FAIRNESS AND BIAS IN DATA ANALYTICS & AI

Dan IANCU

Hands-on Exercise

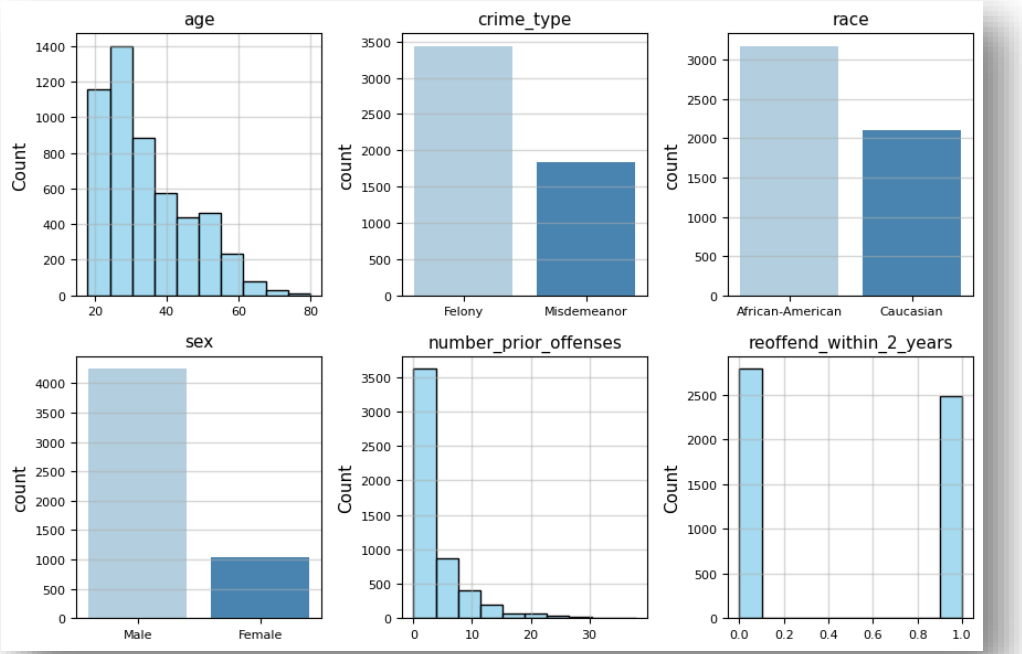
Survey

- Public data obtained by ProPublica
  - 5,278 individuals from Broward County, FL
  - Each got COMPAS score in 2013-4 and a full 2-year recidivism history is available
    - COMPAS score is 1-10, with 10 most risky
    - Also a risk label: “Low Risk” if score <= 4, “High Risk” otherwise
  - *Data far from perfect, but let’s assume it’s representative*

- **Goal:**
  1. **Understand the data** and identify potential **sources of bias**
  2. Build our own AI model – a **Decision Tree** – and assess it for bias
  3. Evaluate **a proprietary/black-box model** (COMPAS) for bias

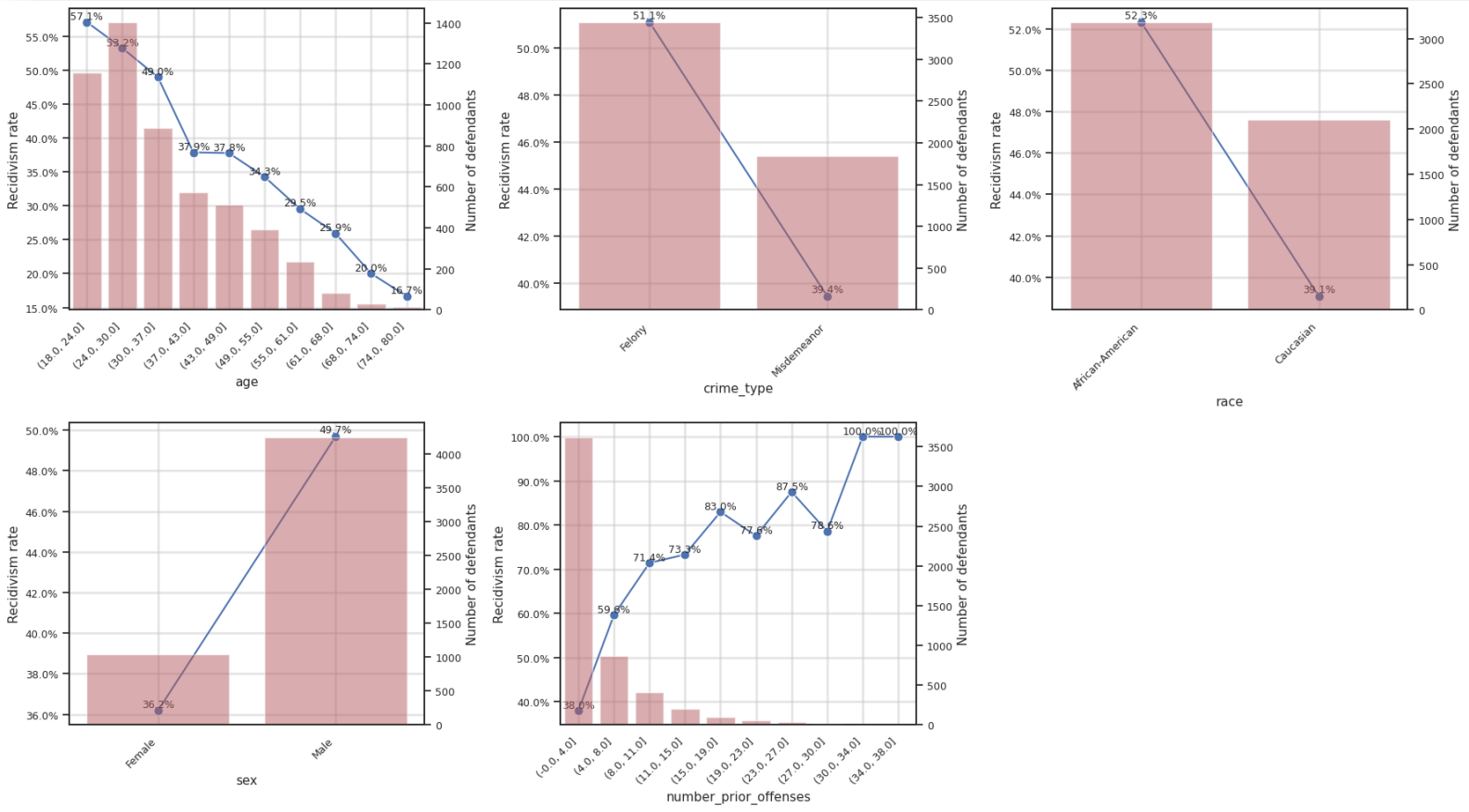
Part 1: Understand the Data

- There are 6 data features
  - ***age*** *:* the defendant's age
  - ***crime_type*** *:* type of crime for most recent arrest; values: **Misdemeanor** or **Felony**
  - ***race*** *:* two values in data, **African American** or **Caucasian**
  - ***sex*** *:* two values in data, **Male** or **Female**
  - ***number_prior_offenses*** *:* total number of prior convictions for the defendant
  - ***reoffend_within_2_years*** *:* whether defendant actually reoffended within 2 years (**1** or **0**)



Part 1: Relationship with (Two-Year) **Recidivism**

Q: How would you characterize the relationship between **recidivism** and these features?
(please answer in the poll!)

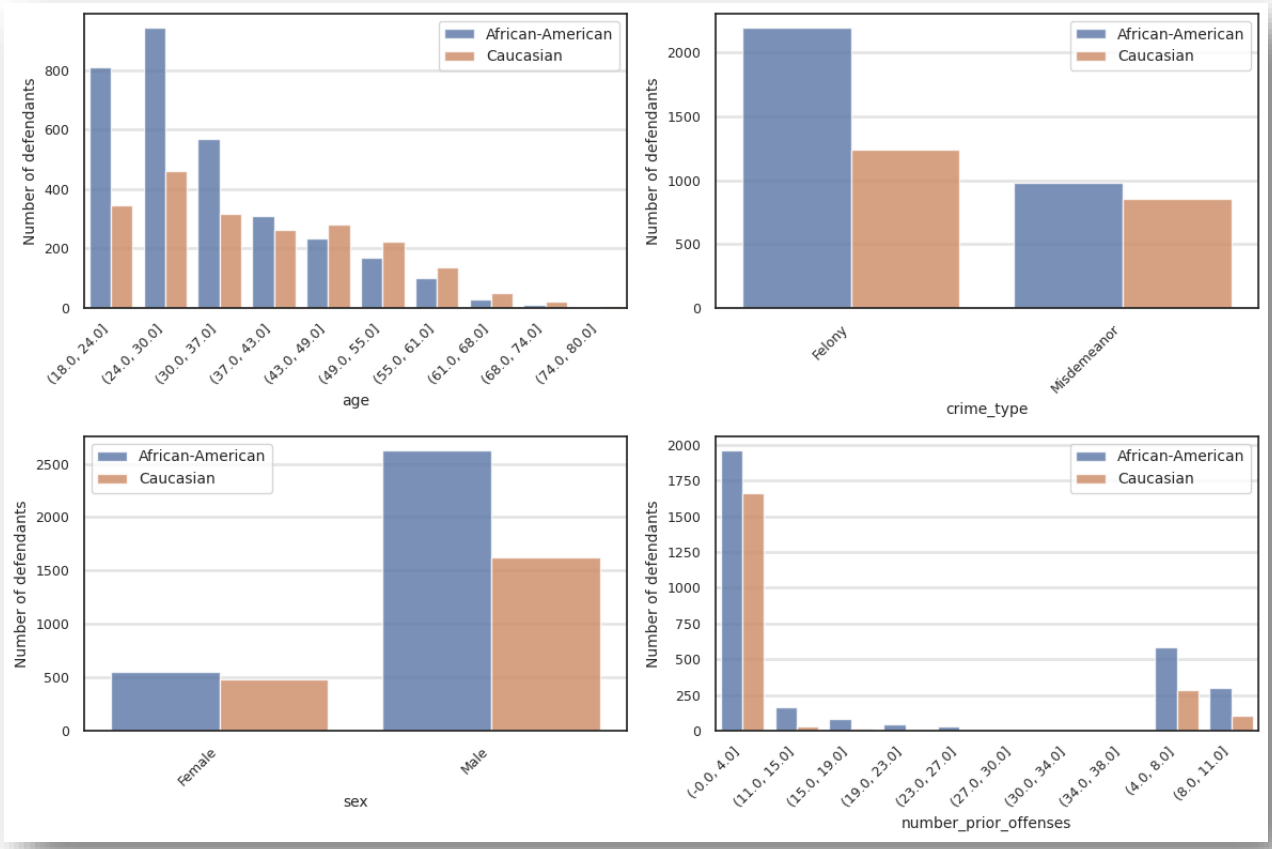


Clarification: The line plot corresponds to the left axis (**recidivism rate**) and the bars correspond to the right axis (**number of defendants**). The recidivism rate is calculated as the fraction of defendants in the respective category who reoffended within two years (i.e., reoffend_within_2_years = 1 in the data).

Example: in the plot corresponding to age, among the defendants with age 18-24 y.o., the recidivism rate was 57.1%.

Part 1: Relationship with **Race**

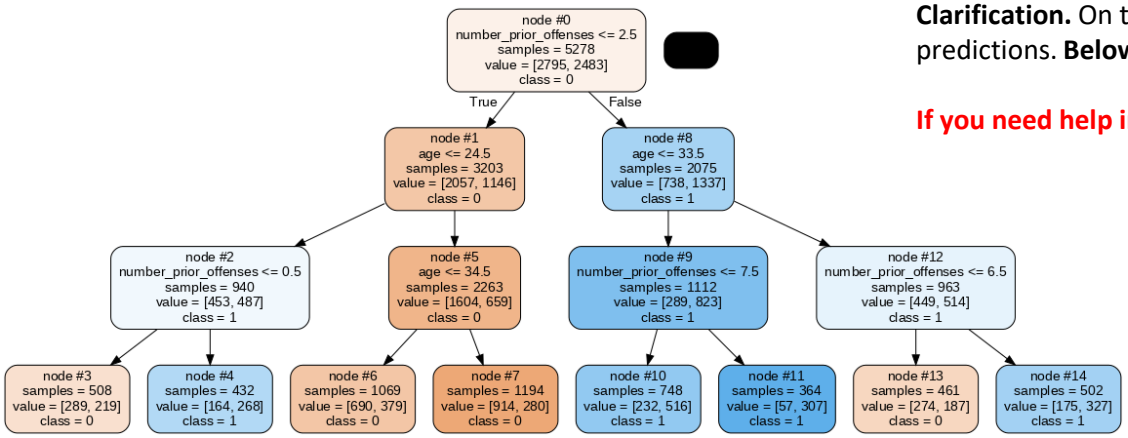
Q: How would you characterize the relationship between **race** and these features?
(please answer in the poll!)



Clarification: The bar plots show the number of defendants from each race in the respective category.
Example: In the plot corresponding to age, among the defendants with age 18-24 y.o., there are 809 African-American and 347 Caucasian defendants.

Part 2. An Interpretable ML Model

- This ML model is trained with all the data, including with protected features such as **race** and **sex**

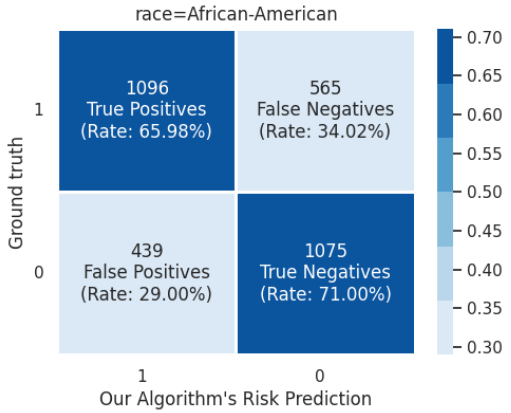


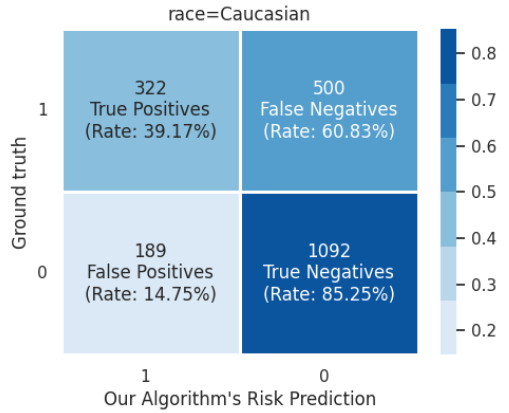
Clarification. On the **left**, you can see the way the model makes predictions. **Below**, you can see the model's performance on the data.

If you need help interpreting these, check the next two pages!

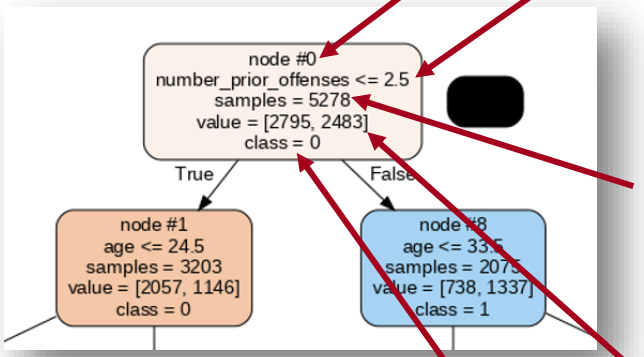
Q: Is this model exhibiting racial bias?

(please answer in the poll!)





How To Interpret the Tree?



- "node #0" is a **unique identifier** for the node
- second line has a logical condition comparing a data **feature** with a certain **threshold**; the left subtree contains all the data where the condition is "**True**" and the right subtree contains the rest ("**False**")

  *Example: in node #0, we check if "*number_prior_offenses <= 2.5*". Data satisfying this are in node #1, the other data are in node #8.*
- **samples** tells how many data samples fall in that node

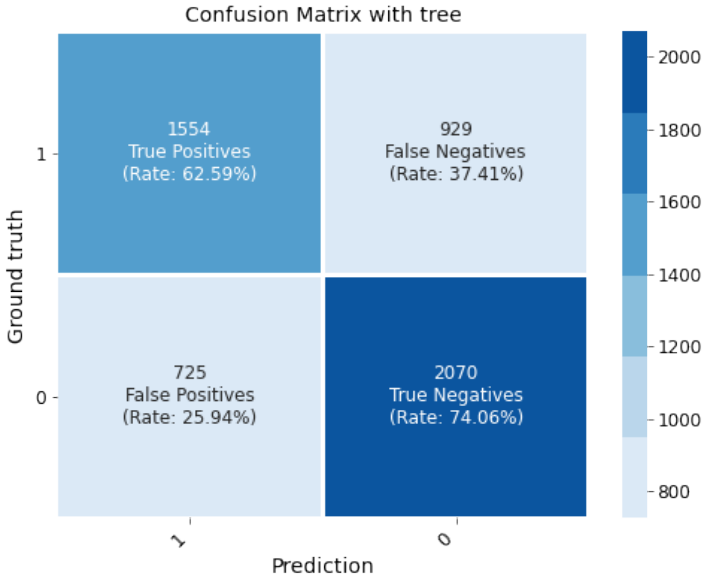
  Example: node #0 contains 5,278 samples (the entire data), node #1 contains 3,203 samples (i.e., all the data satisfying "number_prior_offenses <= 2.5)
- **value** tells how many samples take value 0 and 1 for the predicted target, respectively

  Example: in node #0 we have 2,795 samples with reoffend_within_2_years=0 *and 2,483 samples with* reoffend_within_2_years=1. N*ote that the sum of these equals the total number of samples in node #0, namely 5,278.*
- **class**: what value for the predicted target is the majority one (0 or 1). This is also indicated by the color-coding of nodes: orange means majority 0 and blue means majority 1, and the deeper the color the heavier the majority.

  E.g., in node #0 we have more data with the *reoffend_within_2_years=0 (namely, 2,438 samples) than with reoffend_within_2_years=1 (namely, 2,483), so class=0 to indicate the majority, and the node is colored in a light shade of orange*

How To Interpret the Confusion Matrix?

- Quality of predictions typically summarized with a **confusion matrix**:



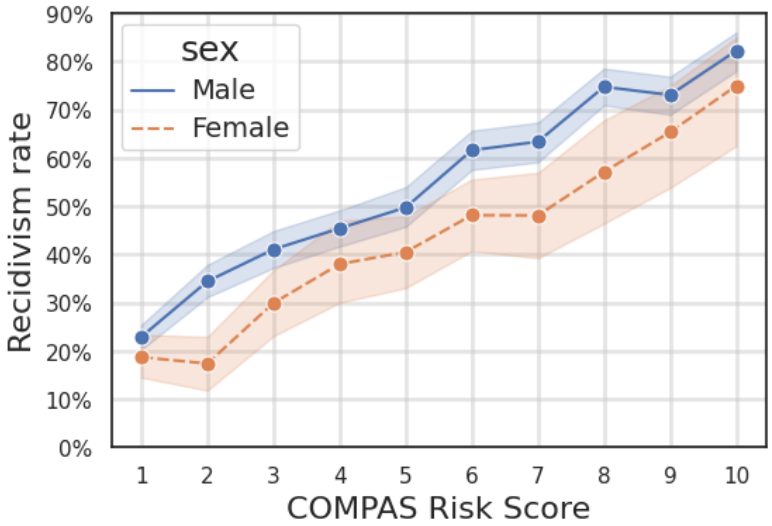
Ground truth: what happened in reality
Prediction: what the algorithm predicted/thought would happen

True: prediction matches reality (the algorithm is right)
False: prediction does not match reality (the algorithm is wrong)
Positive: prediction = 1
Negative: prediction = 0

The **rates** are calculated as a fraction of the total on the row.

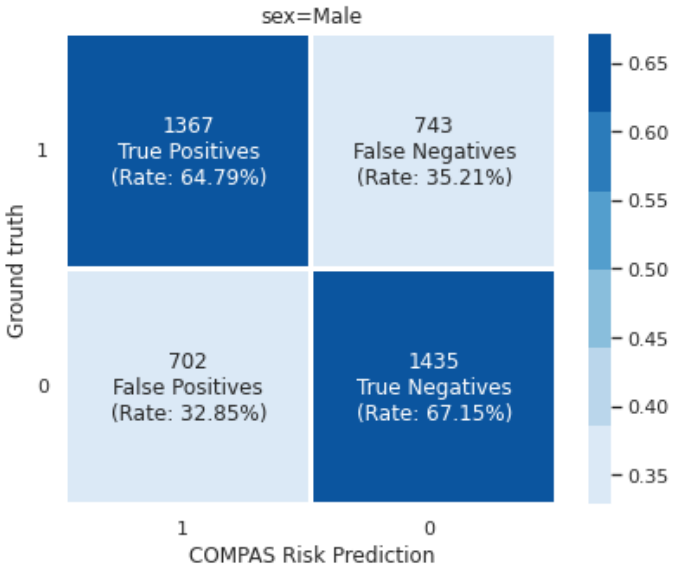
Example. There are 1,554 true positives, meaning this is the number of cases when the algorithm predicts a defendant would reoffend (Prediction=1) and that defendant actually reoffends (Ground Truth=1). Because the total number of defendants who reoffend is 1,554 + 929 (i.e., the total number of cases where Ground Truth=1), the rate of true positives is 1554/(1554+929) = 62.59%.

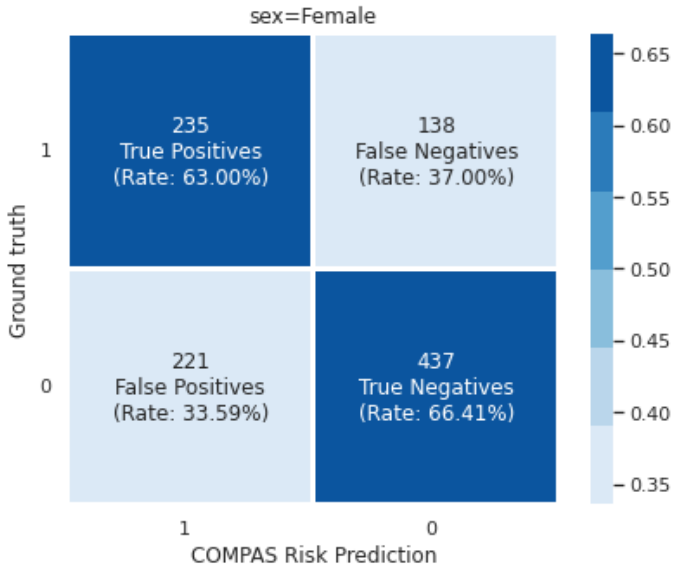
Part 3-a. Analyzing **COMPAS** for **Gender Bias**



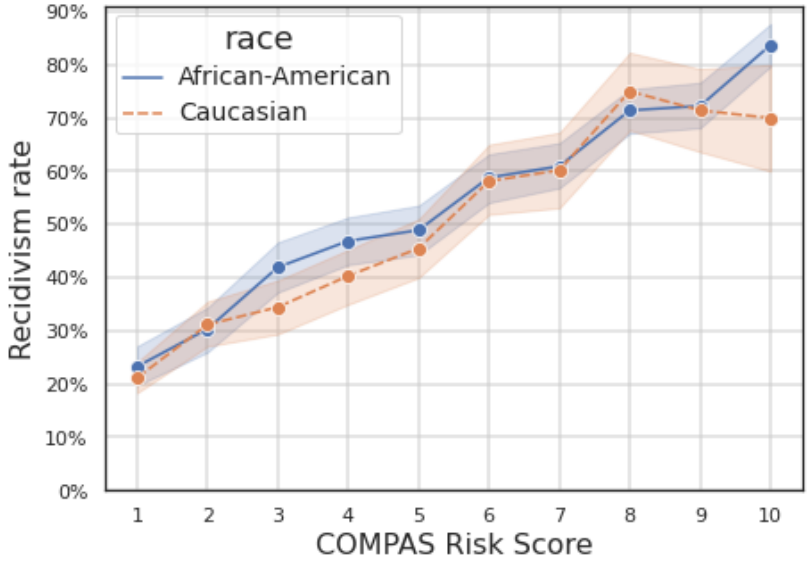
Q: Is COMPAS exhibiting **gender bias**?

(please answer in the poll!)





Part 3-b. Analyzing **COMPAS** for **Racial Bias**



Q: Is COMPAS exhibiting **racial bias**?

(please answer in the poll!)