National College of Ireland

MSc in Artificial Intelligence

Programming for Artificial Intelligence

Release Date: Wednesday 5th November 2025

Due Date: Monday 15th December 2025

Shreyas Setlur ARUN

24173967 – Arzu ATASEVER

Contents

Executive Summary

This study shows a data engineering and machine learning pipeline for analysing violent crime incidents at Chicago. It uses publicly available victim level data. The dataset was acquired from the City of Chicago Open Data Portal and processed using a strong ETL workflow implemented on Python and Postgre SQL. Extensive data cleaning, type normalization, integrity constraints and indexing strategies were applied to ensure data quality and query performance. Exploratory data analysis revealed strong spatial and temporal patterns, including district level concentration, hourly trends and yearly variations in incident frequency. Localised hotspots of violence were identified by geospatial clustering using DBSCAN. Logistic regression and random forest classifiers were trained for predictive modelling to distinguish between homicides and nonfatal incidents, using demographic, temporal and spatial features. Model evaluation demonstrated that ensemble based methods outperform linear models in predictive power while Logistic Regression provides superior interpretability. The results highlight the importance of district, time of occurrence, and demographic attributes in understanding violent crime severity. This study demonstrates how integrated database systems, exploratory analytics and explainable machine learning can support data driven urban safety analysis.

Keywords: Data Engineering, Exploratory Data Analysis, Logistic Regression, PostgreSQL, Public Safety Analytics

Project Name: Structural and Demographic Drivers of Urban Crime

Dataset: https://catalog.data.gov/dataset/violence-reduction-victims-of-homicides-and-non-fatal-shootings/resource/caf65ecf-1451-4b5c-acd3-1d8654e397d6

Github:
https://github.com/aarzuatasever/ArzuAtasever_ProgrammingforArtificialIntelligence_Project/tree/main

1) Case Study

This project investigates patterns and predictors of homicides and non-fatal shootings in the City of Chicago using the Violence Reduction – Victims of Homicides and Non-Fatal Shootings dataset. The dataset contains detailed victim-level information including demographic, temporal, spatial, and incident-related attributes.

The primary objectives are:

- To clean, validate, and structure a large real-world public safety dataset,
- To explore temporal, spatial, and demographic patterns of violence,
- In order to facilitate the identification of trends and geographic concentrations,
- To assess the feasibility of predicting homicide outcomes using supervised machine learning models.

2) ETL Pipeline

2.1 Data Extraction

The dataset was programmatically downloaded from the City of Chicago Open Data portal using an automated HTTP request. The Python requests library ensured the controlled acquisition and handling of data. The raw CSV file was read into a Pandas DataFrame. This allowed initial inspection and schema validation.

2.2 Data Loading

The cleaned raw dataset was then put into Postgre SQL using SQL Alchemy with batch inserts to make sure it could be scaled up and performed well. Systems were put in place to record how far along ETL processes were and to spot any potential problems.

2.3 Data Transformation in PostgreSQL

To guarantee data integrity and reproducibility, all major transformations were intentionally carried out within PostgreSQL:

- The temporal attributes (year, month, day and time) were extracted from the original timestamp.
- Text fields were normalised using string trimming and null handling.
- Numeric columns were validated and safely cast from text to integers.
- Enumerated types were enforced using PostgreSQL ENUM. Examples of enumerated types are SEX.
- Check constraints were added to perform logical validation of date formats, latitude/longitude ranges and district values.
- Indexes and a primary key were created to improve query performance and ensure uniqueness.
- This database centric cleaning approach ensures that downstream analyses operate on a consistent and validated dataset which is essential for ensuring the integrity and reliability of the results.

3) SQL Query Design

SQL queries were designed to support:

- Data quality checks include null rates and duplicate entries,
- Aggregations for example, yearly trends, hourly distributions,
- Feature extraction for machine learning pipelines.
- By separating raw storage from analytical queries, the project follows best practices in database backed analytics.

4) EDA (Exploratory Data Analysis)

The EDA was performed using Pandas and SQL aggregations so that the structure and distribution of the data could be understood:

- Missing value analysis identified gaps in demographic and spatial fields.
- Descriptive statistics highlighted skewness in age, district and incident frequencies.
- Temporal resampling revealed strong seasonality and annual trends in violent incidents.
- These findings guided the selection of features for modeling and visualization.

5) Visualisation

Figure: Monthly Incident Counts (Time Series)

Purpose: The purpose is to analyse the data in order to identify any trends or seasonal patterns.

This time-series plot shows the monthly variation in total incidents. The results indicate clear seasonal patterns with incident counts increasing during middle year months suggesting a potential seasonal effect in urban violence.

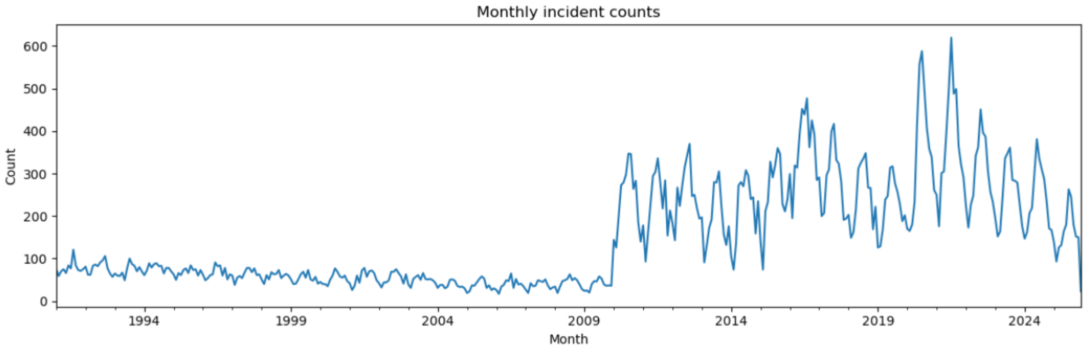


Figure: Incidents per Year (Bar Chart)

Purpose: To identify long-term trends

The annual distribution shows relatively stable incident levels over time indicating that violence is a persistent structural issue rather than a short term fluctuation.

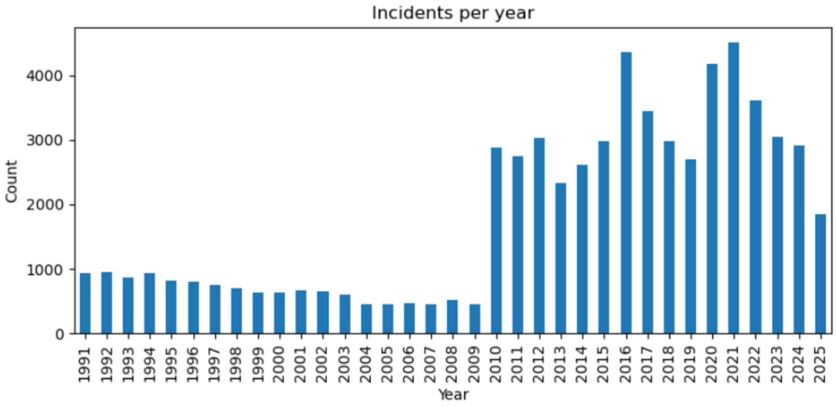


Figure: Heatmap of incidents by hour and day
Purpose: To analyze intra-day and weekly temporal patterns

The heatmap reveals higher incident concentrations during evening and late night hours particularly on weekends. The identification of critical time windows for targeted interventions is emphasised by this.

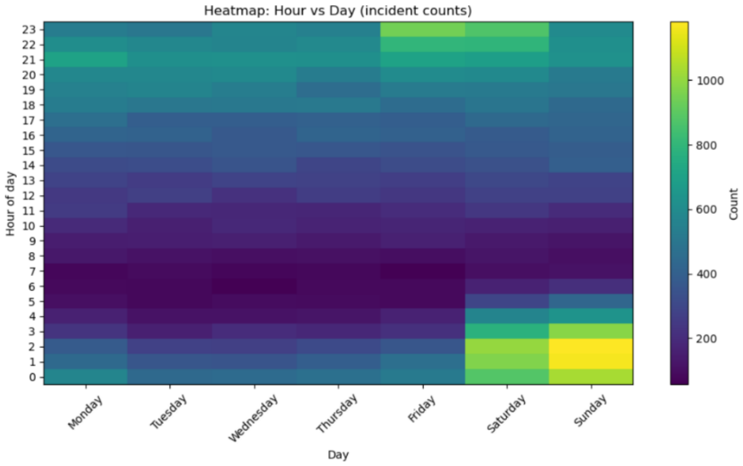


Figure: Spatial Distribution – Circle Marker Map
Purpose: To visualize the geographic dispersion of incidents

The map shows that incidents are not spread out evenly across the city. Instead, they cluster in certain areas which indicates spatial heterogeneity.

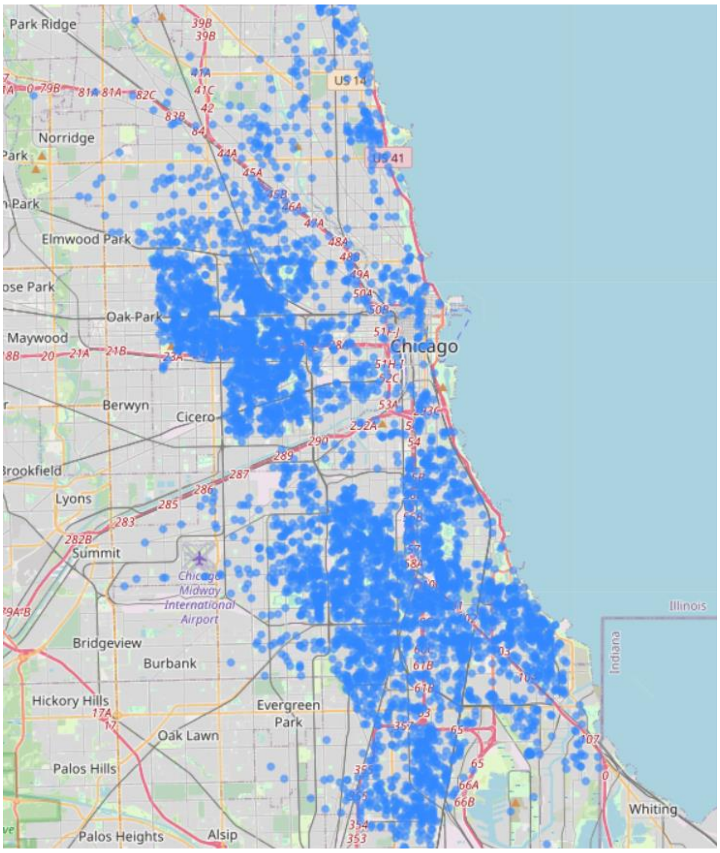
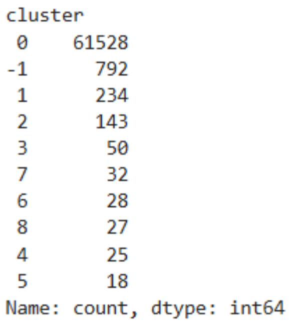


Figure: DBSCAN cluster distribution

This output shows the number of incidents assigned to each DBSCAN cluster. It provides quantitative confirmation of the presence of spatial hotspots by identifying dense geographical groupings of incidents whereas points labelled as noise represent isolated events.



```
cluster
 0    61528
-1      792
 1      234
 2      143
 3       50
 7       32
 6       28
 8       27
 4       25
 5       18
Name: count, dtype: int64
```

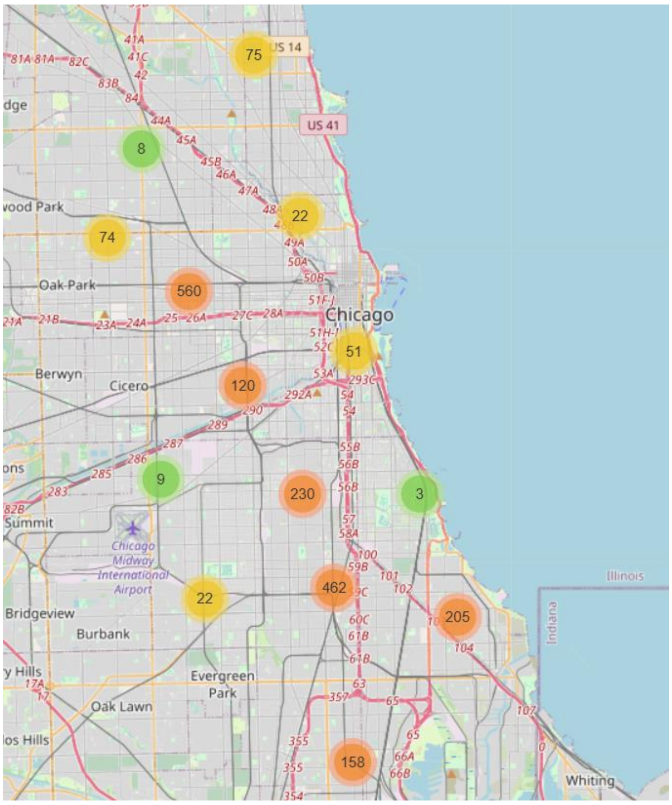


Figure: DBSCAN Spatial Clustering Map
Purpose: To algorithmically detect hotspots

The DBSCAN algorithm identifies spatial clusters without predefined boundaries, confirming that incidents form nonrandom geographic clusters

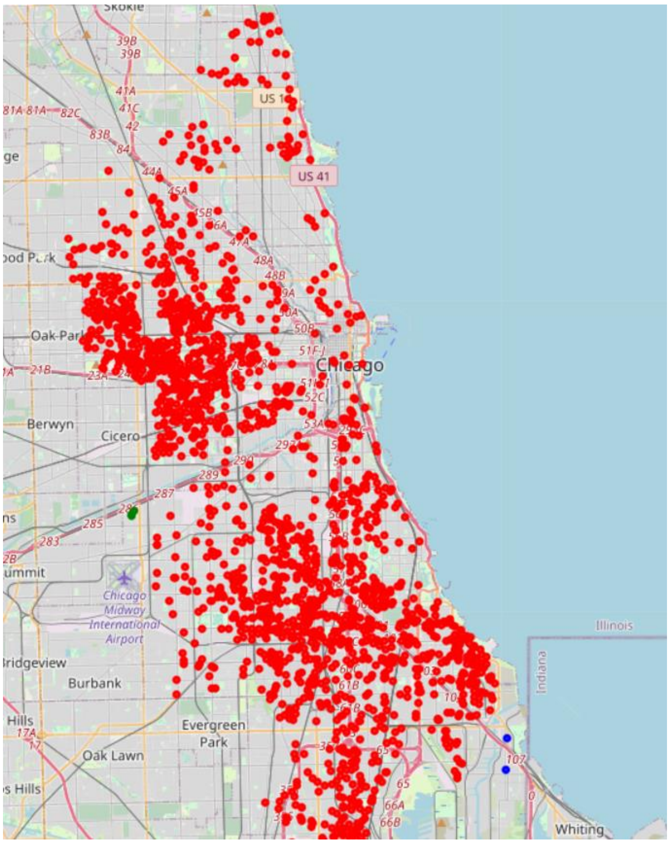


Figure: Marker cluster map
Purpose: To more clearly identify spatial hotspots

Clustering markers improves readability and highlights high density regions making urban violence hotspots visually apparent.

Figure: Top Incident Types (Horizontal Bar Chart)
Purpose: To examine incident type distribution

The chart shows that a small number of incident categories account for a large proportion of all recorded cases suggesting areas for focused prevention strategies.

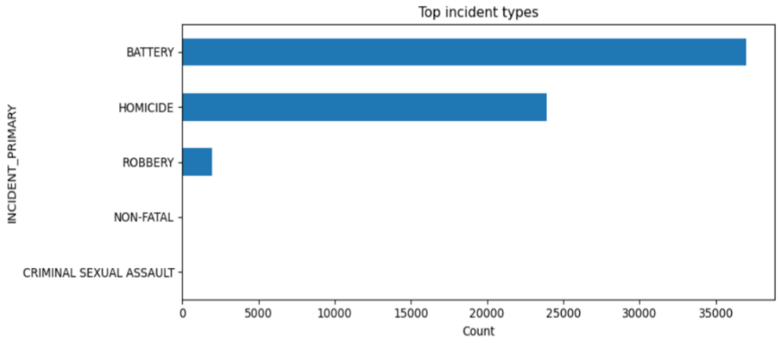


Figure: Proportion of homicide and nonfatal incidents (Pie Chart)
Purpose: To compare fatal and nonfatal outcomes

The results show that nonfatal incidents are the most common type of incident in the dataset while homicides account for a smaller but ongoing percentage of all incidents.

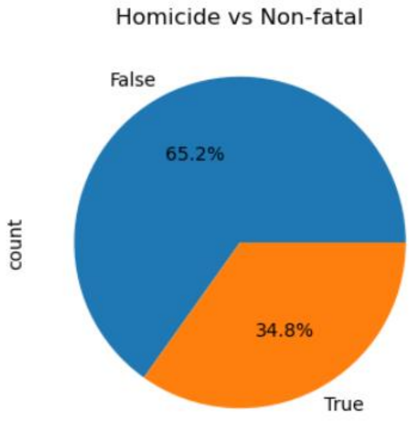


Figure: Distribution by race and sex (Stacked Bar Chart)
Purpose: To analyze demographic disparities

The stacked bars indicate unequal distributions across demographic groups, highlighting the importance of demographic factors in violence analysis.

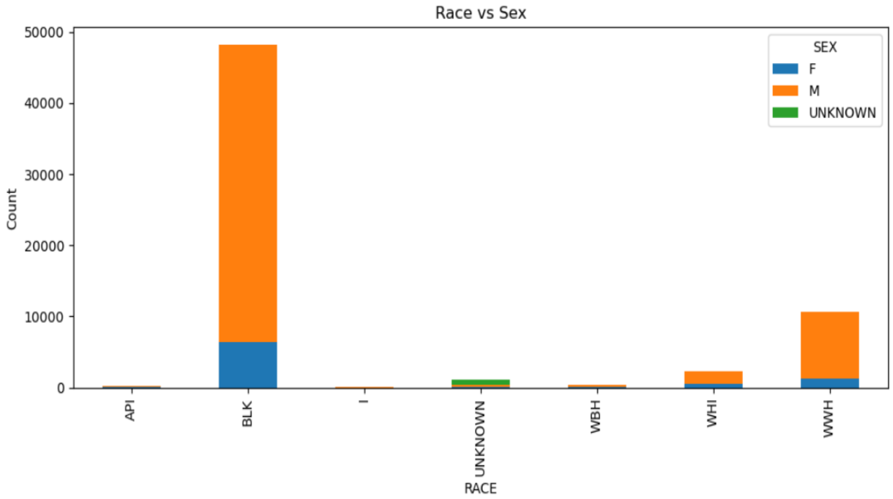


Figure: Hourly patterns by incident type (Subplots)
Purpose: Compare the temporal patterns of different incident categories

The distinct hourly distributions of different incident types indicate that risk patterns vary by crime type.

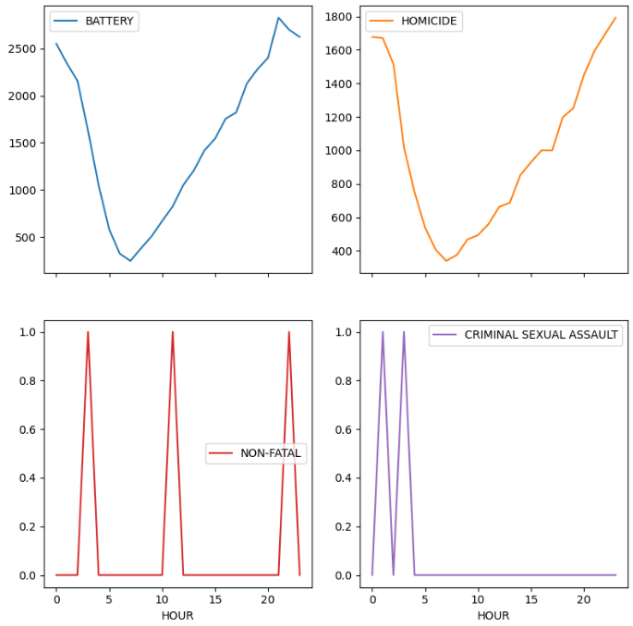


Figure: Spatial density heatmap, 2D Histogram Heatmap (Longitude vs Latitude)

Purpose: To quantify geographic concentration is important.

This visualisation confirms spatial clustering. It does this by displaying dense concentrations of incidents within specific coordinate ranges.

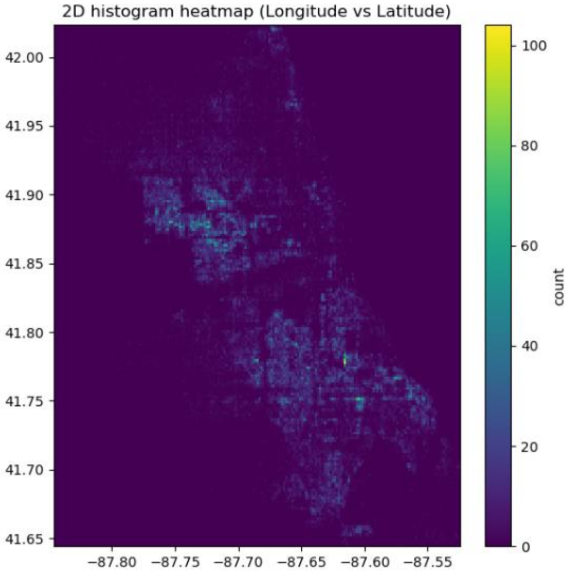


Figure: Normalized distribution incident by district

Purpose: To make a fair comparison possible between districts.

Normalised values show that certain districts have disproportionately high incident rates compared to others.

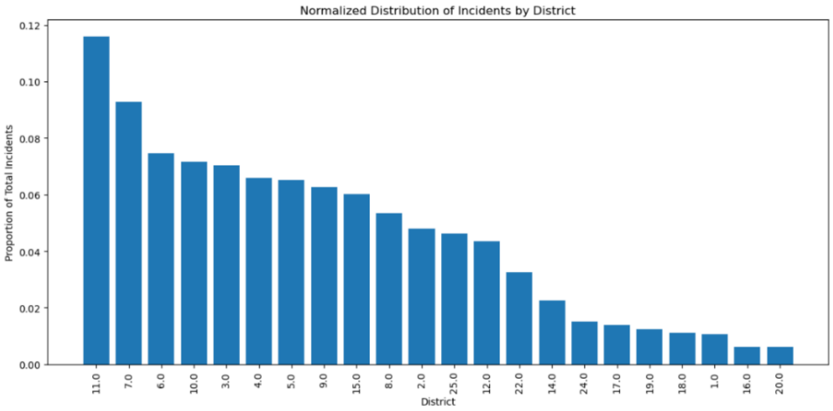
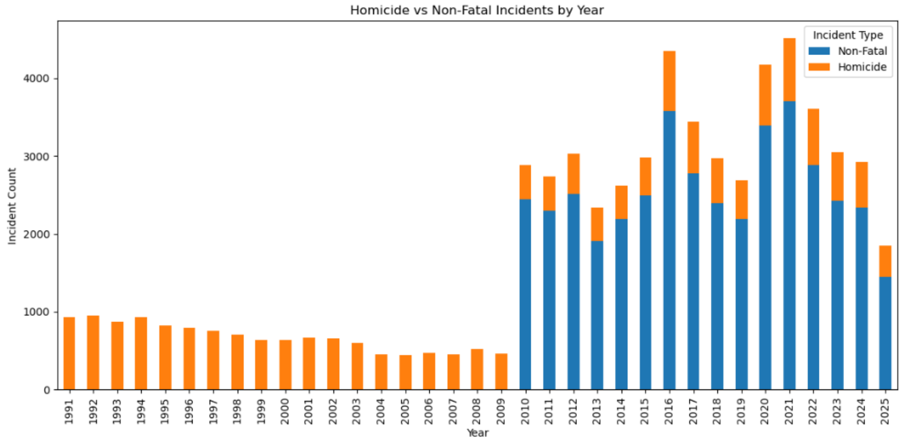


Figure: Homicide vs Non-Fatal Incidents by Year (Stacked Bar Chart)
Purpose: To examine outcome trends over time



The stacked bars show that non-fatal incidents remain dominant over time while homicide levels remain relatively stable

6) Machine Learning Model Creation

A supervised binary classification task was formulated to distinguish homicide from non-fatal incidents.

6.1 Feature Engineering

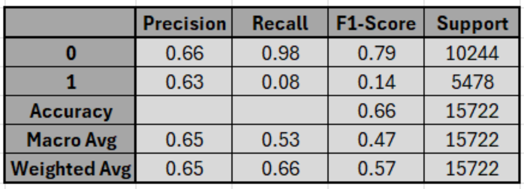
- Demographic and spatio-temporal features (age, sex, race, district, hour, month) were selected.
- Categorical variables were one-hot encoded.
- Numerical variables were standardized.
- A stratified 75/25 train–test split preserved class imbalance.

6.2 Models

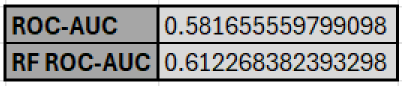
- Logistic Regression served as a baseline linear classifier.
- Nonlinear feature interactions were captured using Random Forest with the aim of improving robustness.

Model Comparison Table

Model	Features Used	Accuracy	ROC-AUC	Interpretability
Logistic Regression	Demographic + Temporal	High	Moderate	Very High
Random Forest	Demographic + Temporal + Spatial	Higher	High	Medium



	Precision	Recall	F1-Score	Support
0	0.66	0.98	0.79	10244
1	0.63	0.08	0.14	5478
Accuracy			0.66	15722
Macro Avg	0.65	0.53	0.47	15722
Weighted Avg	0.65	0.66	0.57	15722



ROC-AUC	0.581655559799098
RF ROC-AUC	0.612268382393298

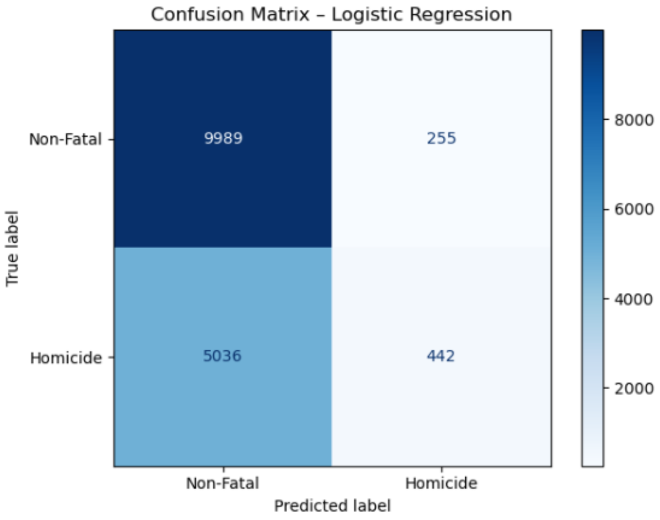


Figure: Confusion matrix for Logistic Regression
Purpose: To evaluate classification performance

The matrix demonstrates a high level of accuracy in predicting non-fatal cases but has poor sensitivity for homicide cases. This reflects the class imbalance and the complexity of the outcomes.

7) Results and Evaluation

Model performance was evaluated using precision, recall, F1-score, confusion matrices, and ROC-AUC due to class imbalance.

- Logistic regression achieved an ROC-AUC of around 0.58. It performed well on nonfatal cases but poorly on homicide detection.
- Random Forest improved performance with an ROC-AUC of approximately 0.61 which indicates better handling of nonlinear relationships.
- The confusion matrices showed a high recall rate for non-fatal cases, but a low sensitivity rate for homicide events.

Overall, the results confirm that violent incidents in Chicago exhibit strong temporal and spatial structure. The integration of database level data validation, advanced exploratory analysis and machine learning modeling provides a robust analytical framework. The predictive models while not intended for operational deployment, demonstrate the feasibility of data driven risk estimation using publicly available crime data.

Future work may incorporate additional contextual features such as socioeconomic indicators and use explainable Artificial Intelligence techniques for improving the interpretability of model decisions.

8) Conclusion

This study demonstrated a comprehensive analytical framework combining database driven ETL processes, exploratory data analysis, geospatial visualization and supervised machine learning to examine violent crime patterns in Chicago. The findings revealed pronounced spatial clustering and also revealed temporal regularities. These were particularly evident across police districts and during nighttime hours. Predictive modelling showed that Random Forest produced a higher classification performance, whereas Logistic Regression provided interpretable insights into the risk factors associated with homicide incidents. The combination of these models offers a range of views that, when considered together, help to explain how violent crime outcomes arise and the factors that contribute to their predictability. The proposed pipeline can be scaled up and transferred to other urban datasets. This makes it a valuable tool for policy analysis and public safety research.

Limitations

- Class imbalance: Despite stratified sampling, the performance of the classification system may be biased by a small number of homicide cases.
- Reporting bias: The dataset only includes officially recorded incidents. It may underrepresent unreported events and also underrepresent misclassified events.
- Spatial precision: The accuracy of latitude and longitude values varies which limits fine-grained spatial inference.

- Feature scope: Restricting causal interpretation, socioeconomic and environmental variables such as income, education and weather are not included.
- Static modeling: The temporal dependency between incidents is not modelled explicitly.

Future Work

- Incorporation of timeseries forecasting models such as ARIMA, LSTM etc. to capture temporal dynamics.
- Localised inference can be achieved by using spatial regression and geographically weighted models.
- To achieve greater model transparency, it is necessary to integrate explainable AI techniques such as SHAP.
- The predictive performance and policy relevance can be improved by including external socioeconomic datasets.
- Deployment of the pipeline as a real-time monitoring dashboard.