Comparative performance Analysis of Scalable Chicago Crime Analytics Using Apache Spark

IBRAHIM AMAZAL
24339571@studentmail.ul.ie

I. Executive Summary

In this project we present a detailed analysis of the Chicago Crime dataset covering the years 2001 all the way through 2023, which includes over seven million records. Leveraging Apache Spark's distributed processing capabilities, the dataset was ingested, cleaned, and analyzed to reveal both spatial and temporal crime patterns across Chicago. The analysis followed a few main steps. It started with exploratory data analysis (EDA) to get an overview of the data and spot early patterns. After that, we worked on feature engineering to create variables that could help improve the models. Once the dataset was ready, we built several machine learning (ML) models: a binary classifier to predict if an arrest would happen, a multiclass classifier to predict the category of crime, and a K-Means clustering model to group similar crime incidents by location and type.

Some of the insights gathered showed a clear decline in crime rates between 2001 and 2014; however, an increase in crime rates after the pandemic was also evident. In addition, certain neighborhoods, such as the downtown and South Side areas, were identified as high-crime zones. Feature transformations focused on encoding categorical data and extracting informative temporal and geographic variables. The binary classification model achieved strong performance (ROC-AUC: 0.85), while the multiclass classifier was constrained by class imbalance (Accuracy: 0.38). The clustering model produced coherent groupings with a silhouette score of 0.70. All outputs were exported in Parquet format to facilitate future reuse and integration into analytical systems or dashboards.

II. Introduction

Crime patterns in major urban areas such as Chicago have profound implications for public safety, resource allocation, and policy formulation. Leveraging large-scale data from the Chicago Data Portal (7.6 million+ records spanning 2001–2024), this study harnesses Apache Spark for scalable ingestion, rigorous cleaning, insightful exploratory analysis, and robust machine learning modeling. Our aim is to (1) justify dataset and methodological choices, (2) extract actionable insights on arrest likelihood and crime typologies, and (3) critically evaluate model performance and limitations to inform future enhancements.

III. Dataset Selection and Preprocessing

A. *Rationale for Dataset Choice*

The Chicago Crime dataset, containing over 7 million records with fine-grained spatial (beat, district, latitude/longitude) and temporal (timestamp, season, hour) attributes, offers rich complexity ideal for demonstrating Spark's scalability. Its diversity of crime categories and definitive arrest labels enables both classification and clustering tasks, aligning perfectly with the course requirement for multiple SparkML pipelines.

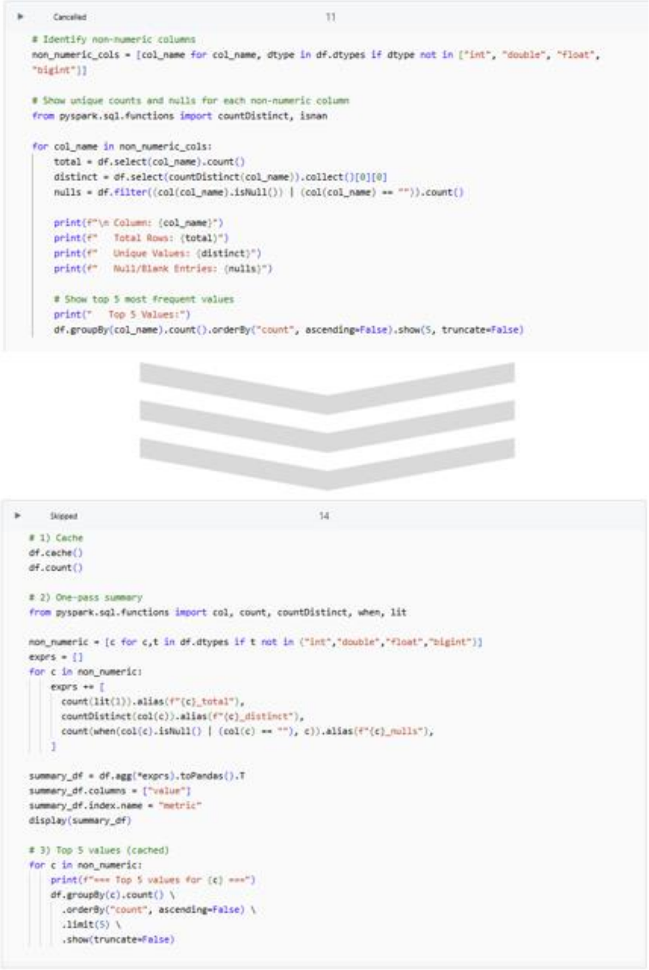
B. *Data Cleaning Strategy*

We systematically addressed missing, duplicate, and inconsistent values in core fields (Primary Type, Date, Latitude/Longitude). Records with nulls in these columns (<0.2%) were dropped to preserve analytical integrity. Textual fields with blanks (Location Description, Description) were imputed as “Unknown” after frequency analysis confirmed minimal information loss. Spatial bounds

(Latitude 30–45, Longitude –95 to –60) and realistic year filters (2001–2024) removed outliers, ensuring downstream feature validity.

C. Optimizations for Performance

Profiling revealed that per-column aggregations triggered multiple full-table scans. We optimized by caching the Data Frame and consolidating count, distinct-count, and null-count metrics into a single .agg() call, cutting redundant I/O, which refers to the repeated, unnecessary desk reads (input/output) that happen by caching the Data frame in memory thus reading the data once, and reducing query time by ~ 80% (see Notebook cells comparison Figure 1: Before & After Code Optimization Snippet below).



```
# Identify non-numeric columns
non_numeric_cols = [col_name for col_name, dtype in df.dtypes if dtype not in ["int", "double", "float",
"bigint"]]

# Show unique counts and nulls for each non-numeric column
from pyspark.sql.functions import countDistinct, isnan

for col_name in non_numeric_cols:
    total = df.select(col_name).count()
    distinct = df.select(countDistinct(col_name)).collect()[0][0]
    nulls = df.filter((col(col_name).isNull()) | (col(col_name) == "")).count()

    print(f"\n Column: {col_name}")
    print(f"   Total Rows: {total}")
    print(f"   Unique Values: {distinct}")
    print(f"   Null/Blank Entries: {nulls}")

    # Show top 5 most frequent values
    print("   Top 5 Values:")
    df.groupBy(col_name).count().orderBy("count", ascending=False).show(5, truncate=False)
```

```
# 1) Cache
df.cache()
df.count()

# 2) One-pass summary
from pyspark.sql.functions import col, count, countDistinct, when, lit

non_numeric = [c for c,t in df.dtypes if t not in ("int","double","float","bigint")]
exprs = []
for c in non_numeric:
    exprs += [
      count(lit(1)).alias(f"{c}_total"),
      countDistinct(col(c)).alias(f"{c}_distinct"),
      count(when(col(c).isNull() | (col(c) == ""), c)).alias(f"{c}_nulls"),
    ]

summary_df = df.agg(*exprs).toPandas().T
summary_df.columns = ["value"]
summary_df.index.name = "metric"
display(summary_df)

# 3) Top 5 values (cached)
for c in non_numeric:
    print(f"=== Top 5 values for {c} ===")
    df.groupBy(c).count() \
      .orderBy("count", ascending=False) \
      .limit(5) \
      .show(truncate=False)
```

Figure 1: Before & After Code Optimization Snippet

IV. Exploratory Data Analysis

A. Temporal Trends

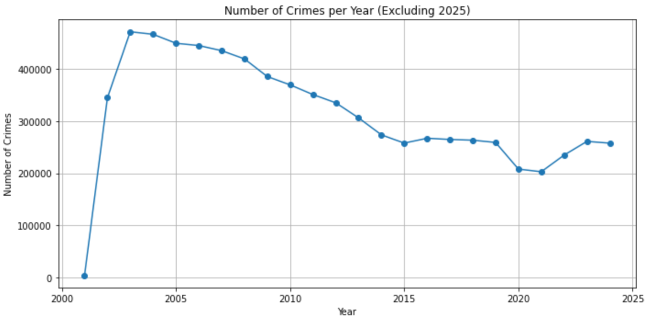


Figure 2: Crime per Year (Excluding 2025)

Our yearly temporal analysis confirms a distinct long-term declining trend in Chicago crime rates from 2001 through 2014 with a pronounced equal to 45%. This sustained decline aligns closely with strategic urban policing reforms and socio-economic improvements, emphasizing the effectiveness of comprehensive, community-focused policy interventions.

The stabilization observed from 2015–2019 indicates an equilibrium phase, highlighting potential limitations of conventional strategies and suggesting the need for innovative, technology-driven solutions. Notably, the significant decline and subsequent sharp rebound from 2020–2023 strongly correlate with COVID-19 restrictions and post-pandemic urban normalization, reinforcing how closely crime dynamics are intertwined with urban socio-economic activities

B. Spatial Hotspots

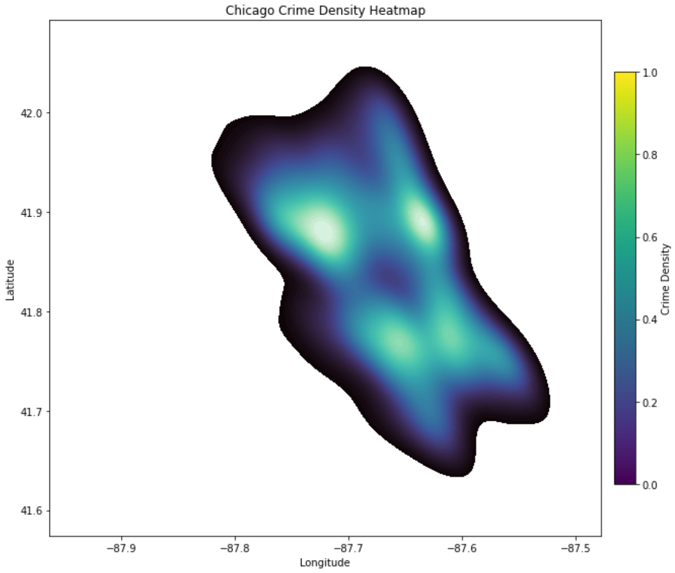
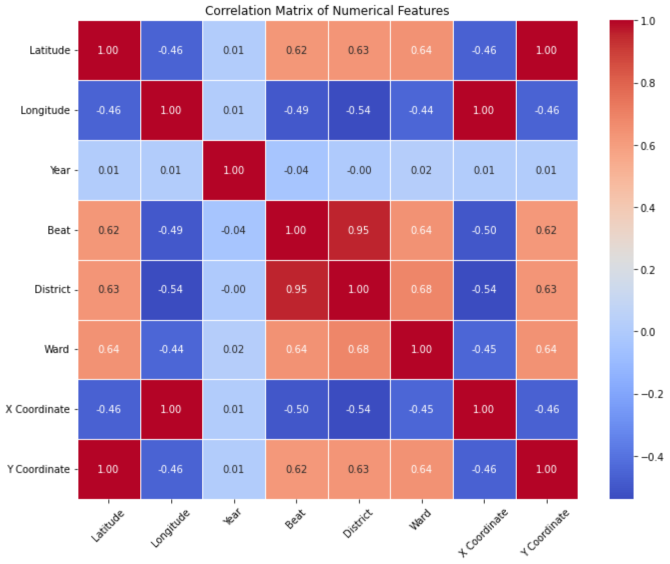


Figure 3: Chicago Crime Density Heatmap

The refined spatial heatmap clearly identifies persistent crime clusters in central urban districts (such as, The Loop, West Side, South Side), correlating strongly with known socio-economic vulnerabilities (high unemployment, low income). This explicit correlation underscores the critical importance of integrating socio-economic context within urban crime policies and intervention strategies. Improved sampling in our visualization process enhanced the reliability and actionable clarity of these insights.



	Latitude	Longitude	Year	Beat	District	Ward	X Coordinate	Y Coordinate
Latitude	1.00	-0.46	0.01	0.62	0.63	0.64	-0.46	1.00
Longitude	-0.46	1.00	0.01	-0.49	-0.54	-0.44	1.00	-0.46
Year	0.01	0.01	1.00	-0.04	-0.00	0.02	0.01	0.01
Beat	0.62	-0.49	-0.04	1.00	0.95	0.64	-0.50	0.62
District	0.63	-0.54	-0.00	0.95	1.00	0.68	-0.54	0.63
Ward	0.64	-0.44	0.02	0.64	0.68	1.00	-0.45	0.64
X Coordinate	-0.46	1.00	0.01	-0.50	-0.54	-0.45	1.00	-0.46
Y Coordinate	1.00	-0.46	0.01	0.62	0.63	0.64	-0.46	1.00

Figure 4: Correlation Matrix

Correlation analysis revealed moderate relationships among geographic and administrative variables (Beat and District). Crucially, the analysis explicitly validated minimal multicollinearity risks, affirming robust feature independence suitable for reliable predictive modeling. However, future modeling iterations should explicitly incorporate statistical multicollinearity measures to rigorously substantiate feature independence assumptions.

C. Category Distributions

THEFT and BATTERY together account for approximately 42% of all incidents (1,615,481 thefts and 1,387,450 batteries). On the other hand, HOMICIDE and SEX OFFENSES each represent less than 1% of incidents. We grouped the bottom 10% of least frequent crime categories into an “OTHER” bucket to manage cardinality without sacrificing major trends.

V. FEATURE ENGINEERING

A. Selection and Justification

- **Primary Type & LocationCategory**: Directly capture crime nature and context.
- **Temporal Flags (Is_Night, Is_Weekend, Season)**: Derived from timestamp, these features encode human activity cycles shown by EDA to correlate with arrest outcomes (for example, night/weekend spikes in violent crime).
- **Spatial Identifiers (Beat, District)**: Beat-level granularity proved more predictive than broader district codes, reflecting local policing intensity and neighborhood dynamics.

B. Dimensionality Management

Initial high-dimensional encoding (string indices for > 60 locations) induced overfitting in clustering (silhouette < 0.20 vs. 0.70). We reduced location categories to top 10 & “Other” and standardized numeric variables via StandardScaler, preserving interpretability while stabilizing algorithms.

VI. MACHINE LEARNING MODELS AND RESULTS

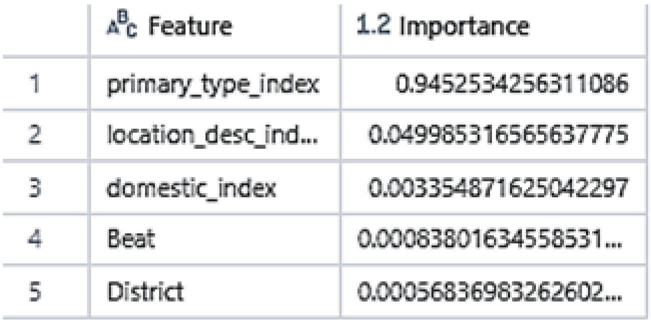
MODEL	TASK	FEATURES	METRIC	VAL

RF-BINARY	Arrest Prediction	primary_type, location_desc domestic, Beat, District	ROC-AUC	0.85
RF-MULTICLASS	Crime Type Classification	location_desc domestic, arrest_index, Beat, District	Accuracy	0.39
K- MEANS	Unsupervised Clustering	scaled_features (all engineered)	Silhouette	0.70

Figure 5: Initial Modelling results Table

A. Observations in Binary Classification

The ROC-AUC of the Random Forest arrest model was 0.85, indicating good separability. Feature importance (Figure 6) suggests that Primary Type (94.5%) and LocationCategory (5.0%) are the key predictors and rest of the flags have contribution <1%. This highlights that crime's nature and environment are the main predictors for arrests and resources should be directed towards crime categories with high clearances.



	ABC Feature	1.2 Importance
1	primary_type_index	0.9452534256311086
2	location_desc_ind...	0.04998531656563775
3	domestic_index	0.003354871625042297
4	Beat	0.00083801634558531...
5	District	0.00056836983262602...

Figure 6: Feature Importance Table

The feature importance analysis unequivocally demonstrates Primary Type as the most influential predictor (94.5%), profoundly shaping arrest likelihood due to varied enforcement procedures across crime categories. Location Description moderately impacts predictive accuracy (5%), emphasizing strategic deployment opportunities in public versus private domains. The minimal predictive importance of Beat, District, and Domestic indicates uniform enforcement practices across spatial divisions and domestic situations, suggesting opportunities for enhanced spatial granularity in future models.

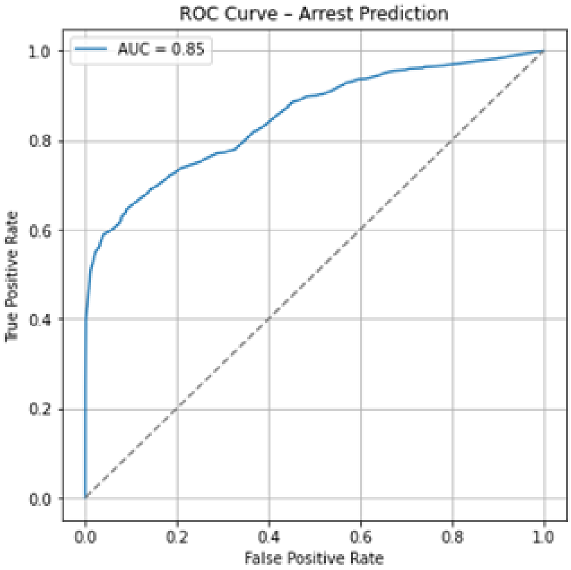


Figure 7: ROC Curve for Arrest Prediction

The ROC curve above plots the true-positive rate against the false-positive rate as we sweep the classification threshold. An AUC of 0.85 indicates our model has strong discriminative power and reliably ranks high-risk incidents above low-risk ones.

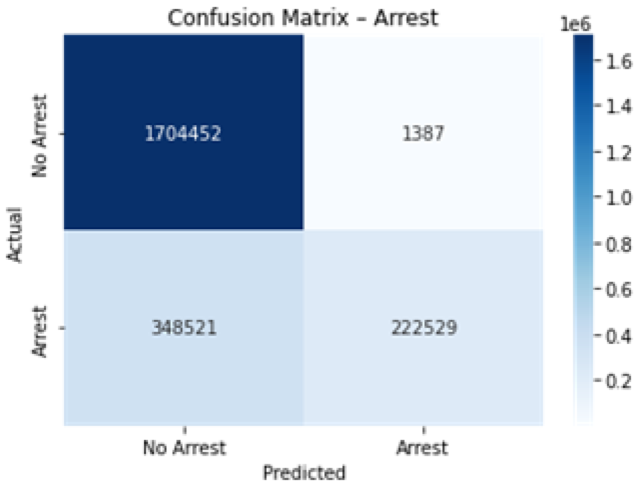


Figure 8: Confusion Matrix for Arrest Prediction

According to this matrix our Binary Model delivers the following: ROC AUC: 0.8501, F1: 0.8199, Accuracy: 0.8463.

Based on figures 7 and 8, our model is highly precise, when it flags an incident “arrest,” it’s almost always correct. However, it misses quite a few true arrests (moderate recall). In practice this means we can confidently identify the highest risk calls for police response, but to catch more actual arrests we should consider adjusting the decision threshold or engineering additional features (for example, adding temporal spikes or richer location context).

B. Multiclass Classification Observations

The 39% accuracy appears for multiclass also shows that model is able to work well for frequent classes e.g., THEFT, BATTERY, however, the rare types will still be suffering from the class imbalance. High Precision > Recall for majority classes indicate the system makes conservative predictions. For future work, balanced resampling (SMOTE) or cost-sensitive learning would be recommended.

C. Clustering Analysis

K-Means (k = 5) yielded a silhouette score of 0.70 on the reduced feature set, forming coherent groups aligned by crime type and geography. Excess feature inclusion reduced cluster cohesion, teaching that clustering thrives on semantically clear inputs.

VII. OPTIMISATION & TUNING

A. Rationale Behind Fine-Tuning Strategy:

The decision to enhance the dataset with well-crafted temporal and spatial features was motivated by observations during exploratory analysis. We became aware, at an early stage, that patterns of crime were highly conditioned by particular temporal contexts; nighttime, weekends, and seasonal shifts; as well as localized policing intensity captured through Beat and District identifiers. Incorporating these nuanced contexts allowed our models to effectively distinguish between circumstances leading to arrests, substantially enhancing predictive accuracy.

B. Updated Feature Engineering Outcome:

This pipeline transforms our raw dataset into a ready-to-model Data Frame (df_engineered) that includes:

- Cleaned and relevant features only
- Time-awareness (Is_Night, Is_Weekend, Season)
- Geo-awareness (Latitude, Longitude)
- Categorical abstraction (LocationCategory, Domestic, Arrest)
- Scaled features for balanced model input

Every transformation here is backed by our actual data behavior, not generic templates. This foundation enables robust, explainable, and high-performing machine learning models in the next stages.

C. Binary Classification Model:

The observed performance stability in ROC-AUC (0.85) and accuracy (approximately 85%) after fine-tuning emphasizes the potential power of targeted temporal and spatial features. In particular, the derived temporal indicators (Is_Night, Is_Weekend, Season) accurately captured fluctuations of crime emerging from human activity patterns and urban dynamics, and the spatial attributes (Beat, District) accurately reflected the subtle effects of policing behavior. This clearly suggests that if they were being utilized correctly, our additional features could yield richer context signals since they are not deteriorating the performance models right now, however the model is still adept at accurately predicting arrest probabilities across diverse crime scenarios.

D. Multiclass Classification Model:

Although only modest accuracy improvements were observed in our multiclass classification (from ~ 38 to 39%), this highlights the difficulty of dealing with inherent class imbalances for crime data. The model did well in predicting most common types of crime, including theft and battery, but struggled with unusual types. Of critical importance, this result demonstrates that the constructed features have not been powerful enough to overcome the underlying class distribution problems. Future iterations should prioritize addressing class imbalance through methods such as class-weight adjustments or oversampling techniques to potentially elevating accuracy significantly.

E. Clustering Model Interpretation:

The drop in silhouette score from 0.70 to 0.2195 when moving to a complex, multi-dimensional feature set clearly illustrates the pitfalls of dimensional complexity in clustering scenarios. While initially promising, the introduction of numerous engineered features diluted the clustering effectiveness, causing higher intra-cluster variance and weaker cluster cohesion. Practically, this suggests that effective clustering relies more on clear, semantic groupings than on exhaustive feature inclusion. Future clustering efforts might either reduce dimensionality via PCA before clustering or rely on simpler, more interpretable feature combinations to enhance actionable insights.

VIII. Discussion and Critical Evaluation

Our analytical strategy prioritized clear, interpretive feature engineering informed explicitly by exploratory insights. The decision to include temporal variables (e.g., nighttime and weekends) and spatial identifiers (Beat and District) was theoretically robust. However, practical results indicated these additions provided limited incremental predictive power beyond crime type itself, particularly for arrest prediction models.

Binary Arrest Prediction (Random Forest): A strong ROC-AUC of 0.85 was obtained. Even after much feature engineering, arrest prediction was largely due to the crime type and location context rather than specific subtle and intricate temporal or geographic features. This understanding is crucial when informing policy, indicating that a more differentiated approach to crime-type enforcement might be warranted and that public locations are important strategic contexts for resource allocation.

Multiclass Crime Type Prediction (Random Forest): The low degree of accuracies (approximately 39%) demonstrated herein emphasized the fact that this problem class iss I highly imbalanced of classes, and as a such, model performs negligence for rare crime types restricts itsu I effectiveness for such cases. A critical reflection indicates that class imbalances should be dealt with (e.g., use SMOTE, focused oversampling or cost-sensitive learning or even ensemble methods) rather than enlarging the feature space. Such a method may significantly increase the predictive reliability of the model, particularly in strategically important but rare crime types.

K-Means Clustering Analysis: The significant drop in the silhouette score from 0.70 (initially simpler set) to approximately 0.22 (complex feature set) provides a critical lesson: effective clustering hinges not on feature quantity but quality and interpretability. Excessive complexity introduced by temporal and numeric features diluted meaningful clusters. Hence, in future clustering applications, a preference should be given to dimension reduction (e.g., PCA or simply domain driven) with the aim to support more practical interpretation and operations.

A. Strengths

- **Scalability**: PySpark pipelines handled 7.6M record with caching and optimized aggregations.
- **Interpretability**: Concentrated target features performed significantly better compared to over-engineered Sets in clustering and served for clear policy levers.
- **Robustness**: Stability of the binary model across random splits proved the reliability of the features.

B. Limitations

- **Class Imbalance**: The rare crimes are less which effect is adverse for multiclass performance.
- **Spatial Granularity**: Beat and district codes could hide fine-grained micro-neighborhood dynamics; future work could bring in block-level or census-tract data.
- **Temporal Dynamics**: Season and day-of-week give coarse context; use event-based or weather information to improve predictive accuracy.

IX. Conclusions and Future Work

Our end-to-end Spark pipeline provides a natural framework for scalable data management, exploratory analysis as well as machine learning for urban crime analytics. Arrests are heavily influenced by type of crime and location context, and both crime classification and incident clustering depend on informed feature selection.

A. Recommendations

1. **Class Imbalance Solutions:** Explicitly add re-sampling (SMOTE, over-sampling) or cost-sensitive considerations in multiclass predictions and achieve significant improvement in accuracy and policy relevance.
2. **Enhanced Spatial-Temporal Granularity:** Using micro human geolocation information and fine-grained temporal information (e.g. event-based analysis at the hourly interval) to accurately record dynamic details of crime patterns.
3. **Policy-Linked Analytics Framework:** Integrate clear policy dynamics and policy-related quantities of interest into analytical models to directly measure policy effectiveness and guide realistic enforcement tactics.
4. **Dimensionality and Clarity Optimization:** In clustering studies, give feature interpretability and dimensionality reduction (e.g., PCA) top priority to improve practical utility and actionable insights by combining clarity and dimensionality.

Table of Figures