

Bias & Fairness in Justice System

Tuan Nguyen



COMPAS in the Criminal Justice System

Motivation:

- U.S. courts use algorithmic risk assessments (like COMPAS) to assist in decisions about bail, sentencing, and parole.
- These decisions directly affect human freedom, so bias is unacceptable.
- The ProPublica investigation (2016) revealed potential racial disparities in COMPAS predictions.

Main Idea:

- Analyze fairness in the public COMPAS dataset:
 - Train classifiers similar to COMPAS's task (predict 2-year recidivism).
 - Measure group disparities in predictions (Black and White defendants).

Summary:

- Reproduced the well-known issue:
 - Higher false positive rates for Black defendants.
 - Higher false negative rates for White defendants.
- Demonstrates how algorithmic bias can reinforce existing inequalities in the justice system.



COMPAS Background

What is COMPAS?

- A proprietary risk assessment tool used in U.S. courts.
- Predicts likelihood of 2-year recidivism.
- Produces a decile score (1–10) used in bail & sentencing decisions.
- Opaque (“black-box”), privately owned (Northpointe).



ProPublica Investigation (2016)

Key findings:

- Investigated COMPAS on Broward County data.
- Black defendants were twice as likely to be mislabeled as “high risk.”
- White defendants more often mislabeled “low risk.”
- Evidence of false positive and false negative disparities.

Source: [ProPublica](#)



Academic Fairness Research

- Chouldechova (2017): COMPAS cannot be simultaneously fair across all metrics.
- Fairness definitions: Demographic parity, equalized odds, equal opportunity.
- Why this matters to justice system: even small bias = human rights risk

Source: <https://pubmed.ncbi.nlm.nih.gov/28632438/>



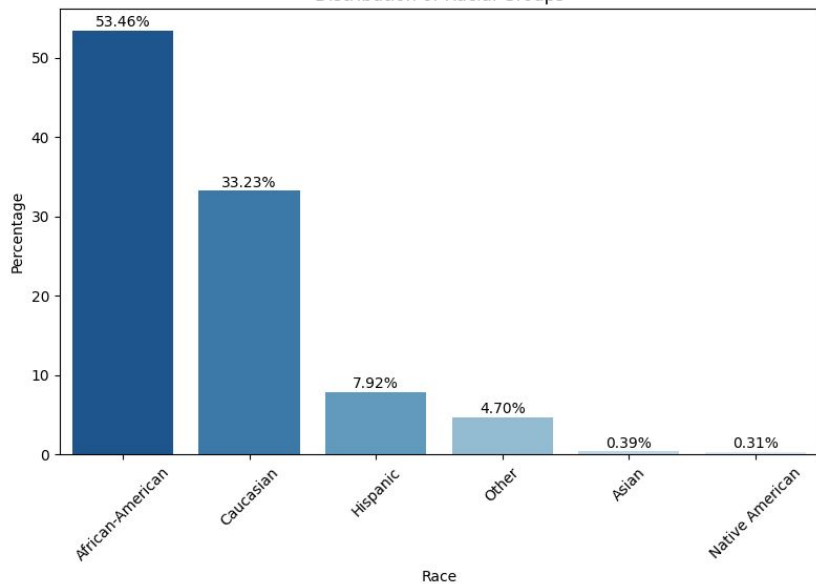
Dataset & Task

- Public COMPAS Recidivism Racial Bias dataset - [Kaggle](#)
- 18316 data points with 52 features
- Prediction target: Two-year recidivism (binary)
- Selected features: age, priors_count, charge_degree, sex, race, etc.
- Preprocessing: cleaning, filtering, dealing with missing data.
- Models trained:
 - Random Forest
- Train/test split: 80/20

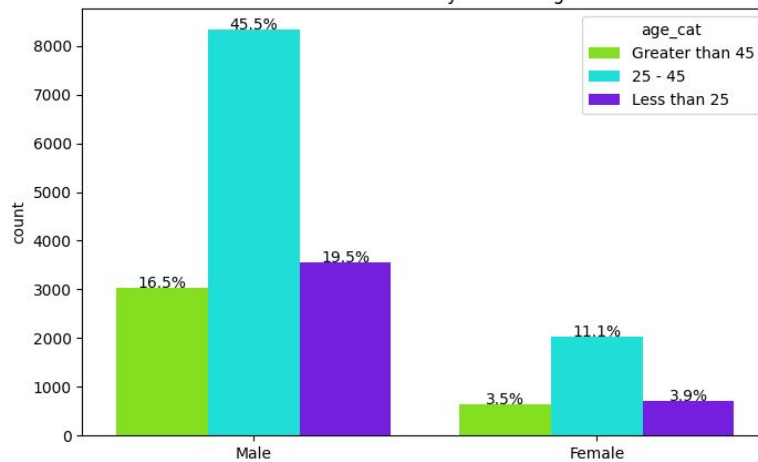


Pre-Analysis

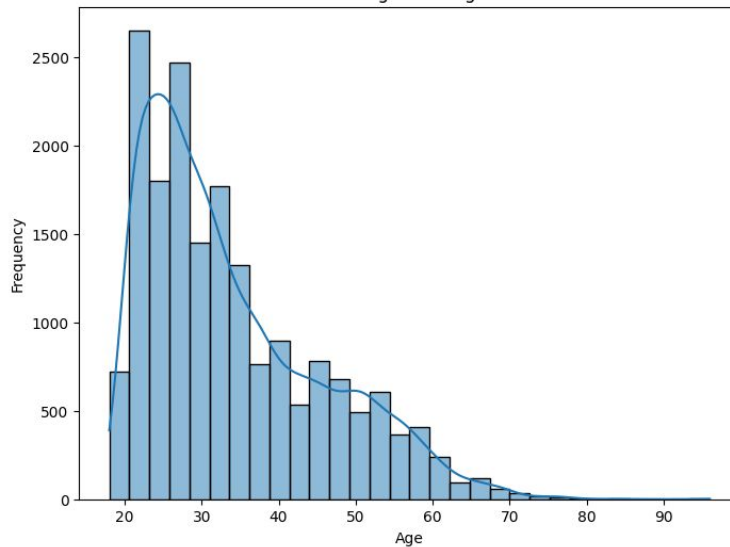
Distribution of Racial Groups



Number of cases by sex and age

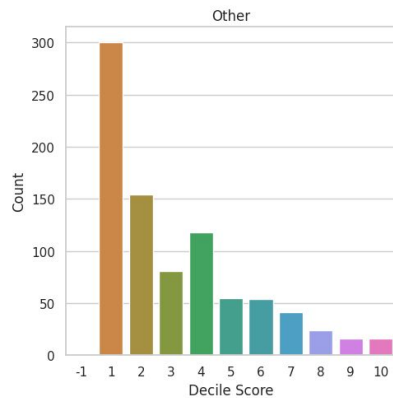
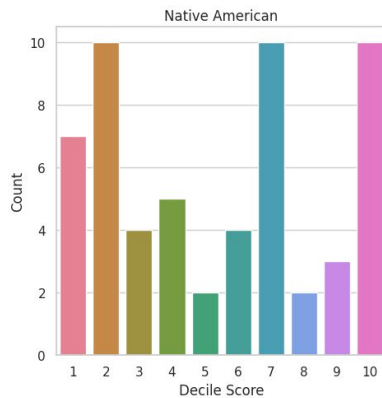
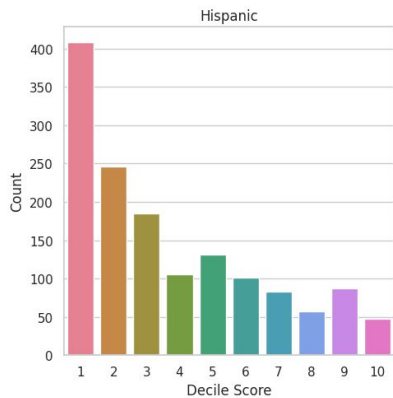
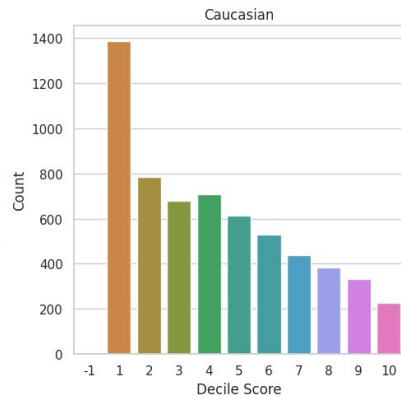
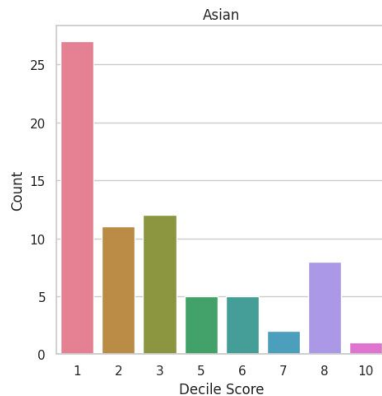
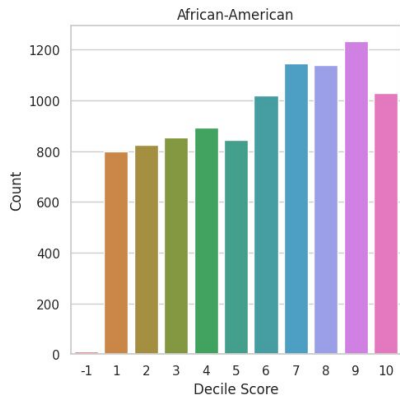


Histogram of Age





Pre-Analysis





Modeling Result

The Random Forest model demonstrates relatively high accuracy across all racial groups, with accuracies ranging from **0.79 to 0.91**. This suggests that the model performs well in predicting outcomes for individuals from diverse racial backgrounds.



Identifying Bias in the Model's

African-American:

- True Positive (TP): 745
- False Positive (FP): $16 + 70 = 86$
- True Negative (TN): $54 + 938 = 992$
- False Negative (FN): $22 + 114 = 136$

Disproportionate Impact: The false positive rate (86) for African-Americans appears relatively high compared to the true positive rate (745), indicating potential bias in false predictions, particularly false positives.

Confusion Matrix for African-American

True labels	Predicted labels		
	0	1	2
0	50	16	4
1	22	745	114
2	6	70	932



Identifying Bias in the Model's

Caucasian:

- TP: 573
- FP: $16 + 70 = 86$
- TN: $54 + 938 = 992$
- FN: $22 + 114 = 136$

Disproportionate Impact: The false positive rate (86) for Caucasians is relatively high compared to the true positive rate (573), indicating potential bias in false predictions, particularly false positives.

Confusion Matrix for Caucasian

True labels	Predicted labels		
	0	1	2
	0	1	2
0	31	28	4
1	11	573	50
2	0	43	478



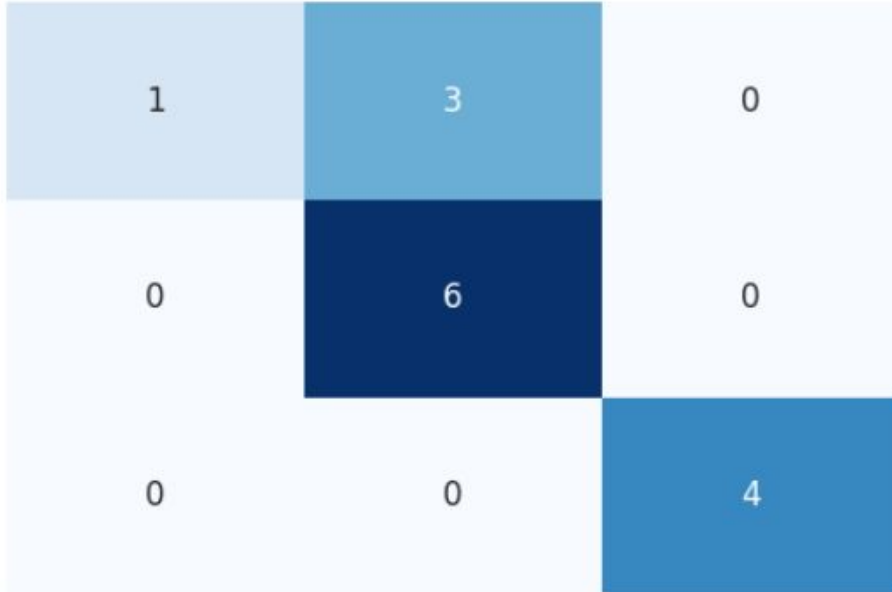
Identifying Bias in the Model's

Asian:

- TP: 6
- FP: 3
- TN: $1 + 4 = 5$
- FN: 0

Disproportionate Impact: The sample size is small, but the model appears to have a low false positive rate, which is a positive sign.

Confusion Matrix for Asian



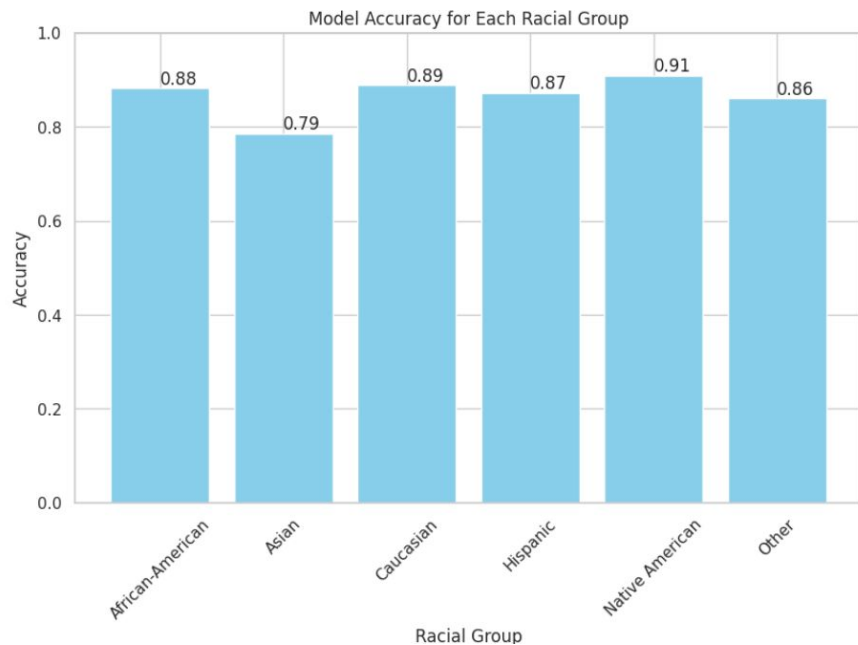
A confusion matrix for the 'Asian' category. The matrix is a 3x3 grid where rows represent 'True labels' (0, 1, 2) and columns represent 'Predicted labels' (0, 1, 2). The cells contain counts of instances. The diagonal elements (1, 6, 4) represent correct classifications. The off-diagonal elements represent misclassifications. The colors of the cells are: (0,0) light blue, (0,1) medium blue, (0,2) very light blue, (1,0) very light blue, (1,1) dark blue, (1,2) very light blue, (2,0) very light blue, (2,1) very light blue, (2,2) medium blue.

True labels	0	1	2
0	1	3	0
1	0	6	0
2	0	0	4
	0	1	2



Racial Group Modelling Visualisation

This visualization allows me to easily compare the performance metrics of the racial group based on their accuracy, the chart also shows that Native American and Caucasian have the best model with the highest value of model accuracy.





Summary

In summary base on the dataset, the analysis suggests potential bias in false predictions (false positives) for African-American, Caucasian, and Other racial groups. Additionally, caution should be exercised when interpreting results for racial groups with small sample sizes, such as Asian and Native American, due to limited data.



Setup System Video

<https://youtu.be/BmMsFNlkyq8>



Reference

COMPAS Dataset: [Kaggle](#)

Propublica - COMPAS Analysis:

<https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>

Chouldechova (2017): <https://pubmed.ncbi.nlm.nih.gov/28632438/>

Random Forest with COMPAS: https://pbiecek.github.io/xai_stories/story-compas.html