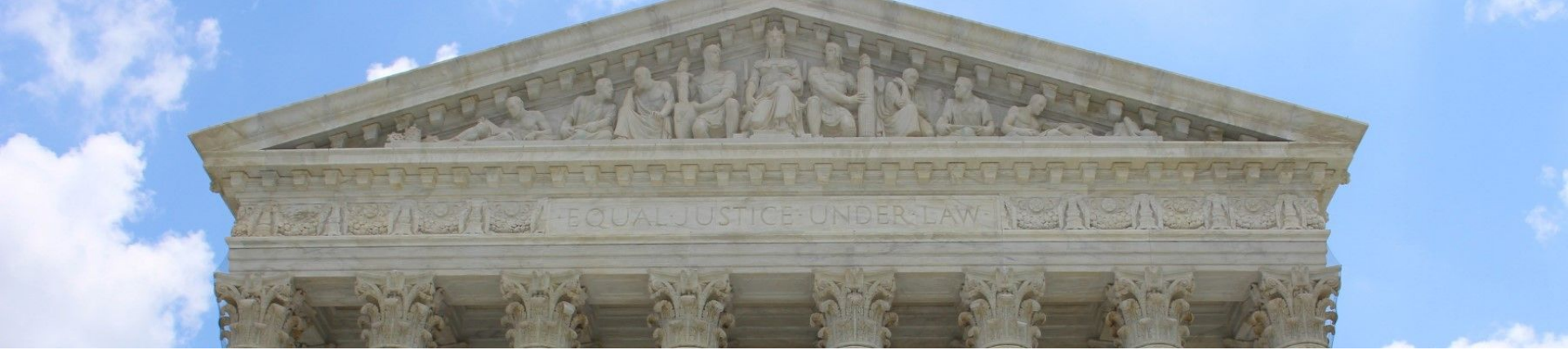


System

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).
  - Use SHAP to explain which features drive "high risk" predictions.

Summary:

- Reproduced the well-known issue:
  - Higher false positive rates for Black defendants.
  - Higher false negative rates for White defendants.
- SHAP shows certain features (e.g., priors count, age, charge degree) contribute significantly and may act as proxies for race.
- 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.
- Used for bail, sentencing, parole decisions.
- Opaque (“black-box”), privately owned (Northpointe).

ProPublica Investigation (2016)

Key findings:

- 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.
- Dataset released publicly → widely used in fairness research.

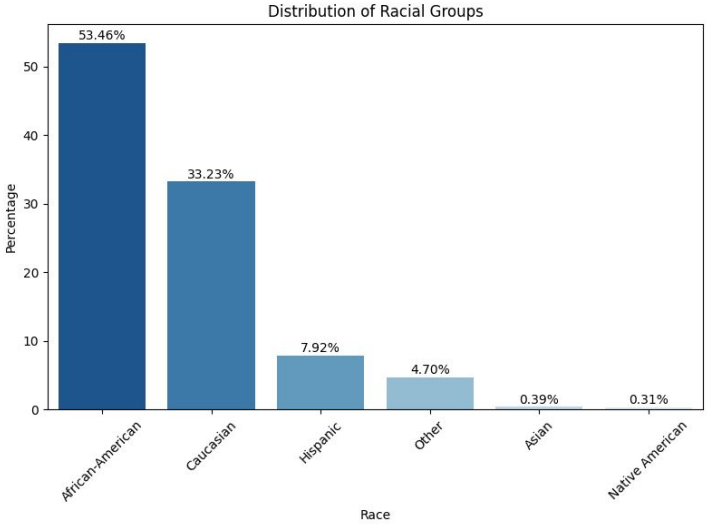
Academic Fairness Research

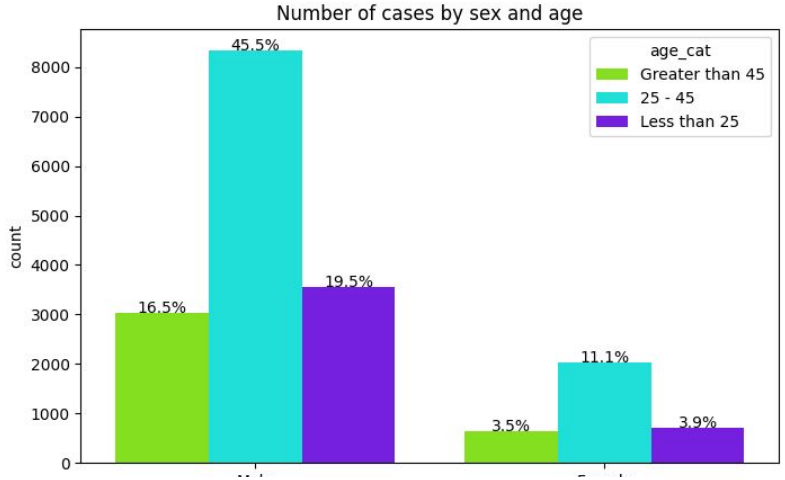
- Kleinberg et al. (2016/2017): fairness impossibility theorem
  - Cannot satisfy calibration + equalized odds simultaneously when base rates differ.
- Chouldechova (2017): COMPAS cannot be simultaneously fair across all metrics.
- Fairness definitions: Demographic parity, equalized odds, equal opportunity.
- Why this matters to justice system: even small bias = human rights risk

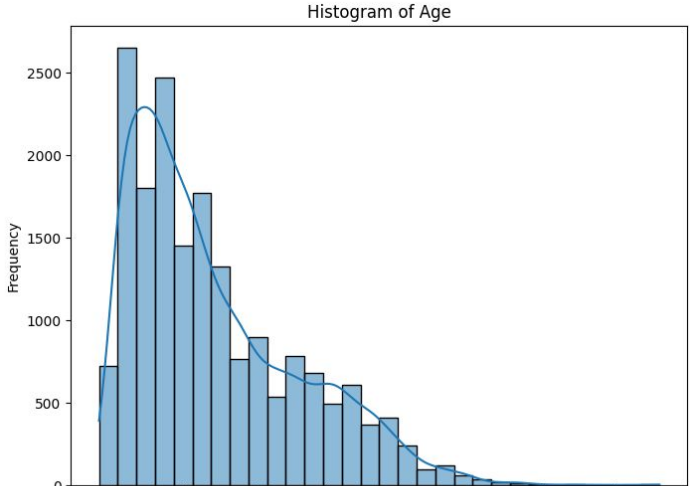
Dataset & Task

- 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: 70/30

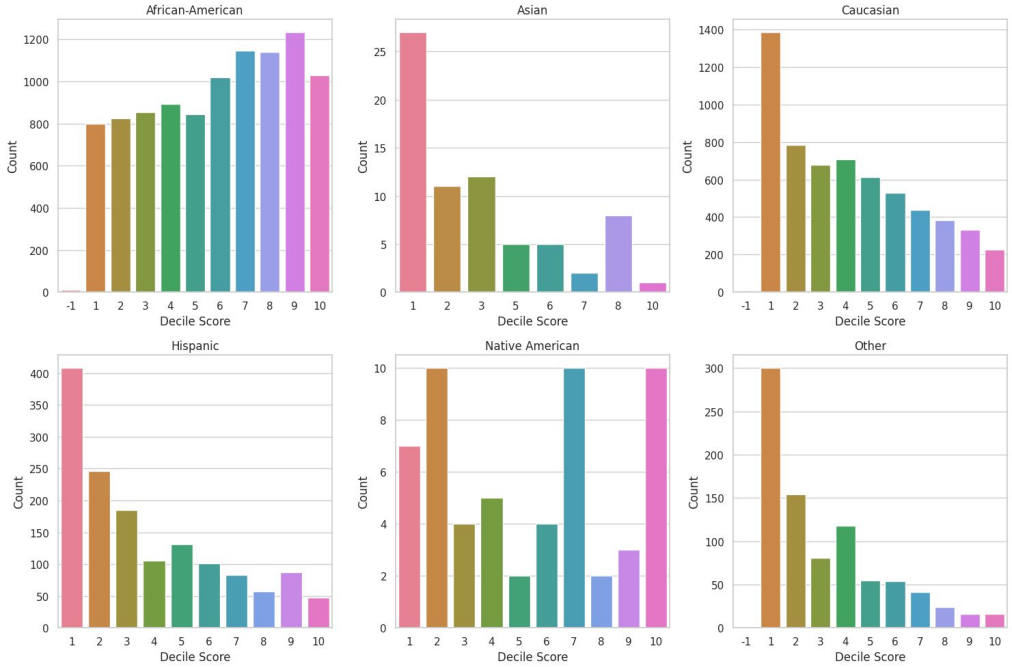
Pre-Analysis







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