Inputs:

- processed_compas.csv — preprocessed COMPAS dataset.
- model_processed_compas.h5 — trained Keras model (binary classifier).

Perturbations:

- Random Gaussian noise (5% scale, 30% chance per feature) is applied to non-sensitive features to avoid overfitting to unrealistic exact duplicates

Baseline Tool:

- Assigns a different race randomly
- Considers a case biased only if the rounded prediction changes.
- Also collects signed prediction differences for Wilcoxon signed-rank test (tests whether the median difference is statistically != 0).

Improved Tool:

- Uses a greedy heuristic to flip the race attribute to the value that maximises prediction difference.
- A case is marked as biased if the absolute difference exceeds a threshold (default: 0.05).
- Performs Wilcoxon test on signed differences to assess systematic directional bias.

Outputs:

- IDI Ratio: proportion of test cases detected as biased.
- Wilcoxon p-value: significance of bias direction.
- Average / Median Signed Difference: used to interpret bias consistency.

The histogram in the report was created using the *signed_differences* array aggregated by the baseline and improved tools. Since both arrays come from different scripts, the graph was created in a new script by just pasting over each of the distributions from the fairness tools.