Developing and Analyzing Fairness in Algorithmic Policing

Thomas Guizzetti

June 27, 2024

Abstract

Law enforcement agencies in the United States use predictive policing algorithms to analyze past crime data and identify high-risk areas where officers are directed to patrol during each shift. This project aims to develop a predictive algorithm using a US criminological dataset and to identify and mitigate biases that may disproportionately affect marginalized communities. The focus is on evaluating the model's performance and fairness both before and after implementing bias mitigation techniques. By addressing these biases, the goal is to understand how a more equitable crime prediction model that provides unbiased insights and predictions can be developed, ultimately contributing to fairer decision-making processes in algorithmic policing.

1 Description

I use the Community and Crime dataset from the URI repository, which includes socio-economic, law enforcement, and crime data for various communities across the United States. Key features of the dataset include demographic data, income levels, employment rates, education levels, and crime statistics. Particular attention is given to identifying how biases towards the Black community may arise and implementing strategies to mitigate these biases.

1.1 Goals

- **Bias Identification**: Analyze the dataset to identify potential biases, especially those impacting the Black community.
- **Bias Mitigation**: Implement techniques to reduce identified biases and enhance fairness in the dataset, such as reweighting, custom loss creation and adversarial biasing.
- **Model Development**: Develop and evaluate crime prediction models before and after bias mitigation.
- **Fairness Evaluation**: Assess the fairness of the models using appropriate metrics and methodologies.

1.2 Deliverables

The deliverables include a comprehensive final report and a GitHub repository containing all the code used in the project.

2 Introduction

2.1 Domain

Algorithmic policing, often referred to as 'crime forecasting,' utilizes mathematical and analytical techniques to predict potential criminal activity. By analyzing vast amounts of data, including historical crime records and social patterns, these algorithms aim to identify areas and individuals at higher risk

of engaging in or being affected by criminal behavior. This predictive approach enables law enforcement agencies to allocate resources more efficiently and proactively address potential threats, ideally leading to a reduction in crime rates and improved public safety. However, it also raises concerns about privacy, bias, and the ethical implications of relying on automated systems for such critical decisions [15].

2.1.1 Perceived benefits of algorithmic policing

The effectiveness of Predictive Policing in crime prevention remains under evaluation, aligning with current research findings. However, insights gleaned from interviews conducted by the CCI consortium with twelve law enforcement agencies (LEAs) currently using, having used, or planning to adopt algorithmic policing suggest that it is generally perceived as a valuable enhancement and extension of existing police strategies and analytical methods [14]. Moreover, it has been noted that Predictive Policing has led to improved internal communication among different police sections and has been viewed as a means to enhance resource efficiency in police force deployment [14].

2.1.2 Potential risks of algorithmic policing

The methodology employed in predictive policing is fraught with challenges due to its narrow focus and the characteristics of the data used. It primarily concentrates on local crimes such as burglaries and muggings, excluding white-collar crimes. This selective attention generates a negative feedback loop where increased police presence in an area escalates crime reports, which in turn may heighten police deployment [7]. Moreover, the reliance on historical data embeds existing societal biases and discrimination into the predictive models, perpetuating these issues if not properly addressed [16]. The quality of the data significantly affects the outcomes of these studies, leading to potentially unreliable or unethical results. Ethical concerns also arise regarding the use of statistical methods for crime prediction. Additionally, the effectiveness of predictive policing remains ambiguous. n fact, it is challenging to distinguish the effects of predictive policing from other concurrent crime prevention efforts, which may actually be responsible for the reduced crime rates [6].

2.2 Data

This study utilizes the Community and Crime dataset from the UCI repository. This dataset amalgamates socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. The primary goal is to develop algorithms for algorithmic policing aimed at pinpointing regions with high crime rates to optimize the allocation of police resources.

Significantly, all numerical data within the dataset have been normalized to a decimal range of 0.00-1.00 using an Unsupervised, equal-interval binning method. This normalization technique maintains the original distribution and skewness of the attributes. For instance, the attribute 'mean people per household' is presented as its normalized value, maintaining relative proportions within each attribute. However, extreme values have been capped: any value exceeding three standard deviations above the mean is set to 1.00, and values below three standard deviations are set to 0.00. It's important to note that this normalization process does not preserve relationships between different attributes, making direct comparisons across different demographic metrics non-meaningful.

2.2.1 Preprocessing

The primary objective of our preprocessing effort is to predict areas with high crime rates using socio-economic data. Our target variable, *ViolentCrimesPerPop*, represents the total number of violent crimes per 100,000 population.

Drawing from recommendations in the literature [4, 8], we eliminate all columns that contain missing values from our dataset. From an initial set of 127 features, we refined our feature selection to enhance the robustness and accuracy of our predictive model. We retained only those features exhibiting a significant correlation (either positive or negative) with the *ViolentCrimesPerPop* exceeding an absolute value of 0.50. To mitigate potential biases, we excluded direct demographic indicators such

as *racepctblack* (percentage of the African American population), *racepctwhite* (percentage of the Caucasian population), *MalePctDivorce* (percentage of divorced males), and *FemalePctDiv* (percentage of divorced females). The final set of features used is detailed in Table 1.

Additionally, supported by methodologies described in related work [11], we defined a binary target variable, *HighCrime*, identifying areas within the top 85th percentile of *ViolentCrimesPerPop*. To monitor demographic fairness and as typically done in the literature [9, 10], we also introduced a sensitive binary attribute, *isBlack*, identifying communities within the top 85th percentile for the *racepctblack* attribute.

All algorithms deployed in this study utilize these refined features, focusing on the binary target for predictions and monitoring algorithmic fairness via the sensitive attribute.

Features	**Description**	**Corr.**
pctWInvInc	percentage of households with investment / rent income in 1989	-0.58
pctWPubAsst	percentage of households with public assistance income in 1989	0.57
PctPopUnderPov	percentage of people under the poverty level	0.52
PctUnemployed	percentage of people 16 and over, in the labor force, and unemployed	0.50
TotalPctDiv	percentage of population who are divorced	0.55
PctFam2Par	percentage of families (with kids) that are headed by two parents	-0.71
PctKids2Par	percentage of kids in family housing with two parents	-0.74
PctYoungKids2Par	percent of kids 4 and under in two parent households	-0.67
PctTeen2Par	percent of kids age 12-17 in two parent households	-0.66
PctIlleg	percentage of kids born to never married	0.74
PctPersOwnOccup	percent of people in owner occupied households	-0.53
Original Target		
ViolentCrimesPerPop	total number of violent crimes per 100K population	
Binary Target		
HighCrime	85% percentile of *ViolentCrimesPerPop*	
Sensitive Attribute		
isBlack	85% percentile of *racepctblack*	

Table 1: Features, target and sensitive variables

3 Methodology

The methodology followed for our analysis is the below:

- **Baseline** Initially, a state-of-the-art model is established to predict the high crime zones given socio-economic data, providing a foundational understanding of the system's performance and potential biases.
- **Fairness analysis of Baseline:** A fairness analysis is conducted on the baseline model to evaluate disparities in predictions across predominantly black communities and non-black communities.
- **Mitigation techniques:**
  - **Reweighting:** Reweighting of the training data is explored to balance for black communities.
  - **Custom loss functions:** Custom loss functions are designed to penalize biased predictions.
  - **Adversarial debiasing:** Strategies such as adversarial debiasing are implemented, training the model against an adversary to detect and mitigate biases.
- **Fairness analysis of mitigation techniques:** Each mitigation technique is rigorously evaluated using fairness metrics to assess its effectiveness in reducing bias while maintaining predictive accuracy. This analysis provides insights into the impact of these techniques on mitigating algorithmic biases across diverse datasets.

4 Models

I implemented four models: one utilizing a state-of-the-art approach and three employing data mitigation techniques to address biases.

4.1 Baseline

I implemented and evaluated several machine learning models using the `scikit-learn` and `xgboost` libraries. The models were chosen based on their popularity and effectiveness in classification tasks, and their relevance in recent literature [5, 1, 2] which have resulted in generally positive results. Below is a summary of the models and the approach used for hyper-parameter tuning and evaluation.

- Logistic Regression: I used the `LogisticRegression` class from `scikit-learn` with a maximum iteration limit of 1000 and a balanced class weight to handle any class imbalance. Hyperparameters tuned for this model included the regularization parameter C.
- Random Forest: The `RandomForestClassifier` from `scikit-learn` was employed, with hyperparameters such as the number of trees (`n_estimators`) and the maximum depth of each tree (`max_depth`) being optimized. The class weight was also set to balanced.
- Gradient Boosting: We implemented the `GradientBoostingClassifier` from `scikit-learn`, tuning the number of boosting stages (`n_estimators`), the learning rate (`learning_rate`), and the maximum depth of the individual regression estimators (`max_depth`).
- Support Vector Classifier (SVC): The `SVC` class from `scikit-learn` was used with a balanced class weight. Hyperparameters such as the regularization parameter C, kernel type (`kernel`), and kernel coefficient (`gamma`) were tuned.
- XGBoost: For this state-of-the-art model, I implemented the `XGBClassifier` from the `xgboost` library. Hyperparameters such as the number of boosting stages (`n_estimators`), the learning rate (`learning_rate`), and the maximum depth of the trees (`max_depth`) were tuned.

4.1.1 Model Selection and Hyperparameter Tuning

I performed a grid search among these models to find the optimal hyperparameters. The grid search was conducted using `GridSearchCV` with a 5-fold cross-validation and the F1 score as the evaluation metric. The best model was selected based on its performance on the F1 score.

4.1.2 Grid-search evaluation

Each model was trained on the training dataset and evaluated on the test dataset. The best model was determined by comparing the F1 scores of the predictions on the test dataset. This approach ensured that the model with the highest F1 score, indicating the best balance between precision and recall, was chosen for the final evaluation.

5 Data Mitigation Techniques

To improve the baseline, I used 3 data mitigation techniques to address biases towards black communities. In particular I used:

5.1 Re-weighting

Data reweighting is a bias mitigation technique used to address imbalances in training data. This approach assigns different weights to samples in the training set based on their importance or the need to balance underrepresented classes or groups. By adjusting the sample weights, the learning algorithm can give more emphasis to certain instances, which helps to ensure that the model does not disproportionately favor the majority class or group. In the provided function, data reweighting is implemented by passing the *sample_weight* parameter to the fit method of the model during training. This parameter is used to specify the weights for each sample in the training set, allowing the model to

account for biases and improve its fairness in predictions. The weights are used in both the hyperparameter tuning process and the final model training, ensuring that the resulting models are evaluated and optimized with bias mitigation in mind.

5.2 Custom Loss Function

Custom loss functions are a bias mitigation technique designed to address unfairness in model predictions. This method modifies the conventional loss function by adding a penalty term that targets bias related to sensitive attributes, in this case race. In my function, the penalty is calculated as the mean absolute difference between the model's predictions and the binary indicator of race. This approach directly addresses potential bias by altering the model's optimization process to reduce disparities. I implemented this in TensorFlow by creating a neural network model that incorporates this custom loss function. The custom loss is applied during model training by specifying it in the loss parameter of the compile method. This adjustment allows the model to not only focus on accuracy but also on minimizing bias, enhancing the fairness of predictions. The model's effectiveness in reducing bias while maintaining predictive accuracy is assessed through performance metrics like the F1 score.

5.3 Adversarial Debiasing

Adversarial debiasing is a bias mitigation approach that leverages the concept of an adversary to reduce bias in machine learning models. In my implementation, I developed two interconnected models using TensorFlow: a main model for predicting crime rates and an adversarial model that attempts to predict the sensitive attribute (race) from the main model's predictions. The main model is trained to minimize its prediction error on the crime rates, while the adversarial model is trained to maximize its accuracy in predicting the sensitive attribute. By iterating training between these models, the main model learns to make predictions that are increasingly invariant to the sensitive attribute, thereby reducing potential bias.

The training process involves a custom loop where both models are updated in sequence: the main model seeks to deceive the adversary by minimizing its ability to predict the sensitive attribute based on the main model's outputs. This adversarial setup is a dynamic and effective way to ensure that the predictive performance of the main model does not inadvertently encode biases related to sensitive attributes. I utilized various metrics, such as accuracy and AUC, to evaluate the efficacy of the debiasing, providing a comprehensive assessment of the model's performance.

6 Evaluation

The state-of-the-art models, along with those adjusted to mitigate bias, were evaluated using a comprehensive set of metrics. The F1 score was employed to assess the accuracy of the models in predicting high-crime areas. Subsequently, several fairness metrics were applied to evaluate the models' fairness, particularly in relation to black communities. These fairness metrics are recognized as standard criteria in the current literature on fairness in machine learning [3, 12, 13].

6.0.1 Statistical Parity

This metric, also known as demographic parity, evaluates whether the proportion of positive outcomes (e.g., an area being predicted as high crime) is consistent across different groups.

$$\text{Positive Class Ratio} = \frac{TP + FP}{TP + TN + FP + FN} \tag{1}$$

$$\text{Negative Class Ratio} = \frac{TN + FN}{TP + TN + FP + FN} \tag{2}$$

Each metric is evaluated separately for the black community and the non-black community (abbreviated as n-black) to assess fairness. The metrics are presented as follows for reference:

$$\text{Positive Class Ratio for Black Communities} = \frac{TP_{black} + FP_{black}}{TP_{black} + TN_{black} + FP_{black} + FN_{black}} \tag{3}$$

$$\text{Positive Class Ratio for Non-Black Communities} = \frac{TP_{nblack} + FP_{nblack}}{TP_{nblack} + TN_{nblack} + FP_{nblack} + FN_{nblack}} \quad (4)$$

6.0.2 Equality of Opportunity

According to this metric, members of each group who exhibit similar behavior should receive equal treatment in terms of positive outcomes. This aligns with the concept of conditional procedure accuracy equality [3].

$$\text{Positive Class Opportunity} = \frac{TP}{TP + FN} \quad (5)$$

$$\text{Negative Class Calibration} = \frac{TP}{TP + FP} \quad (6)$$

6.0.3 Calibration

Calibration requires that the proportion of correct predictions be the same for each class within each group. This means the ratio of true positives (TP) to the total positive predictions (TP + FP) and the ratio of true negatives (TN) to total negative predictions (TN + FN) should be equivalent across all groups [3, 12].

$$\text{Positive Class Calibration} = \frac{TP}{TP + FP} \quad (7)$$

$$\text{Negative Class Calibration} = \frac{TN}{TN + FN} \quad (8)$$

6.0.4 False Rate

This metric assesses the proportion of incorrect predictions for each class relative to the total predictions. To meet this criterion, the ratio of false positives (FP) to the total positive predictions (TP + FP) and the ratio of false negatives (FN) to the total negative predictions (TN + FN) should be balanced between the groups.

$$\text{False Positive Rate} = \frac{FP}{TP + FP} \quad (9)$$

$$\text{False Negative Rate} = \frac{FN}{TN + FN} \quad (10)$$

6.0.5 Treatment Equality

Treatment equality demands that the ratio of errors in positive and negative predictions be uniform across all groups. Specifically, the ratio of individuals erroneously classified as recidivists (FP) to those mistakenly classified as non-recidivists (FN) should be equal for different demographic groups. This criterion ensures that no group is disproportionately affected by the system's misclassifications.

$$\text{Positive Treatment Equality Ratio} = \frac{FP}{FN} \quad (11)$$

$$\text{Negative Treatment Equality Ratio} = \frac{FN}{FP} \quad (12)$$

6.0.6 Positive and Negative Disparity Ratios

The Positive Disparity Ratio quantifies the relative likelihood of an outcome being predicted as positive for one group compared to another, adjusting for group sizes and prediction errors. The Negative Disparity Ratio similarly measures the relative likelihood of a negative prediction outcome between groups. Ideally, both ratios should be close to 1, indicating fair and balanced outcomes across groups without favoring one over another. Ratios deviating significantly from 1 suggest potential biases in the prediction model, warranting further investigation or adjustments to ensure equitable treatment.

For all fairness metrics discussed, including Statistical Parity, Equality of Opportunity, Calibration, False Rate Parity, and Treatment Equality, the calculation of a disparity ratio plays a central role. This ratio is essential for evaluating the fairness of an algorithm: it allows us to measure and compare the impact of the algorithm's predictions on different groups. By analyzing these ratios for each fairness metric, we can ascertain whether an algorithm behaves equitably across various dimensions of data. A ratio close to 1 for any given metric indicates that the algorithm is performing fairly with respect to that metric, providing a balanced outcome that does not disproportionately benefit or disadvantage any group. Conversely, ratios that significantly deviate from 1 indicate areas where the algorithm may be unfair, highlighting the need for adjustments to enhance its fairness. This systematic assessment using disparity ratios is crucial for certifying that algorithms are not only effective but also just and unbiased in their predictions.

7 Results and Analysis

The evaluation of various bias mitigation methods applied to a baseline policing algorithm, which predicts high crime areas based on socio-economic data, reveals significant disparities in performance and fairness metrics between Black and Non-Black communities. The methods analyzed include the baseline model, Reweighting, a Custom Loss Function, and Adversarial Debiasing.

7.1 Performance Metrics

As shown on Table 2, for the Black community, the Baseline method shows a high True Positive (TP) rate of 46.59% but suffers from a high False Positive (FP) rate of 44.32%. In contrast, the Non-Black community benefits from a high True Negative (TN) rate of 82.00% with a lower FP rate of 10.37%. This indicates a bias towards incorrectly predicting high crime in Black communities compared to Non-Black ones.

Reweighting slightly reduces TN rates in both communities, with an increase in FP rates, notably in the Non-Black community to 18.79%, suggesting a deterioration in specificity. The Custom Loss Function method demonstrates an extreme shift, reducing FP to 0% in the Black community but at the cost of increasing the False Negative (FN) rate dramatically to 44.32%. This method seems to prioritize reducing wrongful high crime predictions at the expense of missing actual high crime instances. Adversarial Debiasing achieves almost no FP and TP rates but increases FN significantly, especially in the Black community (12.35%), indicating a conservative approach that avoids false alarms but fails to alert on real issues.

Method	**Black Community**				**Non-Black Community**			
	TP	TN	FP	FN	TP	TN	FP	FN
Baseline	46.59%	7.95%	44.32%	1.14%	5.09%	82.00%	10.37%	2.54%
Reweighting	46.59%	5.68%	46.59%	1.14%	5.48%	73.58%	18.79%	2.15%
Custom Loss Function	3.41%	52.27%	0.00%	44.32%	0.59%	91.39%	0.98%	7.05%
Adversarial Debiasing	0.00%	87.65%	0.00%	12.35%	0.19%	86.10%	0.19%	13.51%

Table 2: Performance Metrics for Baseline and Bias Mitigation Methods

7.2 Fairness Metrics

We use the positive and negative disparity ratios explained earlier to measure fairness. Remember that the further a value is from 1, the less fair the algorithm is concerning that metric. As shown

on Table 3, the Baseline method exhibits moderate statistical parity but poor equality of opportunity and treatment equality across communities. Reweighting shows improvements in treatment equality with values reduced to 4.7 for both positive and negative aspects but worsens equality of opportunity, especially negative (7.3).

The Custom Loss Function stands out in terms of fairness metrics, achieving nearly ideal scores in treatment equality and the lowest false rates among all methods. This method shows a strong balance between reducing bias and maintaining equitable treatment across communities. However, it is important to note that while this method minimizes false rates, it does so potentially at the cost of missing true crime predictions as indicated by its high FN rates.

Adversarial Debiasing achieves perfect scores in some fairness metrics (statistical parity, equality of opportunity for positive predictions, and treatment equality), but these are accompanied by 'NaN' values in some metrics, given the lack of FP and TP rates for this method.

Method	**Stat. Parity**		**Eq. Of Opp.**		**Calib.**		**False Rate**		**Treat. Eq.**	
	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg
Baseline	5.9	9.3	1.5	5.8	1.6	1.1	0.73	0.24	9.6	9.6
Reweighting	3.8	11.0	1.4	7.3	2.2	1.2	0.65	0.17	4.7	4.7
Custom Loss Function	2.2	1.0	0.93	0.99	2.7	1.7	0.0	0.16	0.0	0.0
Adversarial Debiasing	0.0	1.0	0.0	1.0	nan	0.99	nan	1.1	0.0	0.0

Table 3: Fairness Metrics (Disparity Ratio) for Baseline and Bias Mitigation Methods

7.3 Detailed Metric Comparison

Table 4 offers a comprehensive overview of performance metrics across different bias mitigation methods compared to Baseline results. Notably, the Baseline method demonstrates the highest F1 score of 0.5583, which suggests a balanced approach between precision and recall, making it effective in managing both the relevance and retrieval of the model. However, its accuracy, while robust at 0.8230, is not the highest, which is observed in the Custom Loss Function method at 0.8664. This indicates that while the Custom Loss Function is very effective at correctly identifying true negatives and positives overall, it significantly lacks in terms of recall (0.0741), reflecting its poor performance in identifying true positives relative to the actual positives cases.

The Reweighting method shows a better balance in recall (0.8519) than other methods, indicating its strength in identifying most positive cases but at the cost of precision, leading to a relatively lower F1 score of 0.4808 due to a higher rate of false positives. In contrast, Adversarial Debiasing, despite having a high accuracy of 0.8648, has the lowest F1 score (0.0241) and recall (0.0123), underscoring its ineffectiveness in correctly predicting positive cases, likely indicating an overly conservative model that avoids false positives but fails to capture sufficient true positives.

Method	**F1 Score**	**Accuracy**	**Precision**	**Recall**
Baseline	0.5583	0.8230	0.4214	0.8272
Reweighting	0.4808	0.7513	0.3350	0.8519
Custom Loss Function	0.1304	0.8664	0.5455	0.0741
Adversarial Debiasing	0.0241	0.8648	0.5000	0.0123

Table 4: Performance Metrics for Baseline and Bias Mitigation Methods

8 Conclusion

This analysis highlights the trade-offs between maintaining high predictive accuracy and achieving fairness in algorithmic predictions. While the Custom Loss Function and Adversarial Debiasing methods show promising fairness improvements, they raise concerns about their practical utility due to high FN rates. The Baseline and Reweighting methods, while more balanced in predictive performance, still exhibit concerning fairness gaps, especially in the treatment of Black communities. This underscores

the need for continued refinement of bias mitigation techniques in predictive policing algorithms to enhance both fairness and effectiveness.

Potential improvements to this study and how to improve policing algorithms include:

- Hybrid Approaches: Combining the strengths of different methods could yield better overall performance. For instance, integrating the sensitivity of the Custom Loss Function with the broader accuracy of the State-of-the-art method might balance the reduction in false positives without significantly impacting the true positive rate. Hybrid models can leverage the robustness of one method to offset the weaknesses of another.
- Enhanced Feature Engineering: Improving the selection and processing of input variables could reduce inherent biases in the data before they affect the model's outputs. This includes more rigorous preprocessing to identify and mitigate sources of bias such as socio-economic factors disproportionately affecting one community over another.
- Regular Audits and Updates: Continuously monitoring the model's performance and fairness metrics over time and across different demographic groups can help identify biases as they emerge. Regular audits allow for timely adjustments to the model or its training data, keeping the system as fair and accurate as possible.
- Increased Transparency: Implementing explainable AI techniques can help stakeholders understand how decisions are made, which is crucial for maintaining trust and accountability. Transparent models make it easier to identify when and why biases occur, facilitating more targeted interventions.
- Community Feedback Integration: Engaging with communities affected by predictive policing can provide valuable insights that improve model fairness. Feedback from community members can help identify overlooked biases and validate the model's outputs against real-world outcomes.
- Advanced Statistical Techniques: Utilizing more sophisticated statistical methods to assess and adjust fairness metrics could lead to better outcomes. Techniques such as propensity score matching or Bayesian hierarchical models could refine assessments of the model's impact on different populations.

By implementing these improvements, predictive policing algorithms can become more equitable and effective, ensuring they serve all segments of the community justly while maintaining high standards of predictive accuracy.

References

[1] Md Awal et al. "Using Linear Regression to Forecast Future Trends in Crime of Bangladesh". In: (May 2016). DOI: 10.1109/ICIEV.2016.7760021.

[2] Vedhadharshan B et al. "Crime Analysis Using Linear Regression". In: (Dec. 2022), pp. 1–4. DOI: 10.1109/CISCT55310.2022.10046478.

[3] R. Berk et al. "Fairness in Criminal Justice Risk Assessments: The State of the Art". In: *Sociological Methods & Research* 50.1 (2018), pp. 3–44. DOI: 10.1177/0049124118782533. URL: https://doi.org/10.1177/0049124118782533.

[4] Toon Calders et al. "Controlling Attribute Effect in Linear Regression". In: (Dec. 2013), pp. 71–80. DOI: 10.1109/ICDM.2013.114.

[5] Bruno Cavadas, Paula Branco, and Sérgio Pereira. "Crime Prediction Using Regression and Resources Optimization". In: (Sept. 2015). DOI: 10.1007/978-3-319-23485-4_51.

[6] J. van Dijk, A. Tseloni, and G. Farrell. "Introduction". In: *The International Crime Drop. New Directions in Research*. Ed. by Jan van Dijk, Andromachi Tseloni, and Graham Farrell. Crime Prevention and Security Management. United States: Palgrave Macmillan, 2012.

[7] A. Ferguson. "Policing Predictive Policing". In: *Washington University Law Review* 94.5 (2017), pp. 1109–1189. URL: https://openscholarship.wustl.edu/law_lawreview/vol94/iss5/5.

[8] Hoda Heidari et al. “Fairness behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making”. In: (2018), pp. 1273–1283.

[9] Faisal Kamiran, Indrė Zliobaitė, and Toon Calders. “Quantifying Explainable Discrimination and Removing Illegal Discrimination in Automated Decision Making”. In: *Knowledge and Information Systems* 35.3 (2013), pp. 613–644. DOI: https://doi.org/10.1007/s10115-012-0584-8.

[10] Toshihiro Kamishima et al. “Fairness-Aware Classifier with Prejudice Remover Regularizer”. In: *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*. 2012, pp. 35–50. DOI: https://doi.org/10.1007/978-3-642-33486-3_3.

[11] Michael Kearns et al. “Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness”. In: (2018), pp. 2564–2572.

[12] J. Kleinberg, S. Mullainathan, and M. Raghavan. “Inherent Trade-offs in the Fair Determination of Risk Scores”. In: *arXiv preprint arXiv:1609.05807* (2016). URL: https://arxiv.org/abs/1609.05807.

[13] F. Lagioia, R. Rovatti, and G. Sartor. “Algorithmic Fairness Through Group Parities? The Case of COMPAS-SAPMOC”. In: *AI & Society* 38 (2023), pp. 459–478. DOI: 10.1007/s00146-022-01441-y. URL: https://doi.org/10.1007/s00146-022-01441-y.

[14] Maximilian Querbach, Marian Krom, and Armando Jongejan. “Review of State of the Art: Predictive Policing”. In: (Jan. 2020).

[15] Rashida Richardson, Jason Schultz, and Kate Crawford. “Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice”. In: 94 (2019). Available at SSRN: https://ssrn.com/abstract=3333423, p. 192.

[16] A. Shapiro. “Reform predictive policing”. In: *Nature* 541.7538 (2017), pp. 458–460.