Algorithm Bias in Correctional Offenders Management Profiling for Alternative Sanctions (COMPAS)

William W. Winters

Anderson College of Business and Computing

Regis University

MSDS 640: Ethics, Privacy, and Social Justice in Data Science

Dr. Ghulam Mujtaba

August 22, 2025

Abstract

In recent years the use of algorithms to predict recidivism has grown in the criminal justice system. Writing for ProPublica Angwin et al. (2016) published a study that found the Correctional Offenders Management Profiling for Alternative Sanctions (COMPAS) is racially biased against non-Caucasian offenders. The analysis examined more than 10,000 criminal defendants in Broward County, Florida. The authors published the methods they used to determine this conclusion along with source code and datasets used in the analysis (Larson, 2016); however, there are still controversies over its findings (Wikipedia, 2025). Other analyses have been conducted on validating the accuracy of the algorithms used by COMPAS with various results. This study will use the datasets made available by ProPublica to validate if there is bias in COMPAS algorithms or not. Using the ProPublica data, I was able to verify African Americans are more likely to receive a higher recidivism risk score than other ethnic or racial groups.

Algorithm Bias in Correctional Offenders Management Profiling for Alternative Sanctions (COMPAS)

Introduction

Ho et al. (2024) conducted a study that determined judicial risk assessment tools have a significant impact on a judge's decision-making process and while these tools are advertised as being objective many studies indicate that racial bias has been introduced into the algorithms used by these tools that favor Caucasian offenders over other racial groups. One tool that receives a lot of attention is the Correctional Offenders Management Profiling for Alternative Sanctions (COMPAS) risk assessment system produced and supported by NorthPointe (now Equivant) software. The findings of these studies are mixed with a few pointing out that COMPAS is also gender and possibly age biased, or its predictions favor jailing over release (Engel et al., 2023). Using the dataset obtained from the ProPublica study, I aim to confirm or refute if racial bias exists in the COMPAS algorithm. In addition, a review of current literature on COMPAS, algorithm bias, and the ethical and legal ramifications of using risk assessment algorithms will be conducted. Research into this domain is significant since algorithmic bias can lead to unjustified incarceration of individuals based on membership in certain groups and an erosion of trust in the legal system.

Methodology

For this study a pronged approach will be used. The first one will conduct a statistical analysis of the ProPublica dataset to prove or disprove ProPublica's findings. In the second prong, the dataset will be used to train and test a machine learning (ML) model to determine if such an approach is feasible and more accurate than the COMPAS algorithm when predicting two-year recidivism. Throughout this paper, items and conclusions from the literature review will be reference when appropriate.

Descriptive statistics

After evaluating the datasets available in the ProPublica repository, I selected one that includes a flag indicating whether a defendant committed recidivism within two years of release. This dataset was chosen to allow evaluation and modeling of the two-year recidivism feature. The original dataset consisted of 7,214 records and 53 features. Data type mismatches were handled by converting the columns to the correct data type and missing values were addressed by imputing data when possible and removing features with mostly missing values. After cleaning, the dataset contained 7,214 records and 28 features. The ProPublica source code contained a comment about a `decile_score` of negative 1 (-1) indicating the authors could not find any COMPAS data for that record; therefore, it should be dropped, which was accomplished.

Basic exploratory data analysis (EDA) was conducted to identify any trends within the data and to identify any obvious patterns. Categorical features were examined for any imbalances that may skew the analysis and model training. The only item found was the number of individuals in the *African American* racial group outnumbered the other racial groups. The *Asian* and *Native American* ethnic groups were underrepresented in the dataset, so they were combined into the *Other* group. The table below provides the racial group counts.

Table 1

Racial Group Counts

Racial Group	Number of Individuals
African American	3,696
Caucasian	2,454
Hispanic	637
Other	427

Analysis

In the COMPAS system, the risk score assigned to a defendant is referred to as the *decile score*, which ranges from 1 to 10. This score indicates the likelihood of the individual not appearing in court or reoffending if released into the community. A lower decile score corresponds to a lower risk. Using Bayesian descriptive statistics, I will present an overview of the data used in this analysis and conduct an ANOVA (Analysis of Variance) test to determine whether the mean decile score differs significantly across the racial groups. Furthermore, I will select a machine learning model (ML) and train it to predict the two-year recidivism rate for the populations included in this study.

Modeling

A recent study by Engel et al. (2023) suggested the algorithm used by COMPAS may have an accuracy rate as low as 68% when predicting the likelihood of recidivism within two years. To investigate this further, I aim to use the COMPAS dataset provided by ProPublica and train a machine learning model on it. The goal is to predict the two-year recidivism rate and determine whether a publicly accessible model can achieve an accuracy rate comparable to that of the original algorithm.

For this modeling effort three widely used and publicly available classification models for binary classification tasks were evaluated: Logistic Regression, Support Vector Machines (SVM), and Random Forest Classifiers. Specifically, I used the following implementations from the `scikit-learn` libraries:

- `sklearn.linear_model.LogisticRegression()` for logistic regression
- `sklearn.svm()` for support vector machines
- `sklearn.ensemble.RandomForestClassifier()` for random forest classification

These models were chosen because they are commonly used in binary classification problems, easy to implement, and provide easy to interpret results.

The dataset was cleaned using standard data cleaning methods, and a correlation map was created to identify features that were weakly or strongly correlated to the target class (`two_year_recid`). Features deemed to be too weak or strongly correlated were removed from the dataset, categorical features were enumerated using a label encoder mechanism, and dates converted to Unix Epoch time. Train and test datasets were created from the dataset using an 80/20 split and all values were scaled using a standard scaler algorithm. The scaled data was used to train and test the logistic regression and SVM models. Random Forest Classifiers do not require scaled data, so it was trained using the original train/test datasets. Results from each model were acceptable and all performed within a few percent of each other. The table below shows the results.

Table 2

Model Scores

Model	Train	Accuracy	TPR	FNR
Logistic Regression	0.88	0.89	0.92	0.56
Support Vector Machine	0.90	0.89	0.96	0.54
Random Forest Classifer	0.91	0.89	0.99	0.52

Since the performance scores of the models being evaluated were remarkably close, a K-Fold cross-validation test was conducted using Negative Mean Squared Error (NMSE) as the evaluation metric. In this context, the model that yields the score closest to zero is typically considered the best performer. The table below presents the results of the K-Fold cross-validation test:

Table 3

NMSE Scores and Standard Deviations

Model	NMSE	Standard Deviation
Logistic Regression	-0.296990	0.121847
Support Vector Machine	-0.296833	0.014517
Random Forest Classifier	-0.101370	0.007385

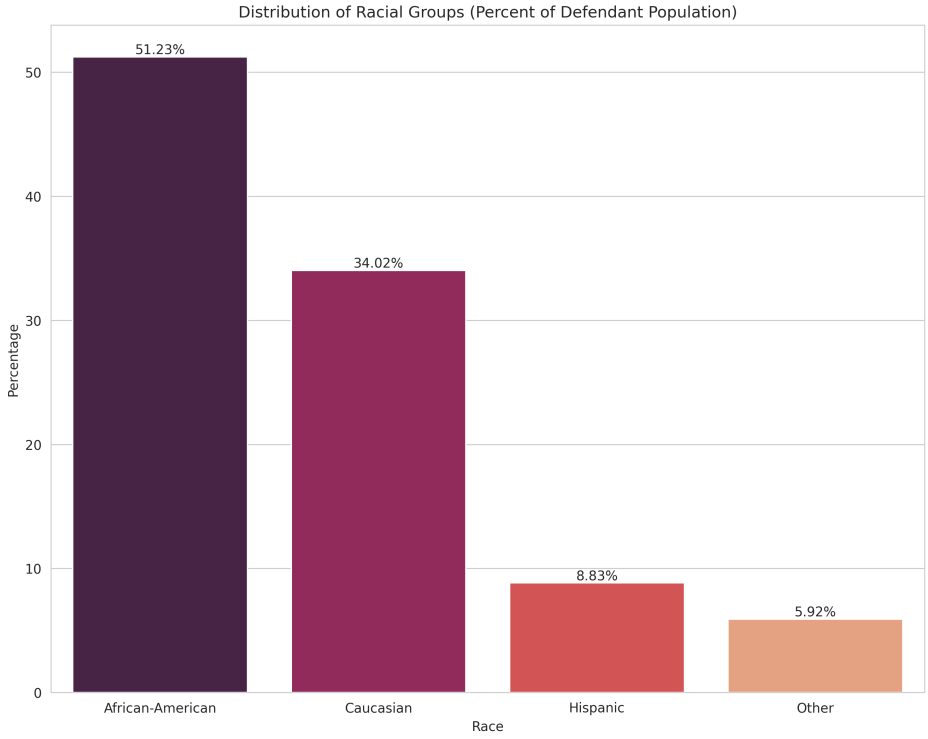
The Random Forest Classifier model was selected for further evaluation and testing based on its NMSE score. According to Engel et al. (2023), the AUC (Area Under the Curve) has become the preferred measure of accuracy because it provides a single, threshold-agnostic value that summarizes the model's overall performance across all possible thresholds. As such, the AUC for the Random Forest Classifier was calculated, and the result was an AUC of **0.97**. This indicates the model makes highly accurate predictions from the test data.

Results

My analysis found that recidivism is higher within the African American population included in this study; however, this group also made up most of the defendant population. After moving Asians and Native Americans into the Other group, the percent of the different racial groups in this study were calculated and are illustrated in the figure below.

Figure 1

Distribution of Racial Groups (Percent of Population)



African Americans in this study comprise approximately 17% more of the defendant

population than the next largest group, Caucasians. The reasons behind this disparity are beyond the scope of this paper. However, I reviewed the demographics of Broward County, Florida, where this data was collected and according to the U.S. Census Bureau (2025), African Americans made up 28.6% of Broward County's population in 2016, while they accounted for over 51% of defendants in this study. In contrast, Caucasians comprised nearly 61% of the county's population but represented only about 34% of defendants. While the COMPAS algorithm may contain biases, it is essential to recognize that the disparity observed in this study begins well before individuals enter the judicial system.

The purpose of this analysis is to verify or refute the findings of ProPublica's study on COMPAS, specifically its claims about bias. One key metric used by ProPublica to support its assumption of bias was the distribution of decile scores. According to ProPublica's 2016 report, African Americans tend to have higher decile scores compared to other racial groups Angwin et al. (2016). After re-examining the data, I found that this pattern holds true. The table below presents the mean decile score for each racial group in my analysis.

Table 4

Mean Decile Scores

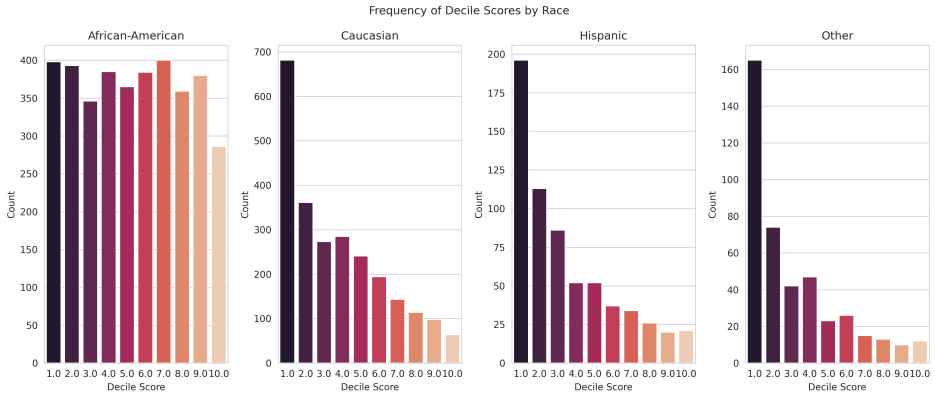
Racial Group	Mean	Standard Deviation
African American	5.37	2.83
Caucasian	3.74	2.60
Hispanic	3.46	2.60
Other	3.08	2.48

To determine whether the differences in decile scores between African Americans and other groups are statistically significant; I used a statistical test called one-way ANOVA (*Analysis of Variance*). Before performing the test, I normalized the decile scores using the *Min-Max* method to ensure that all scores were on the same scale. The results of the one-way ANOVA test showed a P-value below 0.05 (`P-value = 8.089670291495388e-161`), which indicates that the null

hypothesis can be rejected. This means that it is statistically unlikely that the observed differences in mean decile scores between African Americans and other groups are due to chance alone. Therefore, I can conclude with confidence that the means of two or more groups are statistically different. The graph below illustrates the distribution of decile scores across different racial groups, revealing a clear pattern: African Americans tend to receive higher decile scores compared to other racial groups. This graph provides a visual representation of the data, making it easier to understand and compare.

Figure 2

Frequency of Decile Score by Race



The findings from my analysis align with those reported by Angwin et al. (2016) in ProPublica, providing further evidence that supports their observations.

Modeling results

While the Random Forest Classifier performed well, further evaluation of it indicated it may be slightly over fitting the data. To counter this a `GridSearchCV()` algorithm was applied to the model to identify the optimal hyperparameters to use in tuning it. After applying the recommended hyperparameters from the `GridSerachCV()` algorithm, the model was trained again and it performed around the same as before, but the indications of over fitting were eliminated and its TPR increased from 0.91 to 0.99. It's before and after scores are displayed below.

Table 5

Model Metrics Before and After Tuning

Metric	Before	After
Train Score	1.00	0.91
Accuracy Score	0.91	0.89
TPR	0.95	0.99
FNR	0.55	0.52

In addition to the scores above, the model also achieved an AUC of 0.97. All of the observed metrics confirm the exceptional performance of this model and suggests a widely available machine learning model can be trained to accurately predict the two-year recidivism rate. Notably, this model's performance is comparable to or even exceeds that of the original COMPAS one.

Discussion and Conclusions

My analysis of the Broward County, Florida's COMPAS data confirms the pattern identified by Angwin et al. (2016) and COMPAS does have a bias toward awarding higher decile scores to African Americans; however, this group of individuals enter the judicial system in far greater numbers than other racial groups. Based on U.S. Census Bureau (2025) data, African Americans comprised a little less than 29% of the county's population in 2016 but accounted for over 51% of the defendant population. While the COMPAS algorithm may contain biases, it is important to node the disparity begins well before the person enters the judicial system.

In a study by Engel et al. (2023), they found that COMPAS is not necessarily biased toward any one racial or gender group but is more likely to recommend jailing a defendant who would not have recidivated, instead of recommending other options such as bail, electronic monitoring and so on. The authors also highlighted that COMPAS decile scores are calculated in relation to the group the defendant belongs to. These groups may consider gender, and whether a defendant has been in prison, on parole, in jail or on probation (Engel et al., 2023). They continue their study by

recommending a method to remove the bias from the algorithm using a publicly available machine learning model.

Many studies have found bias in the algorithms used in judicial risk assessment tools; however, Ho et al. (2024) argues that using AI tools can reduce gender bias since human judges tend to be more lenient with female offenders and the tools are not. While the authors noted that these tools help alleviate gender-based disparity, they did find evidence of racial bias favoring White offenders over Black ones with the recommendations they make.

This case study confirmed racial bias in the COMPAS algorithm and identified that African Americans enter the judicial system at greater numbers than other groups. A review of current literature supports this study's conclusions and highlights the limitations of using algorithms in the judicial context. Other studies have also identified additional biases such as gender and age. My study was very narrow in its scope and only examined racial bias and two-year recidivism predictions. Further research should be conducted to identify other biases and their effects on judicial decision-making and recidivism predictions.

References

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

Engel, C., Lindhardt, L., & Schubert, M. (2023, Dec). Code is law: how COMPAS affects the way the judiciary handles the risk of recidivism. *Artificial Intelligence and Law*, *33*(2), 383-404. doi: 10.1007/s10506-024-09389-8

Ho, Y.-J. I., Jabr, W., & Zhang, Y. (2024). AI Enforcement: Examining the impact of AI on judicial fairness and public safety. doi: 10.2139/ssrn.4533047

Larson, J. (2016). *compas-analysis.* https://github.com/propublica/compas-analysis. (Accessed: 2025-07-12)

U.S. Census Bureau. (2025). *Census bureau quickfacts: Broward county, florida.* Retrieved from https://data.census.gov/table/ACSCP1Y2016.CP05?q=broward+county,+florida+demographics&y=2016

Wikipedia. (2025). Compas (software). Retrieved from https://en.wikipedia.org/wiki/COMPAS_(software)