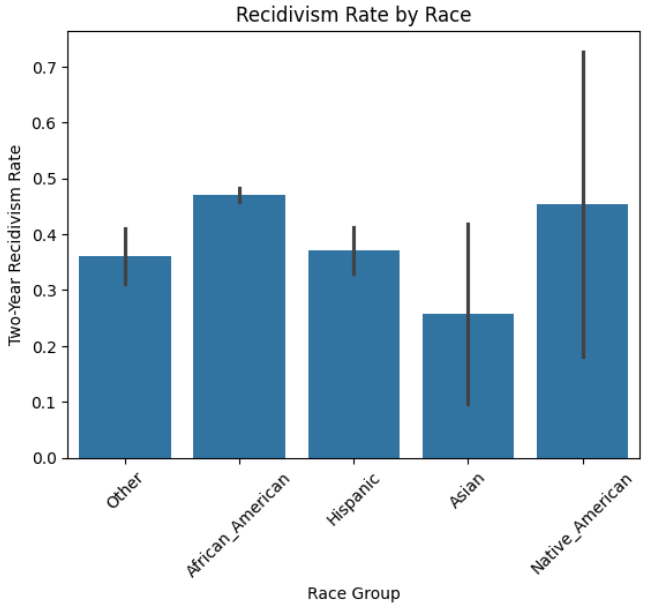

Figure 2: Recidivism Rate by Race

Significant variation is observed across all racial groups. African American defendants show the highest two-year recidivism rate (47.0%), while Asian defendants exhibit the lowest (25.8%). Hispanic defendants re-offend at a rate of 37.1%, while Native Americans at 45.5% and individuals categorizes as Other at 36.2% A Chi-square test confirmed that these differences are statistically significant, suggesting that recidivism outcomes are not independent of race (Figure 2).

Male defendants re-offend at a higher rate (47.9%) compared to female defendants (35.1%). This disparity indicates that gender is also associated with unequal recidivism outcomes.

Younger defendants demonstrate substantially higher recidivism rates. Individuals below the age of 25, re-offend at a rate of 55.9%, compared to 46.5% for those aged 25-45, and 32.0% for those above the age of 45. This pattern reveals a clear negative association between age and likelihood of recidivism.

A heatmap combining race and age groups reveals that disparities are magnified when attributes intersect. For example, African American defendants below 25 years old exhibit the highest risk of re-offending, exceeding 55%. This highlights the importance of examining multiple demographic factors simultaneously (Figure 3).

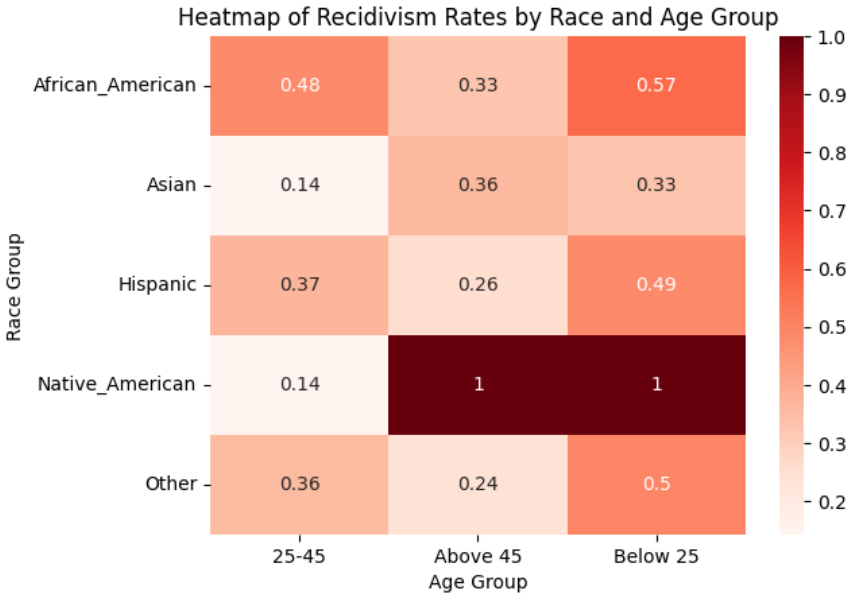

Figure 3: Intersectional Analysis of recidivism rates by demographic subgroups

Number of Priors by Group

The variable *Number_of_Priors* captures the count of previous offenses committed by each defendant. Since prior criminal history is strongly correlated with recidivism outcomes, its distribution across demographic groups is extremely important for fairness analysis.

Marked differences are observed in the mean number of priors across racial groups. African American defendants have the highest average (3.46), compared to Hispanics (2.10), individuals categorized as Other (1.75), and Asians (1.35). Native American defendants display an even higher average (5.18), although this estimate is based on very few cases (n=11). These disparities suggest that criminal history data may disproportionately penalize specific groups (Table 1).

Race Group	Count	Mean	STD	Min	25%	50%	75%	Max
African American	5278.0	3.46	4.88	0.0	0.0	2.0	5.0	38.0
Asian	31.0	1.35	2.24	0.0	0.0	0.0	2.0	9.0
Hispanic	509.0	2.10	3.69	0.0	0.0	1.0	2.0	26.0
Native American	11.0	5.18	7.11	0.0	0.5	2.0	6.0	22.0
Other	343.0	1.75	3.41	0.0	0.0	0.0	2.0	31.0

Table 1: Number of priors distribution across race groups

Male defendants exhibit a higher mean number of priors (3.52) than female defendants (2.09). This gap indicates that gender is also associated with differences in prior criminal history.

Correlation Analysis

To understand relationships between variables and assess potential sources of bias, correlation analysis and statistical tests were conducted.

A correlation heatmap of numeric features reveals several important associations. The variable *Number_of_Priors* shows a moderate positive correlation with the target *Two_yr_Recidivism* (r ≈ 0.29), confirming that prior offenses are predictive of re-offending. Additionally, the race indicator for African Americans is moderately correlated with both *score_factor* and recidivism outcomes. Other features, such as gender, exhibit weaker correlations (Figure 4).

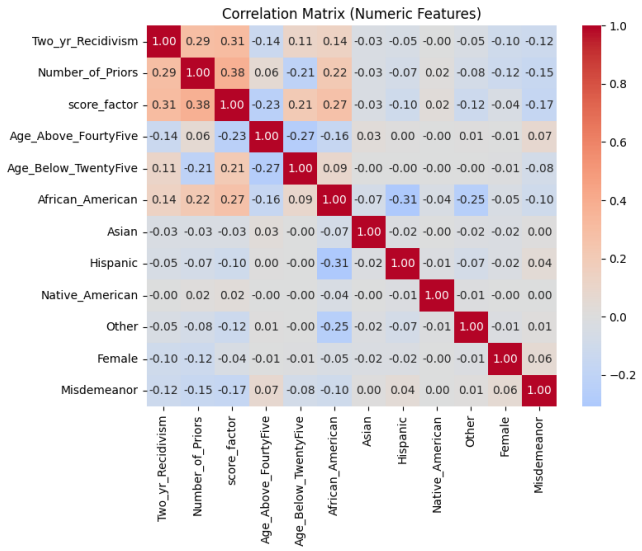

Figure 4: Correlation matrix of numeric features

Overall, these results confirm that recidivism is not only influenced by prior criminal history but is also unevenly distributed across sensitive attributes such as race, gender, and age. Statistical evidence demonstrates that these disparities are significant and likely to propagate into predictive modeling if not mitigated.

Identification of potential biases

Based on the Exploratory Analysis, one of the key potential biases that were found lies in the imbalance of class representation across the dataset. Certain classes such as positive versus negative sentiment or specific demographic groups are disproportioned and the model could generalize effectively on minor classes. This imbalance can manifest in metrics like accuracy, being deceptively high while performance on underrepresented groups remains poor.

The dataset may present underrepresentation of subgroups, specific age ranges, regions, or contextual variations. Some of the biases that were identified are:

- Models may learn patterns mainly from African American group, producing unreliable predictions for minorities.
- African Americans & Native Americans have higher recidivism rates than other groups.
- Female predictions may be less reliable due to underrepresentation.
- Males flagged as higher risk more often
- Younger defendants are much more likely to be labeled as recidivists

Without addressing these disparities, the deployed model risks reinforcing preexisting biases in the data.

Training of an initial model.

Two baseline models were trained to predict two-year recidivism: Logistic Regression and a Decision Tree classifier. Both models were trained using a set of predictive and demographic features, like prior offenses, age groups, gender, race, and charge type. The dataset was split into 70% training and 30% testing with stratification to preserve some class balance.

The logistic regression model achieved an overall accuracy of 68.2%, with a precision of 0.68 and a recall of 0.57. This indicated that the model tended to misclassify positive cases (re-offenders) more frequently than negative ones. In parallel, the decision tree configured with a depth of 5 achieved similar values, with an accuracy of 67.8%, precision of 0.67, and recall of 0.58. Its error distribution was comparable, showing difficulties in correctly identifying recidivist cases. Both models produced moderate performance at the global level, but further analysis revealed uneven results across the different demographic subgroups.

Evaluation of the model segmented by sensitive groups

To assess fairness, the initial models were evaluated across gender, age group, and race. The evaluation of the baseline Logistic Regression model by sensitive groups is summarized in Table 2. While overall accuracy remained moderate (0.68), subgroup analysis revealed notable disparities. For instance, recall was high among younger defendants (0.73 for those below 25) but much lower for older individuals (0.39 for those above 45). Similarly, the model achieved higher recall for African American defendants (0.68) compared to Hispanic (0.40) and female defendants (0.39). These disparities indicate that the model systematically favors certain demographic groups over others.

Group	Accuracy	Precision	Recall	F1-score	Notes
Global	0.68	0.68	0.57	0.62	Baseline overall performance
Female=0 (Male)	0.67	0.68	0.60	0.64	Higher recall
Female=1 (Female)	0.72	0.68	0.39	0.49	Lower recall for women
Age: 25–45	0.67	0.68	0.54	0.60	Moderate
Age: Above 45	0.75	0.66	0.39	0.49	Poor recall
Age: Below 25	0.66	0.69	0.73	0.71	High recall
African American	0.68	0.70	0.68	0.69	Best balance
Hispanic	0.68	0.73	0.40	0.52	Low recall
Other	0.69	0.61	0.40	0.48	Low recall
Asian / Native Am.	–	–	–	–	Too few samples, unstable results

Table 2: Performance metrics of the baseline Logistic Regression model.

Proposed mitigation strategies

1. Feature selection for sensitive attributes such as race and sex were excluded from training to prevent direct bias injection but kept for fairness evaluation afterward.
2. Rebalancing strategies with *class_weight*="balanced" to counteract class imbalance and oversampling techniques like SMOTE.
3. Normalization with *StandardScaler* and categorical variables. Also, encoding to ensured that scale differences did not amplify bias in logistic regression models.

Comparison of the model before and after mitigation

Before mitigation	After mitigation

• Logistic Regression:
  ○ Racial groups like Hispanics and "Other" had significantly worse recall and F1 than African Americans.
• Decision Tree:
	• Logistic Regression:
	  ○ Group disparities narrowed slightly. Female recall improved.
	  ○ African Americans maintained strong performance, while recall for

○ Similar disparities: underperformance for women, Hispanics, Native Americans, and some age subgroups. • About Bias: ○ The models generalized better for majority groups (males, African Americans, 25–45) and underperformed for minorities (females, Hispanics, Native Americans, above 45).	Hispanics remained weak but more consistent. • Decision Tree: ○ Subgroup metrics showed slightly more consistency • About Bias: ○ Gender gap reduced somewhat, especially in recall for females. ○ Age group disparities narrowed, with "Above 45" performing closer to the global average. ○ Racial disparities persisted, especially for very small subgroups (Native American, Asian) — mitigation techniques like reweighting cannot fully solve the problem without more representative data.

Opinions, conclusions, and reflections on responsible use of AI

The analysis revealed that the models reproduced existing social biases, particularly across race, age and gender. While the global accuracy of the models seemed acceptable, subgroup analysis exposed deep inequities that accuracy alone could not capture. Although mitigation strategies improved some results, they did not eliminate systemic bias, underscoring the importance of transparency and accountability in the responsible use of AI.

Future work should prioritize fairness aware modeling techniques and balanced datasets to reduce structural bias. AI predictions must be complemented with human oversight to avoid automated discrimination and reframe algorithms as decision-support tools. Finally, continuous auditing and interdisciplinary collaboration are essential to ensure ethical deployment and to keep models aligned with broader social issues.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Ofer, D. (2017). *COMPAS recidivism racial bias*. Kaggle. https://www.kaggle.com/datasets/danofer/compass

Whitepaper. (2017). *FairML: Auditing Black-Box Predictive Models*. Fast Forward Labs. https://blog.fastforwardlabs.com/2017/03/09/fairml-auditing-black-box-predictive-models.html

Annexes

GitHub Repository: https://github.com/MelissaPerez09/fair-ai-compass

GitHub Page: https://melissaperez09.github.io/fair-ai-compass/