Fairness Analysis of Gender and Age Disparities in the COMPAS Recidivism Dataset

Irmak Erkol
Computer Science and Engineering, Sabancı University
irmak.erkol@sabanciuniv.edu.tr

I. Introduction

Artificial Intelligence (AI) systems are used widely in sensitive areas such as criminal justice. AI helps to make decisions about bail, sentencing, and parole in criminal justice. The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) risk assessment tool is one such system, which is meant to figure out how likely it is that a defendant will break the law again. It is essential to make sure these systems are fair for all demographics since the effects of the decisions have critical effects on people's lives and freedom.

This paper examines fairness disparities in the COMPAS recidivism prediction system, which means it is predicting how likely it is that a defendant will break the law again. We focused on gender and age as protected attributes. In fairness analysis, protected attributes (like gender and age) include all groups within those categories, not just historically disadvantaged ones. While women have historically faced discrimination in many contexts, in this particular algorithmic system, our analysis showed that men actually face higher risks of being incorrectly classified as likely to recidivate. That is why we analyze all groups within a protected attribute.

Using a dataset compiled by ProPublica [1], we analyze how the system's predictions vary across these demographic groups and evaluate whether they meet established fairness criteria. We compute four key fairness metrics: Statistical Parity (Selection Rate), True Positive Rate Parity, False Positive Rate Parity, and Positive Predictive Value Parity.

II. Methodology

A. Dataset Description

The COMPAS dataset contains information about 7,214 defendants in Broward County, Florida. The assessments were made starting in 2013 and ending in 2014. The dataset includes demographic information, criminal history, the COMPAS risk score, and whether the defendant actually recidivated (reoffended) within two years. We chose the COMPAS dataset because it represents a real-world algorithmic system with direct consequences for individuals within the criminal justice system. Also, the dataset contains multiple demographic attributes, which enabled us to make fairness analyses across different protected groups. Also I was interested in this topic because women tend to get more harm from bias and discrimination in general, but for this case, it was men who get more harm.

Our analysis focused on two protected attributes, which are gender and age. The dataset includes male and female defendants with under 25, 25-35, 35-45, and over 45 years of age. The target variable is 'two_year_recid', which indicates whether the defendant recidivated within two years (1) or not (0).

B. Data Preprocessing

We simplified the dataset using ProPublica's approach [1] by keeping only cases where the COMPAS screening happened within 30 days of arrest, removing cases with missing or invalid recidivism information, excluding cases with "Other" charge types (keeping only felonies and misdemeanors), removing cases with missing COMPAS scores and finally creating binary gender variables and dividing age into four groups (under 25, 25-35, 35-45, over 45).

Before we built the prediction model, we examined the dataset and found that 35.15% of female defendants recidivated compared to 47.95% of male defendants, showing a gender disparity of 12.80 percentage points in the actual observed recidivism rates in the real-world data, rather than the model's predictions.

We also found a strong pattern in the actual recidivism data across different age groups. The youngest defendants (under 25 years old) had the highest reoffending rate at 55.15%. As age increased, reoffending became less common - 49.89% for ages 25-35, 37.61% for ages 35-45, and only 31.72% for those over 45 years old. This creates a difference of 23.43 percentage points between the youngest and oldest groups. This pattern shows that age is strongly related to actual reoffending rates in our dataset, even before any algorithmic predictions. This information about age and gender disparities in the real world is important because it will help us to compare the disparity between our model's predictions and disparities in the real world.

C. Model Development

We trained a logistic regression model to predict the likelihood of recidivism using the following features: age, gender, number of prior criminal offenses, and whether the current charge is a felony or misdemeanor. The data was split into training (70%) and testing (30%) sets. The logistic regression model achieved an overall accuracy of 68% on the test set, with a precision of 66% and recall of 57% for the positive class (recidivism).

III. Results

We calculated four key fairness metrics for each demographic group. For each metric, we calculated the disparity as the difference between the maximum and minimum values across groups to quantify the level of unfairness. We used the fairness metrics as defined in Kleinberg et al. [2] and implemented using approaches similar to those in the Fairlearn library [3].

A. Gender-based Fairness Analysis

Our analysis revealed important disparities between male and female defendants across all fairness metrics, as shown in Table I.

TABLE I
FAIRNESS METRICS BY GENDER

Metric	Female	Male	Disparity
Statistical Parity	0.0935	0.4603	0.3668
True Positive Rate	0.1985	0.6434	0.4449
False Positive Rate	0.0315	0.2995	0.2680
Positive Predictive Value	0.7879	0.6536	0.1343

The gender-based analysis revealed that male defendants were predicted to recidivate at a rate 36.68 percentage points higher than female defendants, which is quite larger than the 12.80 percentage point difference in actual recidivism rates. Male recidivists were much more likely to be correctly identified (64.34% TPR) compared to female recidivists (19.85% TPR), resulting in a 44.49 percentage point disparity in true positive rates.

Perhaps most concerning, non-recidivist males faced a false positive rate of 29.95%, nearly 10 times higher than the 3.15% false positive rate for females. This means that non-recidivist males were far more likely to be wrongly classified as future recidivists. Interestingly, when the model did predict recidivism for females, it was more accurate (78.79% PPV) than when it predicted recidivism for males (65.36% PPV).

B. Age-based Fairness Analysis

Age groups exhibited even larger disparities across most fairness metrics, as shown in Table II.

TABLE II
FAIRNESS METRICS BY AGE GROUP

Metric	¡25	25-35	35-45	¿45	Disparity
Statistical Parity	0.5740	0.4218	0.2966	0.1527	0.4213
True Positive Rate	0.6940	0.5948	0.4696	0.2970	0.3970
False Positive Rate	0.4353	0.2394	0.2028	0.0935	0.3418
Positive Predictive Value	0.6481	0.7238	0.5567	0.5660	0.1671

The age-based analysis revealed a clear gradient effect, with younger defendants receiving much higher risk predictions than older defendants. Defendants under 25 were predicted to recidivate at a rate of 57.40%, compared to just 15.27% for defendants over 45, resulting in a disparity of 42.13 percentage points—nearly twice the difference in actual recidivism rates between these groups (23.43 percentage points).

Similarly, the model identified 69.40% of young recidivists correctly, but only 29.70% of older recidivists, resulting in a 39.70 percentage point disparity in true positive rates. The false positive rate exhibited a dramatic gradient, decreasing from 43.53% for the youngest defendants to just 9.35% for the oldest, a disparity of 34.18 percentage points. This means that non-recidivist young defendants were 4.7 times more likely to be incorrectly flagged as future recidivists compared to older non-recidivists.

The positive predictive value was highest for the 25-35 age group (72.38%) and lowest for the 35-45 age group (55.67%), with a maximum disparity of 16.71 percentage points.

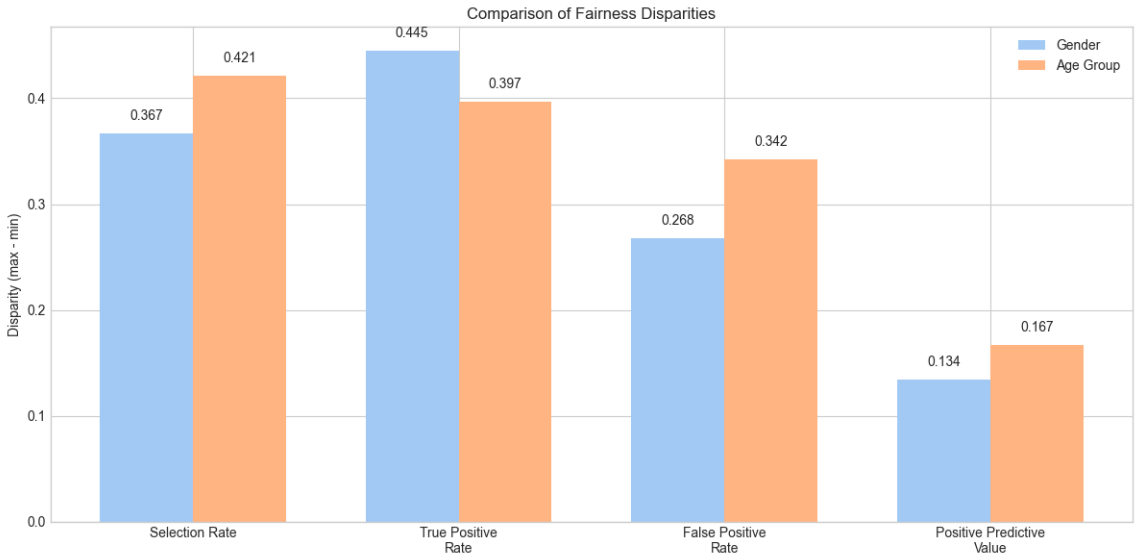

Fig. 1. Comparison of fairness disparities across gender (blue) and age groups (orange). The y-axis shows the difference between maximum and minimum values for each metric within each protected attribute.

C. Comparison of Disparities

To better understand which protected attribute exhibited the greatest fairness concerns, we compared the disparities across metrics, as shown in Fig. **??**.

Age-based disparities were larger than gender-based disparities for Statistical Parity (42.13 percentage points vs. 36.68 percentage points) and false positive rate (34.18 percentage points vs. 26.80 percentage points). Conversely, the true positive rate showed a larger disparity for gender (44.49 percentage points) than for age (39.70 percentage points). Both attributes showed concerning levels of disparity across all metrics, with most exceeding 20 percentage points.

IV. Discussion

A. Interpretation of Findings

Our findings show serious fairness problems with the algorithm. The system takes real differences in recidivism rates and makes them much worse. It predicts much higher risk for men and young people than their actual reoffending rates justify. This creates harmful consequences, especially through false positives where innocent men and young people are wrongly labeled as high-risk, potentially leading to longer sentences or denied bail. The consistent pattern where risk predictions decrease with age shows systematic age-based discrimination. Most concerning is that these biases likely compound for young men, who face discrimination based on both their gender and age simultaneously. As Mehrabi et al. [5] note, these supposedly ”neutral” algorithms often magnify biases through feedback loops, while Roselli et al. [4] highlight how they reinforce existing inequalities in the justice system. Widder et al. [6] further warn that such systems reproduce intersectional biases present in society.

These results show the algorithm produces much larger disparities for men than women. While men do reoffend slightly more often in reality (about 13% more), the algorithm predicts men will reoffend at a rate nearly 37 percentage points higher than women. This means the algorithm amplifies the gender difference almost 3 times beyond what exists in reality.

The system correctly identifies male reoffenders 64% of the time but female reoffenders only 20% of the time, meaning it misses most women who actually reoffend. However, the most troubling finding is that among people who don’t reoffend, men are wrongly flagged as ”high risk” nearly 10 times more often than women. This means innocent men face much harsher consequences from this system. When the system does predict a woman will reoffend, it’s usually right (79% accuracy), but less accurate for men (65%).

The algorithm shows strong disparities against young people. While younger people do reoffend more often in reality, the algorithm exaggerates this difference drastically. It predicts 57% of people under 25 will reoffend compared to only 15% of those over 45, a difference twice as large as the real-world difference.

The algorithm is most accurate at identifying young reoffenders (69% correct) but misses most older reoffenders (only 30% correct). Most concerning is that among people who don’t actually reoffend, young people are 4.7 times more likely to be falsely labeled high-risk compared to older people. This means innocent young people face much harsher treatment from this system.

This comparison shows that both age and gender create serious fairness problems in this algorithm, but in slightly different ways. Age creates bigger disparities in who gets labeled high-risk overall and in false accusations of innocent people. Gender creates bigger disparities in correctly identifying actual reoffenders. Both types of disparities are severe, with most differences exceeding 20 percentage points, which is far beyond what's acceptable for a fair system.

Together, these findings suggest young men likely face the most severe discrimination from this algorithm, being subjected to much higher rates of both correct and incorrect high-risk classifications. These disparities demonstrate that predictive algorithms in criminal justice can perpetuate and amplify systemic biases, raising serious ethical concerns about their use in high-stakes decision-making contexts.

B. Potential Mitigation Strategies

To make the algorithm fairer, we recommend several practical solutions. First, we could fix the training data by giving more importance to female reoffenders and young non-reoffenders, this way we can balance out the skewed predictions as suggested by Mehrabi et al. [5]. Second, we could build fairness constraints directly into the model's training process, making it optimize for both accuracy and demographic fairness simultaneously, following approaches outlined by Roselli et al. [4]. Third, we could adjust the risk thresholds differently for each group, requiring stronger evidence before labeling young people or men as high-risk, an approach described in Kleinberg et al. [2]. Fourth, we should carefully examine which features might indirectly relate to gender or age and consider removing them to prevent hidden biases. Finally, these algorithms should never make decisions alone. Human judges should use them as just one factor among many, considering individual circumstances that algorithms can't capture, as recommended by Widder et al. [6]. Together, these approaches could significantly reduce the unfair treatment we identified while maintaining the system's ability to assess recidivism risk.

C. Challenges Encountered

Our fairness analysis faced several significant challenges. We discovered trade-offs between different fairness measures, where improving one metric often worsened others, confirming Kleinberg et al.'s [2] mathematical proof that certain fairness criteria cannot be satisfied simultaneously. The COMPAS dataset itself presented limitations since it is coming from just one Florida county and potentially not representing wider judicial patterns. More fundamentally, using recidivism as our target variable was problematic since it depends on re-arrest, which may reflect the same biased policing practices we were trying to evaluate, creating what Mehrabi et al. [5] describe as "feedback loops" in algorithmic bias. We also encountered a persistent tension between improving fairness and maintaining predictive accuracy, as attempts to reduce demographic disparities typically lowered the model's overall performance. Finally, the technical implementation required high attention to detail when calculating metrics across multiple demographic groups to avoid drawing incorrect conclusions about the extent of algorithmic bias.

V. Conclusion

Our analysis shows clear fairness problems in how the AI system predicts recidivism for different groups. The algorithm treats men and young people much more harshly than women and older people, with differences far greater than actual recidivism patterns justify. Most troubling is how often the system wrongly labels innocent men and young people as "high-risk", potentially changing their treatment in the justice system.

As Roselli et al. point out, "managing bias in AI systems requires both technical solutions and organizational processes" [4]. We found similar challenges to those described in Mehrabi et al.'s research, which shows that "bias can be introduced at any stage of the AI/ML pipeline" [5]. Our results also support what Widder et al. argue - that supposedly "objective" algorithms actually reinforce existing social inequalities when we don't carefully examine their data and methods [6].

These findings show why we must test algorithms for fairness before using them in high-stakes situations like criminal justice. Without proper testing and fixes, these systems can make existing biases worse and harm vulnerable communities.

Future research should look at how these biases might affect people who belong to multiple groups (like young men), test different solutions to reduce bias, and create ongoing monitoring systems as society changes. As algorithms increasingly influence decisions that affect people's freedom and opportunities, making them fair isn't just a technical issue, it's a basic ethical responsibility.

References

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias," ProPublica, May 23, 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *Proc. 8th Innov. Theor. Comput. Sci. Conf.*, 2016.

[3] "Common fairness metrics," Fairlearn. [Online]. Available: https://fairlearn.org/v0.10/user_guide/assessment/common_fairness_metrics.html

[4] D. Roselli, J. Matthews, and N. Talagala, "Managing bias in AI," in *Companion Proc. 2019 World Wide Web Conf.*, 2019, pp. 539-544.

[5] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1-35, 2021.

[6] D. G. Widder, M. Whittaker, and S. M. West, "Why 'open' AI systems are actually closed, and why this matters," *Big Data & Society*, vol. 11, no. 1, 2024.