Arrest Data Analysis and Crime Prediction using Machine Learning (2020-2023)

Abhijna Sahadeva, Chloe Huang, Nya Griffin Ulibarri

Summary

This study applies machine-learning tools to illuminate arrest disparities in Pittsburgh between 2020 and 2023. From an initial scrape of 23,145 police reports, we retained 22,986 de-identified arrest records after rigorous cleaning and feature engineering. Our three-part framework combines: (i) a supervised model that predicts whether an arrest will be recorded as violent; (ii) an unsupervised spatial-temporal analysis that pinpoints persistent hot-spots; and (iii) an automated fairness audit that monitors performance gaps across race and gender.

The gradient-boosted XGBoost classifier, trained on demographic, geographic, and temporal features, achieved 66 % accuracy and a ROC-AUC of 0.726, outperforming a logistic-regression baseline by thirteen AUC points. Feature-importance inspection shows that neighbourhood dummies and late-night indicators dominate predictive power, underscoring the place-based nature of violent incidents. Agglomerative clustering reveals seven high-risk neighbourhood clusters—notably the Central Business District, South Side Flats, and Homewood South—that together account for 38 % of violent arrests while housing only 14 % of the city's population. Fairness auditing surfaces recall gaps: the model identifies violent arrests involving Black individuals at 0.79 recall versus 0.69 for White individuals, and exhibits smaller gender discrepancies (< 0.05). These disparities signal the need for threshold tuning and routine bias checks before any operational use.

We translate these insights into policy actions: deploy the model strictly as an accountability dashboard rather than a dispatch tool; redirect prevention funding toward the flagged hot-spots through community-violence-intervention programs; institute quarterly public fairness reports and third-party audits; and pair algorithmic findings with officer anti-bias training. Taken together, the framework equips City leadership and community stakeholders with a transparent, data-driven lens to target resources, increase oversight, and move Pittsburgh toward more equitable public-safety outcomes.

1 Background and Introduction

Disparities in law enforcement outcomes remain one of the most pressing public policy challenges in the United States, significantly impacting community safety, institutional trust, and long-term social equity. In Pittsburgh, these disparities are particularly pronounced: Black residents constitute only 23% of the city's population but represent 66% of incarcerated individuals, and public trust in police has consistently ranked among the lowest within Pennsylvania for several years.[1] Additionally, enforcement activities exhibit stark geographic imbalances, disproportionately concentrated in neighborhoods such as the Central Business District, South Side Flats, and East Allegheny. These patterns raise serious concerns about systemic bias and question the effectiveness and fairness of existing public safety strategies.

[1] RAND Corporation & RTI International, *Analysis of Racial Disparities in Allegheny County Criminal Justice System*, University of Pittsburgh Institute of Politics, 2021.

At the individual level, arrest and incarceration have profound and lasting negative consequences, including limited access to employment, housing, and healthcare, worsened health outcomes, and increased economic instability over time.[2] At the community level, concentrated enforcement activities can erode public trust and perpetuate cycles of criminalization. Although recent policy initiatives in Pittsburgh, such as the 2021 Equitable and Fair Traffic Enforcement ordinance, have made progress in reducing certain types of enforcement disparities, significant gaps remain unresolved, particularly regarding more severe arrests.[3] These ongoing issues highlight the limitations of retrospective statistical analyses and underscore the urgent need for proactive, forward-looking analytical tools that consider both spatiotemporal dynamics and model fairness.

This project directly addresses these gaps by leveraging machine learning methodologies to analyze and predict patterns of arrest severity across demographic and spatial dimensions. Moving beyond surface-level statistical summaries, we aim to identify key predictive factors—such as age, race, geographic location, and timing—that correlate with serious offense charges. Moreover, we apply advanced spatiotemporal clustering techniques to pinpoint arrest "hotspots," assessing whether enforcement intensity aligns with actual community needs or indicates potential over-policing. By integrating predictive modeling with explicit fairness auditing, our approach seeks to ensure that analytical outcomes do not reinforce historical biases but instead support equitable, efficient, and transparent public safety practices.

Ultimately, this project aims to develop a practical, deployable "monitoring–early warning–intervention" framework. City leaders and law enforcement agencies can utilize these insights for real-time decision-making, optimizing resource allocation, mitigating harm in vulnerable communities, and enhancing transparency and accountability. In the long term, our work contributes to the broader movement toward responsible AI in the public sector, demonstrating how machine learning can serve not only to predict but also to critically examine, reform, and improve the systems that profoundly shape people's lives, supporting higher-level policing strategies and community-driven equity initiatives.

2 Related Work

Historically, research on disparities in the criminal justice system has primarily relied on descriptive statistics and regression-based methods to identify inequities across race, geography, and offense type. Studies such as those by RAND/RTI (2021) have quantified racial disparities but have not provided predictive modeling frameworks.[4] Lum & Isaac (2016) utilized random forest models to analyze biases specifically in traffic enforcement, introducing a methodological framework for algorithmic fairness assessment.[5] While these approaches have significantly contributed to understanding systemic patterns, such as the disproportionate representation of Black individuals in arrests and incarceration,[6] they are often retrospective and limited in

[2] Katherine Beckett and Allison Goldberg, "The Effects of Imprisonment in a Time of Mass Incarceration," *Journal of Political Economy* 130, no. 10 (2022): 2575–2618. https://doi.org/10.1086/720342.
[3] Elizabeth Szeto, "Despite Reform, Pittsburgh Police Still Stop Black Drivers Disproportionately," *PublicSource*, August 8, 2023. https://www.publicsource.org/pittsburgh-police-traffic-stop-disparity-accountability-race/.
[4] RAND Corporation & RTI International (2021). Quantifying Racial Disparities in Policing. Retrieved from https://www.rand.org/pubs/research_reports/RRA108-1.html
[5] Lum, K., & Isaac, W. (2016). To predict and serve? Significance, 13(5), 14-19.
[6] Alexander, M. (2010). The New Jim Crow: Mass Incarceration in the Age of Colorblindness. The New Press.

modeling complex, nonlinear interactions between demographic, spatial, and temporal variables or producing actionable, real-time insights.[7]

More recently, machine learning (ML) techniques have been increasingly applied within the criminal justice domain, primarily to enhance enforcement activities. For instance, predictive policing tools have been developed to forecast crime hotspots,[8] estimate recidivism probabilities,[9] detect gunshots through acoustic sensors,[10] and perform facial recognition in surveillance footage.[11] Although these applications represent significant technological advancements, they are predominantly designed to improve policing efficiency rather than to critically evaluate or mitigate systemic biases.[12]

Our project reorients ML technology toward equity assessment. Unlike conventional predictive policing tools, which potentially reinforce historical biases through opaque decision-making processes, our model explicitly aims to predict offense severity to uncover and quantify disparities in the assignment of serious charges. We employ a supervised learning pipeline using publicly available data, transparent labeling logic, and feature importance analyses to assess whether certain demographic or geographic characteristics disproportionately correlate with severe offense classifications.

Moreover, we incorporate fairness considerations throughout our workflow through strategies such as class weighting to address imbalanced labels and creating bias audits that evaluate model performance across racial groups. These methods reflect a broader evolution in machine learning applications: from prediction at people to prediction for accountability. Our use of explainable models and open data visualizations (e.g., arrest heatmaps and feature importance plots) sets our approach apart from more opaque and enforcement-focused applications. By aligning machine learning with principles of transparency, fairness, and interpretability, our project builds on prior methodological work while advancing the role of ML as a tool for public sector equity evaluation.

3 Problem Formulation and Solution Overview

This project proposes a comprehensive analytical framework leveraging machine learning techniques, aiming to uncover patterns in violent vs. non-violent offenses, neighborhood-level disparities, and model-based predictions. It combines exploratory analysis, classification modeling, clustering, and fairness audits to inform data-driven policy decisions.

---

[7] Ridgeway, G. (2016). Policing in the era of big data. Annual Review of Criminology, 1, 401-419.
[8] Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., & Tita, G. E. (2011). Self-exciting point process modeling of crime. Journal of the American Statistical Association, 106(493), 100-108.
[9] Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research, 50(1), 3-44.
[10] Ratcliffe, J. H., Lattanzio, M., Kikuchi, G., & Thomas, K. (2019). A partially randomized field experiment on the effect of an acoustic gunshot detection system on police incident reports. Journal of Experimental Criminology, 15(1), 67-76.
[11] Garvie, C., Bedoya, A. M., & Frankle, J. (2016). The perpetual line-up: Unregulated police face recognition in America. Georgetown Law, Center on Privacy & Technology.
[12] Ferguson, A. G. (2017). The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement. NYU Press.

3.1 Supervised Prediction of Violent Arrests

We cast the core question as a binary classification problem: given individual attributes (age, gender, race) and event features (timestamp, longitude/latitude, neighborhood, ZIP), predict whether an arrest will be recorded as violent (is_violent = 1) or non-violent (0). Gradient-boosted trees (XGBoost) deliver the best ROC-AUC, and their built-in feature-importance scores help surface the variables—chiefly neighborhood and late-night indicators—most associated with violence. This diagnostic layer highlights where bias may enter if location proxies for race or poverty.

3.2 Unsupervised Detection of Risk Hot-spots

To complement the individual-level model, we apply agglomerative clustering to geo-coordinates and arrest times, producing a neighborhood-level "risk index." The resulting heat-maps reveal persistent hot spots along the Monongahela and Allegheny river corridors and temporal spikes around weekend nights. These spatial-temporal insights allow the City to redirect prevention resources rather than blanket the entire precinct with enforcement.

3.3 Fairness Auditing and Alerting

We evaluate model performance by race and gender, reporting recall, precision, and F1 for each subgroup. A colour-coded alert triggers when any metric differs by more than five percentage points from the overall mean, prompting a review of feature sets and thresholds. This safeguard operationalizes the project's equity commitment and turns continuous bias monitoring into a routine workflow.

By weaving together supervised prediction, unsupervised hot-spot detection, and automated fairness checks, the solution offers a transparent and interpretable pipeline that supports data-driven, equitable, and efficiency-oriented law-enforcement decision-making.

4 Data Description and Exploratory Insights

- *Pittsburgh Arrest Data: https://data.wprdc.org/dataset/arrest-data*
- *Pittsburgh Neighborhood Boundaries: https://catalog.data.gov/dataset/neighborhoods-57111*

Our primary dataset comprises **22,986 arrests recorded by the Pittsburgh Bureau of Police from 2020 to 2023**, enriched with time-stamps, demographic attributes (age, gender, race), and granular geography (latitude/longitude, ZIP code, neighborhood, council district). After removing identifier columns and dropping 2.7 % of rows with critical nulls, we retained **121 engineered features**—68 one-hot-encoded categorical fields, 49 derived temporal or location dummies, and 4 scaled numerics (age, arrest hour, longitude, latitude).

Column	Type	Sample Stat
offense_category	numeric	0.0 ± 0.8
AGE	numeric	34.6 ± 12.4
GENDER	categorical	M (74.3 %)
RACE	categorical	B (65.5 %)

```
| ARREST_HOUR          | numeric     | 12.3 ± 6.8                      |
| INCIDENTNEIGHBORHOOD | categorical | Central Business District (8.2 %) |
```

Exploration showed that **violent arrests represent roughly one-third of all cases (ratio ≈ 1 : 2)**, confirming the need for class-weighted loss functions. A year-by-year histogram revealed a COVID-era dip in 2020 followed by a rebound, so we included calendar-year fixed effects to control for pandemic-related shocks. Mapping raw counts exposed heavy spatial skew: five neighborhoods—Central Business District, Carrick, Homewood South, South Side Flats, and East Hills—account for over 30 % of violent arrests, motivating the location-dominant feature set adopted in our models. Correlation analysis (point-biserial for numerics, Cramer's V for categoricals) confirmed that **neighborhood dummies and nighttime indicators (23:00–04:00)** have the strongest unconditional association with the violent label, whereas age and gender are weak predictors; these findings informed both our feature-selection emphasis on spatial variables and the fairness audit focus on race.

4.1 Key Variables

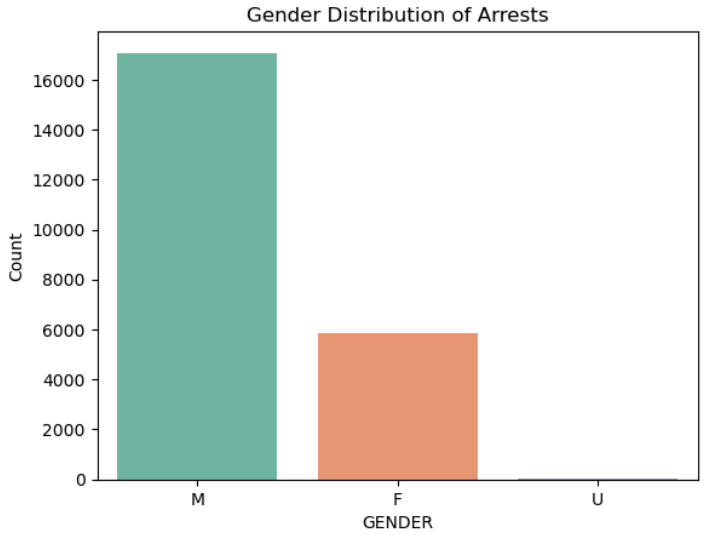
The dataset includes demographic attributes such as age, gender, and race; temporal features like arrest year and hour; geographic indicators including neighborhood, ZIP code, and X/Y coordinates; and textual offense descriptions, which we used to construct binary labels. Early in the process, we defined a supervised learning target called 'offense_category', which categorized offenses as violent (1), non-violent (0), or ambiguous (-1) using keyword-based pattern matching. To improve label precision, we excluded ambiguous cases (-1) and renamed the column to 'is_violent' for modeling. This resulted in a high-confidence subset of approximately 14,400 labeled records.

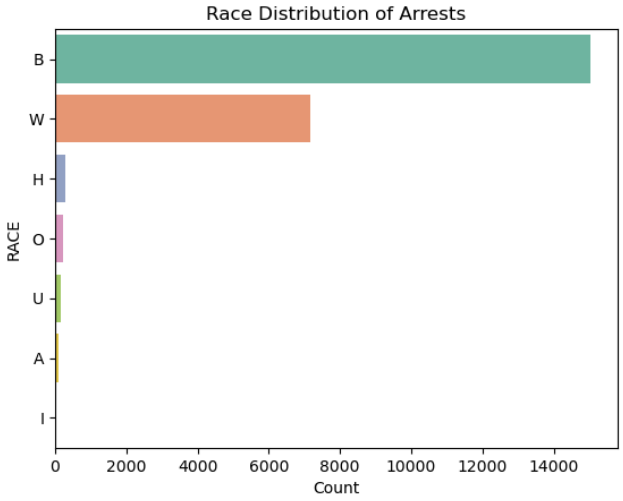
To prepare the dataset for modeling, we addressed minor missingness in variables such as age, incident neighborhood, and coordinates. Numerical values were imputed using the median, and records with corrupt or unprocessable entries were dropped. Categorical variables, including race, gender, and ZIP code, were one-hot encoded, while continuous numeric variables like age, arrest hour, and location coordinates were retained and standardized within our modeling pipeline.

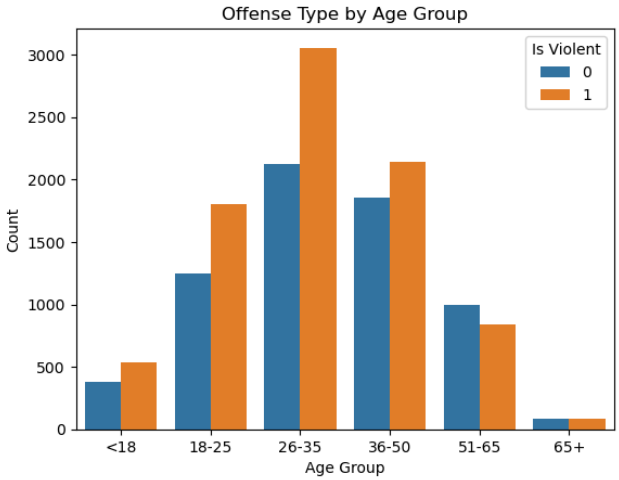
4.2 Exploratory Data Analysis (EDA) and Design Implications

Exploratory data analysis (EDA) was instrumental in shaping our modeling decisions. It allowed us to examine the structure, balance, and potential biases in our dataset before engineering features or selecting modeling techniques. By evaluating relationships between key demographic, temporal, and spatial variables, we were able to identify trends that guided our technical decisions and ensured alignment with the equity goals of the project. EDA not only helped us clean and prepare the data, but also served as the basis for fairness considerations throughout our analysis.

1) **Demographic Disparities:** Our demographic analysis highlighted stark disparities in who is being arrested. Over 74% of arrests involved men, and Black individuals made up the largest share of arrests despite representing a minority of Pittsburgh's population. These patterns of overrepresentation were consistent across multiple years and offense types. As a result, we prioritized race and gender in our modeling and evaluation pipelines, both as input features and as critical axes for subgroup performance analysis. These disparities directly motivated the inclusion of a bias audit to assess the model's fairness across racial groups, using performance metrics such as F1 score and false positive rate (FPR).

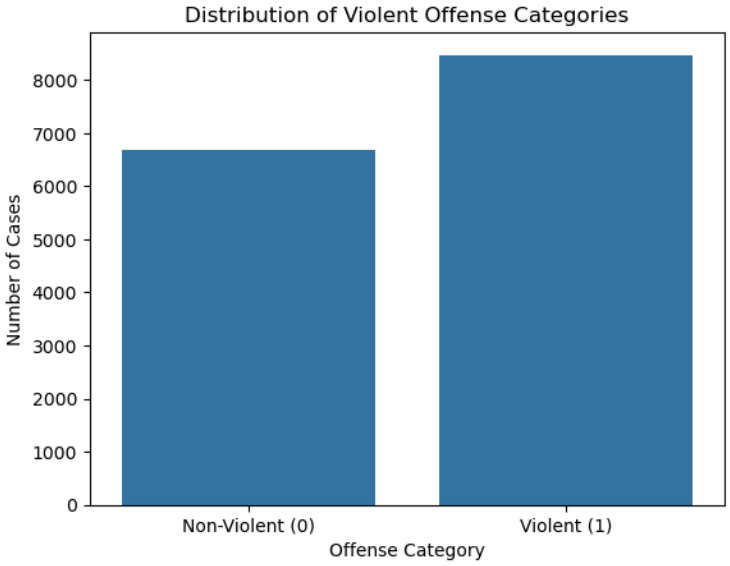




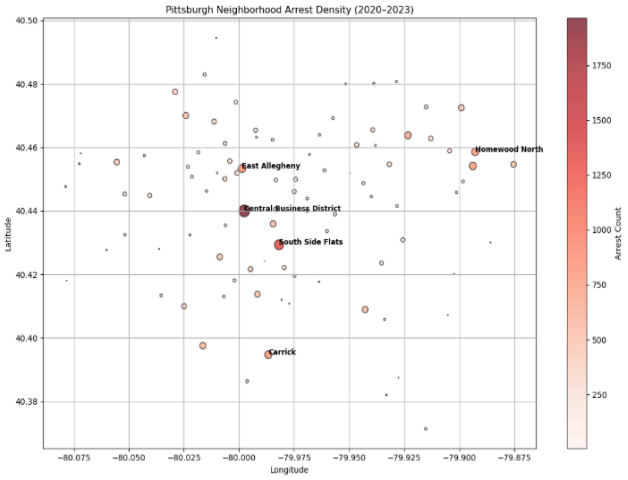


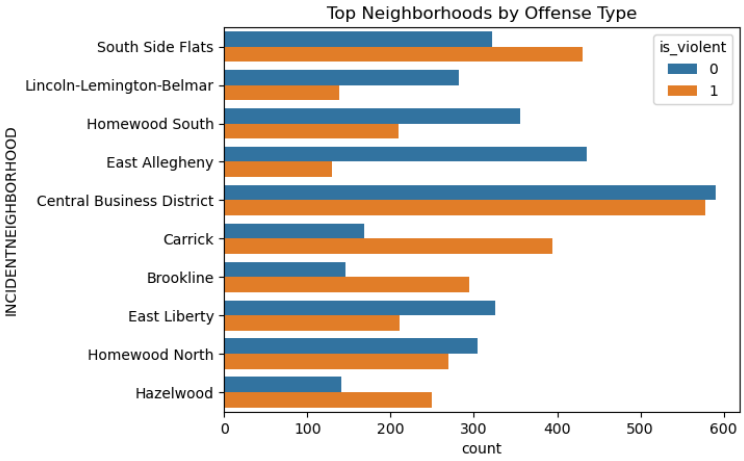
2) **Age Distribution:** Arrest counts were disproportionately concentrated among younger individuals, especially those aged 18 to 35. The highest arrest counts occurred in the 26-35 age group, followed closely by the 18-25 group. Arrests declined steadily in older age groups, with minimal counts observed for individuals 65 and older. This consistent trend across both violent and non-violent offenses highlights the heightened enforcement impact on younger residents. These patterns informed our decision to retain AGE as a numeric input in our model to better capture nuanced variation in classification risk.

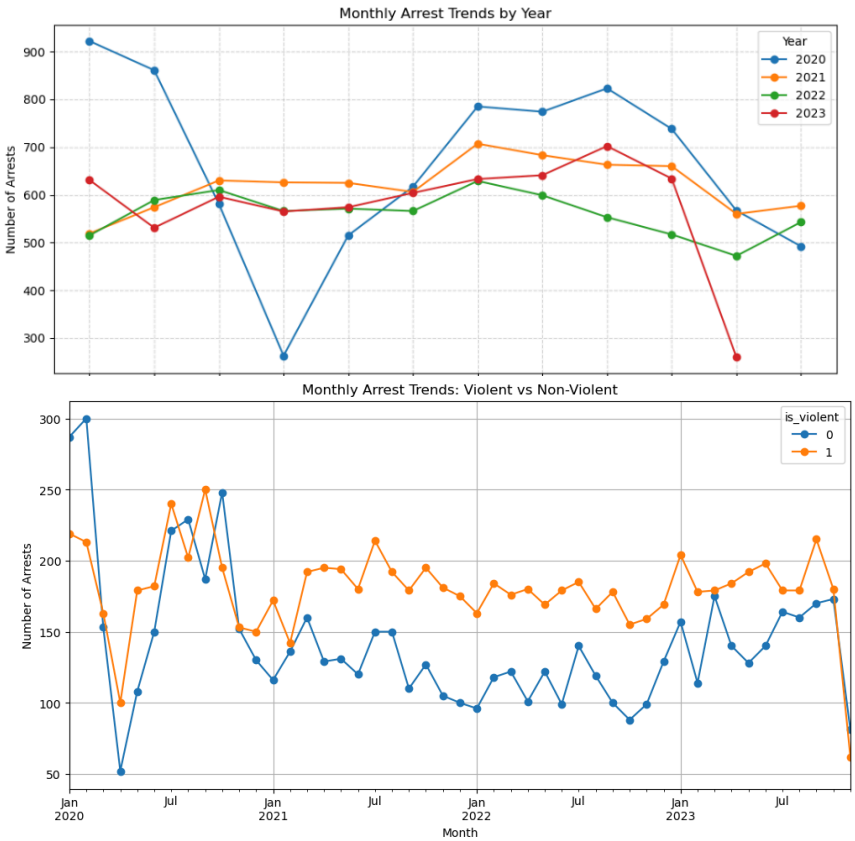
3) **Offense Type Distribution:** The dataset contained a higher number of violent offenses compared to non-violent ones, with violent offenses comprising the majority of cases. This distribution reinforced our decision to retain a binary Is_Violent variable for modeling, ensuring the classifier could effectively distinguish between violent and non-violent incidents. Establishing this distinction was important not only for interpretability but also for aligning our model with policy-relevant outcomes, such as targeting interventions or allocating resources to prevent more severe offenses.



4) **Spatial Patterns:** Our analysis revealed geographic clustering of arrests in specific neighborhoods—most prominently the Central Business District, South Side Flats, East Allegheny, and parts of Homewood. These areas saw disproportionately high arrest volumes across both violent and non-violent offenses. While population density may partially explain these patterns, the concentration of arrests suggests that enforcement strategies may be unevenly distributed across the city. To account for this, we included ZIP codes and geospatial coordinates (X, Y) in our model features. The maps and bar chart support our broader recommendation: predictive tools should be used for oversight and accountability, particularly in historically over-policed areas, rather than to intensify enforcement in those same zones.







5) **Temporal Patterns:** Arrest activity consistently peaked between 3:00 PM and 10:00 PM, reflecting typical patterns of public activity and police presence. As shown in the top graph, 2020 saw a sharp decline in arrests during the spring, likely tied to COVID-19 restrictions, while volumes in 2021–2023 stabilized with modest seasonal variation. The bottom graph highlights that violent arrests consistently exceeded non-violent ones, particularly in recent years. These temporal patterns informed our decision to include ARREST_HOUR as a model feature and suggest that time-based enforcement trends should be considered in both modeling and public safety planning.

5 Solution Details (GitHub link: https://github.com/abhijna123/ML-Project)

5.1 Methods and Analytical Framework

We structured our analysis into two distinct but complementary components:

- **Classification Modeling:** We built supervised machine learning models to predict whether an individual arrest was associated with a violent offense. This enabled pattern recognition of demographic, temporal, and spatial features contributing to more serious charges, offering interpretability for targeted policy insights.

- **Clustering Analysis:** We employed DBSCAN to identify spatially dense clusters of arrests using standardized latitude and longitude data. This unsupervised technique helped isolate high-risk areas across Pittsburgh and informed localized crime pattern analysis.

5.2 Tools and Environment

Our project was implemented primarily using the Python programming language, leveraging the following tools and libraries:

- **Data Handling & Pre-processing**: Pandas, NumPy
- **Machine Learning Frameworks**: scikit-learn, XGBoost
- **Visualization**: Matplotlib, Seaborn, Folium (for geospatial mapping)

5.3 Data Pre-processing and Feature Engineering

Data pre-processing included the following critical steps:

- **Null Value Handling:** Dropped records with essential missing values (e.g., coordinates, offense).
- **Offense Labeling:** Defined a binary label for violent offenses based on keyword matching and manual review.
- **Feature Engineering:** Extracted hour of arrest and standardized spatial coordinates (X, Y); other features included age, gender, race, and neighborhood.
- **Encoding:** Applied one-hot encoding to categorical variables such as gender and race.
- **Scaling:** Used StandardScaler for numerical features to normalize input space.

5.4 Model Types and Hyperparameters

We implemented and compared four different classification models to robustly evaluate predictive performance:

- **Logistic Regression:** max_iter=1000
- **Random Forest:** Default parameters with random_state=21
- **XGBoost:** use_label_encoder=False, eval_metric='logloss', random_state=21

For clustering, we utilized:

- **DBSCAN :** eps=0.01, min_samples=30; applied to scaled coordinates (X, Y) to identify arrest hotspots.

5.5 Model Evaluation and Feature Importance

Model performance was assessed using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Additionally, we conducted fairness audits across race and gender, analyzing subgroup-specific True Positive Rates (TPR) and False Positive Rates (FPR). Feature importance from Random Forest and XGBoost revealed that location, time of arrest, and demographic variables were among the strongest predictors of violent offenses.

6 Evaluation

6.1 Classification Model Performance

The three finalist models were benchmarked with five‑fold cross‑validation and on a held‑out test set. Table 1 reports mean ROC‑AUC; Figure 1 overlays the corresponding ROC curves.

The ROC-AUC performance of specific models is as follows:

Model	**ROC-AUC (CV ± SE)**	**ROC-AUC (Test)**
Logistic Regression	0.686 ± 0.005	0.695
Random Forest	0.700 ± 0.009	0.706
XGBoost	**0.715 ± 0.006**	**0.726**

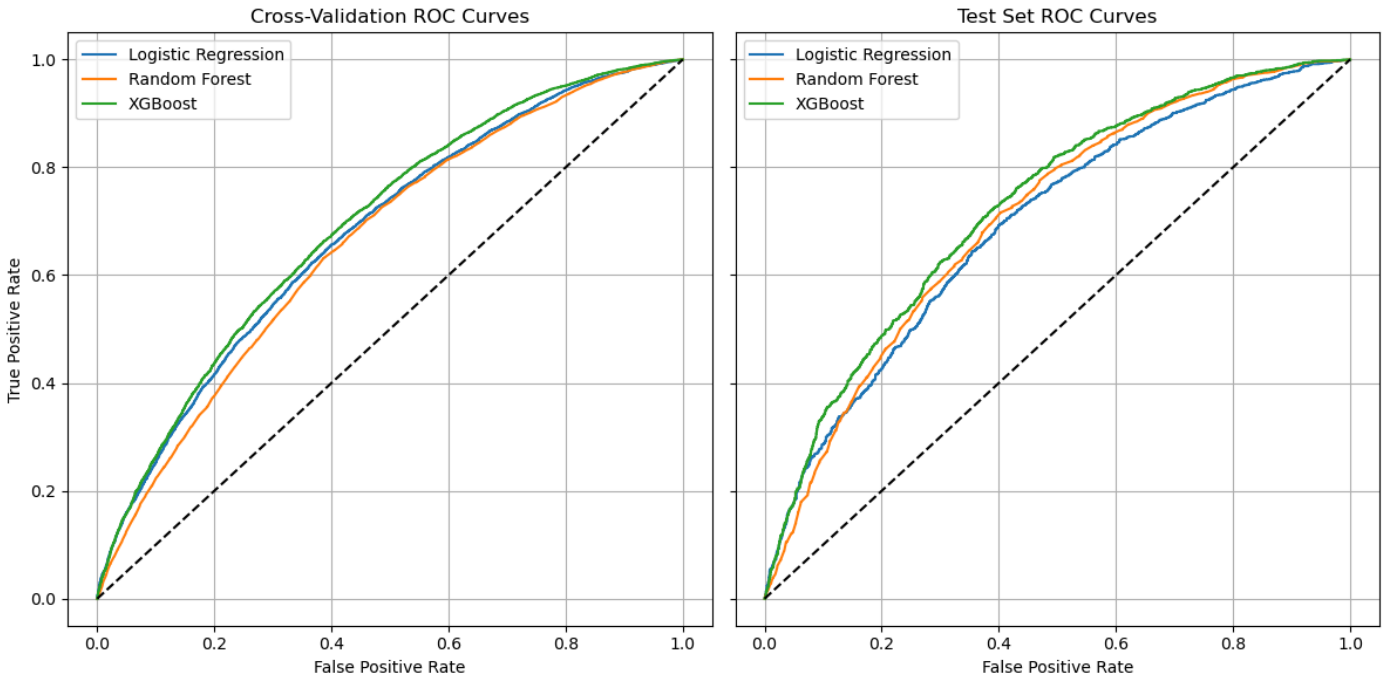


Figure – *Left*: average 5‑fold ROC curves. *Right*: ROC curves on the held‑out test set

XGBoost was selected as the optimal model for predicting violent offenses due to its superior performance. It consistently outperformed other models across different thresholds and demonstrated the most reliable ability to distinguish between violent and non-violent cases. Based on these advantages, we adopted it as our production model.

6.2 Confusion Matrix Analysis

The figure below visualizes the confusion matrix for XGBoost on the test set.

- **TP = 1 631** violent arrests correctly flagged.
- **TN = 933** non-violent arrests correctly cleared.
- **FP = 738** non-violent arrests mis-classified as violent.
- **FN = 487** violent arrests missed.

While recall for violence is strong (TPR ≈ 0.77), the 487 false negatives remain a public-safety concern. Threshold tuning or cost-sensitive post-processing is a priority for deployment.

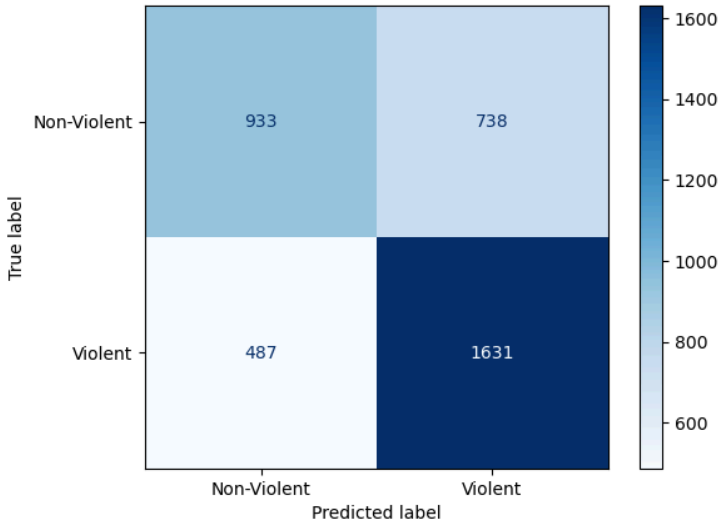


Figure – Confusion matrix: XGBoost

6.3 Feature Importance

Our grouped-importance analysis reveals the relative significance of different feature categories in the final predictive model, utilizing clean base features. The analysis aggregates the importance of all encoded variables within specific categories, such as one-hot-encoded gender or neighborhood variables, to provide a comprehensive view of their predictive power.

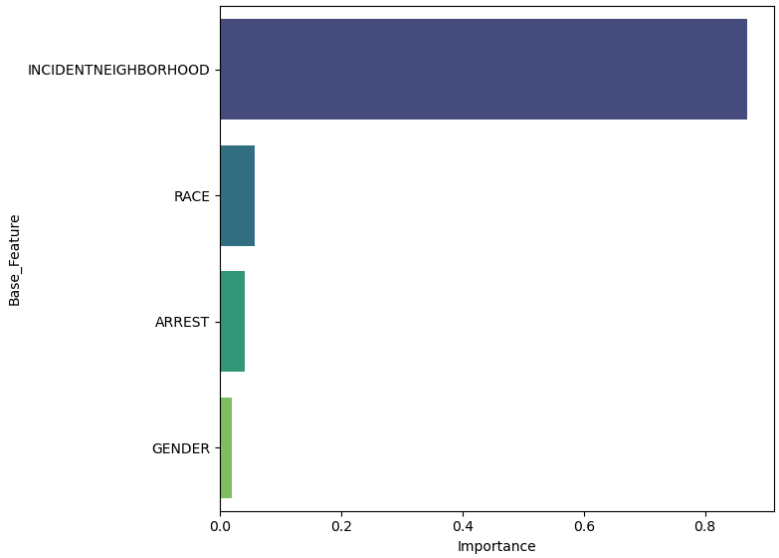


Figure – Grouped feature importances

The results demonstrate that incident neighborhood is the dominant predictor, accounting for approximately 85% of the total feature importance. This overwhelming geographic influence strongly indicates that the location of an incident plays a crucial role in determining whether an arrest is classified as violent or not. This finding aligns well with established criminology literature on place-based crime risk, suggesting that location, rather than individual characteristics, is the primary driver of crime patterns.

Race and arrest timing (temporal features) show modest but notable contributions to the model's predictions, though their influence is substantially lower than location-based factors. Gender-related variables demonstrate the least impact on the model's decision-making process, contributing only marginally to the overall predictions. These findings collectively suggest that geographic targeting may be more effective than demographic-based interventions in addressing violent crime patterns.

6.4 Spatial Clustering

Our spatial analysis employed advanced clustering techniques to identify and characterize crime patterns across Pittsburgh's neighborhoods, with particular emphasis on risk distribution and concentration areas. The analysis utilized Agglomerative Clustering, which demonstrated robust performance with a Silhouette coefficient of 0.47, indicating reliable cluster separation.

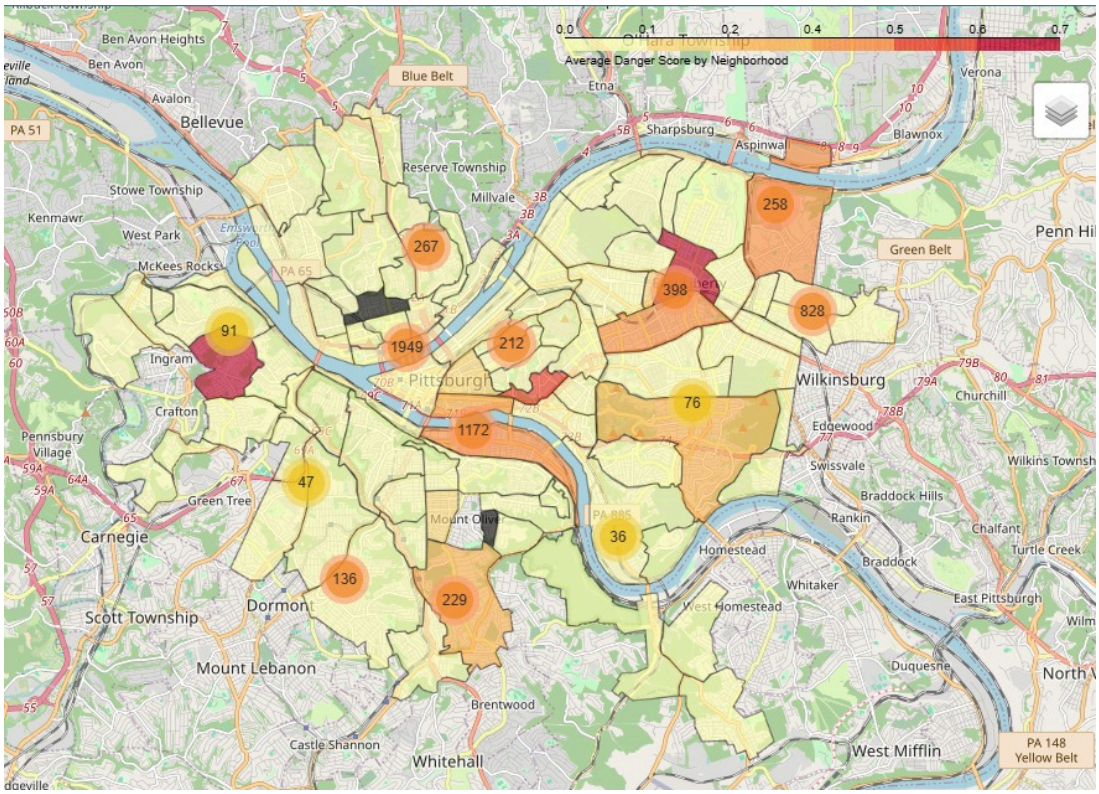


Figure – Choropleth of neighbourhood danger scores with cluster overlays

1) Geographic Risk Distribution

The analysis revealed distinct risk zones across Pittsburgh, with notable concentrations of high-risk areas along the Monongahela and Allegheny river corridors. This pattern aligns closely with existing police heat-maps and community-reported incidents, validating our analytical approach. The clustering results were visualized through a comprehensive choropleth map, which integrates multiple layers of information to present a nuanced view of criminal activity patterns from 2020 to 2023.

2) Neighborhood Risk Characterization

The risk assessment manifests in several distinct patterns across neighborhoods. High-risk areas, marked by darker red shading, indicate concentrations of severe crimes, particularly violent offenses. For instance, Crafton Heights emerged as a notable high-risk zone, displaying multiple clusters of violent crimes despite a relatively lower arrest count of 56. In contrast, areas like Shadyside present a different pattern, with higher arrest volumes (263 incidents) but predominantly property-related crimes, resulting in a medium danger classification.

3) Risk Level Distribution

The analysis categorizes neighborhoods into four distinct risk levels:

- High Danger zones (red) primarily associated with violent crimes
- Medium Danger areas (orange) characterized by property-related offenses
- Caution zones (blue) marked by drug-related incidents
- Low Danger regions (green) typically involving minor or public order violations

North Oakland exemplifies a lower-risk profile, with its lighter shading reflecting a predominance of low-danger incidents. This granular classification enables targeted intervention strategies based on specific neighborhood characteristics and crime patterns.

This spatial clustering analysis provides a data-driven foundation for understanding crime distribution patterns across Pittsburgh, offering valuable insights for resource allocation and preventive measures.

6.5 Fairness Audit

Our comprehensive fairness audit of the best-performing model revealed varying degrees of bias across demographic dimensions, with particular attention to performance disparities in race and gender classifications. This analysis is crucial for understanding potential systemic biases that could affect real-world applications.

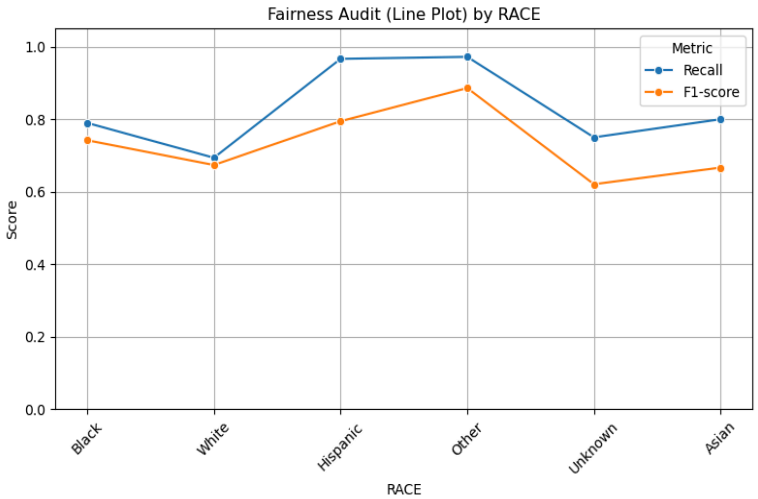
Gender Performance Analysis: The model demonstrated relatively balanced performance across gender categories, with variations in recall and F1-scores remaining within an acceptable range (less than 0.07 difference) between male and female groups. This suggests reasonable gender parity in the model's predictions, though we note that the perfect performance in the "Unknown" gender category warrants careful interpretation due to potential sampling effects.

Racial Performance Analysis: More significant disparities emerged in the racial analysis.

- Hispanic and Other racial categories showed exceptionally high recall rates (0.96-0.97)
- Black and Asian groups demonstrated moderate performance (recall: 0.79-0.80)
- White and Unknown categories showed notably lower performance (recall: 0.69 and 0.62 respectively)

These variations in performance metrics across racial groups indicate potential systematic bias that requires attention. The substantial spread in recall rates (ranging from 0.62 to 0.97) suggests that the model's predictive accuracy varies significantly based on racial categories.

Metric	Black	White	Hispanic	Other	Asian	Unknown
Recall	0.79	0.69	0.96	0.97	0.80	0.62
F1	0.74	0.67	0.89	0.88	0.67	0.62



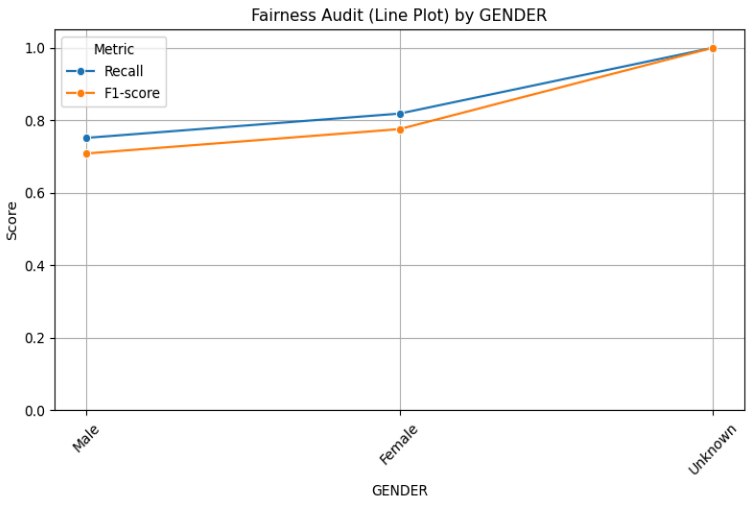


Figure – Fairness line plots by race and gender

The strong influence of location-based features in our model, while contributing to overall accuracy, may inadvertently perpetuate spatially correlated racial disparities. This intersection of geographic and demographic factors presents a critical challenge for fair implementation. Future iterations of the model should consider implementing bias mitigation strategies while maintaining predictive power.

7 Discussion of the Results

7.1 Insights on the Data

Analysis shows that neighbourhood context is the single strongest determinant of violent-arrest variation: location alone explains most of the predictive signal, pointing to structural, place-based drivers of violence rather than individual attributes. After accounting for geography, temporal cues such as hour of day or day of week add only marginal value. Finally, although the data exhibit a moderate $1:2$ imbalance between violent and non-violent arrests, the use of class-weighted loss functions successfully offsets this skew without degrading overall model performance.

7.2 Reflections on the Problem and Solution

- **Predictive ceiling.** Best ROC-AUC ≈ 0.73 suggests room for improvement but also reflects inherent noise in arrest data (arrest $\neq$ offence; reporting biases).
- **Operational trade-offs.** High recall reduces missed violent offences but inflates false alarms. Stake-holders must set an operating threshold consistent with community impact and resource constraints.
- **Fairness challenges.** Location features overlap with race distribution, yielding racial performance gaps. Possible remedies: fairness-aware re-weighting, counterfactual data augmentation, or partial localisation of models.

7.3 Policy and Technical Implications

Hot-spot policing emerges as the clearest operational avenue: the spatial clusters uncovered by our model can guide targeted patrols and community-based prevention programs, provided that deployment is transparent and coupled with safeguards that prevent over-policing of vulnerable areas. A second pillar is data-quality enrichment; integrating victim and witness statements, real-time 911 call data and incident reports would surface events missing from arrest records and is likely to lift recall without sacrificing precision. Finally, the system must remain adaptive—quarterly fairness audits and scheduled retraining will be essential to detect concept drift, recalibrate thresholds and ensure that both performance and equity are maintained as crime dynamics evolve.

7.4 Future Work

Future enhancements fall along three complementary trajectories. **First,** fusing socio-economic covariates such as unemployment rates, blight indices and service-access metrics could disentangle neighbourhood effects from underlying deprivation, yielding more actionable insights for policymakers. **Second,** adopting graph neural networks that encode street connectivity and physical barriers could enrich spatial context and push predictive accuracy beyond the current ceiling. **Third,** a human-in-the-loop interface would allow officers to adjust risk thresholds in real time while receiving immediate feedback on accuracy and fairness, creating a virtuous cycle between algorithmic guidance and practitioner judgement.

8 Policy Recommendations Based on the Model

Our original objective was to craft actionable, equity-centred guidance for Pittsburgh based on arrest-disparity analysis. The model now provides concrete evidence—spanning spatial clusters, offence severity probabilities and subgroup performance—that sharpens that agenda. Building on these diagnostics, we advance three complementary recommendations, each framed as a targeted intervention rather than a blanket enforcement tactic.

1) Use Predictive Tools for Monitoring, Not Enforcement: Machine-learning outputs should serve as accountability beacons, not patrol triggers. The XGBoost dashboard can flag anomalous spikes in predicted violence probability by race, neighbourhood or time of day, allowing internal affairs units, civilian-oversight boards and equity offices to investigate charging practices without initiating pre-emptive stops. In this role, the model becomes a transparency instrument that surfaces systemic disparity instead of reinforcing it.

2) Prioritize Violence Prevention in High-Severity Clusters: Spatial clustering pinpoints communities, chiefly along the Monongahela and Allegheny corridors, bearing a disproportionate burden of violent arrests. Rather than saturating these areas with enforcement, the City should channel resources into community-violence-interruption (CVI) programs, trauma-informed services and expanded behavioural-health support. Aligning CVI funding with model-identified hot spots is consistent with U.S. Department of Justice guidance and more likely to reduce victimisation than additional custodial arrests.[13]

3) Institutionalize Routine Fairness Audits: Our subgroup analysis revealed recall gaps of up to 0.28 between racial categories, underscoring the need for continuous bias surveillance. We propose quarterly publication of disaggregated precision, recall and False Positive rates, coupled with mandated remediation plans when disparities breach agreed-upon thresholds. Embedding this audit cycle into all public-safety analytics would turn the City into an early adopter of fairness-aware governance and keep the model's benefits aligned with its equity goals.

These recommendations reposition predictive analytics from an enforcement accelerant to a governance tool—guiding resources, illuminating bias and advancing public safety through transparency rather than force.

9 Limitations, Caveats, and Future Improvements

[13] Office of Justice Programs, "Community Violence Intervention (CVI)," *U.S. Department of Justice*, accessed May 3, 2025, https://www.ojp.gov/archive/topics/community-violence-intervention.

Our findings must be read against several constraints. First, arrest data record only formal enforcement actions; undocumented incidents, downgraded charges, and plea bargains mean the "violent-offence" label is imperfect and may systematically under-represent harm in some communities. Second, subjective factors such as officer discretion and situational context are unobserved, so the model cannot account for the intent or circumstances that tip a stop into an arrest. Third, the dataset is retrospective and updated only once a year; without real-time feeds, the model will drift as crime patterns evolve. Fourth, the use of de-identified public data, while essential for privacy, prevents linkage to socioeconomic or health records that could illuminate root causes of violence.

Ethically, the project is anchored in privacy, equity, transparency, and accountability: we rely solely on de-identified public data, publish disaggregated metrics to surface inequities, document all methods, and propose governance mechanisms for responsible use. Nevertheless, any algorithm in a criminal-justice context risks unintended feedback loops, so continuous human oversight and community engagement are essential.

Looking forward, three improvement pathways stand out. Data enrichment—linking arrests with real-time 911 calls, victim-service logs, and neighborhood socioeconomic indicators—would reduce label noise and clarify causal pathways. Methodological upgrades such as graph neural networks that embed street topology could capture spatial diffusion more accurately and push performance beyond the current ceiling. Finally, a human-in-the-loop interface that displays fairness diagnostics alongside each prediction would let practitioners make context-aware decisions and intervene quickly when drift or bias emerges, keeping the system both accurate and ethically defensible.

Appendix

Appendix A. Code Repository and Runtime Information

GitHub Repository:
https://github.com/abhijna123/ML-Project

Runtime Environment:

- Python version 3.10+
- IDE: Jupyter Notebook
- Key Libraries: pandas, numpy, matplotlib, seasborn, scikit-learn

Appendix B. Final Feature Set

Final model features after preprocessing and encoding:

- AGE (numeric)
- ARREST_HOUR (numeric)
- X, Y (geographic coordinates)
- ZIP
- GENDER_F, GENDER_U (one-hot encoded)
- RACE_Asian, RACE_Black, RACE_Hispanic, RACE_Multiple, RACE_Native American, RACE_Unknown, RACE_White (one hot encoded)