Investigating Bias in AI-based Risk Estimation

1. Introduction: -

1.1. Input Data Utilized in the Algorithm and Its Application

1.1.1. Background: -

The COMPAS dataset was collected and analyzed by ProPublica with the main purpose of evaluating the fairness of the COMPAS algorithm, a risk assessment tool used in the U.S. criminal justice system to predict recidivism. This dataset includes information on 7,214 defendants, focusing on predictions of committing a crime again within two years of the initial assessment.

This dataset served as a foundation for training and validating the COMPAS tool, designed to predict the likelihood of recidivism. The tool's predictions have been utilized by judges to make critical decisions regarding bail, sentencing, and parole, significantly influencing judicial outcomes and the administration of justice.

1.1.2. About the COMPAS Dataset: -

This algorithm assesses risk using a range of input data. It incorporates demographic features such as race (categorized into groups like Black, White, Hispanic, etc.), sex (Male or Female), and age (recorded both as a continuous variable and in categories such as below 25, 25–45, and above 45).

It also considers criminal history, including the number of juvenile felonies and misdemeanors (juv_fel_count, juv_misd_count), other juvenile charges (juv_other_count), and the total count of prior offenses (priors_count).

And finally, The algorithm predicts the outcome variable *is_recid*, which indicates whether an individual reoffends within two years. Additionally, it generates a **COMPAS score** (decile_score), a risk assessment rating ranging from 1 (low risk) to 10 (high risk).

1.2. Ethical Discoveries in Literature

Over time, the algorithm gained a lot of attention and public scrutiny following an investigation by ProPublica, which alleged that its results exhibited racial bias. According to the report, the algorithm appeared to disproportionately overestimate the likelihood of recidivism for Black defendants while underestimating it for White defendants, raising concerns about fairness and equity in its application.

ProPublica's analysis highlighted significant racial disparities in the COMPAS algorithm's predictions. Black defendants were disproportionately more likely to be incorrectly classified as high risk, while White defendants were more frequently misclassified as low risk. These findings suggest systemic biases in how the algorithm assesses recidivism risk across different racial groups.

1.2.1. Ethical Concerns: -

The use of the COMPAS algorithm raises several critical ethical issues. One major concern is that COMPAS is a private system, so its methods and how it makes decisions are hidden from the public. This lack of transparency makes it challenging for independent researchers, policymakers, and even stakeholders to assess or enhance the algorithm's fairness and accuracy. Without access to the algorithm's code or training data, questions about the validity of its predictions and potential biases remain unanswered, risking public trust.

A deeper ethical challenge lies in the inherent trade-offs within fairness metrics. For example, striving to ensure that individuals from all groups have the same likelihood of being incorrectly classified as high risk may accidently create disparities in the rate at which truly high-risk individuals are misclassified as low risk. These trade-offs are often unavoidable due to statistical principles, such as the "impossibility theorem" in fairness, which states that it is mathematically infeasible to simultaneously satisfy multiple fairness criteria when base rates of recidivism differ across groups.

Such concerns underscore the complexity of operationalizing fairness in predictive algorithms, especially in high-stakes applications like criminal justice. They compel us to critically examine not just the outcomes of these tools but also the societal values and priorities embedded in their design and deployment.

Given that, The dataset and its analysis have raised broader concerns about the implications of using AI in sensitive areas like criminal justice. One key issue is the risk of automation bias, where human decision-makers may place excessive trust in algorithmic recommendations, potentially overlooking their limitations or inaccuracies. Moreover, the use of AI in high-stakes environments underscores the critical need for interpretability and accountability, ensuring that these systems are transparent, understandable,and closely monitored to ensure fairness and justice.

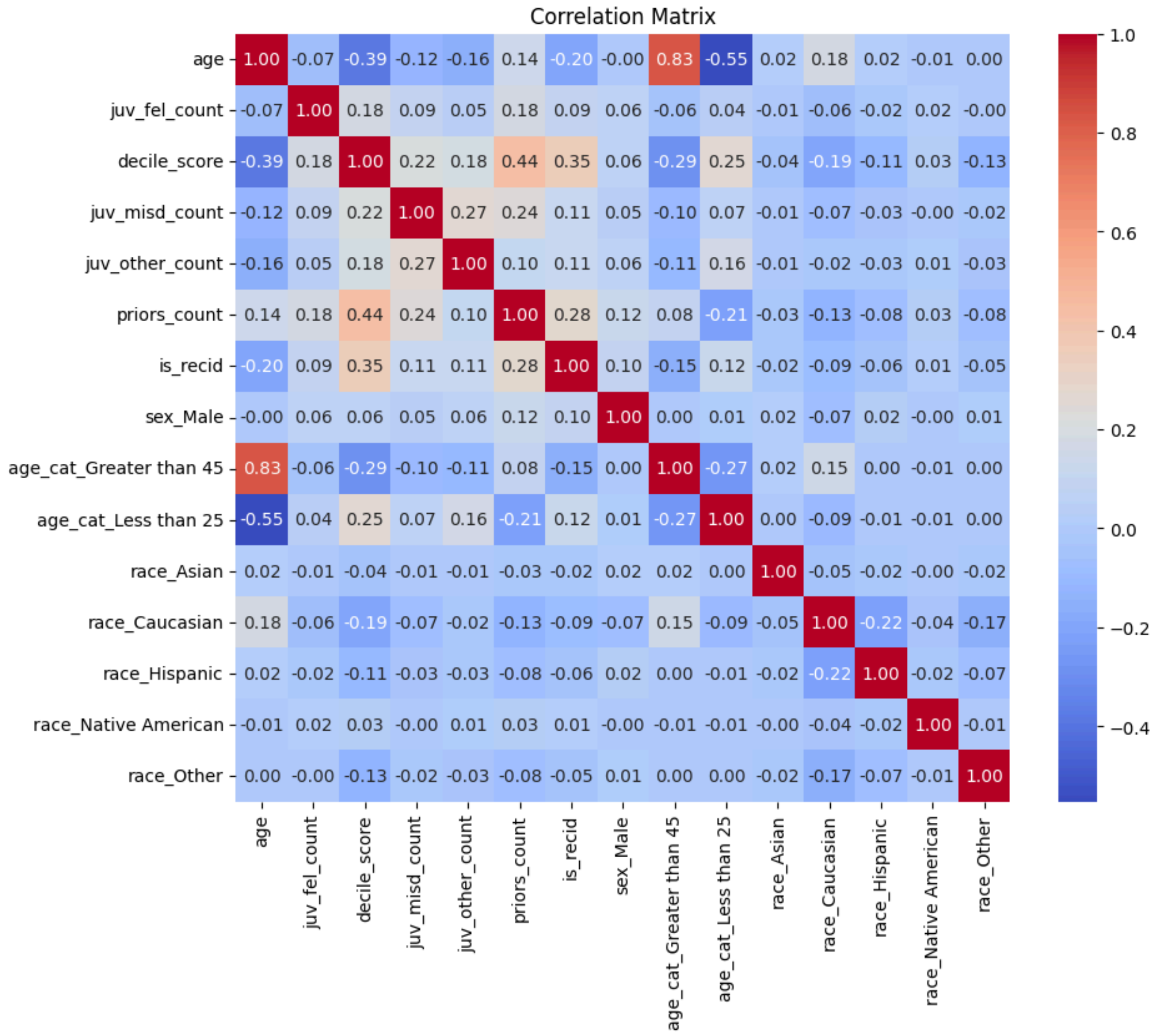
1.2.2. Professionalism and Ethical Principles: -

Ethical frameworks for AI, such as those outlined by professional organizations like IEEE and ACM, stress the importance of fairness, accountability, and transparency in the development and deployment of AI systems. Similarly, the EU AI Act mandates stringent risk assessments for high-risk AI tools, including those used in criminal justice, to ensure responsible and ethical use.

1.2.3. Bias Mitigation Techniques: -

As we will see further in this report, the COMPAS dataset turns out to be a victim of numerous biases. To address algorithmic bias, researchers have developed techniques such as reweighting and adversarial debiasing, which aim to reduce disparities in datasets and improve fairness in predictions. These methods are critical for ensuring that AI systems like COMPAS perform equitably across different demographic groups.

2. Discussion of Dataset



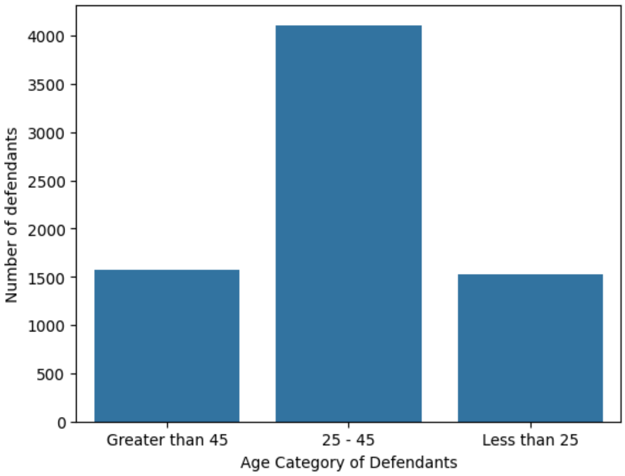
Analyzing the feature correlation matrix above, we can find out that there are strong positive correlations between age and age_cat_Greater than 45 (r = 0.83), reflecting that age aligns well with its categorical grouping, and between priors_count and decile_score (r = 0.44), indicating that individuals with more prior offenses tend to have higher decile scores. Moderate positive correlations are observed between is_recid and decile_score (r = 0.35), suggesting that higher decile scores are moderately associated with recidivism, and between priors_count and is_recid (r = 0.28), showing that individuals with more priors have a greater likelihood of reoffending. Negative correlations include age_cat_Less than 25 with both priors_count (r = -0.21) and decile_score (r = -0.25), indicating that younger individuals generally have fewer prior offenses and lower decile scores.

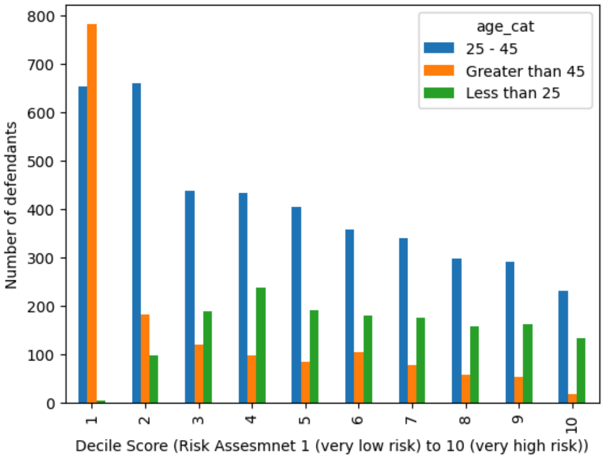
3. Comparison of Predictions Across Demographics

3.1. Age: -

Defendants aged 25–45 represent the most prevalent group in the dataset, followed by those over 45 and then those under 25. When examining risk scores, individuals in the 25–45 age group consistently dominate across all decile scores (1 to 10). This indicates their disproportionate representation in the dataset. High-risk scores (7–10) are relatively rare but appear across all age groups. However, younger individuals under 25 are somewhat less represented in these higher-risk categories.

A critical insight is that the dominance of individuals aged 25–45 in the dataset may influence the algorithm's predictive patterns. Since the majority of data comes from this age group, the algorithm might develop biases that affect its accuracy when assessing risk for younger or older individuals. These biases could result in skewed predictions that do not fully account for the nuances of underrepresented age groups.

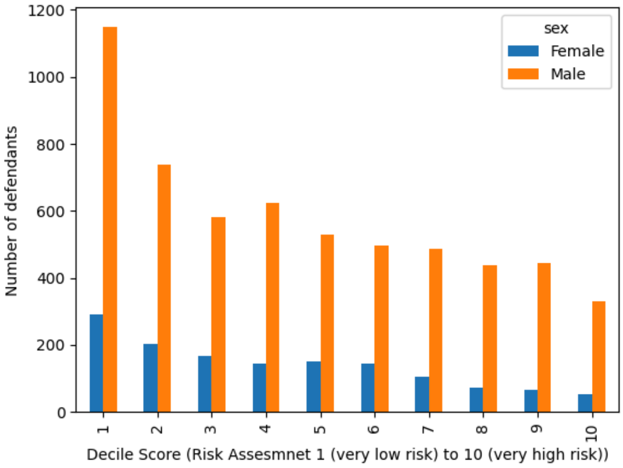


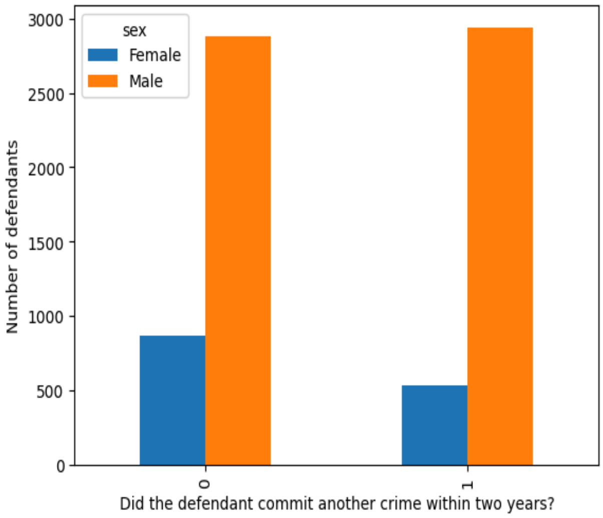


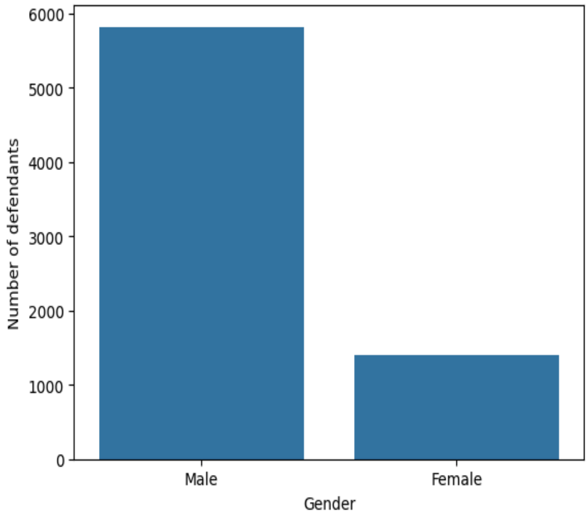
3.2. Gender: -

Males dominate the dataset, outnumbering females across all decile scores, with the largest gap seen at decile 1 (low risk). This trend aligns with males forming the majority of individuals who reoffend, reinforcing their higher risk predictions. The distribution shows that males are more consistently represented in every risk category, which could skew the dataset toward their patterns of behavior. Meanwhile, females, being underrepresented, may not have sufficient data for accurate predictions, raising concerns about fairness and potential misclassification in their risk assessments.

This gender imbalance highlights potential biases in the risk assessment model. The overrepresentation of males might lead to inflated risk estimates for them, while the smaller sample size of females could result in lower prediction accuracy and unreliable outcomes. Addressing these disparities is crucial to ensuring that risk assessments are both equitable and accurate, potentially requiring adjustments to better account for female-specific patterns and improve model performance across genders.



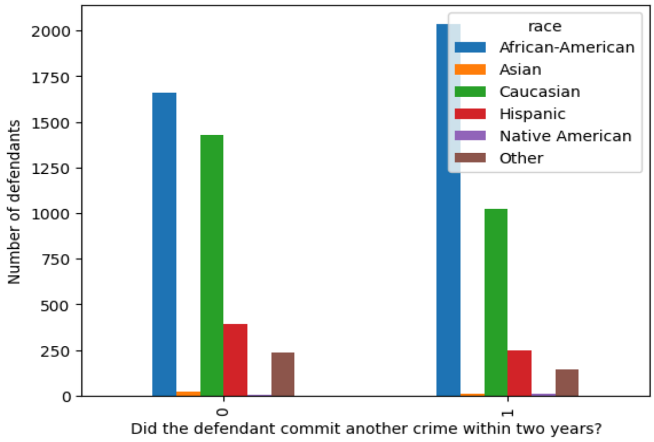


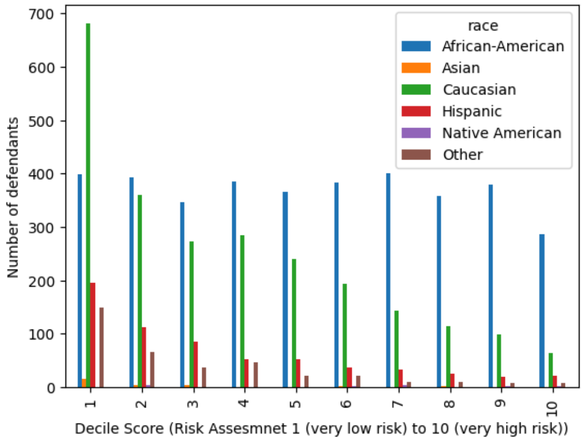


3.3. Race: -

African-Americans are the most represented group in both recidivism and non-recidivism categories, followed by Caucasians. However, recidivism among African-Americans is more evenly distributed across risk levels, while Caucasians show a stronger concentration in the non-recidivist category. Smaller racial groups, such as Asians, Native Americans, and Others, are underrepresented, which limits the algorithm's ability to accurately predict outcomes for these populations. This racial imbalance could lead to biased outcomes, with African-Americans potentially receiving disproportionately higher risk scores, reflecting underlying racial bias in the model.

The distribution of decile scores further highlights racial disparities, with African-Americans consistently receiving higher risk scores compared to other racial groups. Caucasians and Hispanics are more concentrated in the lower-risk categories (decile scores 1–3), while African-Americans are more evenly spread across all risk levels. The underrepresentation of smaller racial groups in the dataset reduces the model's ability to make accurate predictions for these demographics. This imbalance in both the recidivism and decile score distributions suggests that the model may be incorporating historical biases, leading to unfair and inaccurate risk assessments, particularly for African-Americans and smaller racial groups.

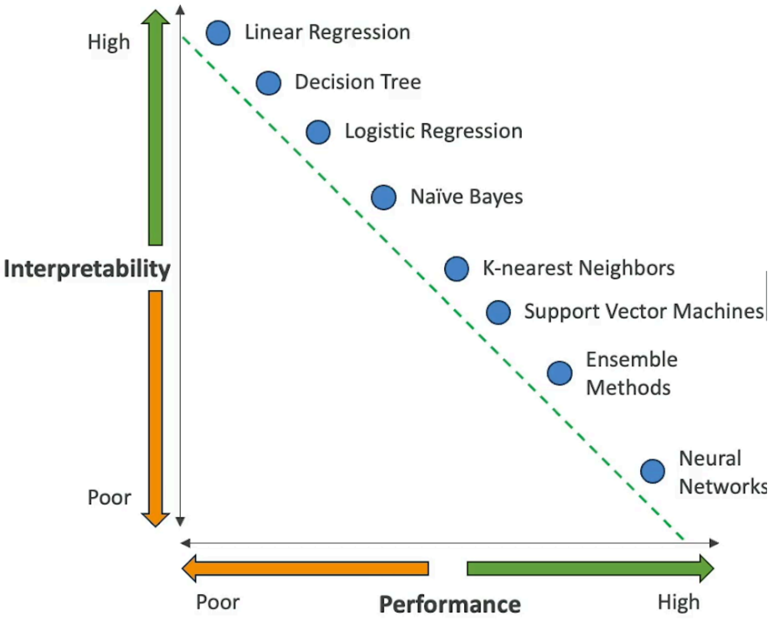




4. Experiments and Discovered Biases: -

4.1 Predictions based on recorded recidivism: -

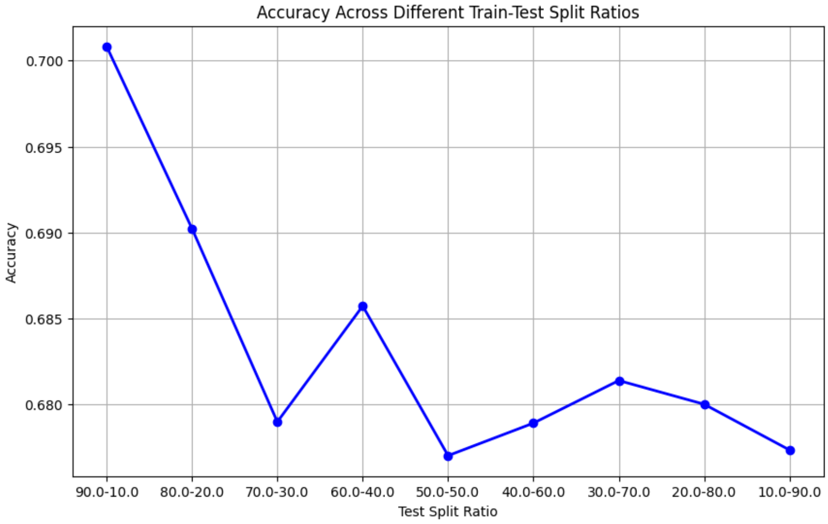
Now that we have understood the COMPAS dataset, we want to now focus on selecting the most suitable model for such a dataset. To do so, we should first assess the performance and interpretability of the common benchmark algorithms available: -



We can notice from the figure above that models like linear regression and decision trees are highly interpretable but offer moderate performance, making them ideal for such applications requiring transparency. As we move towards more complex models like ensemble methods and neural networks, performance improves significantly, but interpretability declines due to their complexity and "black-box" nature.

Considering the binary classification in the COMPAS dataset and the importance of transparency and interpretability in the model's outcomes, Logistic Regression has been selected as the chosen model.

After running the dataset through logistic regression with various train-test split ratios, we observed the following performance outcomes:



The best accuracy achieved was just above 70%, which is good but not great.

4.2. Discovered Biases: -

The model's performance across different demographic groups and examination of the dataset reveal several important considerations regarding bias, and it is essential to critically evaluate whether and how biases are influencing its decision-making.

Examined Biases in the COMPAS Dataset	Explanation
Racial Bias	Analysis consistently shows that the COMPAS algorithm tends to assign higher risk scores to Black defendants compared to White defendants, even when the likelihood of reoffending is similar. This reflects systemic disparities in how different racial groups are treated within the criminal justice system and how those disparities are encoded in the dataset.
Sample Bias	The dataset may overrepresent certain demographic groups, such as individuals aged 25–45, leading to skewed predictions that do not generalize well to underrepresented groups like older adults or younger individuals. This imbalance can cause the algorithm to be less accurate or fair for these minority subgroups.
Historical Bias	Since the dataset reflects decisions and actions taken within a historically biased criminal justice system, it inherently contains the prejudices of past policies and practices. For example, disparities in arrest rates, sentencing, or incarceration for different demographics may inadvertently influence risk predictions.
Outcome Bias	The algorithm relies on rearrest data as a proxy for recidivism. However, rearrest rates are not always an accurate measure of criminal behavior, as they can be influenced by factors like policing practices and socioeconomic conditions, which

	disproportionately affect marginalized communities.
Label Bias	The ground truth used to train the COMPAS algorithm, whether someone reoffended, is based on human judgments and systemic processes that may themselves be biased. For example, decisions about arrests, charges, and convictions can be influenced by race, gender, and socioeconomic status, introducing bias into the labels used for training.
Overgeneralization Bias	Because the dataset pools individuals from various demographic and social contexts, the algorithm may overgeneralize patterns that apply to the majority population, failing to account for the specific circumstances of smaller subgroups.
Gender-Based Bias	The model overestimates non-recidivism for females and underpredicts recidivism due to their lower representation, while for males, it shows higher recall for recidivism but struggles with accurately predicting non-recidivism, likely due to overrepresentation in the dataset.
Race-Based Bias	For African-Americans, the model may tend to overpredict recidivism, reflecting potential historical biases in the data, while for Caucasians and Hispanics, it may underpredict recidivism, suggesting insufficient balance in how recidivism is modeled for these groups.
Representation Bias	The model performs poorly for underrepresented groups like Asians, achieving unreliable predictions due to their small sample size, highlighting issues of insufficient data diversity.

4.3 Cause of Bias: -

4.3.1. Systemic Inequalities: -

These biases reflect broader systemic inequalities within the criminal justice system, including discriminatory practices and policies that disproportionately affect certain racial and demographic groups. These historical disparities shape how data is collected and interpreted, leading to skewed risk assessments.

4.3.2. Data Imbalance: -

Certain demographic groups, such as individuals aged 25–45, may be overrepresented in the dataset, while others, like older adults or younger individuals, are underrepresented. This imbalance results in predictions that do not generalize well across the entire population, leading to unfair treatment of minority subgroups.

4.3.3 Historical Prejudices: -

The dataset encapsulates the decisions and actions of a historically biased criminal justice system, which includes uneven arrest rates and sentencing practices. These historical influences perpetuate existing inequalities and embed them into the predictive algorithms, reinforcing rather than mitigating biases.

4.3.4 Proxy Measures: -

The reliance on rearrest data as a proxy for recidivism introduces inaccuracies because rearrest rates can be influenced by policing practices, socioeconomic factors, and systemic discrimination, particularly affecting marginalized communities. This reliance can distort the true relationship between criminal behavior and risk assessment.

4.3.5 Human Judgments in Labeling: -

The labels used to train the COMPAS algorithm are based on human judgments and systemic processes that may be biased. Factors such as race, gender, and socioeconomic status can influence decisions regarding arrests and convictions, leading to biased labels that skew the model's learning process.

4.3.6 Overgeneralization: -

By pooling individuals from diverse demographic and social contexts, the algorithm may overgeneralize patterns observed in the majority population. This lack of specificity can lead to the failure to recognize the unique circumstances and risk factors associated with smaller or marginalized subgroups.

4.3.7. Gender Disparities: -

The differing representation of genders in the dataset can lead to biases, where the model overestimates non-recidivism for females due to their lower numbers while being more accurate for males, who are overrepresented. This gender imbalance affects the algorithm's ability to accurately predict recidivism across all groups.

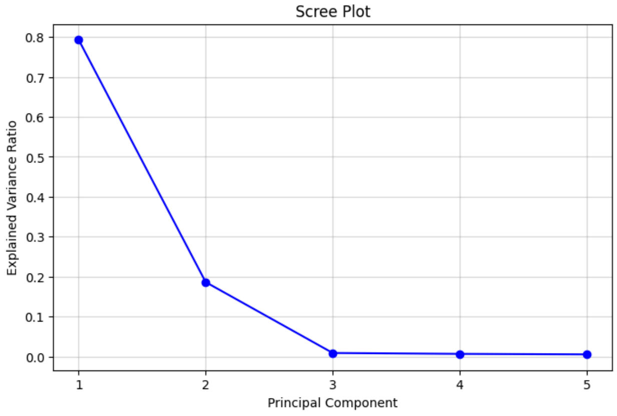
4.3.8. Racial Disparities: -

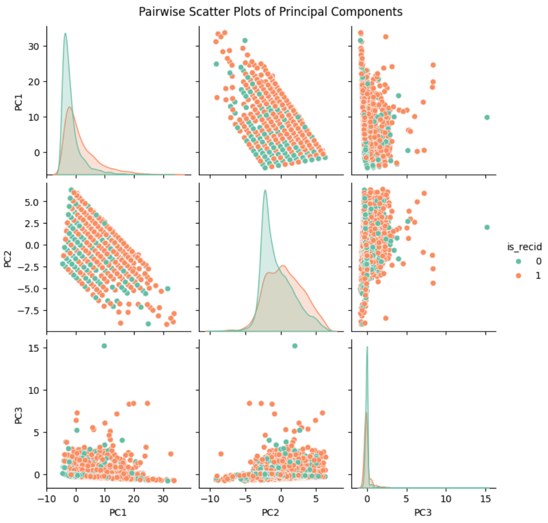
The model's tendency to over predict recidivism for African-Americans and underpredict it for Caucasians and Hispanics suggests a lack of balance in how recidivism is modeled for these groups. These racial disparities can stem from historical biases encoded in the data, influencing the model's outputs.

5. Debiasing: -

To achieve a de-biased model, I have implemented and evaluated a data de-biasing strategy by first removing the protected features of age, sex, and race from the dataset. This approach aimed to eliminate direct biases associated with these sensitive attributes, allowing the model to focus on other relevant features for predicting recidivism. After evaluating the model performance on this debiased dataset.

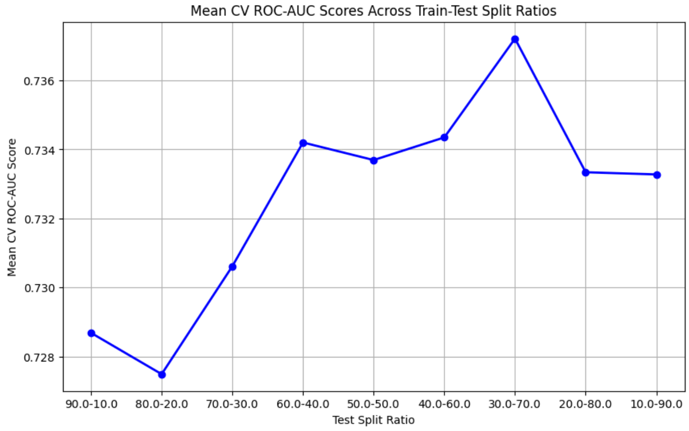
Additionally, I proposed and implemented a second de-biasing method using Principal Component Analysis (PCA) on the debiased dataset. This strategy is particularly effective because PCA helps to reduce dimensionality while capturing the essential variance in the data, thereby mitigating the influence of remaining biases and noise. By transforming the data into a new set of orthogonal features (principal components), PCA not only enhances model interpretability but also promotes a more generalized understanding of the underlying data structure, potentially leading to improved model performance.



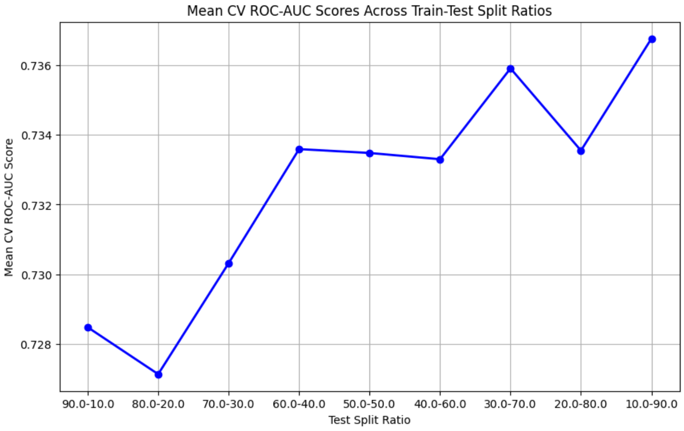


This dual approach—removing protected features and applying PCA—addresses both direct and indirect sources of bias, ensuring a more equitable assessment of recidivism risk. Implementing PCA further refines the model by emphasizing the most informative aspects of the data while minimizing the effects of residual biases associated with the original feature set. This holistic de-biasing strategy enhances the fairness and accuracy of predictions, ultimately contributing to more responsible and ethical decision-making in the context of predictive policing and criminal justice assessments.

With the help of PCA, we can easily examine that the pairwise scatter plots of the principal components (PC1, PC2, and PC3) show significant overlap between the two classes (is_recid = 0 in orange and is_recid = 1 in teal), indicating limited class separability in this reduced feature space. While PC1 contributes the most variance and shows a sharp density peak, PC2 and PC3 exhibit wider spreads with similar overlaps. The scatter plots reveal triangular and clustered patterns but no clear separation between classes, suggesting that PCA alone may not be sufficient for distinguishing these groups and that additional modeling or transformations may be needed for improved classification.



Debiased model accuracy before PCA



Debiased model accuracy after PCA

On examining the model's accuracy across various train-test split ratios before debiasing and PCA, we can see that the trend indicates that accuracy decreases sharply as the test set size increases. For instance, the model achieves its highest accuracy of 0.701 at a 90-10 split (90% training data, 10% test data), but the performance drops significantly as the test split grows larger. At a balanced 50-50 split, accuracy is just 0.678, and it continues to decline with larger test sets. This pattern suggests the model is overfitting to the training data, performing well on smaller test sets but struggling to generalize as the test size increases. Overfitting often arises from biases in the dataset or noisy features that cause the model to focus on irrelevant patterns rather than learning generalizable ones.

Whereas, on examining the model's performance, measured using the mean CV ROC-AUC score, after applying debiasing and PCA, we can notice that the results demonstrate a significant improvement, as the ROC-AUC scores are now higher and more consistent across all train-test splits. For example, the performance at extreme test splits like 10-90 improves to 0.737, compared to the steep decline seen in the first graph. This stability indicates that debiasing helped remove data imbalances or misleading patterns, while PCA reduced redundant or noisy features, allowing the model to focus on the most informative components. As a result, the model generalizes better and maintains a more robust performance regardless of the train-test split ratio, addressing the overfitting issue observed earlier.

6. **Conclusion: -**

The ethical appropriateness of AI recidivism prediction hinges on whether the systems can balance fairness, accuracy, and transparency while mitigating biases that disproportionately impact vulnerable groups. The findings highlight meaningful progress toward de-biasing, with strategies like removing sensitive features (age, sex, and race) and applying Principal Component Analysis (PCA) to reduce indirect biases and noise. These steps demonstrate a commitment to enhancing equity and fairness in predictions, which is crucial for ethical application in criminal justice.

However, limitations persist. While the debiased models showed improved generalizability and consistency, the overlap between classes in the PCA-reduced feature space suggests challenges in separating high-risk and low-risk individuals effectively. This lack of separability raises concerns about potential inaccuracies in decision-making, especially in high-stakes scenarios such as sentencing or parole. Furthermore, the residual effects of historical biases embedded in non-protected features and systemic inequalities in the training data remain a concern, as these may perpetuate inequities despite de-biasing efforts.

The reliance on AI systems to make critical judgments about human behavior must also be scrutinized for transparency and accountability. If predictions remain opaque to stakeholders or fail to acknowledge inherent uncertainties, they risk reinforcing automation bias—where users place undue trust in algorithmic outcomes. To ethically justify AI recidivism predictions, it is essential to pair technical improvements with robust oversight mechanisms, interpretability, and a clear framework for addressing unintended harms. Without these safeguards, the risk of exacerbating existing biases and inequalities outweighs the potential benefits of such predictive tools.

7. Source Code: -

https://colab.research.google.com/drive/17WXeBn2_wMbVy_NgjeOb2e9cihpAD-OU?usp=sharing

8. References: -

ProPublica (2024) *COMPAS Recidivism Risk Score Data and Analysis*. Available at: https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis

ProPublica (2024) *How We Analyzed the COMPAS Recidivism Algorithm*. Available at: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

van Slooten, K. (2019) *ProPublica's COMPAS Data Revisited*. arXiv. Available at: https://ar5iv.labs.arxiv.org/html/1906.04711

S. Maarek, "Aws ai practitioner certified," n.d. [Online]. Available: https://www.udemy.com/course/aws-ai-practitioner-certified/?couponCode=ST19MT121224

Brookings Institution (2024) *Algorithmic Bias Detection and Mitigation: Best Practices and Policies to Reduce Consumer Harms*. Available at: https://www.brookings.edu/articles/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Aimode.org (2024) *12 Types of AI Bias Explained in Detail*. Available at: https://aimode.org/types-of-ai-bias/