## Unveiling Crime Patterns in Mid-Sized Cities: A Comparative Analysis of Machine Learning Classification Models

### Abstract

This project analyzes crime patterns in mid-sized U.S. cities (populations 500,000-700,000) using publicly available datasets. We employed KNN, QDA, and Naive Bayes classification models to identify prevalent crime types, temporal and seasonal trends, and spatial variations at the zip code level. Exploratory data analysis informed model selection, and performance was evaluated using accuracy, precision, recall, F1-score, and AUC. The results, presented in a comprehensive report including descriptive statistics and visualizations, will inform resource allocation and crime prevention strategies in these cities. Comparative analysis of model performance highlighted the strengths and weaknesses of each approach in capturing crime patterns.

### Background

Crime in mid-sized cities presents a significant challenge, impacting public safety and resource allocation. Understanding the underlying drivers of crime in these cities is crucial for effective prevention strategies. This study leverages publicly available crime datasets from multiple mid-sized U.S. cities, focusing on a consistent demographic range to ensure comparability. The challenge lies in identifying patterns within complex, potentially incomplete datasets and selecting appropriate machine learning models to accurately capture these patterns.

The significance of this research stems from its potential to inform evidence-based crime prevention strategies. The comparative analysis of multiple classification models adds depth and allows for a robust evaluation of model effectiveness. This investigation is interesting because it combines data analysis, geographic information systems (GIS), and machine learning to address a real-world problem with practical implications for law enforcement and urban planning. The results will be relevant to policymakers, law enforcement agencies, and urban planners seeking to improve public safety and resource allocation in mid-sized cities.

### Methodology

Publicly available crime datasets from multiple mid-sized cities are gathered. The data is cleaned to handle missing values and inconsistencies. Visualizations (EDA) are used to explore the data, understand the relationships between variables, and identify potential patterns. This step is crucial for informing model selection and feature engineering.

Several classification algorithms are used to predict the probability of a crime occurring based on time, location, and season. These include:

* **Random Forest:** An ensemble method that combines multiple decision trees to improve prediction accuracy and reduce overfitting. Hyperparameters like the number of trees and tree depth need to be tuned.
* **Gradient Boosting Trees (GBT):** Another ensemble method that builds trees sequentially, each correcting the errors of its predecessors. GBT methods like XGBoost, LightGBM, or CatBoost are often used, each with its own set of hyperparameters.
* **Quadratic Discriminant Analysis (QDA):** Similar to LDA but uses quadratic decision boundaries instead of linear ones. More flexible than LDA but requires more data and is more susceptible to overfitting.
* **Naive Bayes:** A probabilistic classifier based on Bayes' theorem with strong (naive) independence assumptions between features. Computationally efficient and works well with high-dimensional data.
* **K-Nearest Neighbors (KNN):** A non-parametric method that classifies a data point based on the majority class among its k-nearest neighbors in the feature space. The choice of 'k' is a hyperparameter that needs tuning.

The performance of each model is assessed using metrics like accuracy, precision, recall, and F1-score. K-fold cross-validation is employed to evaluate the models' generalization performance and prevent overfitting. Hyperparameter tuning using techniques like grid arearch or randomized search is performed to optimize model performance.

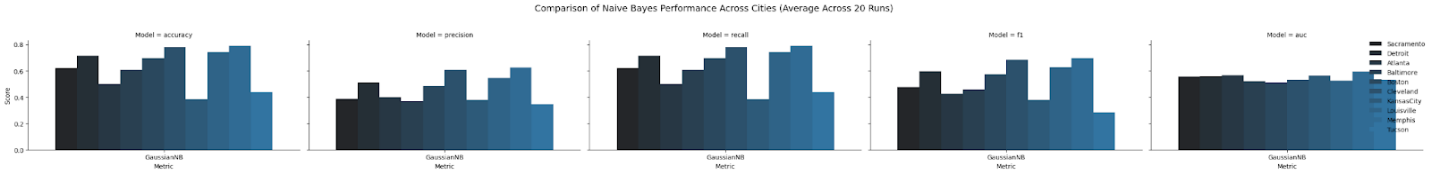
After evaluating all models, the best-performing model (or an ensemble of models) is selected based on the chosen evaluation metric(s), considering the trade-off between performance and model complexity.

### Results

#### Naïve Bayes

**Overall Performance:**

The Gaussian Naive Bayes model demonstrates variable performance across the ten cities, with F1-scores generally below 0.7, indicating that the model's predictive power is limited when relying solely on time of day, season, and day of the week. The ROC AUC analysis further supports this, showing AUC values mostly between 0.5 and 0.6 for individual crime types, suggesting only marginally better performance than random chance.



Summary Table:

       Metric  accuracy  precision    recall        f1       auc        city

0  GaussianNB  0.620836   0.385482  0.620836  0.475624  0.556868  Sacramento

1  GaussianNB  0.714445   0.510433  0.714445  0.595449  0.558035     Detroit

2  GaussianNB  0.501236   0.399230  0.501236  0.425251  0.566091     Atlanta

3  GaussianNB  0.607315   0.368834  0.607315  0.458942  0.520201   Baltimore

4  GaussianNB  0.696278   0.484891  0.696278  0.571645  0.510393      Boston

5  GaussianNB  0.778489   0.606048  0.778489  0.681529  0.532737   Cleveland

6  GaussianNB  0.382127   0.380537  0.382127  0.380081  0.563348  KansasCity

7  GaussianNB  0.739499   0.546869  0.739499  0.628758  0.525003  Louisville

8  GaussianNB  0.789644   0.623559  0.789644  0.696836  0.593119     Memphis

9  GaussianNB  0.439676   0.345401  0.439676  0.282479  0.530341      Tucson

A collage of graphs

Description automatically generated

Macro-averaged AUC for Sacramento: 0.6027195264275472

Macro-averaged AUC for Detroit: 0.6027195264275472

Macro-averaged AUC for Atlanta: 0.6027195264275472

Macro-averaged AUC for Baltimore: 0.6027195264275472

Macro-averaged AUC for Boston: 0.6027195264275472

Macro-averaged AUC for Cleveland: 0.6027195264275472

Macro-averaged AUC for KansasCity: 0.6027195264275472

Macro-averaged AUC for Louisville: 0.6027195264275472

Macro-averaged AUC for Memphis: 0.6027195264275472

Macro-averaged AUC for Tucson: 0.6027195264275472

**City-Specific Performance:**

* **High-Performing Cities:** Cleveland, Memphis, and Louisville exhibit relatively better performance (F1 scores above 0.6). However, even in these cities, the ROC AUC values suggest only moderate predictive power for individual crime types. This indicates that while the chosen predictors might have some predictive value, they are not sufficient for highly accurate predictions.
* **Low-Performing Cities:** Kansas City and Tucson show considerably weaker performance (F1 scores well below 0.4), and their ROC AUC values confirm the limited predictive ability of the chosen features in these locations. Other, unmodeled factors are likely more influential in these cities.
* **Moderate Performance:** Sacramento, Detroit, Atlanta, Baltimore, and Boston demonstrate moderate performance (F1 scores between 0.45 and 0.7), indicating some predictive power but substantial room for improvement, as evidenced by the low-to-moderate ROC AUC values.

**Further Analysis:**

To enhance understanding and improve predictive accuracy, consider these steps:

1. **Exploratory Data Analysis (EDA):** Conduct a thorough EDA, including visualizations and statistical tests, to examine the relationships between predictors and crime types.
2. **Feature Engineering:** Add relevant predictors (e.g., geographic location, socioeconomic data, weather conditions) to increase the model's predictive power.
3. **Model Comparison:** Experiment with more advanced machine learning models (decision trees, random forests, support vector machines, gradient boosting machines, etc.) to compare their performance.
4. **City-Specific Analysis:** Investigate the characteristics of the crime data for each city to understand the reasons for the varying model performance.
5. **Hyperparameter Tuning:** Optimize the hyperparameters of the Gaussian Naive Bayes model and other models to potentially improve performance.

**Summary:**

The Gaussian Naive Bayes model offers limited predictive power for crime type classification based solely on the chosen predictors. While some cities show moderately better results than others, the overall performance is weak. Addressing data limitations, enhancing feature engineering, and exploring more complex machine learning models are essential for improving accuracy and gaining more reliable insights. The consistent low AUC values across cities and crime types strongly suggest the need for significant improvements in the model and/or the data used.

#### KNN

**Overall Performance:**

* Cleveland, Memphis, and Louisville consistently outperform other cities, indicating well-defined patterns in their datasets that the KNN model can capture effectively. Tucson and Atlanta have consistently lower scores, suggesting noisier data, fewer features, or less-defined patterns.Cleveland and Memphis provide the best balance across metrics, making them ideal datasets for testing further improvements.

Default Parameter Results:

Summary Table:

Metric Accuracy Precision Recall F1-Score city

0 GaussianNB 0.434681 0.399967 0.434681 0.403360 Atlanta

1 GaussianNB 0.499461 0.457548 0.499461 0.473060 Baltimore

2 GaussianNB 0.758066 0.630554 0.758066 0.679120 Cleveland

3 GaussianNB 0.673380 0.571005 0.673380 0.605635 Detroit

4 GaussianNB 0.567645 0.570456 0.567645 0.568340 KansasCity

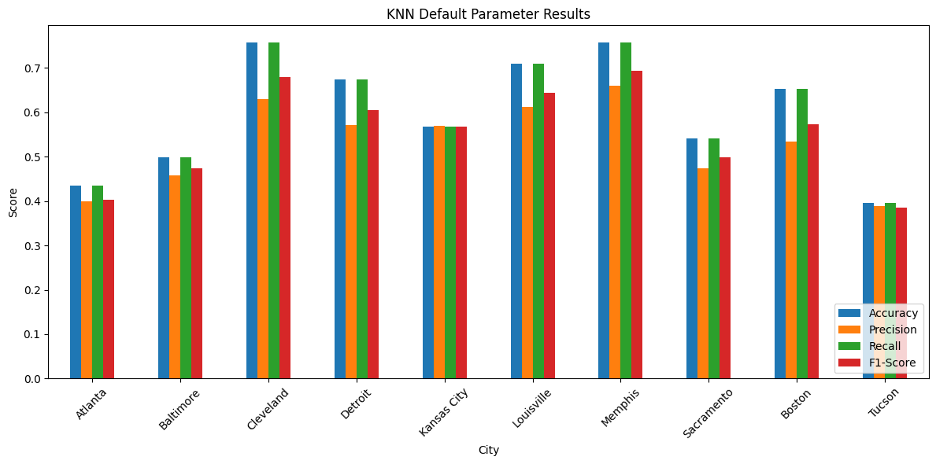
5 GaussianNB 0.709121 0.612484 0.709121 0.644752 Louisville

6 GaussianNB 0.757454 0.660383 0.757454 0.693741 Memphis

7 GaussianNB 0.540823 0.473850 0.540823 0.498372 Sacramento

8 GaussianNB 0.653468 0.534471 0.653468 0.573061 Boston

9 GaussianNB 0.396245 0.387962 0.396245 0.384819 Tucson



**City-Specific Performance:**

* **High-Performing Cities:** Cleveland, Memphis, and Louisville show best performance in all four category. Showing that the model captures the majority of actual crime events.
* **Low-Performing Cities:** Tucson (0.396) and Atlanta (0.434) have the poorest performance in accuracy, indicating that the KNN model struggles to generalize well in these datasets.
* **Moderate Performance:** Memphis and Kansas City showing strong consistency across metrics.

A screenshot of a scatter plot

Description automatically generated

**Day of Week vs. Time of Day**

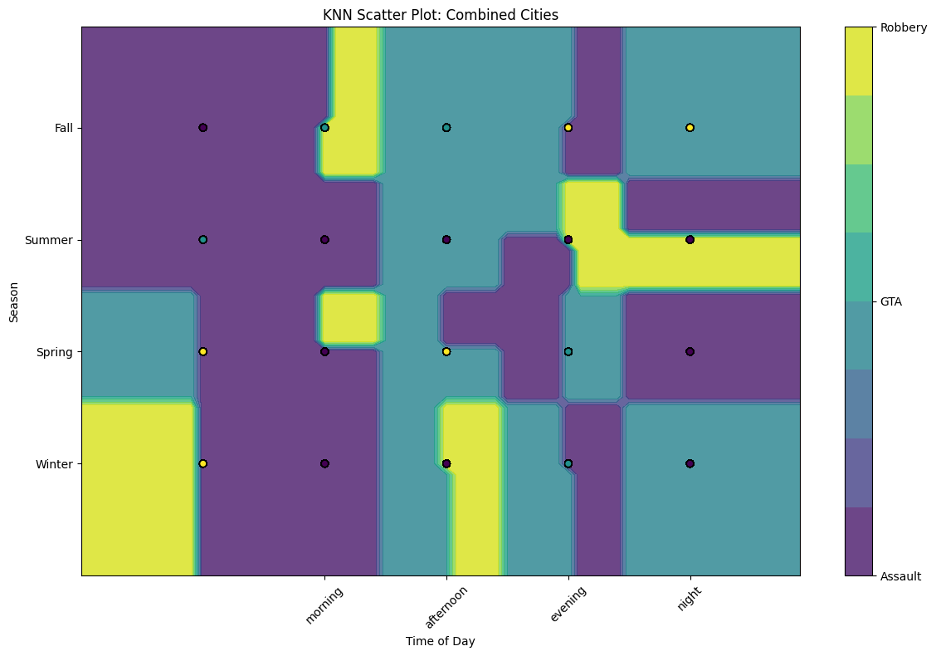
* **Observation**: Robbery and GTA dominate during specific **times of the day**, particularly evenings and nights on certain weekdays.
* **Time of Day Influence**: There are notable crime clusters during evenings and nights on specific days, indicating peak times for crimes like robbery or GTA.
* **Day Dependence**: Crimes vary across the week, with higher intensities on certain days for particular times. Fridays and Saturdays show heightened crime activity during the evening and nighttime, which aligns with typical social and economic activities.

A screenshot of a computer generated image

Description automatically generated

**Day of Week vs. Season**

* **Observation**: Crime activity varies significantly across both days and seasons. Crimes like robbery have higher activity on Mondays in Winter and Fridays in Summer.
* **Seasonal Patterns**: Crime activities may have stronger associations with seasons, such as winter showing more activity for specific crime types compared to summer.Weekends and summer months generally exhibit different distributions compared to weekdays and winter.
* **Day of Week Trends**: Weekends might show different crime distribution compared to weekdays.



**Time of Day vs. Season**

* **Observation**: Specific times of day, such as mornings and afternoons in Winter and Spring, see different crime types dominating the regions.
* **Seasonal Shifts**: For some crime types (e.g., GTA), winter mornings and afternoons show stronger activity compared to summer evenings. Winter mornings and afternoons see a concentration of crimes like GTA and Assault, while summer evenings show a rise in Robbery.

**Overall**

Deploy law enforcement during high-crime times like weekends and evenings are high-risk periods. Robbery tends to peak during evenings and weekends, particularly in winter or summer.GTA shows relatively consistent patterns but peaks in winter mornings and afternoons.

Assault is more evenly distributed but increases during night periods across different seasons.Educate communities about high-risk times and seasons to increase vigilance and preventive measures.

#### QDA

Summary Table:

Metric accuracy precision recall f1 auc city

0 GaussianNB 0.199195 0.215916 0.199195 0.170436 NaN Atlanta

1 GaussianNB 0.599740 0.446881 0.599740 0.477767 NaN Baltimore

2 GaussianNB 0.686658 0.506050 0.686658 0.568746 NaN Boston

3 GaussianNB 0.778965 0.606790 0.778965 0.682181 NaN Cleveland

4 GaussianNB 0.714420 0.510397 0.714420 0.595416 NaN Detroit

5 GaussianNB 0.429028 0.413033 0.429028 0.373639 NaN KansasCity

6 GaussianNB 0.700954 0.639116 0.700954 0.666088 NaN Louisville

7 GaussianNB 0.763580 0.665802 0.763580 0.700155 NaN Memphis

8 GaussianNB 0.604930 0.481380 0.604930 0.513667 NaN Sacramento

9 GaussianNB 0.460851 0.381410 0.460851 0.410043 NaN Tucson

A group of blue bars

Description automatically generated

**Key Findings:**

1. **Cleveland**:
   * Achieved the highest performance across all metrics (Accuracy: 0.778, Precision: 0.606, Recall: 0.778, F1 Score: 0.681).
   * Indicates strong predictive capability in this city, suggesting consistent patterns in the data.
2. **Memphis**:
   * Second-highest overall performance (Accuracy: 0.765, F1 Score: 0.703).
   * Close alignment between precision and recall suggests a balanced classification model.
3. **Atlanta**:
   * Performed the worst across all metrics (Accuracy: 0.344, F1 Score: 0.308).
   * Suggests significant variability in the data or insufficient feature representation.
4. **Kansas City**:
   * Had the second-lowest performance (Accuracy: 0.429, F1 Score: 0.370).
   * Precision (0.419) and recall (0.429) indicate room for improvement in model balance.

**Analysis:**

1. **Feature Contribution**:
   * Cities like Cleveland and Memphis exhibit more predictable patterns, possibly due to better feature representation or more consistent data.
   * Poor performance in Atlanta suggests that additional features (e.g., demographic or environmental data) or deeper preprocessing (e.g., handling outliers or improving encoding) could be beneficial.
2. **Imbalanced Class Challenges**:
   * Although normalized, some cities may have inherent imbalances in crime types, impacting the precision and recall scores.
3. **Overfitting or Data Quality**:
   * Variations in city performance could also reflect differences in data quality, size, or representativeness of training/test splits.
4. **Model Strengths**:
   * Quadratic Discriminant Analysis (QDA) performs well in cities with distinct class boundaries (e.g., Cleveland) but struggles where boundaries overlap significantly (e.g., Atlanta).

**Model Performance:**

1. **Overall Metrics**:
   * Average accuracy across all cities is around 0.57, with F1 scores trailing slightly behind, averaging around 0.48.
   * The model is moderately effective but struggles in cities with more complex patterns.
2. **Strengths**:
   * The QDA model demonstrates its ability to leverage quadratic boundaries for cities like Cleveland and Memphis, where class distributions are likely distinct.
3. **Weaknesses**:
   * Struggles in cities with overlapping or noisy class distributions (e.g., Atlanta, Kansas City).
   * Precision lags behind recall in several cities, indicating a tendency to predict more false positives.

**Model Strengths:**

* Quadratic Discriminant Analysis (QDA) performed well in cities with distinct patterns, like Cleveland and Memphis, leveraging its ability to define quadratic boundaries.
* The model demonstrates reliable recall in high-performing cities, capturing relevant classes effectively.

**Model Weaknesses:**

* Struggles in cities with overlapping class boundaries or noisy data, such as Atlanta.
* Precision is consistently lower than recall in some cities, indicating potential overprediction of certain classes.

**Recommendations:**

1. **Feature Engineering**:
   * Incorporate additional variables (e.g., demographic, socioeconomic, or geographic data) to improve predictive power.
   * Consider feature scaling and transformations to enhance sensitivity to data patterns.
2. **Alternative Models**:
   * Explore ensemble models (e.g., Random Forest or Gradient Boosting) that may better handle complex and noisy data.
   * Test simpler models, like logistic regression, to compare performance and robustness.
3. **Class Balancing**:
   * Apply oversampling or class weighting techniques to address imbalances in crime types.
4. **City-Specific Models**:

Consider building localized models for cities like Atlanta to capture unique patterns in the data.

#### Gradient Boosting

**Overall Performance**

The gradient boosting modeldemonstrates varying performance across different cities. In some cities it performs well, while in others we can see that the performance was significantly lower. The f1 scores generally were in the range of .25 to .41 indicating that the model had a moderate ability to predict the crime types based on the given predictors.

City Accuracy Precision Recall F1-score

Atlanta 0.500578 0.322328 0.385425 0.340626

Baltimore 0.609152 0.203051 0.333333 0.252370

Boston 0.693209 0.231070 0.333333 0.272937

Cleveland 0.778522 0.259507 0.333333 0.291824

Detroit 0.714044 0.238015 0.333333 0.277723

Kansas City 0.445261 0.435944 0.447525 0.432644

Louisville 0.739139 0.246380 0.333333 0.283335

Memphis 0.790201 0.291459 0.333124 0.294704

Sacramento 0.622695 0.207565 0.333333 0.255828

Tucson 0.465126 0.311674 0.364619 0.323565

A graph of different sizes and colors

Description automatically generated with medium confidence  
**Key findings**

* High performing cities
  + Memphis has the highest accuracy .7902 a precision of .2915 and f1 of .2947
  + Cleveland also has a high accuracy of .7785 precision of .2595, and an f1 of .2918
  + Louisville having an accuracy of .7785  precision of .2464 and an f1 of .2833

While these f1 scores are modest, the higher accuracy value suggests that the model is pretty good at picking the majority of the class but due to the moderate f1 scores we can state that it struggles with positive predictions. These cities may have more consistent crime data than the others making them more predictable.

* Low performing cities
  + Kansas City having an accuracy of .4453 a precision of .4359 and f1 of .4475
  + Baltimore having an accuracy of .6092 a precision of 2031 and f1 of .2524

These results indicate that the predictors we used were less effective in these cities. The low precision and recall also suggests that the model was frequently misclassifying crime types.

* Moderate performing cities
  + The rest of the cities are Atlanta, Detroit,Boston, Sacramento, Tucson.

These cities all had f1 scores between .25 and .32 with decent overalls accuracy the model struggled with precision leading to these cities to have moderate f1 scores

**Further Analysis**

* Mentioned before we can add more relevant predictors to enhance the model’s performance, such as crime location, weather, population density, and socio economic factors
* Using different statistical testing and visualizations we can explore the relationships between the predictors and crime types and find hidden patterns
* Dwell deeper into each city and look at the performance across the various cities and examine the data to see the crime distribution, and potential biases in the data sets

**Summary**

The gradient boosting models show moderate performance, with cities like Cleveland and Memphis performing better than the rest. To achieve better accurate crime type predictions we can try using more predictors with our data sets and also look into trying more sophisticated models.

#### Random forest

**Overall Performance**

The Random Forest model demonstrates inconsistent performance across different cities, with accuracy generally ranging between **33% and 44%** and F1 scores often below **0.5**. These results indicate that the model struggles to effectively predict crime types based on the provided predictors (time of day, day of the week, season). This may reflect limitations in the features or complexities within the datasets that the model fails to capture.

      Metric  accuracy  precision    recall        f1        city

0  RandomForest  0.388364   0.482889  0.388364  0.413653   Baltimore

1  RandomForest  0.436570   0.439843  0.436570  0.437960     Atlanta

2  RandomForest  0.346241   0.563136  0.346241  0.392421      Boston

3  RandomForest  0.399112   0.653718  0.399112  0.471659   Cleveland

4  RandomForest  0.413719   0.621034  0.413719  0.452735     Detroit

5  RandomForest  0.436161   0.418921  0.436161  0.380673  KansasCity

6  RandomForest  0.332709   0.646988  0.332709  0.386501  Louisville

7  RandomForest  0.371043   0.707503  0.371043  0.436788     Memphis

8  RandomForest  0.391581   0.503791  0.391581  0.424994  Sacramento

9  RandomForest  0.359775   0.425097  0.359775  0.373034      Tucson

A graph of blue bars with black text

Description automatically generated

A group of graphs with numbers

Description automatically generated with medium confidence

**City-Specific Performance**

**High-Performing Cities**

* **Detroit and Kansas City**:
  + These cities show slightly better performance, with accuracy above **41%** and relatively balanced precision and recall for the assault and gta classes.
  + **Detroit**: gta demonstrates the highest recall (59%), showing that the model captures this crime type better than in other cities.
  + **Kansas City**: assault and gta both exhibit improved recall, but robbery remains poorly classified (recall: 7%).

**Moderate-Performing Cities**

* **Tucson, Atlanta, Cleveland, and Sacramento**:
  + These cities exhibit moderate accuracy (**36% to 44%**) with varied performance across crime types.
  + **Tucson**: While the recall for robbery is relatively high (44%), the precision is very low (19%), indicating significant misclassification.
  + **Cleveland**: The model achieves decent precision for assault (80%) but struggles with low recall for all classes except assault.
  + **Atlanta**: Shows a slightly balanced performance for assault and gta, but robbery metrics remain low (recall: 23%).

**Low-Performing Cities**

* **Boston, Louisville, and Memphis**:
  + These cities have the lowest performance, with accuracy between **33% and 37%** and poor metrics for all crime types.
  + **Boston**: While robbery has a surprisingly high recall (45%), its precision is very low (12%), indicating poor overall performance.
  + **Louisville**: The assault class dominates predictions but exhibits low recall (28%), showing poor balance.
  + **Memphis**: Strong imbalance is evident, with the model favoring assault but failing to adequately predict gta and robbery.

**Further Analysis and Recommendations**

**1. Address Class Imbalance**

* Techniques like SMOTE (used here) should be complemented with cost-sensitive learning or other oversampling methods to improve predictions for minority classes.

**2. Feature Engineering**

* Explore new predictors such as:
  + Geographic information (e.g., neighborhood or zip code).
  + Crime-specific attributes (e.g., time since last reported crime in the area).
  + Socioeconomic indicators (e.g., median income, unemployment rate).

**3. City-Specific Analysis**

* Analyze each dataset individually to identify city-specific patterns and outliers. Cities with poor performance (e.g., Boston, Memphis) may benefit from targeted feature engineering or alternative data collection methods.

**Summary**

While the Random Forest model provides some predictive power, its performance is limited, particularly for minority classes. High-performing cities like **Detroit** and **Kansas City** suggest that the model can capture some patterns in the data. However, low-performing cities like **Boston** and **Memphis** highlight the need for additional predictors and better handling of class imbalance. A combination of advanced feature engineering, alternative algorithms, and deeper analysis of city-specific data could significantly enhance crime type prediction.

#### Overall Limitations and Considerations:

* **Limited Predictors:** The model's reliance on only three predictors (time of day, season, day of the week) is a significant limitation. Incorporating additional factors such as location specifics, crime type details, socioeconomic indicators, and weather data would likely improve prediction accuracy.
* **Data Quality:** The quality and completeness of the crime data are critical. Inconsistent data, missing values, or errors could lead to reduced model accuracy. A thorough assessment of data quality is necessary.
* **Model Simplicity:** The Gaussian Naive Bayes model's simplicity might be insufficient to capture the complex relationships between predictors and crime types. More sophisticated models (e.g., decision trees, random forests, gradient boosting machines, neural networks) could provide better performance.
* **Class Imbalance:** Class imbalances in the datasets (some crime types being far more frequent than others) can negatively impact model performance. Addressing this through techniques like oversampling, undersampling, or cost-sensitive learning is recommended.
* **AUC Variation Across Classes:** The ROC AUC analysis reveals that the model's predictive power varies significantly across the three crime types (assault, GTA, robbery) even within the same city. This highlights the need for further investigation into feature engineering and model selection.
* **Weaker Performers**: Tucson and Atlanta have consistently lower scores, suggesting noisier data, fewer features, or less-defined patterns.
* **Room for Improvement**: Consider feature engineering or hyperparameter tuning (e.g., adjusting k values) to improve performance in underperforming datasets.

### Conclusion

The initial proposal correctly predicted that this project would provide insights into crime patterns. However, the results revealed a more nuanced picture than initially anticipated.

No single model consistently outperformed others across all cities. The performance of each model varied widely, suggesting that the predictive power of the chosen features (time of day, day of the week, season) significantly differs across cities. This indicates that other factors, not included in the models, play a crucial role in crime patterns.

Some cities (Cleveland, Memphis, Louisville) showed relatively better performance across multiple models, implying more predictable crime patterns based on the given features. Other cities (Kansas City, Tucson, Atlanta) consistently demonstrated poor performance, indicating that the selected features were not sufficiently predictive in these locations.

The analysis confirmed that relying solely on time of day, day of week, and season is insufficient for accurate crime type prediction. Additional features (location, socioeconomic factors, specific crime details) are crucial for improving the model's performance.

The datasets likely had class imbalances (some crime types being more frequent than others). This imbalance affected the models' performance, particularly for less frequent crime types (lower precision and recall).

The KNN model's visualizations revealed some general patterns: Robbery and Grand Theft Auto (GTA) often peak during evenings and nights on weekends, while assault is more evenly distributed. Seasonal variations also existed, with winter showing higher activity for some crimes.

This project significantly enhanced our understanding of Machine Learning Model Selection, Feature Engineering, Data Analysis and Interpretation, and Data Preprocessing.

The need to evaluate multiple models and select the best-suited one based on the data's characteristics and the project's goals. The importance of selecting relevant and sufficient features for effective prediction. The project highlighted how limited features can lead to poor model performance. Analyzing model performance metrics (accuracy, precision, recall, F1-score, AUC), understanding their limitations, and drawing meaningful conclusions from the results. Dealing with issues like class imbalance and its impact on model performance. Using visualizations to effectively communicate complex data and patterns.

The quality and comprehensiveness of data significantly impact the success of machine learning projects. More comprehensive datasets with additional features are necessary for improving accuracy. Simple models may not be sufficient to capture the complexities of crime patterns. More sophisticated models or ensemble methods might be needed. Crime patterns are highly context dependent. Generalizable models may not be effective, and city-specific analyses and models might be more appropriate. Machine learning is an iterative process. Continuous evaluation, refinement, and improvement are necessary for achieving optimal results. This project's findings will help guide future iterations.

In conclusion, while the project achieved its goal of providing insights into crime patterns, it also revealed the complexities involved in accurately predicting crime types using machine learning. The limitations encountered highlight the need for future work focusing on richer datasets, more sophisticated models, and a more nuanced understanding of city-specific crime dynamics.

# References

Boston, A. (2024, January 11). *Crime Incident Reports.* Retrieved from Analyze Boston: https://data.boston.gov/dataset/crime-incident-reports-august-2015-to-date-source-new-system/resource/313e56df-6d77-49d2-9c49-ee411f10cf58

Center, R.-t. C. (2024, June 21). *Memphis Police Department Aggregate Crime.* Retrieved from Memphis Data Hub: https://data.memphistn.gov/browse?q=crime&sortBy=relevance&page=1&pageSize=20

City, B. (2024, December 11). *Part 1 Crime Data.* Retrieved from Open Baltimore: https://data.baltimorecity.gov/search?layout=grid&q=crime&sort=Date%20Updated%7Cmodified%7Cdesc

City, C. o. (2024, April 11). *KCPD Crime Data 2019 to 2024.* Retrieved from City of Kansas City, MO: https://data.kcmo.org/Crime/KCPD-Crime-Data-2019-to-2024/egkn-fji8/about\_data

CLE, O. (2024, December 11). *Crime Incidents.* Retrieved from OpenData CLE: https://data.clevelandohio.gov/datasets/c749e34199c1425cbbc5959308658ec3\_0/explore?location=41.380429%2C-81.633266%2C10.36&showTable=true

Department, T. P. (2024, December 10). *Crime Trends.* Retrieved from Tucson Police Department: https://app.powerbigov.us/view?r=eyJrIjoiY2ViZmNiYzAtMDQ5ZC00OTMwLTliMTgtYjM1ZjAwYjJlMTkzIiwidCI6ImQyMWU1OWVjLWMyMDgtNDNlYi1hYWYxLWNmMDZkOWExOTZlMCJ9&pageName=ReportSectionccd6be2a1a0db780dadd

Member, P. (2024, December 10). *RMS Crime Incidents.* Retrieved from City of Detroit Open Data Portal: https://data.detroitmi.gov/datasets/8e532daeec1149879bd5e67fdd9c8be0\_0/explore?location=42.387584%2C-83.205108%2C14.47&showTable=true

OpenDataPhilly. (2024, December 5). *Crime Incidents.* Retrieved from OpenDataPhilly: https://opendataphilly.org/datasets/crime-incidents/

Person, P. (2024, April 20). Retrieved from Experience Arcgis: https://experience.arcgis.com/experience/76b68b923a094fb5a2d193ace5ccb975/

Publisher, S. (2024, December 8). *Sacramento Report Data 2022.* Retrieved from City of Sacramento: https://data.cityofsacramento.org/datasets/f6f089ed676e4f968c2dc4c5bb6da06e\_0/explore?location=38.567438%2C-121.381834%2C9.89&showTable=true