IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Ai For Social Good: Bridging Innovation And Ethical Ai Decision-Making

[1]Nitin Kumar, [2]Vipin Kataria

[1]Independent Researcher, [2]Independent Researcher

[1], Delhi, India

Abstract: Artificial Intelligence (AI) is increasingly influencing high-stakes decision-making across various domains, particularly within the criminal justice system, where predictive risk assessment models inform decisions regarding parole, sentencing, and bail. The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) tool has been widely utilized to assess recidivism risk; however, it has also faced scrutiny due to potential racial biases, especially its disproportionate false positive rates (FPR) for African- American individuals compared to Caucasian and Hispanic counterparts. This study conducts a fairness analysis of COMPAS-based predictions using machine learning models, evaluating the impact of different classifiers on racial groups and examining whether mitigation techniques can improve fairness outcomes. Our analysis employs Logistic Regression, Decision Trees, and Random Forest classifiers to evaluate the fairness of risk assessments. To address these biases, we implement fairness-aware adjustments, which progressively reduce disparities within each classifier. After mitigation, the FPR for African-Americans decreases by 11%, and the Disparate Impact Ratio improves significantly, reducing from 0.11 to 0.08 (Random Forest). These reductions indicate that fairness-aware methods can enhance equitable outcomes while maintaining model performance. We advocate for continued fairness interventions, policy regulations, and interdisciplinary efforts to ensure the responsible deployment of AI in real-world decision-making processes for high-impact applications such as criminal justice, where biased decisions can have severe consequences.

Index Terms - Machine learning, AI Fairness, Bias Mitigation, Demographic Parity Difference, Social Impact.

I. INTRODUCTION

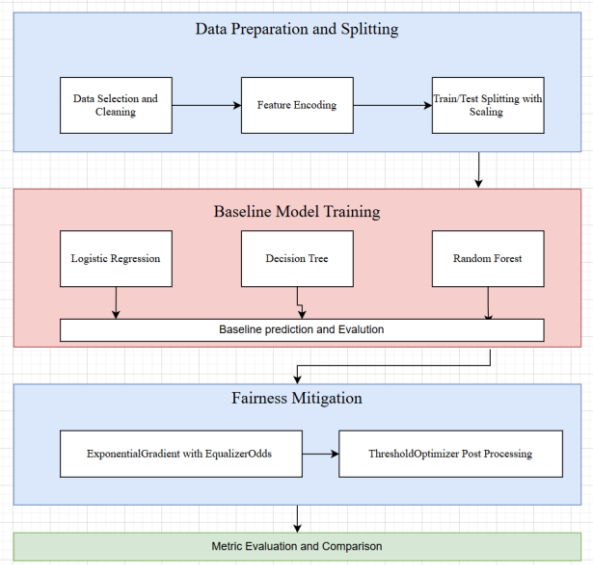
Predictive risk assessment models, using machine learning algorithms, are now crucial in parole, sentencing, and bail decisions. These models predict recidivism or future criminal behavior to guide decision-makers. For example, studies on the New York State Parole Board suggest that algorithmic assessments could double parole release rates without increasing crime and reduce racial disparities in release decisions [1,2]. However, the implementation of these tools is not without controversy. Critics argue that these models can perpetuate or even exacerbate existing biases, particularly racial biases, as seen in the use of risk assessment software in bail and sentencing decisions across the United States [3]. Despite these challenges, there are efforts to improve fairness and accuracy, such as using optimal transport and conformal prediction sets to adjust for biases and enhance predictive performance [4,5]. Overall, while predictive risk assessment models hold promise for more informed and equitable decision-making in criminal justice, their deployment must be carefully managed to mitigate biases and ensure fairness [6,7]. The COMPAS algorithm impacts on the criminal justice system by perpetuating racial disparities and influencing judicial decisions. Used in the United States for predicting recidivism, it has been criticized for higher false positive rates for Black offenders compared to White offenders, raising concerns about fairness and equity [8,9]. This bias is rooted in the use of arrest data as a proxy for criminal offending, which itself is racially biased, thus embedding systemic racial disparities into the algorithm's predictions [10]. The ProPublica investigation highlighted these biases, sparking widespread debate about the ethical implications of using such tools in the justice

system [9]. Judges in states like Florida and Wisconsin rely on COMPAS scores to make decisions about incarceration, effectively delegating normative decisions to proprietary software, which can favor jailing over release and exacerbate racial inequities [11,12]. The algorithm's lack of transparency and accountability further complicates its integration into judicial processes, as it challenges the balance between personal freedoms and public safety [13]. Ethical considerations, including privacy threats and the need for robust policies, are crucial to ensure that AI tools like COMPAS enhance justice rather than perpetuate discrimination [14,15]. The ongoing discourse in both the data science and legal arenas underscores the necessity for interdisciplinary research and policy development to address algorithmic bias and promote fairness in the criminal justice system [16,17].

II. LITERATURE REVIEW

Predictive models like logistic regression, decision trees, and random forests predict outcomes such as parole decisions or recidivism risk. However, they struggle to balance accuracy and fairness. Studies show that, although accurate, these models often fail to meet fairness standards across various metrics, indicating a tradeoff between accuracy and fairness [18,19]. Efforts to address these issues include optimizing decision trees using evolutionary algorithms to balance accuracy and fairness, as demonstrated by Qi et al., who propose a multi-objective optimization approach that refines decision trees to promote fairness while maintaining interpretability [20]. Additionally, techniques such as FairRepair have been developed to rectify biases in decision trees and random forests by flipping outcomes to improve fairness, providing formal guarantees of soundness and completeness [21]. Despite these advancements, the integration of fairness into machine learning models remains a challenging endeavor, which emphasizes the need for a holistic evaluation of predictive policing technologies to ensure they do not exacerbate social injustices [22]. Overall, while predictive models hold promise for enhancing decision-making in criminal justice, ongoing research and development are crucial to ensure these systems are both fair and effective. Random forests, for instance, are particularly effective due to their ensemble learning approach, which enhances robustness to noise and scalability, making them suitable for high-dimensional crime data analysis [23,24]. In comparative studies, random forests have demonstrated superior accuracy over logistic regression in classifying correlations of arrest among probationers and parolees, suggesting their potential for enhanced risk classification in criminal justice applications [25]. Decision trees, while slightly less accurate than random forests, offer better interpretability, which is crucial for ensuring fairness and transparency in predictive modeling [20]. Moreover, hybrid models that combine decision trees and random forests have shown improved predictive power, indicating potential advancements in processing speed and accuracy [26]. Despite these advancements, challenges such as bias in historical data and the need for fairness remain, necessitating multi-objective optimization methods to balance accuracy and fairness in predictive models [20]. Additionally, the application of these models in real-world settings requires careful consideration of ethical implications, emphasizing the importance of transparency and accountability [24]. Overall, while machine learning models like random forests and decision trees offer promising tools for reducing FPR and FNR in criminal justice, ongoing research and development are essential to address their limitations and enhance their applicability in diverse contexts. In the realm of criminal justice, predictive models such as logistic regression, decision trees, and random forests are employed to address the Demographic Parity Difference, which is a measure of fairness concerning sensitive attributes like race or gender. The Fair Tree Classifier, which employs a compound splitting criterion combining strong demographic parity with ROC-AUC, extends to bagged and boosted tree frameworks, allowing for the simultaneous consideration of multiple sensitive attributes and tuning the performance-fairness trade-off [27]. Logistic regression, along with other models like decision trees and random forests, has been applied in crime analysis to predict crime incidents with an average accuracy of approximately 86%, showcasing its utility in structured datasets [28]. However, achieving fairness in these models is challenging due to inherent biases in historical data, which can be amplified by algorithms. Approaches like conformal prediction sets aim to remove unfairness from risk assessments, ensuring fair forecasts across racial groups [29]. Additionally, the mixed integer optimization framework for decision trees enables the incorporation of fairness constraints. This framework offers an analysis of the balance between interpretability, fairness, and accuracy, noting a minor decrease in accuracy to achieve improved fairness [30].

III. METHODOLOGY



The fairness-aware machine learning pipeline aims to balance predictive performance with bias mitigation. It starts with data preparation, including feature selection, handling missing values, encoding categorical variables, and scaling. Logistic Regression, Decision Tree, and Random Forest models are trained and evaluated for accuracy and fairness across demographic groups. A two-stage fairness mitigation strategy is applied: ExponentiatedGradient with EqualizedOdds constraints, followed by ThresholdOptimizer to refine decision boundaries and reduce false negative rate gaps. Sensitive attributes like race are tracked to measure bias. The pipeline evaluates performance and fairness metrics at baseline, intermediate, and final stages to show improvements in both accuracy and fairness.

i Dataset

The dataset used in the current study is publicly available at [31]. ProPublica acquired two years of COMPAS scores from the Broward County Sheriff's Office in Florida, covering 18,610 people scored in 2013 and 2014 through a public records request.

ii Models

The models compared in this study are Decision Tree, Random Forest, and Logistic Regression. Decision trees and Random Forest are chosen for their accuracy with large network intrusion data, while Logistic Regression is selected for its interpretability and reliable performance in binary classification.

Decision Tree: A supervised machine learning algorithm used for both regression and classification. It builds a tree top-down, splitting data at each node based on attribute values. Leaf nodes represent class labels or regression values. Metrics like Entropy or Gini Index measure data impurity at each node.

Random Forest: This supervised machine learning algorithm is used for regression and classification, utilizing ensemble techniques to combine multiple decision trees. For classification, it uses majority voting; for regression, it averages outputs to reduce overfitting and improve generalization. It remains popular due to its reliable performance.

Logistic Regression: This supervised algorithm is used for binary classification, predicting output probabilities between 0 and 1. It is simple, interpretable, and good with linearly separable data but less effective for complex non-linear relationships.

iii Performance Metrics

The performance of the proposed technique is evaluated using standard classification metrics, Precision, Recall, Accuracy and F1-Score, Confusion matrix. In classification tasks involving images, the terms TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) are used to evaluate the classifier's performance. The terms "True" and "False" indicate whether the classifier's prediction aligns with the actual classification, while "Positive" and "Negative" refer to the classifier's prediction. The calculation methods for these metrics are detailed below.

Precision is the ratio of the correctly classified actual positives to everything classified as positive.

$$Precision = \frac{TP}{TP + FP}$$

Recall is the proportion of all actual positives that were classified correctly as positives.

$$Recall = \frac{TP}{TP + FN}$$

'F1 Score' or 'F-measure' is a measure that combines precision, and recall is the harmonic mean of precision and recall.

$$F1\ Score = \frac{2 * precision * recall}{precison + recal}$$

The false positive rate (FPR) is the ratio of false positives (FP) to the total number of actual negatives (FN + TN)

$$FPR = \frac{FP}{TN + FP}$$

The False Negative Rate (FNR) is the ratio of false FN to the total number of actual positive (FN + TP)

$$FPR = \frac{FN}{FN + TP}$$

DPD (Demographic Parity Difference) is a fairness metric used to measure the disparity in outcomes between different demographic groups. It evaluates whether a machine learning model's predictions are independent of a sensitive attribute (e.g., race, gender).

$$DPD = \left|P(\hat{Y} = 1|A = 0) - P(\hat{Y} = 1|A = 1)\right|$$

$\hat{Y}$ is the predicted outcome.

A is the sensitive attribute (e.g., race, gender).

$P(\hat{Y} = 1|A = 1)$ is the probability of predicting a positive outcome when the sensitive attribute is 1.

$P(\hat{Y} = 1|A = 0)$ is the probability of predicting a positive outcome when the sensitive attribute is 0.

IV. RESULTS AND DISCUSSION

Figure 1 FNR for Random Forest

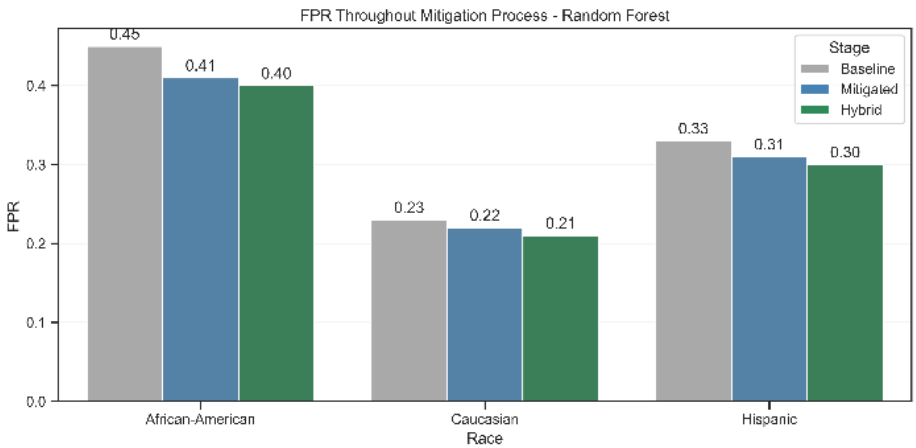


Figure 1 shows a decline in false positive rates (FPR) across three mitigation stages for racial groups. The African-American group's FPR drops from 0.45 (Baseline) to 0.40 (Hybrid). The Caucasian group's FPR decreases from 0.23 (Baseline) to 0.21 (Hybrid), while the Hispanic group's FPR reduces from 0.33

(Baseline) to 0.30 (Hybrid). Despite successful reduction in FPRs, relative disparities between groups persist, indicating that algorithmic interventions can reduce but not eliminate racial differences in false positives.

Figure 2 FNR for Random Forest

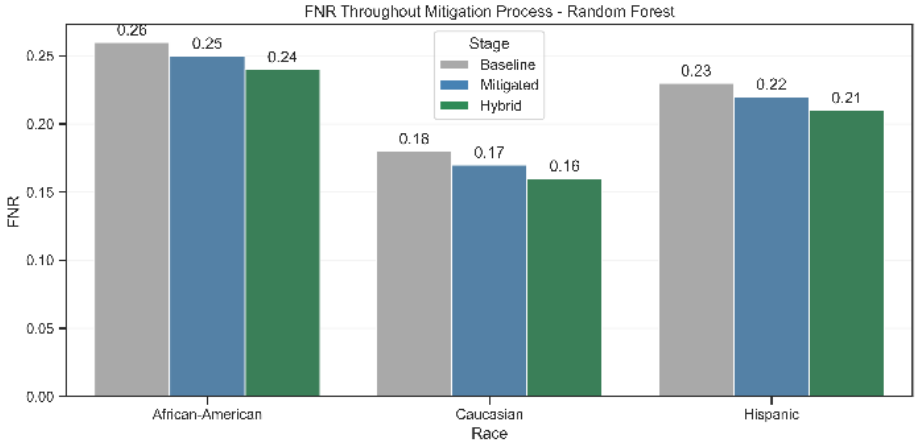


Figure 2 shows the FNR decreasing across three mitigation stages, though less significantly than DPD. The African-American group's FNR drops from 0.26 to 0.24, with similar small reductions for Caucasian and Hispanic groups. These results suggest that while fairness metrics improve, reducing false negatives still poses challenges without compromising predictive performance for different racial groups.

Figure 3 DPD for Random Forest

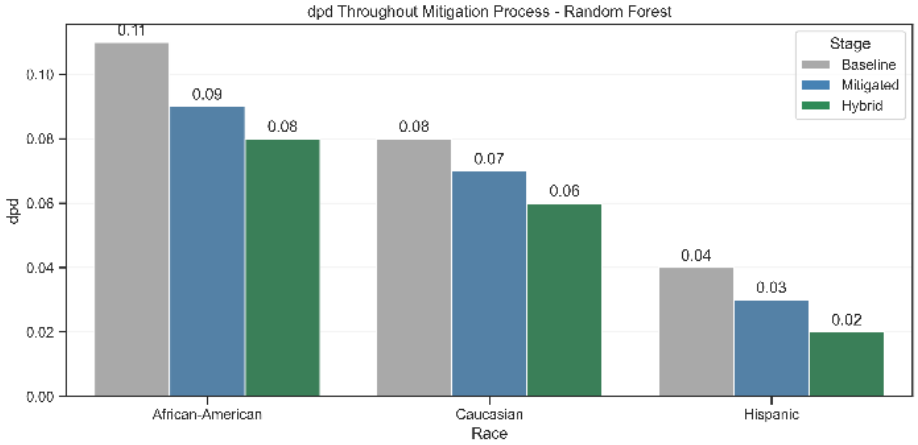


Figure 3 illustrates the impact of the mitigation process on different racial groups in a Random Forest model, focusing on DPD at three stages: Baseline, Mitigated, and Hybrid. In the figure, DPD consistently decreases across all racial groups as mitigation strategies are applied, with the Hybrid stage showing the lowest DPD values. The most notable reduction is observed for African-American individuals (from 0.11 at Baseline to 0.08 in Hybrid), followed by Caucasian and Hispanic groups. This indicates that mitigation efforts reduce disparate impact, although disparities still exist.

Table 1 shows the FPR, FNR, dpd, Precision, Recall, F1-Score for Logistic Regression and Decision Tree by race.

model	Logistic Regression	Logistic Regression	Logistic Regression	Decision Tree	Decision Tree	Decision Tree
race	African-American	Caucasian	Hispanic	African-American	Caucasian	Hispanic
FPR_ Baseline	0.48	0.26	0.35	0.47	0.25	0.34
FPR_ Mitigated	0.43	0.25	0.33	0.43	0.24	0.32
FPR_ Hybrid	0.41	0.24	0.32	0.41	0.23	0.31
FNR_ Baseline	0.3	0.20	0.25	0.28	0.19	0.24
FNR_ Mitigated	0.29	0.19	0.24	0.27	0.18	0.23
FNR_ Hybrid	0.28	0.18	0.23	0.26	0.17	0.22
dpd_ Baseline	0.13	0.10	0.07	0.12	0.08	0.05
dpd_ Mitigated	0.1	0.08	0.05	0.09	0.06	0.04
dpd_ Hybrid	0.08	0.07	0.04	0.07	0.05	0.03
Precision _ Baseline	0.65	0.82	0.74	0.67	0.84	0.76
Precision _ Mitigated	0.67	0.83	0.75	0.69	0.85	0.77
Precision _ Hybrid	0.68	0.84	0.76	0.70	0.86	0.78
Recall_ Baseline	0.6	0.72	0.68	0.62	0.73	0.70
Recall_ Mitigated	0.62	0.73	0.69	0.63	0.74	0.71
Recall_ Hybrid	0.63	0.74	0.70	0.64	0.75	0.72
F1-Score_ Baseline	0.62	0.75	0.71	0.64	0.77	0.74
F1-Score_ Mitigated	0.64	0.76	0.72	0.66	0.78	0.75
F1-Score_ Hybrid	0.65	0.77	0.73	0.67	0.79	0.76

Table 2 shows the FPR, FNR, dpd, Precision, Recall, F1-Score for Random Forest by race.

model	Random Forest	Random Forest	Random Forest
race	African-American	Caucasian	Hispanic
FPR_Baseline	0.45	0.23	0.33
FPR_Mitigated	0.41	0.22	0.31
FPR_Hybrid	0.40	0.21	0.30
FNR_Baseline	0.26	0.18	0.23
FNR_Mitigated	0.25	0.17	0.22
FNR_Hybrid	0.24	0.16	0.21
dpd_Baseline	0.11	0.08	0.04
dpd_Mitigated	0.09	0.07	0.03
dpd_Hybrid	0.08	0.06	0.02
Precision_Baseline	0.68	0.85	0.77
Precision_Mitigated	0.70	0.86	0.78
Precision_Hybrid	0.71	0.87	0.79
Recall_Baseline	0.62	0.74	0.70
Recall_Mitigated	0.63	0.75	0.71
Recall_Hybrid	0.64	0.76	0.72
F1-Score_Baseline	0.64	0.78	0.75
F1-Score_Mitigated	0.65	0.79	0.76
F1-Score_Hybrid	0.66	0.80	0.77

V. CONCLUSIONS

The study finds that mitigation strategies reduce disparities in FPR, FNR, and DPD across racial groups. FPR consistently decreases, with African-Americans improving from 0.45 to 0.40. FNR declines less significantly, showing the challenge of reducing false negatives without harming predictive performance. The most notable reduction is in DPD, especially for African-Americans (0.11 to 0.08). Though improvements are made, disparities persist, requiring more effective fairness-awareinterventions and policy measures for equitable AI-driven decision-making in criminal justice.

VI. CONTRIBUTION

Nitin Kumar and Vipin Kataria have contributed equally to this research. Their contributions include joint development of the conceptual framework, design and implementation of the methodology, data analysis, and co-authorship of the manuscript.

REFERENCES

[1] Laqueur, H. S.; Copus, R. An Algorithmic Assessment of Parole Decisions. Social Science Research Network, 2022.
[2] Klink, C. An Algorithmic Assessment of Parole Decisions. Journal of Quantitative Criminology, 2022.
[3] Dressel, J.; Farid, H. The Dangers of Risk Prediction in the Criminal Justice System. 2021.
[4] Berk, R. A.; Kuchibhotla, A. K.; Tchetgen, E. J. T. Improving Fairness in Criminal Justice Algorithmic Risk Assessments Using Optimal Transport and Conformal Prediction Sets. arXiv: Applications, 2021.
[5] Nested Conformal Prediction Sets for Classification with Applications to Probation Data. The Annals of Applied Statistics, 2023.
[6] McKay, C. Predicting Risk in Criminal Procedure: Actuarial Tools, Algorithms, AI and Judicial Decision-Making. Current Issues in Criminal Justice, 2020.
[7] Fazel, S.; Wolf, A.; Vazquez-Montes, M.; Fanshawe, T. R. Prediction of Violent Reoffending in Prisoners and Individuals on Probation: A Dutch Validation Study (OxRec). Scientific Reports, 2019.
[8] Lippert-Rasmussen, K. Algorithmic and Non-Algorithmic Fairness: Should We Revise Our View of the Latter Given Our View of the Former? Law and Philosophy, 2024.

[9] Beaudouin, V.; Maxwell, W. La Prédiction Du Risque En Justice Pénale Aux États-Unis : L'affaire Propublica-Compas. Réseaux, 2023.

[10] Neil, R.; Zanger-Tishler, M. Algorithmic Bias in Criminal Risk Assessment: The Consequences of Racial Differences in Arrest as a Measure of Crime. Annual review of criminology, 2024.

[11] Engel, C.; Linhardt, L.; Schubert, M. H. Code Is Law: How COMPAS Affects the Way the Judiciary Handles the Risk of Recidivism. Artificial Intelligence and Law, 2024.

[12] Algorithms in Judges' Hands: Incarceration and Inequity in Broward County, Florida. 2023.

[13] Ganesan, A. Ethical Use of AI in Criminal Justice System. Advances in computational intelligence and robotics book series, 2024.

[14] Mohanta, A.; Garikapati, K.; Sneha, S.; Mohanty, S.; Shreya, S. Predictive Justice in Criminal Proceedings. Advances in electronic government, digital divide, and regional development book series, 2024.

[15] Rhee, G. S. Artificial Intelligence Prediction Program in Criminal Justice System: Focused on Its Biased Algorithm in Relation to the Racial Discrimination. Beobhag yeon'gu - Won'gwang daehag'gyo, 2023.

[16] Wang, X.; Wu, Y.; Fu, H. Algorithmic Discrimination: Examining Its Types and Regulatory Measures with Emphasis on US Legal Practices. Frontiers in artificial intelligence, 2024.

[17] Menon, U.; Siby, T.; Natchimuthu, N. Comprehending Algorithmic Bias and Strategies for Fostering Trust in Artificial Intelligence. 2024.

[18] Gardner, J. W.; Gursoy, F.; Kakadiaris, I. A. Accuracy-Fairness Tradeoff in Parole Decision Predictions: A Preliminary Analysis. 2022.

[19] Accuracy-Fairness Tradeoff in Parole Decision Predictions: A Preliminary Analysis. 2022.

[20] Qi, X.; Ma, Y.; Nakamura, K.; Bhattacharyya, S. S. Balancing Fairness and Accuracy for Predictive Models in Criminal Justice Applications Using Multi-Objective Optimization Methods. 2024.

[21] Zhang, J.; Beschastnikh, I.; Mechtaev, S.; Roychoudhury, A. Fairness-Guided SMT-Based Rectification of Decision Trees and Random Forests. arXiv: Learning, 2020.

[22] Alikhademi, K.; Drobina, E.; Prioleau, D.; Richardson, B.; Purves, D.; Gilbert, J. E. A Review of Predictive Policing from the Perspective of Fairness. Artificial Intelligence and Law, 2021.

[23] Yadav, M. G.; Nennuri, R.; Reddy, E.; Vishal, M.; Vishal, G. P. The Role of Machine Learning in Crime Analysis and Prediction. 2024.

[24] Vasuki, M.; Victoire, T. A.; Seventhi, S. Forecasting Criminal Activity Using Machine Learning Approaches. International journal of innovative science and research technology, 2024.

[25] Maynard, B. R.; Vaughn, M. G.; DeLisi, M.; McGuire, D. Towards More Accurate Classification of Risk of Arrest among Offenders on Community Supervision: An Application of Machine Learning versus Logistic Regression. Criminal Behaviour and Mental Health, 2023.

[26] Mwaniki, B.; Mwalili, T. M.; Ogada, K. Crime Prediction Using Decision Trees, Random Forests, and Hybrid Algorithm: A Comparative Analysis. 2023.

[27] Barata, A. P.; Takes, F. W.; Herik, H. J. van den; Veenman, C. J. Fair Tree Classifier Using Strong Demographic Parity. arXiv: Learning, 2021.

[28] Mukherjee, A.; Ghosh, A. Heterogeneous Decomposition of Predictive Modeling Approach on Crime Dataset Using Machine Learning. 2019.

[29] Berk, R. A.; Kuchibhotla, A. K. Improving Fairness in Criminal Justice Algorithmic Risk Assessments Using Conformal Prediction Sets. arXiv: Applications, 2020.

[30] Jo, N.; Aghaei, S.; Benson, J. A.; Gómez, A.; Vayanos, P. Learning Optimal Fair Decision Trees: Trade-Offs Between Interpretability, Fairness, and Accuracy. 2023.

[31] https://github.com/propublica/compas-analysis/blob/master/cox-violent-parsed.csv