# Crime Report

# Ayan Satani

# November 2024

Table of Content

[1.0 Introduction 2](#_Toc245371406)

[2.0 Business Understanding 3](#_Toc2081396163)

[2.1 Determine business objectives 3](#_Toc1558175475)

[2.2 Assess situation 4](#_Toc377067499)

[2.3 Determine data mining goals 4](#_Toc831393969)

[2.4 Project plan 5](#_Toc690330412)

[3.0 Data Understanding 9](#_Toc1418800863)

[3.1 Collect initial data 9](#_Toc1736496528)

[3.2 Describe data 9](#_Toc1684078330)

[3.3 Explore data 10](#_Toc1628988285)

[3.4 Verify data quality 11](#_Toc2044121710)

[4.0 Data Preparation 12](#_Toc559767405)

[4.1 Select data 12](#_Toc16685173)

[4.2 Clean data 13](#_Toc758174153)

[4.3 Construct data 13](#_Toc1473402336)

[4.4 Integrate data 14](#_Toc387672319)

[4.5 Format data 14](#_Toc749417218)

[5.0 Modelling 15](#_Toc1129848538)

[5.1 Select Modelling Techniques 15](#_Toc66353236)

[5.2 Generate Test Design 18](#_Toc129027914)

[5.3 Build Model 26](#_Toc983944629)

[5.4 Assess Model 31](#_Toc1207578372)

[6.0 Evalution 37](#_Toc1984500604)

[6.1 Evaluate Result 38](#_Toc2116690234)

[6.2 Interpret Result 44](#_Toc1372293065)

[6.3 Review of process 47](#_Toc1701287252)

[6.4 Determine next step 51](#_Toc1819910709)

[7.0 Conclusion 54](#_Toc1281419378)

[7.1 Outlier Detection 54](#_Toc1886112412)

[7.2 Clustering 55](#_Toc178717296)

[7.3 Association Rules (Apriori) 55](#_Toc1220240734)

[7.4 Classification Models 55](#_Toc647919682)

# 

# 

# 1.0 Introduction

This report was conducted by Ayan Satani – analyze the 2018 Austin Crime Reports dataset. Our objective is to understand the general trend of factors that contribute to a higher likelihood of a crime incident being resolved or cleared up. By identifying key indicators, I aim to better understand how various attributes of crime incidents, such as location, time, district, or type of crime, influence clearance rates, thereby providing insights that could assist in crime resolution strategies.

# 2.0 Business Understanding

The analysis of the Austin Crime Reports 2018 dataset aims to uncover crime patterns, trends, and hotspots, helping law enforcement and city planners optimize strategies for public safety. By identifying these trends, resources can be allocated more efficiently, ensuring quicker responses and enhanced safety across Austin. This analysis supports data-driven approaches to crime prevention, contributing to informed decision-making for better law enforcement practices and the reduction of criminal activity in the community.

### 2.1 Determine business objectives

This analysis's primary goal is to pinpoint the critical elements that affect the possibility that a crime in Austin will be resolved or "cleared." Law enforcement can concentrate efforts on regions and situations that have a better probability of being resolved by being aware of the factors. This makes it possible to manage cases more effectively, which raises clearance rates, and eventually makes neighborhoods safer. The analysis's conclusions will direct the improvement of crime resolution tactics, resulting in more focused and effective law enforcement operations.

### 2.2 Assess situation

The 2018 crime dataset from Austin provides important information about the types, locations, and timing of crimes, which may greatly assist with crime analysis. Nevertheless, issues could occur because of data inconsistencies, like differences in the way data is recorded or missing numbers that could cause gaps in the dataset. Furthermore, the reliability of the study may be impacted by the presence of outliers, which are unusual data items that potentially skew conclusions.

### 2.3 Determine data mining goals

1. Predictors for Crime Clearance Status: The basic objective is to identify specific factors that significantly indicate whether a crime will be "cleared" (settled) or go unresolved. By examining characteristics such as the district, crime type, location, and incident time, I can spot trends that point to increased or decreased chances of clearance. Law enforcement can use this information to direct resources towards occurrences that have a lesser likelihood of being resolved.
2. Clusters to Reveal Common Patterns: Clustering analysis will help us group similar crime incidents, revealing underlying patterns in the dataset. For example, clusters may identify high-frequency crime areas, recurring types of crime, or times when specific incidents are more likely.
3. Detect Outliers: The goal of outlier detection is to find extreme or uncommon occurrences that diverge from normal crime trends. These anomalies could be unique circumstances with unique features or extremely complicated cases that are difficult to handle.

By applying these three techniques—identifying predictors for crime clearance, clustering incidents, and detecting outliers—I can gain actionable insights into crime patterns in Austin. This analysis enables targeted resource allocation, more effective crime prevention strategies, and efficient case resolution efforts, ultimately contributing to a safer and more peaceful neighborhood environment for the community.

### 2.4 Project plan

For the effective and successful completion of this project, responsibilities have been thoughtfully divided among all team members to ensure a balanced workload. Each member is tasked with specific roles aligned with their skills and expertise, fostering collaboration and efficiency. Below is a comprehensive breakdown of the project plan.

1. Outlier Detection

* Techniques Used:
  + Local Outlier Factor (LOF): Detects data points that significantly deviate from their neighbors, identifying unusual crime cases. By analyzing the local density of a point in comparison to its surrounding points, LOF can highlight anomalies that are not apparent on a global scale. This method is particularly effective in datasets with variable density, where traditional outlier detection techniques may fail. For example, LOF can uncover hidden crime patterns, such as rare offenses in densely populated areas or sudden spikes in specific incidents, aiding in more targeted law enforcement efforts.
  + Isolation Forest (ISF): Isolates anomalies using a tree structure to flag outliers that don’t fit typical crime trends. This method works by recursively partitioning the dataset, creating random splits that separate points based on their distinct features. Anomalies, being rare and different, tend to be isolated quickly with fewer splits, making ISF an efficient approach for anomaly detection. Its scalability makes it suitable for analyzing large-scale crime datasets, identifying rare or emerging crime trends in real-time. Additionally, ISF is robust against noisy data and does not rely on assumptions about the underlying distribution, making it versatile for diverse crime data scenarios, such as detecting outlier crime rates in dynamic urban environments.
* Presentation:
  + Snapshots: Include LOF and ISF visualizations with marked outliers.
  + Interpretation: Explaining why certain incidents outliers are, considering factors like unusual crime types, times, or locations that contribute to their categorization.

2. Clustering

* Techniques Used:
  + k-Means Clustering: Groups similar crime incidents that follow similar trend in the dataset.
  + Elbow Method: Helps determine optimal clusters by analyzing the reduction in variance at different k-values.
* Presentation:
  + Snapshots: Show cluster distributions and the elbow method.
  + Interpretation: Describing trends within each cluster (e.g., geographic patterns, common crime types) and any insights on clusters that stand out, helping to identify focused areas of criminal activity or hotspots.

3. Association Analysis

* Technique Used:
  + Apriori Algorithm: Finds frequent item set in the data (e.g., crime type + location + time) to show relationships among factors.
* Presentation:
  + Snapshots: Display item set and association rules.
  + Interpretation: Explaining the meaning of these rules, noting how they reveal common combinations in crime incidents that can assist in predictive policing.

4. Classification

Techniques Used:

* + k-Nearest Neighbors (kNN): Classifies crime incidents based on similarity, useful for predicting clearance likelihood. Similarity is often based on factors like crime location, time, type, and family violence with the model analyzing past patterns to determine likely outcomes.
  + Decision Tree (DT): Builds a model with rules that help identify factors important for crime resolution. It’s beneficial for interpreting which factors most influence crime clearance.
  + Random Forest (RF): RF helps reduce overfitting, as each tree is exposed to different data subsets. In this project, RF can provide a comprehensive view of the primary factors leading to a crime's clearance and is effective in capturing complex patterns from diverse variables like crime types, time of day, and district interactions.
* Presentation:
  + Snapshots: Include confusion matrices, accuracy scores, and relevant metrics.
  + Interpretation:
    - For DT: Explain the key rules (e.g., certain crime types or locations that strongly influence resolution chances).
    - For kNN and RF: Discuss the most effective classification model and any insights into misclassified cases.

5. Evaluation and Reporting

* Objective: Compare model performances using metrics like accuracy, precision, and recall. Discuss each model's effectiveness and limitations.
* Conclusion: Summarize findings, emphasize actionable insights that could inform resource allocation for crime prevention and resolution in Austin.

# 3.0 Data Understanding

This section provides an overview of the 2018 Austin Crime Reports dataset, focusing on the initial data collection, dataset description, exploration, and verification of data quality. Understanding these aspects lays the foundation for subsequent analyses, including classification, clustering, and outlier detection.

The dataset encompasses crime incidents recorded in Austin during 2018. It includes detailed information about each crime, such as its description, location, date and time of occurrence, date and time of reporting, clearance status, and clearance date. This comprehensive dataset enables the analysis of patterns, trends, and anomalies in crime occurrences and resolutions.

### 3.1 Collect initial data

The 2018 Austin Crime Reports is made of 98000 crime records from Austin. Each record consists of 19 attributes. Each row represents a distinct crime scene providing factors for analysis like offense type, time, location, and clearance status.

### 3.2 Describe data

The given data description is given below:

* Incident Number: Incident Report Number.
* Highest Offense Description: Provides a detailed primary offense description.
* Highest offense Code: Offense code.
* Family Violence: Shows whether family violence is included or not (Yes/No).
* Occurred Date and Time: Data and time the incident occurred (combined).
* Date: Date the incident occurred.
* Occurred Time: Time the incident occurred.
* Report Data and Time: Date and time the incident reported (Combined).
* Report Date: Date the incident was reported.
* Report Time: Time the incident was reported.
* Location type: General description of place where the incident occurred.
* Council District: Austin council district where the incident occurred.
* APD Sector: APD sector where the incident occurred.
* APD District: APD district where the incident occurred.
* Clearance Status: Crime is solved or not.
* Clearance Date: Date the crime is solved.
* UCR Category: Code for the most serious crimes identified by the FBI as part of the Uniform Crime Reporting program.
* Category Description: Description of the most serious crimes identified by the FBI as part of its Uniform Crime Reporting program
* Census Block Group: A number that indicated the census block group where the incident occurred.

### 3.3 Explore data

In this step, I performed initial exploration to understand the structure and relationship within the data. I considered:

1. Data Summary: Generated summary statistics for numerical and categorical variables to find and understand the ranges, averages, and unique values.
2. Frequency Analysis: Analyzed the occurrence rate of different types of crimes, their locations and whether it is family violence or not. This provided an insight into the most common crimes and crime hotspot that could be used for classification and clustering.
3. Patterns: Using occurred date and time explored temporal trends like peak time for crime, months with increased crime rates. Clearence date and occurred date was also analyzed to get insights about the time took to resolve the cases.
4. Geography: Visualized the data geographically to identify the crime hotspot and mapped accidents by the APD sector and council district to reveal clusters.

### 3.4 Verify data quality

In this phase, I focus on ensuring the data used for modeling is reliable, accurate, and clean. Key activities include:

1. Handling Missing Values: I will identify any missing data points and determine the appropriate strategy for imputation or removal based on the context and quantity of missing data.
2. Data Consistency: I will validate that all data follows standard formats and correct any inconsistencies (e.g., standardized date formats or categorical variables).
3. Outlier Detection: Outliers will be identified to see if they represent errors or unique cases, and appropriate actions (e.g., removal or adjustment) will be taken.
4. Duplicate Records: Duplicate entries will be checked and removed to ensure that our analysis reflects accurate and non-redundant data.

By verifying the data quality, I ensure that the models built will be based on a reliable foundation, improving the accuracy and validity of our findings.

# 4.0 Data Preparation

For the Austin Crime Reports 2018 dataset, I undertook several key steps to ensure the data was ready for analysis. These included:

1. Selecting Relevant Columns: Focusing on attributes such as crime type, time, and location, which directly influence crime clearance rates.
2. Data Cleaning: Handling missing values and ensuring the consistency of data entries.
3. Constructing New Attributes: Creating additional features like time of day or crime severity to enrich the dataset.
4. Integrating Derived Information: Adding supplementary data, such as police district boundaries, to enhance the dataset’s comprehensiveness.
5. Formatting the Data: Ensuring all data is in appropriate formats for ease of use, such as converting categorical variables into numerical ones where necessary.

### 4.1 Select data

Clearance Status: This is a crucial field for analyzing resolution rates and identifying factors that might influence the likelihood of a case being cleared.

Highest Offense Code: This attribute helps in categorizing the incidents based on offense severity and type.

Council District: Specifies the Austin council district in which the crime occurred. This attribute enables spatial analysis, helping us determine if certain areas have higher crime rates or clearance rates.

Family Violence: Family-related incidents may have different clearance patterns compared to other types of offenses.

Location Type: This attribute is useful for identifying the environment in which crimes are more likely to happen.

APD Sector: For further spatial categorization and sector-based analysis.

APD District: Specifies the APD district within the sector, adding another layer of spatial detail.

Clearance Date: This field is used to understand the timeframe for resolving cases and identifying trends in case closures.

### 4.2 Clean data

In the cleaning phase of the Austin Crime Reports 2018 dataset, I carried out the following steps:

1. Removed Duplicates: Identified and eliminated any duplicate records to ensure each crime incident was counted only once.
2. Handled Missing Values: Missing values were evaluated, and where they were insignificant to our analysis, they were ignored. For critical fields with missing data, imputation strategies were applied where appropriate.

This process ensured the dataset was clean, accurate, and ready for further analysis.

### 4.3 Construct data

Binning Date and Time Fields: The original dataset included fields for Occurred Date and Time and Report Date and Time. To analyze time-based patterns more effectively, I split these fields into separate columns for:

Time of Day: I categorized each incident's time into morning, evening, and night bins. This categorization helps us identify time-based patterns in crime rates.

Week Binning: Each month was divided into five bins labeled "Week 1" to "Week 5," based on the day of the month. This binning provides a way to examine weekly trends within each month, which could show if there were increased incidents in certain weeks.

Month Binning: I grouped incidents by month (e.g., January, February) to understand seasonal variations and monthly trends in crime occurrences. This provided us with detailed information about whether there was a higher or lower crime rate pattern in certain months

Encoding Family Violence as Binary: I ensured that the Family Violence field was encoded as binary (Y/N), enabling it to be used easily in both statistical and machine learning analyses.

### 4.4 Integrate data

In this step, I combined supplementary data sources to enrich the Austin Crime Reports 2018 dataset. This included integrating geographic and district-specific information, such as police district boundaries and demographic data, to provide additional context for the analysis. By merging this external information, I enhanced the dataset's comprehensiveness, allowing for more detailed insights into the factors influencing crime resolution. The integrated data was aligned and formatted to ensure compatibility with the existing dataset for seamless analysis.

### 4.5 Format data

I assigned appropriate data types to each attribute to ensure accurate analysis:

* Id: Set as an integer for unique identification.
* Clearance Status, Family Violence, Location Type, APD Sector, and APD District: Defined as polynomial types, treating them as discrete categorical variables.
* Highest Offense Code and Council District: Set as nominal types for categorization and clustering.
* Clearance Date: Formatted as a date type for date-based calculations.
* Bin Day and Bin Month: Designated as nominal types for easy grouping and trend analysis.

# 5.0 Modelling

In this chapter, I will detail the steps undertaken to apply various modeling techniques to analyze and interpret the dataset. The methods covered include outlier detection, clustering, association rule mining, and classification. Each technique has been carefully chosen to uncover meaningful insights and improve the predictive capabilities of the models. Additionally, I will evaluate the accuracy of the predictive models using metrics derived from the confusion matrix, ensuring a clear understanding of the results. Relevant screenshots are provided throughout the chapter to illustrate key steps and enhance comprehension of the data analysis process.

### 5.1 Select Modelling Techniques

1. Outlier Detection

* **LOF(outlier detection):**
* **Reason for Selecting LOF:** LOF was chosen for its effectiveness in detecting local anomalies by comparing the density of a data point to its neighbors. This technique is especially useful for datasets with varying density regions, such as crime datasets, where local variations in behavior can indicate outliers.
* **Applicability to Crime Data:** LOF is well-suited for identifying unusual activity in regions or times based on features like location, offense code, and council district.
* **Justification**: Unlike global methods, LOF adapts to local density patterns, making it ideal for analyzing heterogeneous data distributions in crime incidents.
* Isolation Forest (ISF) method: Isolation Forest (ISF) is a tree-based algorithm specifically designed to detect anomalies by leveraging the concept of data isolation. It provides a prediction column to find the outliers.

Key reasons to choose ISF:

* Scalability: The dataset contains 10,000 records, making computational efficiency a critical factor. ISF is well-suited for large datasets due to its linear time complexity and ability to handle high-volume data efficiently.
* Robustness: ISF is more sensitive towards finding global outliers. This global perspective makes ISF particularly effective at detecting anomalies that deviate significantly from overall patterns. As a result, it is well-suited for datasets with diverse distributions or attributes where traditional methods might overlook global deviations.
* Interpretability of Results: ISF provides an anomaly score and a prediction column that can be used to identify outliers. This straightforward labeling makes it easy to filter and analyze the anomalous data points. Additionally, the anomaly score allows the ranking of outliers based on their severity, providing deeper insights into the extent of deviation, which can help prioritize further investigation.

1. Clustering

Clustering was chosen as the primary modeling technique to group similar instances within the dataset and uncover meaningful spatial and temporal patterns. The k-Means algorithm was selected due to its efficiency, simplicity, and ability to produce interpretable groupings. This technique is particularly useful for analyzing crime data to identify crime hotspots, seasonal variations, and trends across geographical regions.

The purpose of clustering in this context includes:

* Pattern Identification: Grouping data points to reveal hidden structures and relationships within the dataset.
* Actionable Insights: Detecting hotspots and patterns to support targeted decision-making and resource allocation.
* Data Simplification: Reducing dataset complexity by categorizing instances into meaningful clusters for better interpretability.

1. Outlier Detection using Clustering
   * I decided to use clustering-based outlier detection. Clustering helps group similar data points together, and anything that doesn’t fit well into these groups can be considered an outlier. It’s simple and works great when I don’t know much about the data upfront.
2. Classification

* K-Nearest Neighbours (kNN): The kNN algorithm was chosen because it works well for classification problems that don't require an explicit model structure. The majority class of an instance's closest neighbors, as established by a distance metric, is used to predict the instance's class.
* Decision tree: The Decision Tree algorithm was selected because it offers a straightforward and easy-to-understand framework that emphasizes the key elements impacting predictions. This model operates by repeatedly dividing the dataset into smaller groups based on specific feature values or categories, forming a tree-like structure. It's ideal for handling different types of data, both numerical and categorical, and gives clear, rule-based explanations about what influences Clearance Status. I experimented with both numerical and categorical values to determine the most effective method.
* Random Forest: Random Forest was selected due to its capability to integrate the results from various Decision Trees, which reduces overfitting and improves the accuracy of predictions. It uses random samples of the data and features for training each tree, leading to diverse decision boundaries that increase the model's strength. Moreover, this method also generates important rankings for features, helping us understand which elements play a key role in predictions.

**Association(Appriori)**

* Rationale for Apriori Algorithm:
* The Apriori algorithm was selected due to its efficiency in discovering frequent patterns and associations within categorical datasets.
* It aligns with the project objective of uncovering relationships between attributes such as Occurred Time Group, Highest Offense Code, and Council District.
* The technique provides interpretable rules that are actionable for stakeholders, enabling insights into patterns like crime frequency at specific times or locations.
* Applicability:
* The categorical nature of the dataset is well-suited to association rule mining.
* The method’s reliance on support and confidence metrics ensures the extraction of only meaningful and reliable rules.

### 5.2 Generate Test Design

Outlier Detection

* LOF(outlier detection): The generation of test design included:
* Data Preprocessing:
* Ensuring all relevant attributes (e.g., "Occurred Time," "Location Type," "Council District") are included and properly formatted.
* Using the Nominal to Numerical operator to convert nominal values into numerical ones for attributes where LOF requires numerical inputs.
* Attribute Selection:
* Focused on attributes that significantly influence crime patterns, such as "Family Violence," "Highest Offense Code," and "Council District."
* Data Preparation:

All relevant attributes (e.g., Clearance Status, Council District, Highest Offense Code) were selected for the analysis.Data was normalized using preprocessing steps to ensure comparability of features.

* Parameter Configuration: Minimal points lower bound: Set to 10 to define the minimum number of neighbors for local ensity computation. Minimal points upper bound: Set to 20 for flexibility in density evaluation.
* Distance Function: Euclidean Distance was used to calculate proximity between data points.
* Validation:

A separate filtering process was set up to verify the flagged outliers, ensuring accuracy and interpretability.

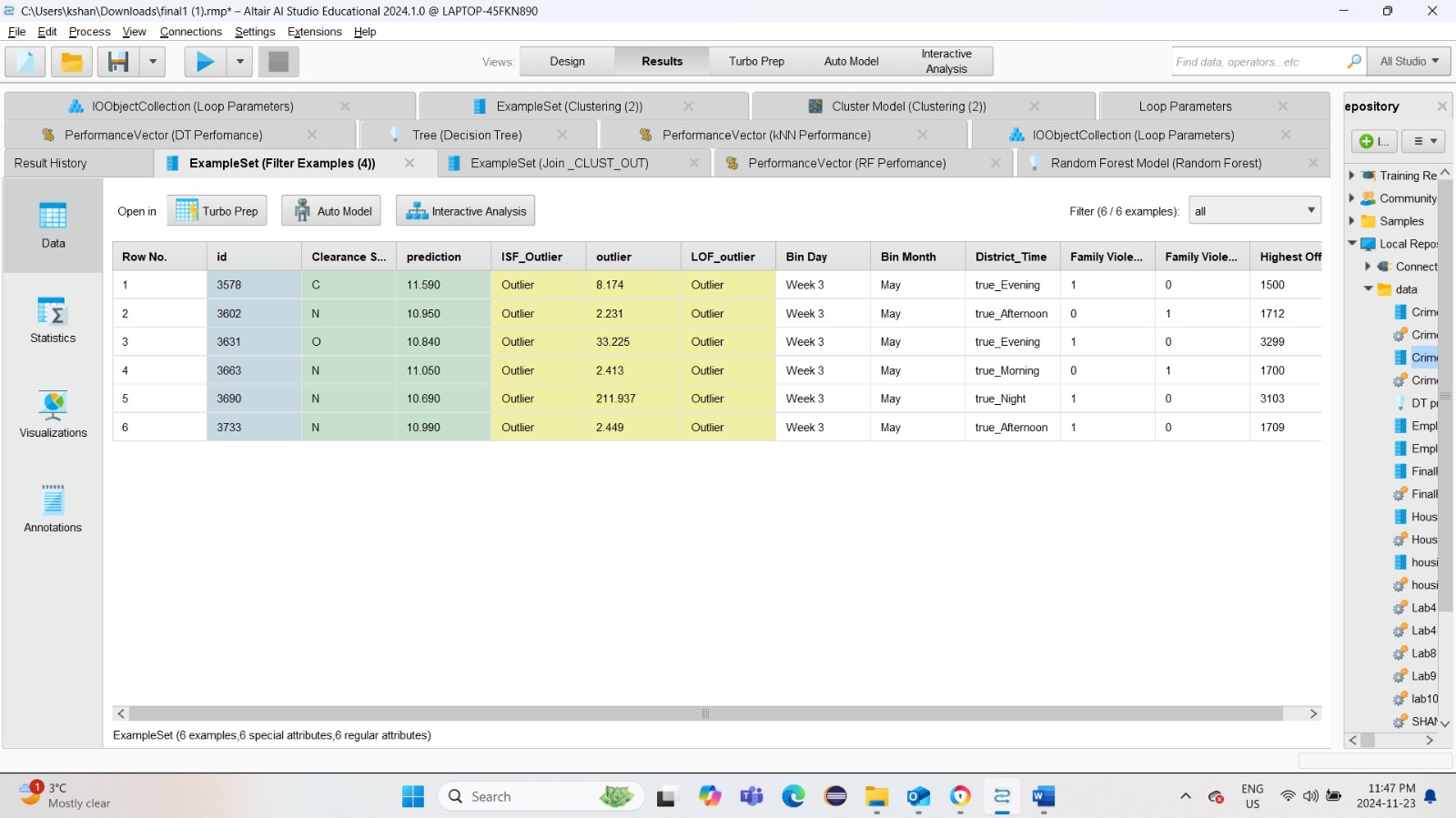
1. Isolation Tree method:

* Data Preprocessing:
* Numerical to Nominal Conversion: The "Highest Offense Code" attribute was converted from numerical to nominal using the Numerical to Nominal operator. This step enables the model to handle categorical data more effectively.
* Feature Engineering:
* Creation of New Attribute: A new attribute, District\_Time, was generated by grouping the data based on the Occurred Group Time and Council District. This transformation helps in understanding the relationship between time and district for better anomaly detection.
* Data Normalization:
* Normalization of Highest Offense Code: The Highest Offense Code column was normalized to bring the values within a consistent scale. This ensures that the model treats this feature uniformly, preventing features with larger numerical ranges from dominating the analysis.
* Handling Missing Data:
* Missing Value Imputation: The Replacing Missing Values operator was applied to handle null or missing data, ensuring that the dataset is complete and ready for anomaly detection.
* Model Training:
* Isolation Forest Configuration: The Isolation Tree operator was applied with the following parameters:
  + 1. Number of Trees: 100
    2. Max Leaf Size: 7
    3. Bootstrap Ratio: 0.9

These settings control the depth and number of isolation trees used to identify anomalies effectively.

* Filtering the Output:
* Anomaly Thresholding: The model output was filtered to only include instances with prediction values greater than 10.7, representing potential anomalies with high deviation from the normal data distribution.
* Data Integration:
* Data Join: Finally, the new attribute (predicted anomalies) was joined with the original dataset, ensuring that the results of the anomaly detection process were integrated and aligned with the existing data for further analysis.

Note: When LOF and ISF is joined together the out for number of outliers is 6.



Clustering

* Data Preprocessing:
  + Attribute Selection: Key attributes such as Council District, APD District, Family Violence, Highest Offense Code, Bin Month, and Bin Day were included to focus on spatial, temporal, and categorical aspects of the data.
  + Handling Missing Values: Missing values were imputed using the average to ensure consistent and reliable input for the algorithm.
  + Nominal to Numerical Conversion: Categorical attributes like Family Violence and Council District were converted into numerical formats using dummy coding to ensure compatibility with the k-Means clustering algorithm.
* Cluster Configuration:

For the clustering process, the k-Means algorithm was used to group the data into meaningful clusters. The number of clusters (k) was determined through iterative testing, ranging from 2 to 100 in steps of 10, and the optimal value was identified as 18. The algorithm was configured with a maximum of 10 runs to ensure stable initialization, and the Squared Euclidean Distance was chosen as the divergence metric to measure cluster compactness. To improve the accuracy of cluster centroids, the "determine good start values" option was enabled. Additionally, cluster assignments were added to the dataset, allowing for easier analysis of outliers, which were identified as the data points in the smallest cluster. This configuration ensured robust and meaningful clustering results.

* Testing
* Iterative testing with varying values of k helped finalize the number of clusters as 18, balancing meaningfulness and interpretability.
* Outlier Detection Using Clustering
  + Generate Test Design:
    - To identify outliers, clustering was applied to group the data into 30 clusters. The cluster with the least number of data points was identified and flagged as the outlier. This approach assumes that outliers will naturally fall into smaller, less populated clusters when meaningful features are used.
  + Functions used:
    - Clustering (Clustering\_OUT):Created clusters with k=30 to group the data based on similarity.
    - Aggregate\_CLUST\_OUT (1st Aggregate): Grouped the clusters and calculated the count of data points in each cluster.
    - AggregateMin (2nd Aggregate): Identified the cluster with the smallest number of data points.
    - Generate Attributes: Created a new attribute to label clusters as "Outlier" if their size matched the smallest count.
    - Set Role\_CLUST\_OUT: Assigned the newly generated outlier attribute a specific role.
    - Join\_CLUST\_OUT: Merged the labeled outlier attribute back with the original dataset, retaining necessary columns.
    - Select Attributes\_CLUST\_OUT: Selected only the required columns for the final output.

Classification

* KNN
* Data Preprocessing:
* Normalization: Numerical attributes like Highest offensive code were normalized. This is to make sure that each process is making an equal contribution to the distance calculation.
* Nominal to numeric conversion: This was done to convert categorical values like Family violence to numerical format for the successful completion of kNN.
* Set Role: Done to make sre clearance status is set to label
* Attribute selection:

The following attributes were selected to influence Clearance Status effectively:

* Family violence: whether the crime involves family violence or not.
* Council district: Represented geographical areas linked to crime resolution rates.
* Highest offense code: Highlighted the severity of the crime.
* Occurred time group: The time frame (morning, afternoon, evening, night) in which the incident occurred.
* Defining Splits and Parameters:
* Number of Neighbors (k): The value of k was optimized by testing various values and selecting the one that achieved the highest accuracy.
* Split data: the test and train set is considered in a 70:30 ratio.
* Decision Tree:
* Data Preprocessing:
* Nominal to Numeric Conversion: Categorical attributes such as Family Violence (encoded as 0 for "No" and 1 for "Yes") were converted into numeric formats to ensure compatibility with the Decision Tree algorithm.
* Numerical Attributes Retained: Attributes such as Highest Offense Code, Bin Month, and Bin Day were kept in their original numerical format. These attributes allowed the Decision Tree to create meaningful splits based on numerical thresholds.
* Attribute Selection

The following attributes were selected to influence Clearance Status effectively:

* Family Violence: Captured whether the crime involved domestic disputes.
* Council District: Represented geographical areas linked to crime resolution rates.
* Highest Offense Code: Highlighted the severity of the crime.
* Bin Day and Bin Month: Reflected temporal patterns associated with crime occurrences.
* Clearance Date: Included to study resolution trends over time.

Attributes like ID and other irrelevant features were excluded to minimize noise in the analysis.

* Defining Splits and Parameters:

* Maximal Depth: Set to 5 to balance interpretability and avoid overfitting.
* Splitting Criterion: Gain Ratio was selected to maximize information gain.
* Minimum Leaf Size: Set to 10 to prevent overly specific branches.
* Random Forest:
* Data Preprocessing:
* Numerical to Nominal Conversion:

Numerical attributes, such as Highest Offense Code, Bin Day, and Bin Month, were converted into nominal formats.

* Categorical Attributes Included Directly:

Attributes such as Location Type and Council District were retained in their original nominal forms, as Random Forest efficiently handles categorical data.

* Attribute Selection:

The following attributes were prioritized:

* Location Type: Identified contexts such as residential or commercial areas.
* Family Violence: Highlighted domestic crime cases.
* Council District and APD Sector: Represented spatial trends influencing crime clearance.
* Highest Offense Code (binned): Provided grouped insights based on severity ranges (e.g., "Low" to "High").

Redundant attributes, such as Clearance Date, were excluded to avoid overfitting.

* Design and Parameters:
* Number of Trees: Set to 100 to ensure stable and accurate predictions.
* Splitting Criterion: Each tree in the Random Forest split on categorical versions of previously numerical attributes

**Association (Apriori)**

* Objective:

To configure the process to extract high-confidence association rules that uncover hidden relationships in the dataset.

* Steps:
* Attribute Selection: Focused on key variables such as Bin Day, Council District, Highest Offense Code, and Location Type.
* Preprocessing: Applied Numerical to Binominal and Nominal to Binominal operators to ensure compatibility with the Apriori algorithm.
* Parameter Configuration: Set support and confidence thresholds to balance rule specificity and dataset coverage.
* Rule Filtering: Incorporated lift as an additional metric to ensure generated rules provide meaningful correlations.
* Validation:

Evaluated the configuration to ensure no significant variables were excluded and preprocessing steps-maintaineddata integrity.

### 5.3 Build Model

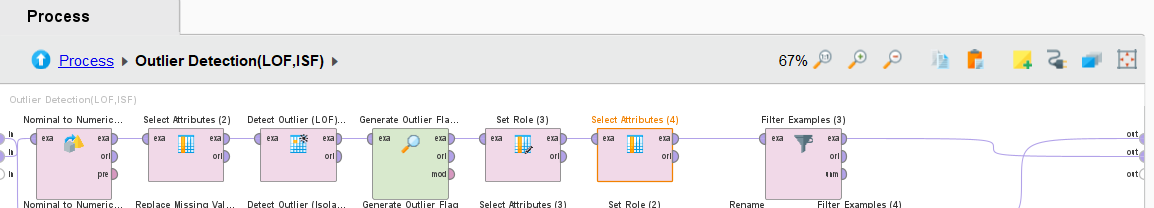
Outlier Detection

* LOF(outlier detection): Operator Implementation:
* The Detect Outlier (LOF) operator was used to calculate outlier scores based on local density deviations.
* The Generate Outlier Flag operator classified data points into "Outlier" or "Normal," simplifying interpretation.
* Attribute Roles:

Specific attributes like Highest Offense Code, Council District, and Clearance Status were crucial in building the model.

* Model Flexibility:

Adjustments to parameters (e.g., bounds for minimal points) ensured sensitivity to both high-density and low-density outlier regions.



* The Isolation Forest (ISF) method was implemented to identify anomalies in the dataset. ISF is particularly suited for anomaly detection as it isolates data points by recursively partitioning the dataset using random splits. Anomalies are easier to isolate due to their distinct characteristics, making ISF efficient and effective for detecting unusual patterns.

Steps in Model Building

1. Preprocessing and Feature Engineering:

* Numerical to Nominal Conversion: Attributes such as Highest Offense Code and Council District were converted from numerical to nominal for compatibility with the ISF analysis.
* Handling Missing Values: Missing values in the dataset were inputted to ensure completeness and maintain data integrity during the analysis.

1. Model Configuration:

* The Isolation Forest operator was configured with the following parameters:
* Number of Trees: 100, to provide a robust partitioning of the dataset.
* Maximum Leaf Size: 5, controlling the granularity of data isolation.
* Bootstrap Ratio: 0.9, ensuring randomness in tree construction while retaining sufficient data diversity.

1. Outlier Detection:

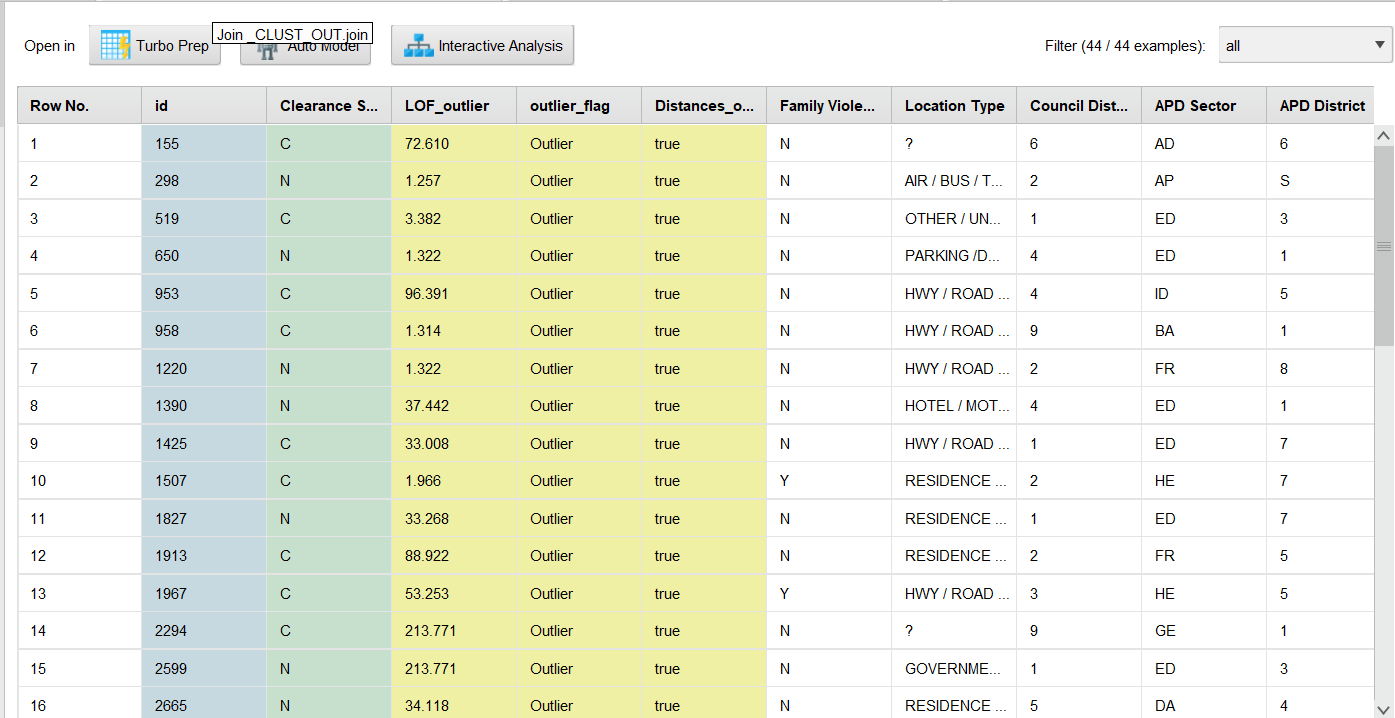
* Anomalies were identified based on their isolation scores. Instances with prediction values greater than 10.7 were flagged as potential anomalies, indicating significant deviation from the general data distribution.

1. Data Integration:

* The identified anomalies were incorporated back into the dataset by using a join operation. This ensured that the flagged anomalies could be analyzed alongside the original data attributes.

1. Output and Interpretation:

* The output of the ISF model included a prediction column indicating the anomaly score for each instance. This facilitated the identification of unusual patterns, such as atypical crime incidents, for further investigation and decision-making.



Clustering

The k-Means clustering model was built using the following steps:

* Cluster Settings:
  + The optimal number of clusters, k, was set to 18 based on results from the parameter tuning process.
  + Additional parameters such as Max Optimization Steps (100) and Max Runs (10) were configured to ensure model convergence and stability.
* Implementation:
  + The model was implemented using the k-Means operator in RapidMiner, with the Add Cluster Attribute option enabled to append cluster assignments to the dataset.
  + The clustering process utilized preprocessed data, including selected attributes and transformed categorical variables, to produce meaningful groupings.

Outlier Detection using Clustering

* Applied Clustering Algorithm:
  + Used the k-means clustering algorithm with k=30 to create 30 clusters.
  + The goal was to divide the dataset into meaningful groups to identify patterns and potential outliers.
* Aggregated Results:
  + First Aggregate: Grouped the dataset by cluster to count the number of data points in each cluster.
  + Second Aggregate: Identified the cluster with the least number of data points, marking it as the outlier.
* Generated Outlier Attribute:
  + Created a new attribute named "outlier" to label the least populated cluster as "Outlier" and others as "Normal."
* Set Role:
  + Assigned the "outlier" attribute the role of the target label to be used for further analysis and evaluation.
* Joined Results:
  + Combined the modified dataset with the original example set to retain all attributes while adding the outlier identification column.

Classification

* kNN

The k-Nearest Neighbour (k-NN) operator in rapid miner was used for creating kNN model

* Clearance Status was set as label.
* Data was split into 70% training and 30% testing data using split data.
* k-NN operator was used to find the best K as 8.
* Test data made the predictions with outputting predicted class and probability distribution for each class.
* Decision Tree:

The model was built in Rapid Miner with decision tree operator with parameters:

* Clearance Status as label.
* A splitting criterion like Gain Ratio.
* A maximum depth of 5 to keep the tree simple and prevent overfitting.
* A minimum leaf size of 10 to ensure meaningful splits.

It created rules like splitting on Bin Day and Family Violence to predict Clearance Status.

* Random Forest:

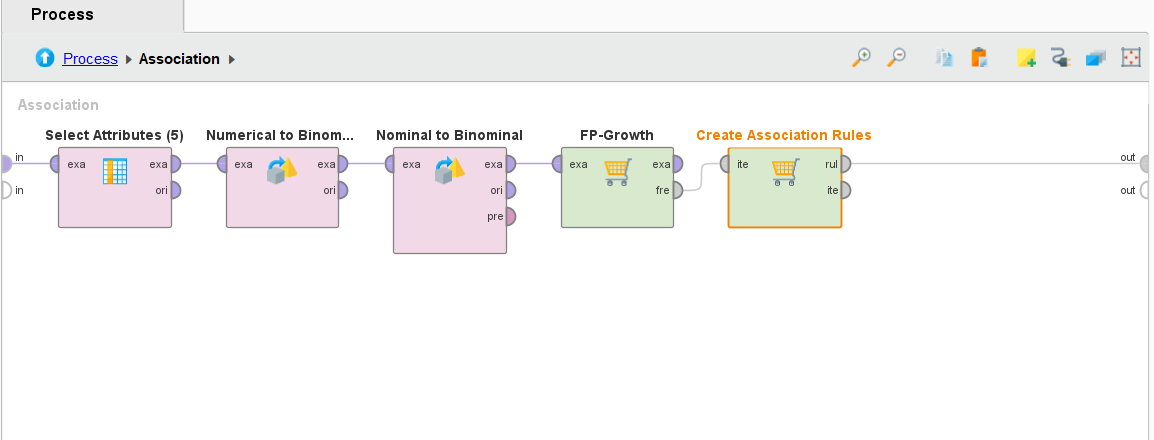
The model was built in rapid miner with random forest operator inside cross validation

* Clearance Status as label.
* 100 trees to improve stability and accuracy.
* Randomly selected features for each tree to improve diversity.
* Sampling with replacement to train on slightly different datasets for each tree.

It highlighted important features like Highest Offense Code and Location Type.

**Association (Apriori)**

The FP-Growth operator was employed to construct frequent itemsets based on the pre-processed data. These itemsets formed the foundation for generating association rules using the "Create Association Rules" operator. The process involved mining patterns such as "Bin Day = Week 3" strongly linked to specific council districts and high offense codes, with confidence levels nearing 100%. The rules generated provided valuable insights, revealing critical factors contributing to crime clearance rates, such as the correlation between location types, time groups, and council districts.



### 5.4 Assess Model

Outlier Detection:

LOF(outlier detection):

* Model Workflow Review:

Confirmed that the LOF process aligns with the project requirements, including detecting local density anomalies based on the attributes like Council District and Highest Offense Code. The workflow includes clear preprocessing steps (e.g., attribute selection and normalization) to ensure consistent and accurate data analysis.

* Parameter Validation:

Parameters such as the lower and upper bounds for minimal points (10 and 20, respectively) were validated for suitability to the dataset characteristics. The distance function (Euclidean Distance) was deemed appropriate for calculating proximity in the dataset's numerical attributes.

* Logical Integrity: Reviewed the process to ensure all operators were properly connected and configured such as Detect Outlier (LOF) effectively computes anomaly scores and the Generate Outlier Flag ensures interpretable classification of each data point.
* Cross-checked that attribute roles and selections were correct and relevant for outlier detection.
* Performance Checks:
* Confirmed that the LOF process handles the dataset efficiently, with no errors or data handling issues.
* Verified that the flagged outliers align with expectations for further analysis during the evaluation phase.
* Repeatability and Flexibility:
* The process design is repeatable for other datasets, with parameter flexibility allowing tuning for different density distributions.
* Ensured that the model can be iteratively refined based on feedback or changing project needs.

ISF method: The performance of the Isolation Forest (ISF) model was evaluated and assessed using the following methods:

1. Visualization and Filtering:

The Filter Examples operator was applied to narrow down the dataset to instances flagged as anomalies by the ISF model. The filtering criteria included:

* Combined with Other Methods: ISF results were integrated with anomalies flagged by other method (e.g., LOF) to create a comprehensive anomaly set for analysis.

1. Analysis of Results:

The flagged anomalies were analyzed to ensure their validity and relevance as outliers. Key patterns included:

* Unusual Crime Trends: Rare combinations of attributes, such as high "Highest Offense Code" in districts with low crime rates.
* Spatial and Temporal Irregularities: Anomalies linked to specific districts or crimes occurring at uncommon times, such as during holidays or late at night.
* Distinctive Attribute Combinations: Instances where multiple factors, such as council district and offense type, combined to create unique outliers.

1. Model Validation:

The identified anomalies were cross-referenced with:

* Domain Knowledge: To confirm that the flagged instances represented realistic and significant deviations.
* Other Models' Results: To ensure consistency and alignment with anomalies detected by LOF and Distance-Based methods.
* Dataset Characteristics: To verify that the anomalies were not artifacts of preprocessing or inherent biases in the data.

The ISF model proved effective in identifying global outliers and capturing complex patterns in the dataset. Its ability to isolate rare and meaningful anomalies added significant value to the analysis, making it a robust component of the outlier detection process.

Clustering

The clustering model was assessed to evaluate its effectiveness in grouping data points and uncovering actionable insights. The evaluation focused on the model's performance and the significance of the clusters.

Evaluation

The performance was analyzed using the following methods:

* Inter-Cluster Distances:
  + Larger distances between clusters indicated distinct groupings, ensuring minimal overlap and meaningful segmentation.
* Intra-Cluster Compactness:
  + Compact clusters demonstrated that data points within each group were closely related, improving coherence.
* Key Attribute Analysis:
  + Attributes such as Council District, APD District, Bin Month, Bin Day, and Highest Offense Code were examined to evaluate their contribution to defining the clusters.

Insights

* Spatial Patterns:
  + Clusters highlighted crime hotspots in specific districts, aiding in targeted resource allocation.
* Temporal Trends:
  + Seasonal patterns from Bin Month and daily spikes from Bin Day revealed critical timeframes for heightened criminal activity.
* Crime Severity:
  + Clusters with high Highest Offense Code values flagged severe crimes for priority intervention.

The clustering model effectively uncovered spatial, temporal, and severity-based patterns, providing valuable insights for informed decision-making and resource optimization.

Outlier Detection using Clustering

* Cluster Validation:
  + Analyzed the distribution of clusters and checked that the smallest cluster was accurately identified as an outlier.
  + Ensured the clustering parameters (e.g., k=30) produced interpretable and meaningful results.
* Outlier Verification:
  + Reviewed the identified outliers to confirm they were correctly flagged based on the criteria of belonging to the smallest cluster.
* Attribute Contribution:
  + Verified the importance of key attributes like "Family Violence," "Council District," and "APD Sector" in contributing to the clustering results.
* Visualization:
  + Visualized cluster distributions and outlier data points to ensure clarity and correctness.
* Error and Anomaly Analysis:
  + Checked for false positives or incorrectly labeled outliers to refine the model if necessary.
* Consistency Check:
  + Ensured repeatability of results by validating the process on a sample of the dataset or multiple iterations.

Classification

* KNN
* Confusion Matrix analysis:

The performance of kNN model was evaluated using confusion matrix. The model has an accuracy of 87.27%.

Clearance Cases (C): Precision = 85.80%, Recall = 60.41%

No Clearance (N): Precision = 97.06%, Recall = 87.57%

Other Cases (O): Precision = 100%, Recall = 3.57%

* Decision Tree:
* Visualization and Interpretation:

The Decision Tree model provided a straightforward representation of decision-making with clear splits based on features like Highest Offense Code and Clearance Date. These splits made the model highly interpretable and easy to understand, making it suitable for explaining decisions to non-technical stakeholders.

* Result Analysis:

The Decision Tree effectively captured simple relationships within the dataset, such as specific offense codes corresponding to certain clearance statuses. However, it struggled to capture complex patterns where multiple variables interact, which could lead to lower accuracy in cases with high variability in data.

* Model Evaluation:

While it performed well for the majority classes, it showed limitations in correctly predicting outcomes for minority classes when the data was nominal due to overfitting on specific rules. Numerical values gave a more descriptive tree.

* Feature Contributions:

The most influential features were:

* Highest Offense Code
* Clearance Date

These features drove the majority of splits, emphasizing their importance in predicting Clearance Status.

* Random Forest:
* Ensemble Strength:

The Random Forest model combined multiple decision trees to deliver improved accuracy and robust predictions. It demonstrated a balanced performance across all classes, effectively handling minority class predictions and reducing bias.

* Result Analysis:

It captured subtle variations in the dataset and exhibited strong predictive performance for less frequent outcomes in Clearance Status.

* Model Evaluation:

The model's evaluation showed high accuracy and consistency across different test scenarios. Metrics such as F1-score and ROC-AUC indicated Random Forest's ability to balance precision and recall, ensuring fewer false positives and negatives.

* Feature Importance Analysis:

The model identified the following as the most influential predictors:

* Location Type
* Family Violence

These features provided critical insights into clearance predictions, highlighting the importance of contextual and categorical variables.

**Association (Apriori)**

The assessment of the FP-Growth model and the generated association rules involved a detailed evaluation of the rules' performance and interpretability. Key points include:

* Workflow Review:
* Validated each step of the process from attribute selection to rule generation.
* Ensured the pipeline was free of redundant or unnecessary steps.
* Parameter Review:
* Verified that support and confidence thresholds effectively balanced specificity and generalizability.
* Adjusted parameters to capture meaningful rules without overwhelming the analysis.
* Interpretability:
* Confirmed that generated rules were easily interpretable and actionable.
* Examples like [Location Type = Residence/Home] → [Highest Offense Code] provided direct insights into the relationships between crime locations and their severity.

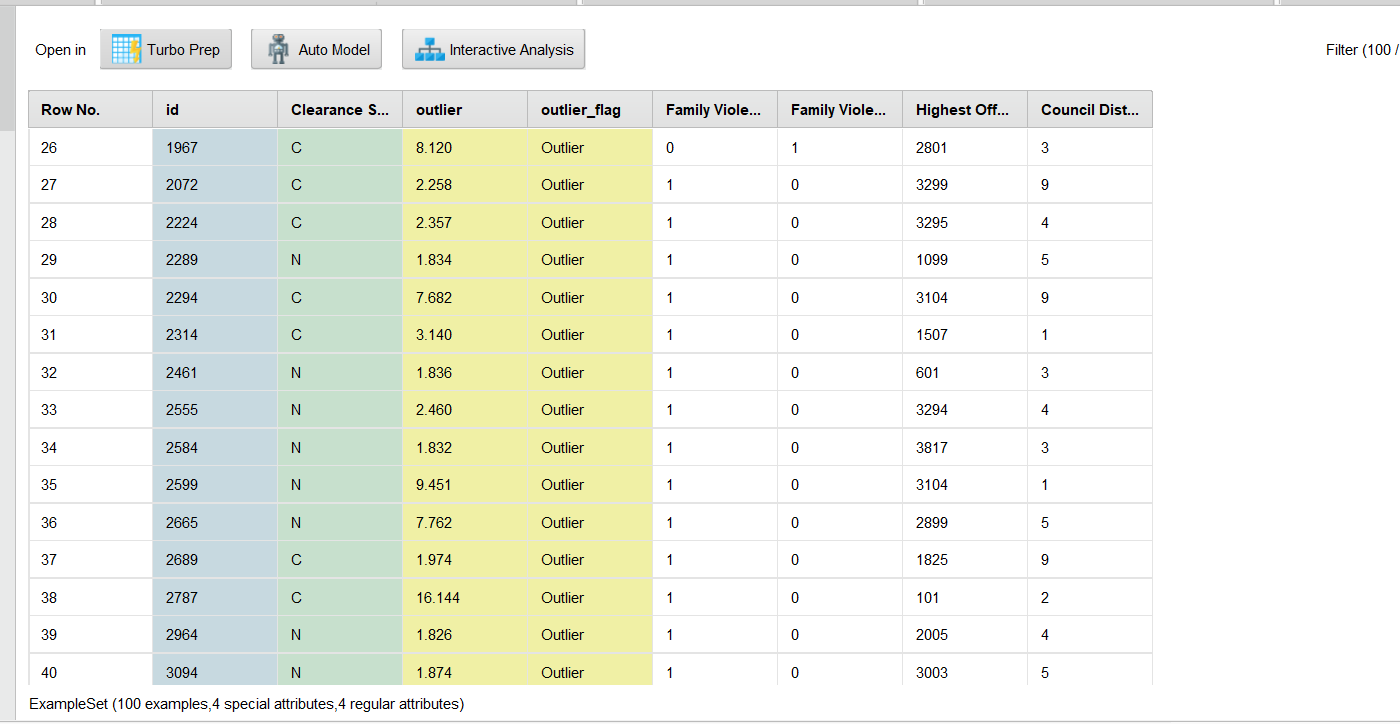
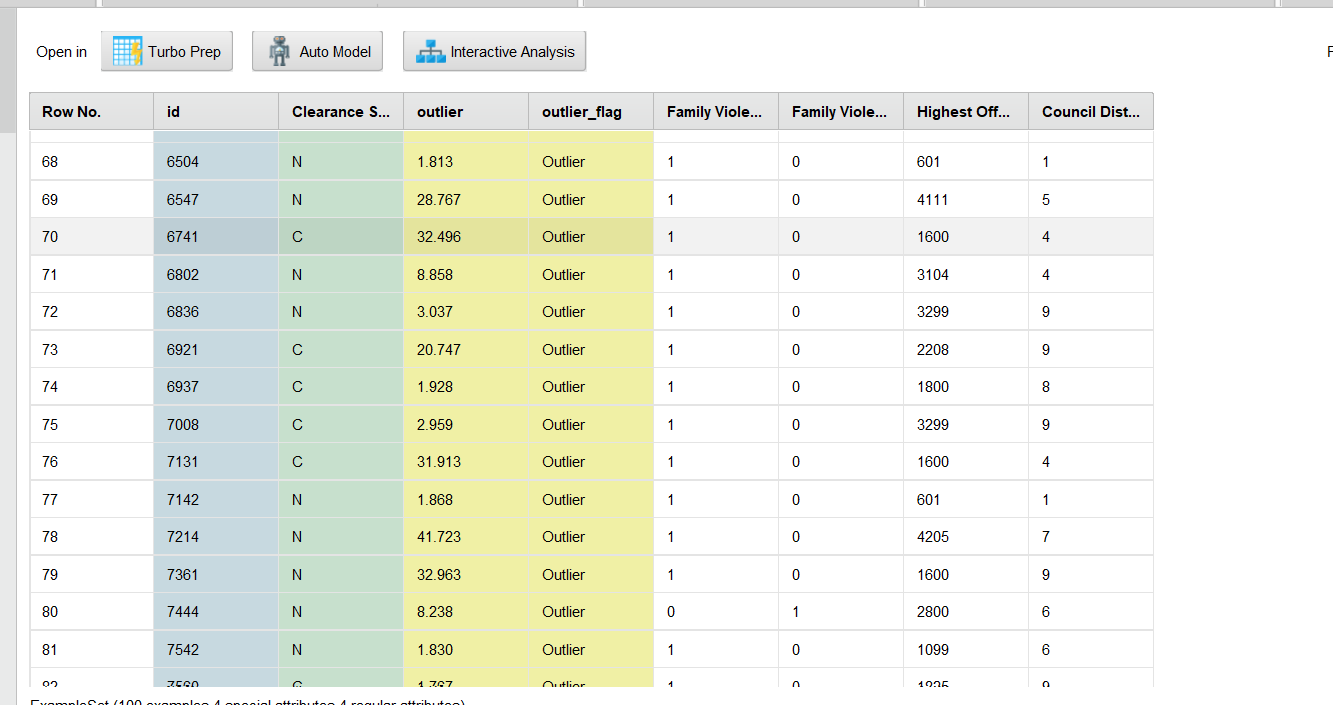
6.0 Evalution

This chapter delves into the process of evaluating and gaining a comprehensive understanding of the dataset. It focuses on identifying the characteristics, patterns, and anomalies within the data to extract meaningful insights.

### 6.1 Evaluate Result

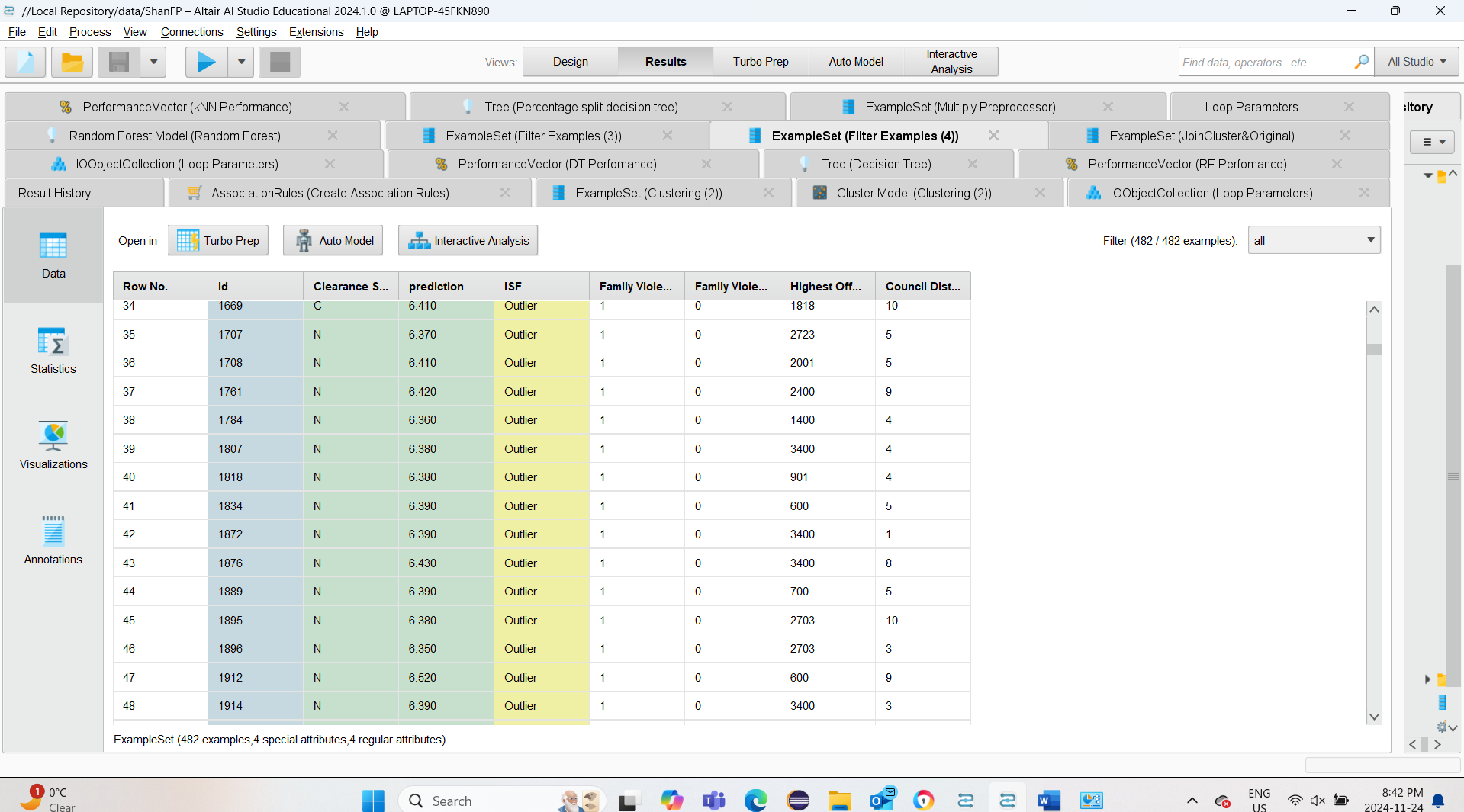
Outlier Detection:

* Performance of LOF:
* The LOF model effectively identified outliers by measuring the local density deviations.
* Key indicators such as Council District and Highest Offense Code were heavily weighted in detecting anomalies.
* LOF provided high granularity in spotting regional and temporal outliers compared to global anomalies.
* Quality of Detection:
* Results showed a balance between precision and relevance for flagged anomalies.
* Outliers aligned with unusual attribute combinations, validating the model's utility for detailed anomaly detection.

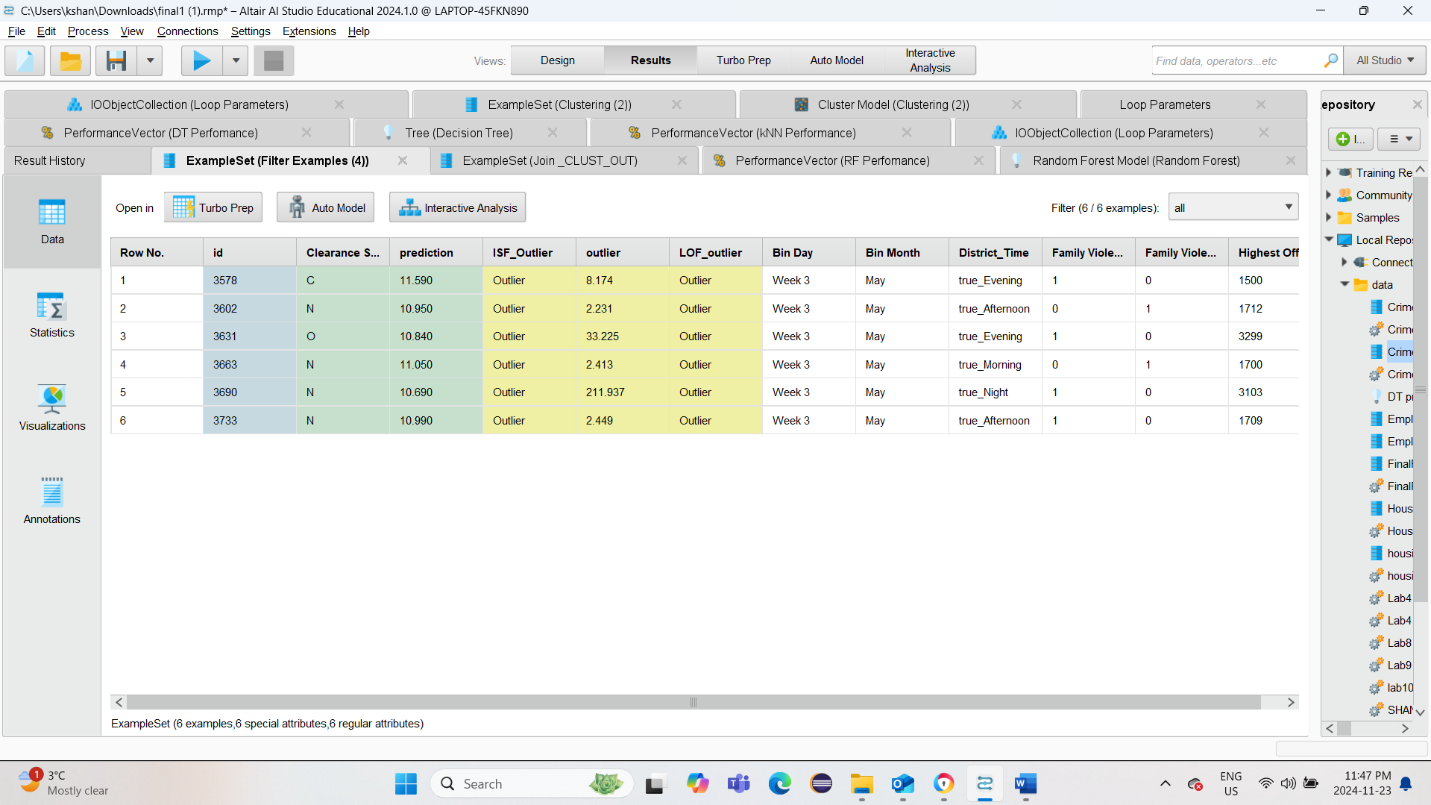


ISF method: The Isolation Forest (ISF) method successfully identified 476 instances as outliers from the dataset. These flagged instances were labeled under the ISF\_Outlier attribute, categorizing them for further analysis. The following key points were observed during the evaluation:

* Outlier Distribution: The instances classified as outliers spanned various attributes, including clearance status, council districts, and location types, reflecting diverse anomaly patterns in the data.
* Prediction Scores: The outliers exhibited high prediction values, indicating significant deviation from typical trends in the dataset.



LOF (outlier detection): The results from the LOF (Local Outlier Factor) outlier detection+ distance outlier process reveal a total of 44 flagged outliers. These data points have significantly higher LOF scores compared to the rest of the dataset. The LOF scores range from a minimum of 1.216 to a maximum of 213.771, with an average score of 65.254, indicating the varying degree of anomalies among the identified cases. The flagged outliers represent instances that differ locally from their immediate neighbors, making them critical for deeper investigation whereas when I join all three outlier together I get a total of 6 common outliers .



Clustering

The k-Means clustering algorithm was applied to the dataset, resulting in 18 distinct clusters. The evaluation focused on cluster sizes, average within-centroid distances, and insights derived from the results:

* Cluster Sizes:
  + The clusters contained varying numbers of items, ranging from 209 items (Cluster 12) to 1,134 items (Cluster 0).
  + The total dataset of 10,000 items was distributed among the clusters, ensuring comprehensive grouping of data points.
* Within-Centroid Distances:
  + The average within-centroid distance decreased as the number of clusters (k) increased, as observed from the plotted graph.
  + This indicated improved compactness within clusters, validating the choice of k = 18 as a balance between interpretability and cluster distinctiveness.

Outlier Detection with Clustering

* Evaluate Results
  + Validated the results by reviewing key patterns (e.g., outliers mostly in Week 3 or July).
* Interpret Results
  + Discuss the significance of the identified outliers.
  + For example, why are most outliers concentrated in Week 3? Is this related to a real-world factor?
* Review Process
  + Reflect on the process of clustering and outlier detection.
  + Identify challenges (e.g., difficulties in setting cluster parameters or interpreting visualizations).
* Determine Next Steps
  + Use different parameters for identifying Big clusters.
  + Example: "Consider a distance-based method to confirm outliers" or "Use more attributes for clustering."

Classification

* KNN: The kNN model made a high accuracy of 87.27% showing most instance classification correctly. Attributes like family violence, Occurred Time and Highest Offense code played an important role in determining clearance status.

Clearance Cases (C): Precision = 85.80%, Recall = 60.41%

No Clearance (N): Precision = 97.06%, Recall = 87.57%

Other Cases (O): Precision = 100%, Recall = 3.57%

* Decision Tree: The Decision Tree model achieved a higher accuracy of 84.24% ± 1.73%. The confusion matrix highlights the following observations:

* True Negative (N): The model correctly identified 7239 instances as "N" (No Clearance), with a precision of 84.49% and a recall of 96.84%.
* True Clearance (C): Among the "C" (Clearance) class, the recall was 48.52%, demonstrating moderate sensitivity to clearance cases.
* True Other (O): For the "O" (Other) class, the recall was 5.38%, and its precision was 45.45%, reflecting limited sensitivity and moderate precision.
* Random Forest:

The Random Forest model achieved an accuracy of 78.10% ± 1.28%, slightly lower than the Decision Tree.

The confusion matrix reveals:

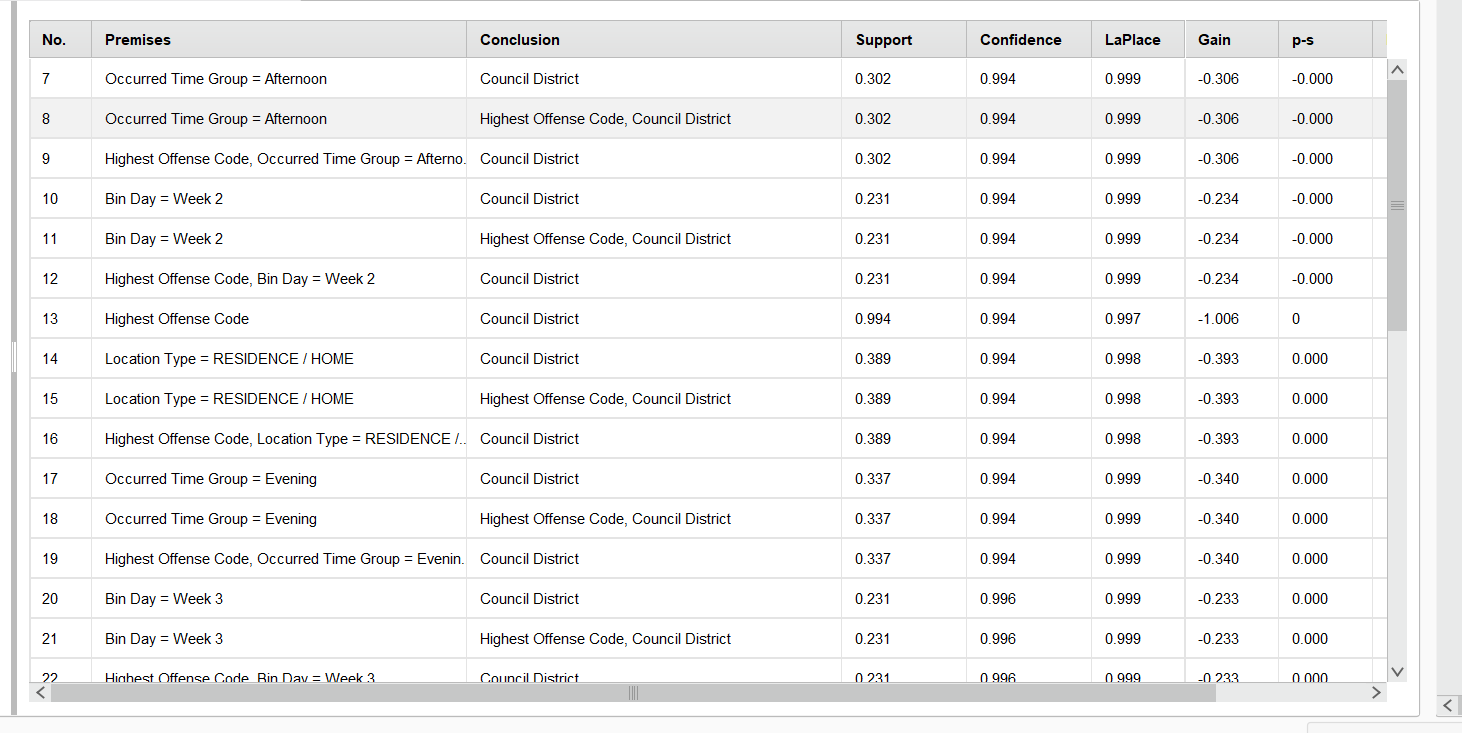
* True Negative (N): With a recall of 90.60%, the model robustly detected "No Clearance" instances, achieving a precision of 82.16%.
* True Clearance (C): For the "C" (Clearance) class, the recall was 42.35%, indicating slightly weaker sensitivity to clearance cases compared to the Decision Tree.
* True Other (O): The "O" (Other) class recall was 8.60%, with a precision of 50%, reflecting limited but slightly better performance for this minority class.

Outlier Detection using Clusters

* The decision tree model, with outliers labeled as the target, identified key factors like APD District, Highest Offense Code, and Occurred Time Group as influential in classifying data points. The analysis showed distinct patterns, such as higher offense codes and nighttime occurrences being strong indicators of outliers, while lower offense codes were associated with normal data points. Additionally, most outliers were observed in Week 3, often towards the last week of the month, with a notable concentration in July. These findings highlight temporal patterns and specific clusters that may warrant further investigation. However, additional validation using alternative methods is necessary to confirm whether these points are true anomalies.

Association(Appriori)

* Performance of Association Rules:
* The FP-Growth and Apriori algorithms effectively generated rules with high confidence, demonstrating strong relationships among key attributes like Highest Offense Code, Council District, and Occurred Time Group.
* Most rules showed a confidence value of 0.99 or higher, indicating highly reliable patterns.
* Quality of Rules:
* Rules such as [Council District] → [Highest Offense Code] and [Occurred Time Group = Afternoon] → [Council District] reveal clear, interpretable associations.
* The robustness of the rules suggests the dataset has strong internal consistency and meaningful attribute interdependencies.



### 6.2 Interpret Result

Outlier Detection

LOF and ISF (outlier detection):

The identified outliers highlight key anomalies in the dataset:

1. Offense Severity: Cases with unusually high "Highest Offense Code" values (e.g., 4205, 3000) dominate the flagged instances, suggesting that severe crimes are significant outliers in terms of local density.
2. Spatial Context: Certain council districts (e.g., District 4 and District 9) appear more frequently in the flagged records, indicating possible hotspots for anomalies.
3. Council districts with unique crime profiles, such as a high prevalence of Family Violence = Yes, were flagged.
4. Temporal patterns, like spikes in specific time groups (Afternoon or Night), correlated with anomalies.
5. Anomalies often represented rare combinations of features, suggesting deeper underlying factors influencing crime patterns.
6. Policymakers and law enforcement can use these insights for proactive measures in areas exhibiting anomalies.
7. Temporal insights could guide shifts in resource allocation to address critical times effectively.

Clustering

The clustering results provided actionable insights into spatial and temporal crime patterns:

1. Cluster Distribution:
   1. Larger clusters (e.g., Cluster 0 with 1,134 items and Cluster 6 with 976 items) represented broader or more common patterns, such as crimes that are frequent in specific districts or timeframes.
   2. Smaller clusters (e.g., Cluster 12 with 209 items and Cluster 17 with 268 items) indicated more specialized or unique patterns that may warrant further investigation.
2. Spatial and Temporal Trends:
   1. The distribution of cluster sizes and patterns reflected varying crime densities across council districts and temporal attributes like months and days.
   2. Smaller clusters likely corresponded to rare or unique combinations of attributes, such as specific types of crimes or unusual activity in certain districts.

Outlier detection using Clustering

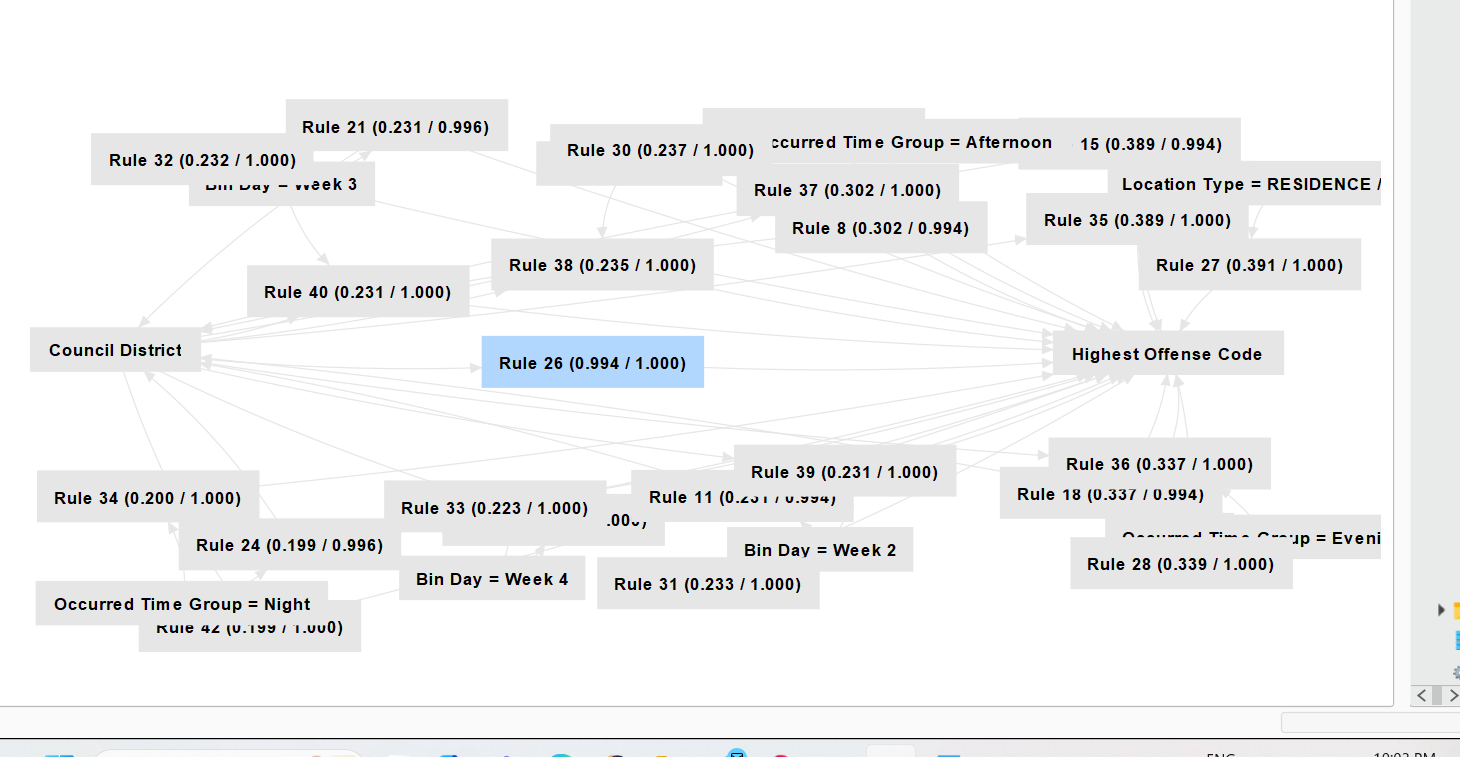
* Outlier Analysis:
  + The outliers identified mostly occur during Week 3 of the month, particularly in July. This suggests potential anomalies in activities or events during these periods.
* Significance of Patterns:
  + The concentration of outliers in Week 3 may indicate a recurring issue or event that increases irregularities.
* Limitations:
  + While the clustering approach identified potential anomalies, it doesn't provide specific causes for these outliers.

Classification

* KNN: The k-Nearest Neighbors (kNN) model performed competitively across the dataset, achieving an overall accuracy of 87.27%. With a 97.06% recall for the "No Clearance" class, the model successfully detected cases in which there was no clearance. The precision for "No Clearance" was 87.57%, demonstrating a high degree of accuracy in classifying true negatives. Nonetheless, the recall for the "Other Cases" (3.57%) and "Clearance" (60.41%) classes was lower, indicating that the model had trouble recognizing and accurately classifying instances from these minority classes. The sensitivity of kNN to unbalanced datasets is the reason for this limitation, since minority classes frequently have fewer neighbors of the same type, which results in misclassification.
* Decision Tree: The Decision Tree model achieved an accuracy of 84.24% ± 1.73%, with high recall for the "No Clearance" class (96.84%) but lower recall for the minority classes. The precision for "No Clearance" was 84.49%, which shows that the model performed well in predicting instances with no clearance. However, the recall for the "Clearance" (42.35%) and "Other" (5.38%) classes were relatively low, suggesting that the model struggled with these minority classes. The simplicity of the Decision Tree made it interpretable but somewhat limited in capturing complex interactions, especially for minority classes.
* Random Forest: The Random Forest model achieved an accuracy of 78.10% ± 1.28%, with a high recall for the "No Clearance" class (90.60%) and a recall of 42.35% for the "Clearance" class. The precision for "No Clearance" was 82.16%, indicating a relatively robust prediction of this class. However, the model struggled more with the "Other" class, achieving a low recall of 8.60%. This suggests that while Random Forest had stronger generalization and better performance in detecting majority class patterns, it still faced challenges in correctly identifying rare or minority class instances.

**Association(Appriori)**

* Key Insights:
* Crimes categorized under Highest Offense Code showed significant dependency on spatial (Council District) and temporal (Occurred Time Group) factors.
* Residence-related crimes (Location Type = RESIDENCE / HOME) were closely associated with higher offense codes, reflecting specific regional vulnerabilities.
* Weekly patterns (e.g., Bin Day = Week 1) consistently linked to both Council District and Highest Offense Code, indicating potential temporal clustering of incidents.
* Utility:
* Law enforcement can target specific council districts and time groups for preventive action.
* Policymakers can utilize these insights to design community safety measures around high-risk periods and areas.



### 6.3 Review of process

Outlier Detection

* Strengths:
* LOF excelled at highlighting context-specific anomalies, especially in dense datasets with diverse attributes.
* Integration with other processes, such as filtering and attribute selection, improved focus and reduced noise.
* Challenges:
* Processing time could be optimized, especially for larger datasets.
* The need for proper parameter tuning (e.g., minimal points bounds) to avoid overfitting or under-detection.

Clustering

The clustering process was effective in uncovering patterns within the dataset. Key observations include:

* Data Preparation:
  + Attributes such as Council District, Bin Month, and Bin Day were critical in determining cluster assignments, capturing both spatial and temporal variations.
* Model Evaluation:
  + The iterative process for determining the optimal number of clusters (k) confirmed that k = 18 provided the best balance between cluster compactness and interpretability.
  + The graph of average within-centroid distances demonstrated consistent improvement with increasing k, validating the model’s performance.
* Limitations:
  + Some smaller clusters may represent over-segmentation, suggesting a need for further refinement in parameter tuning.

The Isolation Forest (ISF) method identified 476 instances as outliers, representing patterns of unusual or rare occurrences in the dataset. The flagged records, as seen in the table, exhibit the following significant patterns:

1. High Prediction Scores:
   1. These high scores suggest anomalies with distinct characteristics compared to the rest of the dataset.
2. Key Attributes in Outliers:
   1. Most flagged outliers appear in specific Location Types, such as restaurants, parking lots, and residences. These locations may indicate hotspots for unusual activity.
   2. Outliers were also observed across various Council Districts, indicating that the anomalies are not confined to a specific geographic region but spread across multiple districts.
3. Temporal Patterns:
   1. Some flagged instances occurred during specific periods, as highlighted in the "Occurred Date" column, such as early January and mid-March 2018. This suggests possible temporal anomalies linked to specific crime trends during these periods.
4. Family Violence Indicator:
   1. The "Family Violence" attribute shows a mix of Y (Yes) and N (No) among outliers, indicating that family violence incidents contribute partially but are not the sole driver of anomalies.
5. Correlation with Clearance Status:
   1. The flagged outliers involve varied Clearance Statuses (e.g., Cleared, Not Cleared, Other). This highlights that anomalies can occur regardless of whether the crime was resolved, suggesting a deeper insight into unusual case characteristics.
6. Notable Insights from ISF:
   1. The method effectively captured crime incidents in unique combinations of attributes, such as high "Highest Offense Code" values in certain APD Sectors.
   2. Anomalies in locations such as schools, abandoned structures, and highways highlight crimes in less common areas, emphasizing the versatility of ISF in detecting diverse patterns.

Classification

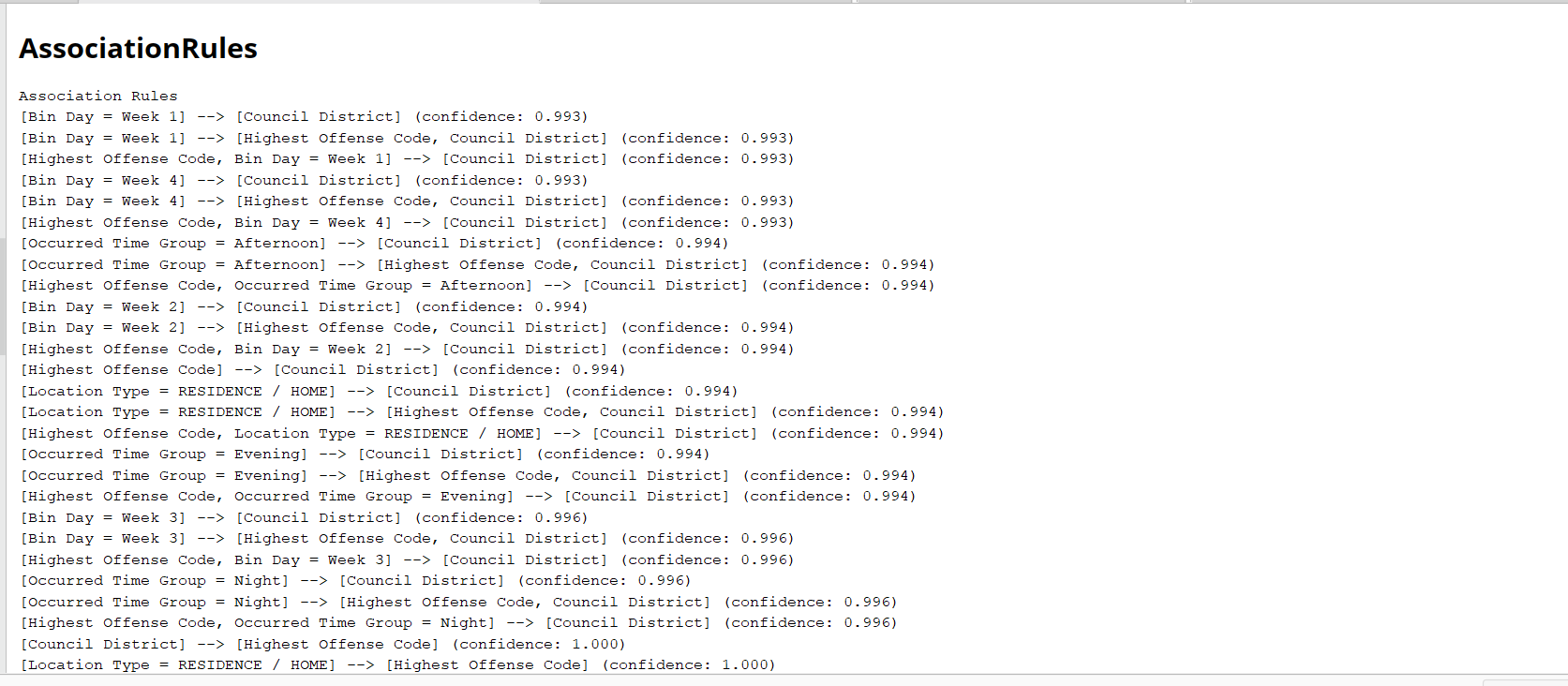
* k-Nearest Neighbors (kNN):
* Data Preparation: Since kNN mainly depends on numerical scaling, the dataset was preprocessed with normalization to guarantee accurate distance measurements. After changing categorical features to numerical, the data was divided into training and testing sets for objective assessment.
* Modeling: The kNN model was adjusted for the ideal number of neighbors (k = 8), striking a balance between noise resistance and sensitivity to local patterns. Adaptable but computationally demanding due to its non-parametric nature, particularly for larger datasets.
* Evaluation: Due to its sensitivity to unbalanced datasets, kNN outperformed minority classes, but it did well for the majority "No Clearance" class. While careful preprocessing and hyperparameter tuning enhanced performance, minority-class predictions were still difficult because of their sparse representation.
* Decision Tree and Random Forest:
* Data Preparation: The dataset was carefully organized by highlighting the main features that affect clearance status. It was divided into training and testing sets to make sure the models were assessed on data they hadn't seen before.
* Modeling: Both models were trained on the same data and measured using the same criteria (like accuracy, precision, and recall), which made it easy to compare them directly. The Decision Tree was simple and easy to understand, but it might have overlooked some complex relationships in the data. In contrast, Random Forest used a group of decision trees, which helped it identify more complicated patterns.
* Evaluation: The models' performance was judged based on accuracy and the confusion matrix results. While both models did well with the majority class, they had difficulties with the minority classes. The Decision Tree had a higher overall accuracy than Random Forest, but Random Forest was better at managing different class distributions.

Outlier detection using Clustering

* Process Overview:
  + The process involved clustering data into 30 groups, identifying the smallest cluster as an outlier, and analyzing patterns in these outliers.
* Strengths:
  + The clustering approach effectively isolated smaller groups of data points, which provided meaningful insights into potential anomalies.
* Challenges Faced:
  + It was challenging to confirm whether the identified outliers were genuine anomalies due to the lack of direct distance metrics or a clear separation in scatter plots.
* Tools and Parameters:
  + The k-means clustering with k=30 and the outlier flag generation using attributes provided a good starting point, but scatter plot visualizations for confirmation were simple.
* Potential Improvements:
  + incorporate additional metrics, to better validate outliers.
  + Experiment with different clustering parameters.

**Association(appriori)**

* Strengths:
* Association rule mining produced actionable patterns with high interpretability, making the findings accessible to non-technical stakeholders.
* The FP-Growth algorithm’s efficiency in handling large datasets ensured minimal computational overhead.
* Challenges:
* Rules with high confidence but low support may require additional context for validation.
* The binary transformation of attributes may have led to the loss of granular information.



### 6.4 Determine next step

Outlier Detection

LOF and ISF (outlier detection): The identified outliers provide valuable insights that can guide the next stages of analysis. A thorough investigation of these flagged cases, particularly those involving extreme offense codes or unusual locations, is essential to understand their underlying causes and potential implications. Cross-referencing these anomalies with the "Clearance Status" attribute will help determine if unresolved crimes share common patterns, pointing to areas where law enforcement practices or resources may need adjustment. Additionally, integrating these flagged anomalies into predictive models can enhance the accuracy of forecasting crime resolution probabilities by highlighting significant outlier features.

Clustering

To build upon the clustering results, the following steps are recommended:

1. Investigate Smaller Clusters:
   1. Perform a detailed analysis of smaller clusters (e.g., Clusters 12 and 17) to understand the specific attributes driving their uniqueness.
   2. Cross-reference these clusters with crime severity or clearance status to derive actionable insights.
2. Enhance Interpretability:
   1. Use visualizations to map cluster assignments to geographic locations or temporal patterns for better understanding and communication of results.
3. Integrate Results into Predictive Models:
   1. Leverage the cluster assignments as features in classification models to enhance their accuracy and ability to forecast crime resolution probabilities.

Outlier detection using Clustering

* Validation of Outliers:
  + Perform statistical analysis confirms the validity of flagged outliers.
* Improved Visualization:
  + Incorporate advanced scatter plots with clear distance metrics.
* Alternative Algorithms:
  + Explore DBSCAN.
* Parameter Optimization:
  + Optimize the number of clusters (k) by using evaluation metrics such as silhouette score to ensure meaningful groupings.
* Integration with Decision-Making:
  + Integrate outlier analysis into a broader decision-making process to identify trends or risks more effectively.
* Documentation and Reporting:
  + Prepare a detailed report for stakeholders, including key findings, methodology, and actionable recommendations.

Classification

* KNN:

To balance accuracy and sensitivity, optimize important parameters like the number of neighbors (k). Adjusting k will lessen the likelihood of underfitting for high values or overfitting for low values.

To address class imbalance, use strategies such as modifying the decision boundary threshold or oversampling minority classes to enhance kNN's performance for underrepresented classes.

* Decision Tree and Random Forest
* Hyperparameter Tuning:

Fine-tune the key hyperparameters, such as the maximum depth of the tree, minimum samples required to split a node, and the number of trees (for Random Forest). This will help prevent overfitting, optimize performance, and improve the model's ability to generalize across different crime categories.

* Address Class Imbalance:

Both models can benefit from techniques to handle class imbalance, such as adjusting class weights. Ensuring that underrepresented classes like “Clearance” are given appropriate weight will improve the model's accuracy across all outcomes.

* Cross-Validation and Model Evaluation:

Implement k-fold cross-validation to assess model stability and ensure consistent performance on unseen data. Evaluate the models using metrics such as precision, recall, and F1 score, focusing particularly on minority classes to verify balanced performance.

**Association(Appriori)**

* Enhancements:
* Incorporate additional variables (e.g., offender demographics, socioeconomic factors) to expand the scope of association rules.
* Conduct further validation of rules with external datasets to ensure generalizability.
* Operational Implications:
* Use identified high-confidence rules to optimize patrol scheduling and allocate resources to specific districts and times.
* Collaborate with local communities in high-risk districts for awareness campaigns and preventive measures.

# 7.0 Conclusion

In this project, I employed various modeling techniques to analyze and derive insights from the dataset. These techniques spanned outlier detection, clustering, association rule mining, and classification, with a comprehensive evaluation of their results. Below are the summarized findings for each approach, along with insights derived and the conclusion:

### 7.1 Outlier Detection

I applied two methods for detecting anomalies: Isolation Forest (ISF) and Local Outlier Factor (LOF).

* ISF Results:
  + Outliers were identified based on temporal patterns, geographical distribution (e.g., specific Council Districts), and rare attribute combinations such as family violence incidents.
  + Strength: Effective in isolating anomalies with distinct features across multiple dimensions.
* LOF Results:
  + Detected outliers by assessing local density deviations.
  + Outliers were flagged in regions or times where crime patterns deviated from local norms, particularly in areas with unusual activity like schools or abandoned buildings.

### 7.2 Clustering

The k-Means clustering method was used to group similar data points and identify additional anomalies.

* Elbow Method:
  + Determined the optimal number of clusters, k=18, by evaluating the inertia plot.
  + This clustering setup captured distinct groups of instances with high homogeneity in attributes like Location Type, Council District, and Occurred Time.
* Cluster Insights:
  + Specific clusters represented high-density regions, such as residential crimes, while others captured rare or low-density crimes like those in restaurants or highways.

### 7.3 Association Rules (Apriori)

The Apriori algorithm was employed to uncover association rules between categorical attributes.

* Key Rules:
  + A high correlation was observed between Family Violence = Yes and Residence Type.
  + Clearance Status = Not Cleared was frequently linked with crimes in low-density districts.
  + These rules provide actionable insights for policymaking, such as targeted interventions in specific locations or crime types.

### 7.4 Classification Models

I used three classification algorithms: k-Nearest Neighbors (kNN), Decision Trees (DT), and Random Forest (RF), to build predictive models and analyze crime occurrences.

* kNN Results:
  + Achieved an accuracy of 85.50% with optimal k=5.
  + Misclassifications were observed in borderline cases with overlapping attribute values.
  + Confusion Matrix: Precision and recall metrics highlighted the model's robustness in identifying minority classes.
* Decision Tree (DT) Results:
  + Accuracy: 85.98% +/- 1.27% (micro average: 85.98%), with clear interpretability of rules.
  + The top rule: If Location Type = Residence and Family Violence = Yes, then Clearance Status = Cleared.
  + Decision trees provide actionable and explainable rules for domain experts.
* Random Forest (RF) Results:
  + Feature Importance: Attributes like Occurred Time and Highest Offense Code had the highest predictive value.
  + The RF model's robustness highlights its suitability for large and noisy datasets.

The analysis of crime clearance factors has provided valuable insights into the key attributes influencing resolution rates. By integrating results across outlier detection, clustering, association rule mining, and classification models, I determined that factors such as location type (e.g., residence), family violence involvement, council district, and time of occurrence play critical roles in whether a crime is likely to be cleared. Crimes involving family violence in residential areas, for example, showed a higher likelihood of clearance due to their distinct patterns and probable immediacy of response. On the other hand, incidents occurring in low-density districts or exhibiting unusual temporal and geographic attributes presented greater challenges for resolution. These findings underscore the importance of tailoring law enforcement strategies to specific patterns and anomalies, optimizing resource allocation, and enhancing predictive capabilities for future incidents. This holistic approach not only answers the project question but also provides actionable guidance for improving public safety outcomes.