TML Course Report

# COMPAS Fairness Analysis

Yuxuan Bai, Jizhong Cui

December 16, 2024

Faculty of Science
University of Helsinki

**Contact information**

P. O. Box 68 (Pietari Kalmin katu 5)
00014 University of Helsinki,Finland

Email address: info@cs.helsinki.fi
URL: http://www.cs.helsinki.fi/

# Contents

HELSINGIN YLIOPISTO – HELSINGFORS UNIVERSITET – UNIVERSITY OF HELSINKI

| Tiedekunta — Fakultet — Faculty | Koulutusohjelma — Utbildningsprogram — Study programme |
|---|---|
| Faculty of Science | |

Tekijä — Författare — Author

Yuxuan Bai, Jizhong Cui

Työn nimi — Arbetets titel — Title

COMPAS Fairness Analysis

Ohjaajat — Handledare — Supervisors

Prof. Indrė Žliobaitė

| Työn laji — Arbetets art — Level | Aika — Datum — Month and year | Sivumäärä — Sidoantal — Number of pages |
|---|---|---|
| TML Course Report | December 16, 2024 | 12 pages, 2 appendix pages |

Tiivistelmä — Referat — Abstract

This study analyzes fairness issues in the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) recidivism prediction system, examining potential racial and age-based biases in its risk assessments. Using a dataset of 7214 defendants from 2013-2014, we implemented an XGBoost classifier and evaluated multiple fairness metrics, including statistical parity, equal opportunity, and disparate impact. Our analysis revealed significant disparities in predictions across demographic groups, with higher false positive rates for African-American defendants and younger individuals. To address these biases, we evaluated two mitigation strategies: resampling techniques (SMOTE and ADASYN) and a fairness-aware model modification (FairXGB). Results demonstrate that SMOTE resampling significantly improved fairness outcomes while maintaining model performance, effectively reducing disparate impact across protected groups. Based on these findings, we recommend implementing SMOTE as a preprocessing step for developing more equitable risk assessment models in criminal justice applications.

**ACM Computing Classification System (CCS)**
Applied computing → Law, social and behavioral sciences
Computing methodologies → Machine learning

Avainsanat — Nyckelord — Keywords

fairness, algorithmic bias, resampling

Säilytyspaikka — Förvaringsställe — Where deposited

Helsinki University Library

Muita tietoja — övriga uppgifter — Additional information

# 1 Introduction

Recently, the State Department of Corrections approached our team with concerns about the fairness of their risk assessment system, COMPAS (Correctional Offender Management Profiling for Alternative Sanctions). Several public defenders had raised questions about potential racial bias in the system's predictions, prompting the department to commission our team to conduct a comprehensive fairness analysis and develop concrete strategies to enhance the system's equity across different demographic groups.

COMPAS, developed by Northpointe, is one of the most widely used algorithmic risk assessment tools in the United States criminal justice system. The system generates recidivism risk scores on a scale of 1 (lowest) to 10 (highest) by evaluating multiple factors including criminal history, substance abuse, and socioeconomic background. While intended to bring greater objectivity to judicial decisions, COMPAS has faced significant scrutiny regarding racial bias, most notably through a 2016 ProPublica investigation [JA] that identified potential disparities in predictions affecting Black defendants.

Our analysis utilizes the "compas-scores-two-years" dataset, which tracks 7,214 defendants from 2013 to 2014. The dataset includes demographic information (age, race, gender), criminal history features (prior count, charge degree), and recidivism outcomes within a two-year period post-release. We frame our investigation as a binary classification task predicting two-year recidivism likelihood, with a focus on quantifying and addressing any disparity in prediction accuracy across demographic groups.

# 2 Base Study

To conduct a comprehensive fairness analysis of COMPAS, we employ a two-phase approach. First, we analyze the dataset's inherent bias by examining statistical parity across demographic groups. Second, we evaluate the predictive model's fairness through four metrics. Based on our empirical findings and the specific context of recidivism prediction, we identify the most relevant metric to guide our subsequent bias mitigation strategies.

For a binary classifier predicting recidivism, we analyze four key fairness metrics across demographic groups. Statistical Parity examines whether risk predictions are equally distributed across groups, measured through both the difference and ratio (disparate impact, DI). Equal Opportunity compares the True Positive Rate across groups, ensuring consistent identification of actual recidivists. Equalized Odds extends this by requiring both equal True Positive Rates and False Positive Rates across groups, addressing both public safety and individual rights concerns. Predictive Parity ensures that high-risk predictions carry the same meaning across groups by comparing positive predictive values.

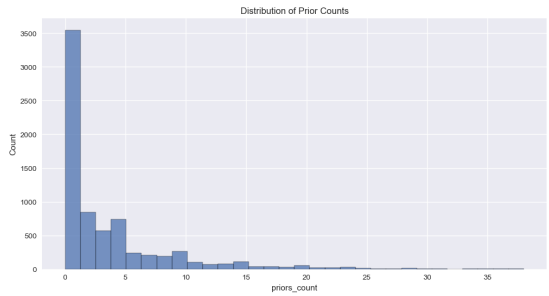
These metrics are particularly relevant in the criminal justice context, as they address different aspects of fairness in recidivism prediction. Statistical Parity identifies systemic bias in risk assessment, Equal Opportunity ensures equitable identification of high-risk individuals across communities, Equalized Odds balances both false positives and false negatives, which have serious consequences, and Predictive Parity maintains the credibility of risk assessments across demographic groups. While it is impossible to satisfy all metrics simultaneously when base recidivism rates differ across groups, understanding these trade-offs is crucial for prioritizing different aspects of fairness in the system.

## 2.1 Dataset Analysis

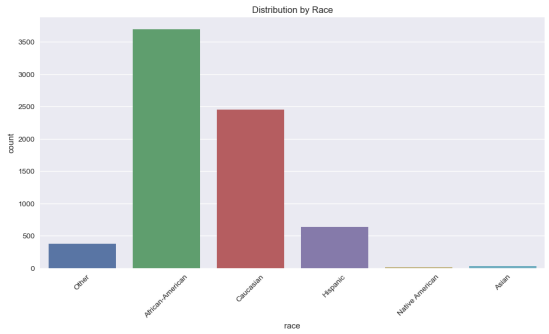
We analyzed the COMPAS dataset focusing on demographic factors (Race, Sex, Age) and criminal history (Prior Count) to identify potential biases in sample distribution and recidivism predictions across demographic groups.

Figure 2.1 shows the key dataset characteristics: prior counts follow a decreasing pattern with minimal impact on predictions; African-American and Caucasian groups dominate

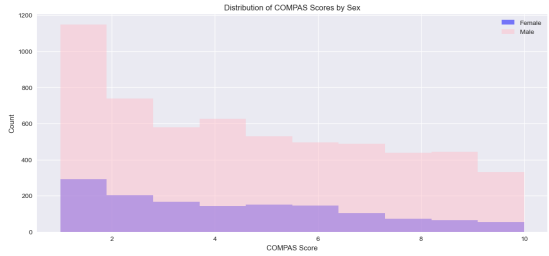
the racial distribution; males outnumber females while maintaining similar score patterns; and the 25-45 age group is overrepresented.



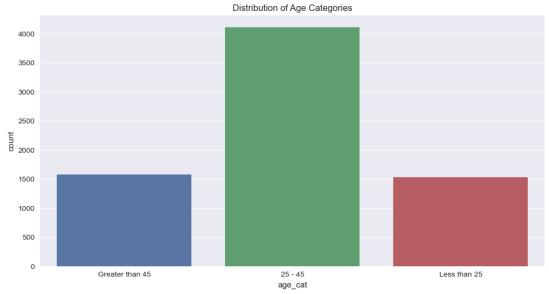

(a) Prior count




(b) Race

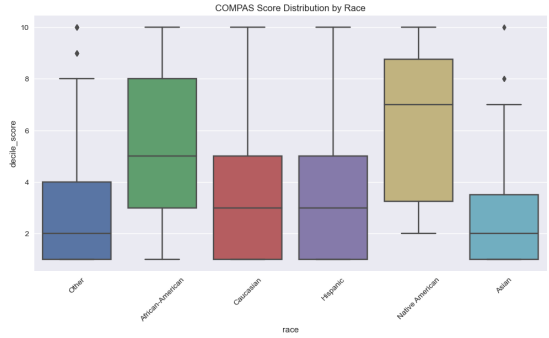



(c) Sex

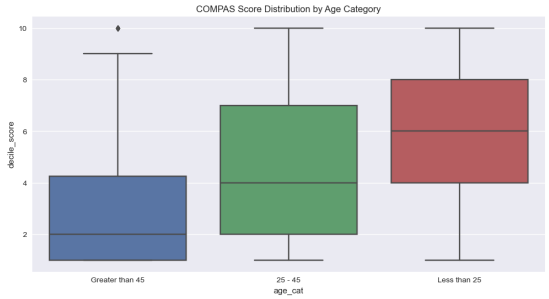



(d) Age

Figure 2.1: Demographic and prior count distributions




(a) Race




(b) Age

Figure 2.2: COMPAS scores by demographic variables

Figure 2.2 reveals disparities in COMPAS scores: African-American and Native American groups show higher median scores than other races, and younger individuals (<25) receive higher scores than older ones (>45). Sex differences in scores were not significant. The difference in scores between males and females is not significant, as shown in FigureA.1.

## 2.2 Classification Results

For our binary classification task of predicting recidivism is_recid, we utilized a focused set of features and implemented an XGBoost classifier configured with a binary logistic objective. Details of the selected features and classifier parameters are provided in the Appendix A. We used 5-fold cross-validation with the F1 score as the evaluation metric. The results of our model are presented in Table 2.1.

| **Class** | **Precision** | **Recall** | **F1-score** | **Support** |
|---|---|---|---|---|
| 0 | 0.71 | 0.72 | 0.71 | 1153 |
| 1 | 0.67 | 0.66 | 0.67 | 1012 |
| **Accuracy** | - | - | 0.69 | 2165 |

Table 2.1: Classification report for recidivism prediction

While the model achieved reasonable accuracy, our primary concern was its impact on fairness across different demographic groups. We analyzed this by comparing fairness metrics between the original dataset and model predictions.

| Group | Class | TPR | TNR | FPR | FNR | PPV |
|---|---|---|---|---|---|---|
| Privileged | Caucasian, Hispanic | 0.476 | 0.819 | 0.181 | 0.524 | 0.620 |
| Unprivileged | African-American | 0.753 | 0.628 | 0.372 | 0.247 | 0.700 |

Table 2.2: Comparison of model performance metrics by race

| Data | ΔStatistical Parity | Disparate Impact | ΔEqual Opportunity | ΔEqualized Odds |
|---|---|---|---|---|
| Original | 0.148 | 1.387 | - | - |
| Predicted | 0.280 | 1.955 | 0.277 | 0.277 |

Table 2.3: Fairness metrics comparison for race

Table 2.2 and 2.3 show racial bias in our model. With Caucasian and Hispanic groups as privileged, and African-American as unprivileged, the model shows higher TPR (0.753 vs 0.476) and FPR (0.372 vs 0.181) for the unprivileged group, indicating bias toward classifying them as "likely to reoffend". The Disparate Impact increase from 1.387 to 1.955 demonstrates bias amplification. Similarly, Table 2.4 and 2.5 reveal age-based bias. Comparing individuals over 45 (privileged) to those under 25 (unprivileged), the model

shows a significant bias toward predicting younger individuals as recidivists, with the Disparate Impact increasing from 2.009 to 2.866.

| Group | Class | TPR | TNR | FPR | FNR | PPV |
|---|---|---|---|---|---|---|
| Privileged | >45 | 0.464 | 0.864 | 0.136 | 0.536 | 0.602 |
| Unprivileged | <25 | 0.767 | 0.464 | 0.536 | 0.233 | 0.697 |

Table 2.4: Comparison of model performance metrics by age

| Data | ΔStatistical Parity | Disparate Impact | Δ Equal Opportunity | Δ Equalized Odds |
|---|---|---|---|---|
| Original | 0.310 | 2.009 | - | - |
| Predicted | 0.442 | 2.866 | 0.303 | 0.400 |

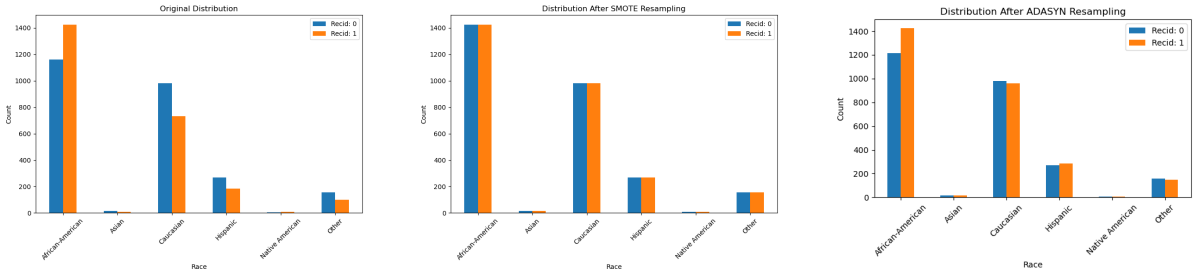
Table 2.5: Fairness metrics comparison for age

# 3 Bias Mitigation Strategies

In Chapter 2, we identified two critical fairness issues in our analysis: (1) an imbalance in sample distribution across different demographic groups, and (2) discrimination amplification by the XGBoost model. This chapter presents two approaches to address these challenges: resampling techniques to balance the dataset and a fairness-aware model that incorporates fairness metrics during training.

## 3.1 Resampling

As shown in Figure 3.1a, the distribution of recidivism outcomes varies significantly across demographic groups, contributing to classification disparities. To mitigate this imbalance, we implement and evaluate two resampling techniques: *SMOTE* and *ADASYN*.




(a) Original (b) SMOTE (c) ADASYN

Figure 3.1: The distribution of positive and negative samples across racial groups of the original and resampled datasets.

*SMOTE* (Synthetic Minority Over-sampling Technique) [Cha+02] addresses class imbalance through the generation of synthetic examples for minority classes by analyzing the feature space and creating new instances that share statistical properties with the original data. For each minority sample, the algorithm identifies its k-nearest neighbors (typically $k$=5), randomly selects one neighbor, and creates a synthetic sample through interpolation along the line segment joining them, which reduces overfitting risk compared to simple oversampling while preserving essential feature relationships and forming more general decision regions.

*ADASYN* (Adaptive Synthetic Sampling) [He+08] extends SMOTE by introducing adaptive sampling density, employing a density-based assessment to determine the required number of synthetic samples for each minority instance based on learning difficulty. The algorithm generates proportionally more synthetic samples for minority instances in challenging regions—where samples are sparse or closely intermixed with the majority class—effectively shifting the classification decision boundary toward difficult examples and providing more nuanced treatment of class imbalance.

The distribution of positive and negative samples across racial groups is shown in Figure 3.1. After applying resampling techniques, SMOTE achieved perfectly balanced class distributions, while ADASYN reduced but did not completely eliminate the class imbalance.

## 3.2 Fairness-aware Model

To prevent discrimination amplification, we propose a novel fairness-aware XGBoost model, FairXGB, that incorporates disparate impact considerations directly into the training process. This approach modifies the gradient updates during training to balance predictive accuracy with fairness constraints.

We augment the standard XGBoost objective function with a fairness penalty term:

$$\text{New Gradient} = \text{Original Gradient} + \lambda_f \times \text{Fairness Penalty}$$

where $\lambda_f$ (fairness weight) controls the trade-off between prediction accuracy and fairness. Based on empirical validation, we set $\lambda_f = 0.2$ as the default value.

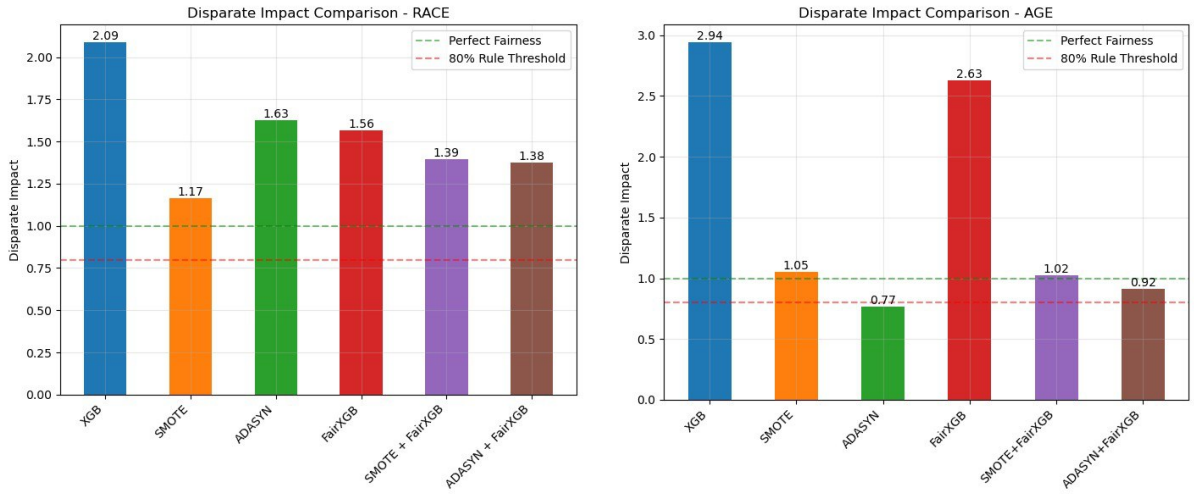
The fairness penalty is computed using the disparate impact (DI) metric after each iteration:

$$\text{Penalty} = \begin{cases} \min(1, |\text{DI} - \frac{1}{\alpha}| \times 0.1) & \text{if DI} \leq \frac{1}{\alpha} \\ \min(1, |\text{DI} - 1| \times 0.1) & \text{if DI} > \frac{1}{\alpha} \end{cases}$$

where $\alpha$ is the fairness threshold, which is 0.8 by default. The factor 0.1 scales the penalty to prevent overshadowing the original gradient and the $\min(1, \cdot)$ operation caps the penalty to maintain training stability. Besides, the DI is clipped to [0.1, 20] to avoid exploding gradients.

## 3.3 Results

We evaluated the effectiveness of three approaches against the baseline: resampling alone, FairXGB alone, and their combination. The disparate impact comparisons in Figure 3.2 demonstrate that resampling, especially SMOTE, effectively reduces disparity to near-ideal levels. While FairXGB shows inconsistent effects on disparate impact when used alone, its combination with resampling demonstrates enhanced discrimination reduction, particularly for age-based discrimination.

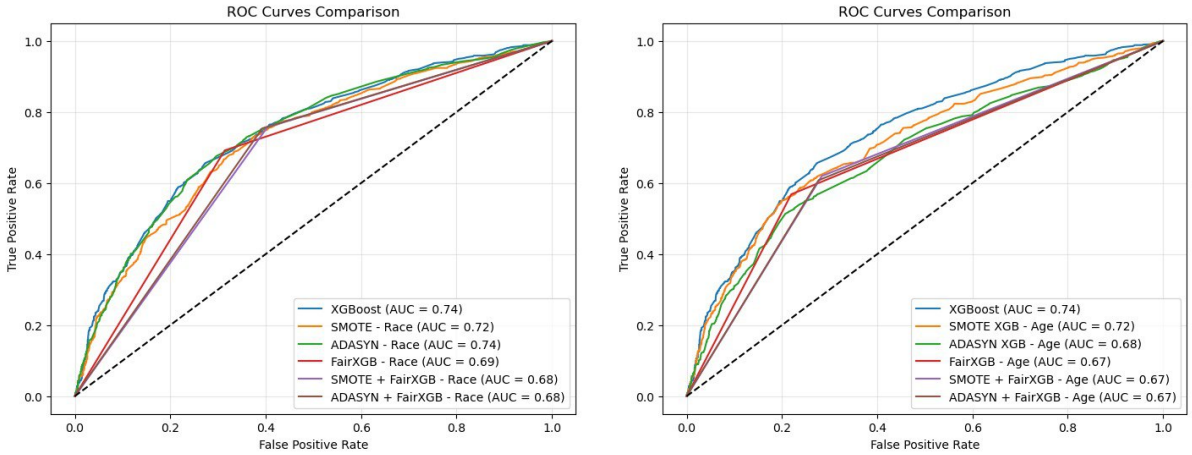



(a) Race (b) Age

Figure 3.2: Disparate impact analysis of the baseline XGBoost model, XGBoost with resampled data, the FairXGB model, and their combination on race and age separately. The ideal value is 1.0.

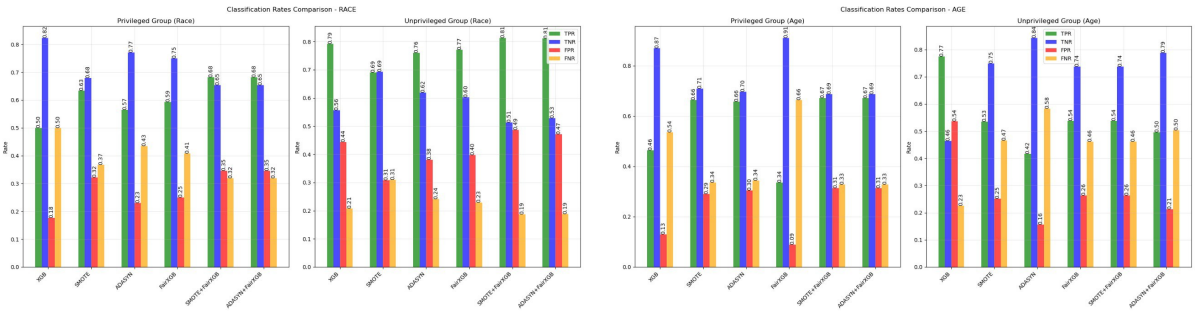
The ROC curves presented in Figure 3.3 indicate that these fairness-enhancing methods maintain model performance while improving fairness outcomes.

Detailed performance metrics in Figure 3.4 reveal that resampling produces two key improvements: it decreases the false negative rate (FNR) for privileged groups while reducing the false positive rate (FPR) for unprivileged groups. These changes indicate more equitable predictions across groups - unprivileged individuals are less likely to receive false predictions of recidivism, while privileged individuals are more likely to be correctly identified when at high risk of reoffending.




(a) Race (b) Age

Figure 3.3: ROC curves for different bias mitigation approaches applied separately to race and age.




(a) Race (b) Age

Figure 3.4: Model performance metrics comparing privileged and unprivileged groups across different bias mitigation approaches.

# 4 Conclusion

Our analysis demonstrates that applying the SMOTE resampling algorithm to the COMPAS dataset before model training significantly improves fairness outcomes while maintaining model performance. This preprocessing approach effectively reduces disparate impact across protected groups, resulting in more equitable predictions without compromising classification accuracy. Based on these findings, we recommend SMOTE resampling as a robust preprocessing step for developing fair risk assessment models on the COMPAS dataset, particularly when balancing the dual objectives of algorithmic fairness and predictive performance.

# Bibliography


[Cha+02] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. "SMOTE: synthetic minority over-sampling technique". In: *J. Artif. Int. Res.* 16.1 (June 2002), pp. 321–357. ISSN: 1076-9757.

[He+08] H. He, Y. Bai, E. A. Garcia, and S. Li. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning". In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. 2008, pp. 1322–1328. DOI: 10.1109/IJCNN.2008.4633969.

[JA] L. K. Jeff Larson Surya Mattu and J. Angwin. *How We Analyzed the COMPAS Recidivism Algorithm*. URL: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (visited on 12/15/2024).

# Appendix A Base Study Details

Figure A.1 demonstrates that while male subjects received marginally higher risk scores on average compared to females, the substantial overlap in score distributions between the sexes suggests that gender alone is not a strong determinant of risk assessment outcomes.

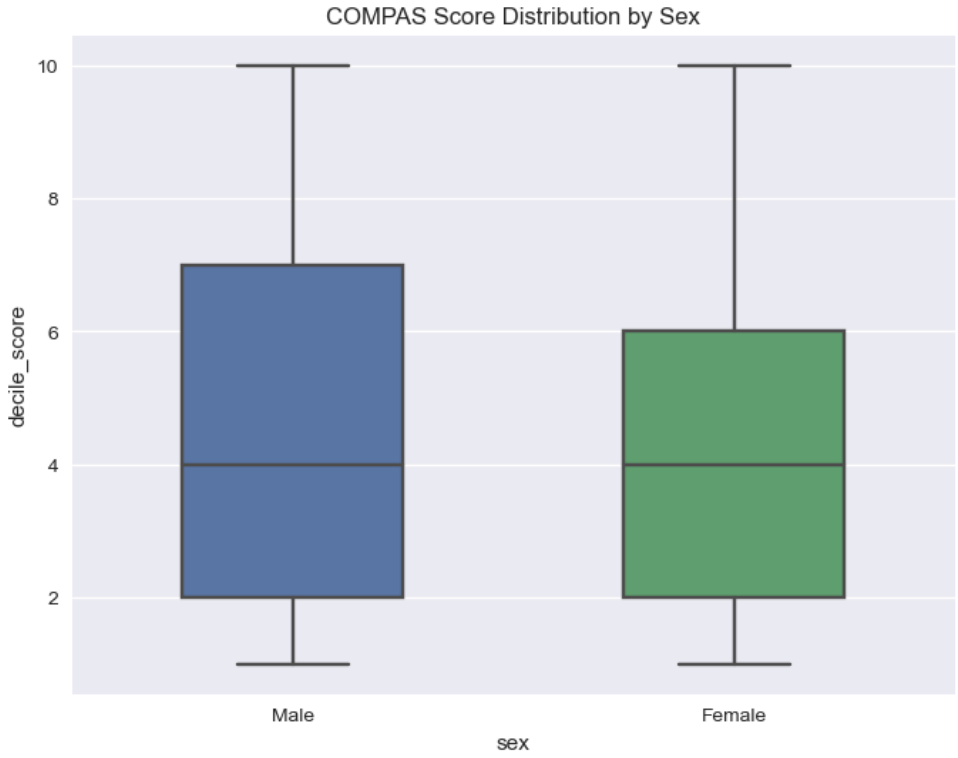



Figure A.1: COMPAS scores of male and female

The predictive model implements binary classification utilizing a feature set comprising demographic attributes (race, sex, age) and criminal history metrics (`c_charge_degree`, `prior_count`, `juv_fel_count`, `juv_misd_count`, `juv_other_count`) as predictor variables. The classification architecture employs an XGBoost algorithm with a binary logistic objective function, with `n_estimators`=200, `max_depth`=3, `learning_rate`=0.1, and complete sampling parameters (`colsample_bytree`=1.0, `subsample`=1.0), while maintaining standard regularization coefficients (`gamma`=0.1, `min_child_weight`=1) to optimize the bias-variance trade-off.

## Appendix B LLM Usage Report

This report used Claude 3.5 Sonnet for text refinement and help with LaTex formatting.