|  |
| --- |
| **ePROJECT DATA SCIENCE**  **Project Title: Analysis of the relationship between crime type and location in Victoria, Australia (2024 Data)** |

### **Developed by**

**Member:**

|  |  |  |
| --- | --- | --- |
| **No.** | **Student Name** | **Student ID** |
| **1** | **Nguyễn Đức Trường Giang** | **Student1612595** |
| **2** | **Trần Ngọc Hiển** | **Student1527792** |
| **3** | **Trần Công Khôi** | **Student1604958** |

**Class No.: DT2407L**

### **I. Project Proposal & Planning**

#### **1. Problem Definition**

**Problem:** Crime is a complex social problem, causing great loss of life and property, affecting public order and people's lives. Law enforcement agencies and local authorities need to have a deeper insight into crime patterns to **allocate resources effectively** , develop **prevention strategies and respond promptly** . Understanding the relationship between crime types and locations is key to addressing this problem in a targeted and effective manner.

**Project Objectives:**

* **Main objective: To explore "interesting" and "non-obvious"** relationships between crime types (Offence Division, Offence Subdivision) and the types of locations where they occur (Location Division, Location Subdivision) based on recorded crime data.
* **Secondary objectives:**
  + Provide **actionable insights** to law enforcement agencies to optimize patrolling, surveillance and security resource allocation.
  + Assist in the development of **targeted crime prevention policies** and community awareness raising in specific areas.
  + Identify **characteristic** crime patterns in different types of locations, thereby supporting urban planning and regional security.

#### **2. Dataset Description**

**Data Source:** Data is collected from **the Crime Statistics Agency of Victoria (Australia) public crime statistics dataset** , specifically "Table 02: Offences recorded by offence type and location type - Jan 2024 to Dec 2024".

**Data characteristics:**

* **Time Range: Data covers events from January 2024 to December 2024, providing a one-year** overview of crime.
* **Main fields (Variables):**
  + Offence Division : The highest level of crime classification (e.g. Crimes against persons, Crimes against property).
  + Offence Subdivision : A more detailed crime group (e.g. Theft, Fraud).
  + Offence Subgroup : The most specific crime segment.
  + Location Division : Top level location division (e.g. Commercial, Residential).
  + Location Subdivision : More detailed location area (e.g. Retail Store, Private Home).
  + Location Subgroup : The most specific location subgroup.
  + Offence Count : The number of reported crimes recorded for each combination.
* **Data volume:**
  + File size: **139 KB** (Excel file).
  + Data size: **4,184 rows x 9 columns** .
* **Data Type:** The data is mainly **categorical** for both Offence Type and Location Type . The Offence Count column is integer data, representing frequency.

#### **3. Methodology Overview**

**Overall Approach:** This project will use **Association Rule Mining** as the main focus to discover hidden relationships between crime types and locations. In addition, to meet the project requirements, Classification, Regression and Clustering techniques will also be applied and evaluated, although they may not be the most optimal methods given the data structure and main objectives.

**Main technique:**

* **Data Preprocessing:** Focuses on cleaning, normalizing, and converting raw data into a **transaction format** suitable for association rule mining algorithms.
* **Association Rule Mining: Apply Apriori (or FP-Growth)** algorithm to find frequent itemsets and infer association rules.
* **Analysis and Interpretation:** Evaluate rules based on metrics like **support, confidence, lift** and draw valuable insights into the relationship between Offence and Location .
* **Complementary techniques: Implement classification, regression and clustering** algorithms to illustrate the diverse applicability of Data Mining and evaluate their effectiveness on given data types.

### **II. Data Wrangling (25 points)**

#### **2.1 Data Collection**

**Data Sources:**

* Data was collected directly from the **Crime Statistics Agency of Victoria - Australia portal** through publicly available statistical files.
* Specifically, the project uses data from **"Table 02: Recorded Offences Visualisation Year Ending December 2024"** , ensuring accuracy and clear origin.

**Data Quality Assessment:**

* **Completeness Check:** Ensure there are no missing (null) values in key classification columns such as Offence Division , Offence Subdivision , Location Division , Location Subdivision . This is necessary to avoid distorting relationships during analysis.
* **Consistency Check:** Evaluate name variations for the same crime or location (e.g., "Retail Store" and "Retail outlet" have the same meaning). If so, standardize.
* **Outlier Check on Offence Count :** While the main columns are categorical, the Offence Count column is numeric. Extreme values (e.g., unusually high or low case counts) will be checked to ensure there are no data entry errors or special events that skew the overall analysis.

#### **2.2 Data Integration & Preprocessing**

**Data Integration:**

* In this project, the data is fetched from a single master table ( Table 02 ). Therefore, the data integration step mainly involves ensuring **that the data is properly loaded from the source (Excel file) into the working environment (Python Pandas DataFrame)** in an efficient manner.
* To accommodate association rule mining, we will perform a data “expansion” step. Specifically, if a data row says there are N Offence A incidents at Location X , we will generate N individual records, each representing a specific crime event with its Offence Division and Location Subdivision . For example, if there are 100 “Drug Crime” incidents in “Residential,” we will generate 100 {Offence: Drug, Location: Residential} records . This is done efficiently using the **Pandas library** .

**Missing Value Handling:**

* **Strategy:** For important classification fields like Offence Division , Offence Subdivision , Location Division , Location Subdivision (used for association rules), if any value is missing, we will **discard those records** .
* **Reason:** Assumingly missing identifier data can distort the relationships found by the association rule algorithm. The omission ensures the integrity of the "items" in transactions.
* **Implementation:** Use df.dropna(subset=['Offence Division', 'Offence Subdivision', 'Location Division', 'Location Subdivision']) .

**Outlier Detection & Treatment:**

* For nominal data ( Offence Type and Location Type ), the concept of outlier does not apply in the traditional statistical sense.
* However, there may be extremely rare crime types or extremely rare location types. These can be considered "Outliers" in the context of frequency.
* **Processing:** Association rule mining algorithms like **Apriori automatically remove items with frequencies lower than a given min\_support threshold** . This helps to automatically remove extremely rare cases, helping to focus on more common and meaningful relationships.

**Data Transformation:**

* **Convert to Transaction format:** Each row in the expanded DataFrame is converted to a list or set containing "items" as **(Offence Division - Location Division)** and **(Offence Subdivision - Location Subdivision) pairs** . For example: ['O:Drug', 'L:Residential'] . Combining both crime and location hierarchies allows for more granular rules.
* **One-Hot Encode (for Association Rules):** To accommodate the mlxtend library's Apriori algorithm , transactions are converted to a boolean matrix format (one-hot encoded) where each column is an item (e.g. O:Drug , L:Residential ) and a value of 1/0 indicates the presence of that item in a transaction. Use the TransactionEncoder in the mlxtend.preprocessing library .

#### **2.3 Exploratory Data Analysis**

**Statistical Analysis:**

* **Descriptive statistics:**
  + Calculate the frequency of occurrence of each Offence Division , Offence Subdivision , Location Division and Location Subdivision .
  + Identify the most common types of crime and the types of locations with the highest crime rates.
  + Calculate simple "ratios": % of each type of crime to total, % of each type of location.
* **Qualitative Analysis: Observe** the frequency distribution of the pairs **(Offence Division - Location Division)** and **(Offence Subdivision - Location Subdivision)** to generate initial hypotheses about possible relationships before applying the algorithm.

**Data Visualization:**

* **Bar Charts:** Displays the frequency of top crime types (Top N Offence Division / Subdivision ) and top location types (Top N Location Division / Subdivision ).
* **Stacked/Grouped Bar Charts:** Illustrate the distribution of Offence Subdivisions within each Location Subdivision or vice versa. For example, the chart shows which types of crimes are most common in "Residential Areas" vs. "Public Places", providing a visual representation of the differences.
* **Heatmap:** Create a co-occurrence matrix between Offence Divisions and Location Divisions (or Subdivisions ). The heatmap visualizes the strength of the relationship, making it easy to identify common pairs.

### **III. Data Analysis & Modeling (35 points)**

#### **3.1 Classification Tasks**

* **Example Objective:** Predict whether a crime is likely to be **a "Drug Offence" (Offence Division: Drug Offences)** based on the Location Subdivision where it occurred. This is a binary classification problem.
* **Prepare data:**
  + Create a binary target column: is\_Drug (1 if Offence Division is 'Drug Offences', 0 otherwise).
  + Encode Location Subdivision using **One-Hot Encoding** to convert categorical values into numerical features suitable for the algorithm.
* **Algorithm 1: Logistic Regression**
  + **Description:** Logistic Regression is a linear classification algorithm that estimates the probability that an instance belongs to a given class. Despite the name "Regression", it is a powerful and easy-to-interpret classification model.
  + **Implementation:** Use sklearn.linear\_model.LogisticRegression .
* **Algorithm 2: Decision Trees**
  + **Description:** Decision Tree is a supervised learning model that generates a decision tree based on rules inferred from data. It is capable of handling both numerical and categorical data, and provides understandable rules.
  + **Implementation:** Using sklearn.tree.DecisionTreeClassifier .

#### **3.2 Regression Analysis**

* **Example goal:** Using additional data from "Table 01: Offences recorded and rate per 100,000 population" (if possible to integrate, otherwise use Offence Count in Table 02 ). The goal is to predict **Rate per 100,000 population** or Offence Count based on features related to Offence Division or Location Division (after encoding) or other aggregate features.
* **Prepare data:**
  + Use the Offence Count column from Table 02 (or Rate per 100,000 population if integrating Table 01 ).
  + Encode categorical variables ( Offence Division , Location Division ) using One-Hot Encoding to use as independent variables.
* **Technique: Simple/Multiple Linear Regression**
  + **Description:** Linear Regression is a supervised learning algorithm used to model the linear relationship between a continuous dependent variable and one or more independent variables.
  + **Implementation:** Use sklearn.linear\_model.LinearRegression .

#### **3.3 Clustering Analysis**

* **Example goal:** Group Location Subdivisions based on their "crime profile" (i.e. the distribution of types of crimes that occur there). For example, areas with similar crime profiles would be grouped into the same cluster.
* **Prepare data:**
  + Create a **frequency matrix** where the rows are Location Subdivisions and the columns are Offence Divisions (or Offence Subdivisions ), with the values being the number/rate of each type of crime in that location. Standardizing this data (e.g. with StandardScaler ) is necessary before clustering.
* **Technique: K-Means (Partition-based)**
  + **Description:** K-Means is a popular unsupervised clustering algorithm. It divides N data points into K clusters, where each data point belongs to the cluster with the closest center.
  + **Implementation:** Use sklearn.cluster.KMeans . Need to determine the optimal number of clusters K (e.g. use **Elbow Method** or **Silhouette Score** for evaluation).

#### **3.4 Advanced Analytics**

**Frequent Pattern Mining & Association Rules (5 points):** This is the main focus of the project and needs to be thoroughly presented.

* **Goal:** Explore interesting and non-obvious "if-then" association rules of the form:
  1. {Type of crime} → {Type of location}
  2. {Location Type} → {Crime Type}
  3. {Crime Type 1, Location Type} → {Crime Type 2}
  4. or more complex combinations.
* **Algorithm:** **Apriori** (implemented using the mlxtend.frequent\_patterns.apriori and mlxtend.frequent\_patterns.association\_rules libraries ).
* **Detailed implementation steps:**
  1. **Data Transformation:** Convert the preprocessed DataFrame (exploded event table) into a "list of lists" format or One-Hot matrix suitable for mlxtend (using TransactionEncoder ).
  2. **Find Frequent Itemsets: Run the** apriori algorithm with selected min\_support parameters (e.g. experiment with values like 0.01 to 0.05 to find a balance between the number and significance of frequent itemsets).
  3. **Association rule generation:** Generate rules from frequent itemsets using association\_rules with parameters metric (e.g. 'confidence' or 'lift') and min\_threshold (e.g. min\_confidence from 0.6 to 0.8 , min\_lift from 1.2 to 1.5 ).
  4. **Filter and sort:** Filter rules based on specific criteria (e.g. highest lift , highest confidence ) to find the most interesting and valuable rules for your business goals. Avoid trivial rules.
  5. **Explanation:** Explain what each metric ( support , confidence , lift ) means and specifically interpret the rules found (e.g. "If a 'Theft' occurs, it is likely (with X% confidence) that it will be in a 'Residential Area', and this co-occurrence is Y times higher than random.")

### **IV. Model Evaluation & Validation (15 points)**

This section should clearly distinguish the evaluation methods for different types of models.

**1. Evaluation for Classification & Regression Models:**

* **Cross-Validation:**
  + Apply **K-Fold Cross-Validation** (e.g. K=5 or 10) to classification and regression models to evaluate the overall performance and stability of the model on different datasets.
  + **Why:** Using Cross-Validation helps avoid overfitting to the training dataset, providing a more reliable estimate of the model's actual performance on unseen data.
* **Performance Metrics:**
  + **Classification:**
    - **Accuracy:** Overall correct prediction rate.
    - **Precision:** The proportion of correct positive predictions out of the total number of positive predictions.
    - **Recall:** The proportion of true positive cases that are correctly identified.
    - **F1-Score:** Harmonic average of Precision and Recall (especially important if there is class imbalance).
    - **AUC-ROC (Area Under the Receiver Operating Characteristic Curve):** Measures the ability to discriminate between classes.
    - *is\_Drug prediction model* , special attention is paid to Precision , Recall and F1-Score for the positive class ( is\_Drug=1 ).
  + **Regression:**
    - **Mean Squared Error (MSE):** Mean squared of the errors.
    - **Root Mean Squared Error (RMSE):** The square root of the MSE, in the same units as the target variable.
    - **R-squared (R2):** Represents the percentage of variance in the dependent variable that is explained by the independent variables. (A negative R2 value indicates that your model is performing worse than simply predicting the mean.)
* **Model Comparison:**
  + Compare the performance of implemented classification algorithms (Logistic Regression vs. Decision Tree) based on selected metrics (Accuracy, F1-Score, AUC).
  + Discuss the advantages and disadvantages of each algorithm in the context of the problem and the type of data used.
  + Explain why supervised learning models might underperform on this dataset (e.g., lack of predictive features, mostly categorical data, and no clear linear relationships).

**2. Evaluation for Clustering Analysis:**

* **Performance Metrics:**
  + **Silhouette Score:** Measures the cohesion and separation of clusters. Values range from -1 to 1, with values closer to 1 being better.
  + **Explanation:** This index helps to determine whether objects in the same cluster are similar or not and whether objects in different clusters are different or not.

**3. Rating for Frequent Pattern Mining & Association Rules:**

* **Performance Metrics:**
  + **Support:** The proportion of transactions that contain both the left and right sides of the rule. Represents the popularity of the rule.
  + **Confidence:** The proportion of transactions containing the left side that also contain the right side. Represents the confidence of the rule.
  + **Lift:** The ratio between the confidence of the rule and the support of the right side. Lift > 1 indicates a positive relationship, Lift = 1 is independence, Lift < 1 is a negative relationship. Lift is the most important index to determine the "interestingness" of the rule.
  + **Explain what each metric means** for assessing the "quality" of a rule, and why Lift is more important than Support and Confidence in finding non-obvious relationships.

### **V. Results & Business Insights (10 points)**

#### **1. Interpretation**

* Present **the most prominent combination rules (e.g. 5-10 strongest/most interesting rules)** found, focusing on those with high Lift and reasonable Confidence .
* **Analysis of the Top 5 Strongest Laws according to Lift:**
  + **Rule 0: (L:26 Justice) → (O:E Justice procedures offenses)**
    - **Support:** 2.07%
    - **Confidence:** 74.11%
    - **Lift:** 4.96 (Very high!)
    - **Interpretation:** "When an incident occurs at **'Location: Justice'** , there is **a 74.11%** chance that it involves **'Type of Offence: Justice procedures'** . This relationship is **nearly five times stronger** than if the two events occurred by chance." This is the strongest rule and makes clear sense: events at a justice location are more likely to involve justice procedures offences.
  + **Rule 1: (O:E Justice procedures offenses) → (L:11 Dwelling - private)**
    - **Support:** 9.55%
    - **Confidence:** 63.85%
    - **Lift:** 1.79
    - **Interpretation:** "If an incident is **a 'Procedural Offense'** , then there is **a 63.85%** chance that it occurred at **a 'Location: Dwelling - private'** ." This is an interesting relationship, suggesting that procedural violations (perhaps involving arrest warrants, law enforcement at private homes) are often associated with private settings.
  + **Rule 2: (L:13 Grounds/surrounding land) → (O:B Property and deception offenses)**
    - **Support:** 8.46%
    - **Confidence:** 88.20%
    - **Lift:** 1.49
    - **Interpretation:** "When the incident occurred at **'Grounds/surrounding land'** , there is an **88.20%** chance that it was **a 'Property and deception offences'** ." This makes sense, as outdoor areas such as yards and sidewalks are more likely to be the site of petty theft, vandalism, or fraud.
  + **Rule 3: (L:33 Retail) → (O:B Property and deception offenses)**
    - **Support:** 13.74%
    - **Confidence:** 83.82%
    - **Lift:** 1.41
    - **Interpretation:** "If the incident occurred at **'Location: Retail'** , then there is **an 83.82%** chance that it was **'Property Crime and Fraud'** ." This is a very common and business-sounding law, typical of shoplifting and fraud.
  + **Rule 4: (L:25 Other Transport) → (O:B Property and deception offenses)**
    - **Support:** 3.00%
    - **Confidence:** 79.66%
    - **Lift:** 1.34
    - **Interpretation:** "When the incident occurred at **'Location: Other Transport'** , there is **a 79.66%** chance that it was **'Property Crime and Fraud'** ." This also makes sense, given that it involves theft on public transport or from a private vehicle (e.g. theft of vehicle parts, theft from vehicles).
* **General comments on the average indexes of the Filtered Law:**
  + **Rules kept: 5:** A small number of rules indicates that your filtering thresholds ( min\_sup=0.02, min\_conf=0.6, min\_lift=1.2 ) are quite tight, or that the data does not have many strong relationships that meet those criteria. This small number of rules makes it easier for you to analyze and interpret.
  + **Avg support: 0.074:** On average, 7.4% of the crime events in your data contain both the left and right sides of the filtered rules. This prevalence is reasonable.
  + **Avg confidence: 0.779:** On average, 77.9% of the time when the left side of these rules occurs, the right side will also occur. This is a very high confidence level, indicating that these rules have strong predictive power.
  + **Avg lift: 2.20:** On average, the right side of a rule is 2.2 times more likely to occur when the left side occurs than when the left side does not occur. This is a very good Lift score, confirming that the relationships found are significant and not random.
  + **Overall assessment:** These figures show that the filtered laws are **very strong and meaningful** . Although the number of laws is not large, their quality is very high in terms of popularity, reliability and level of connection. The content of the laws is also meaningful and reasonable in the context of the crimes.

#### **2. Business Recommendations**

**Overall comments on model performance (Important):**

* **Poor performance of supervised learning models:** Whether it is classification or regression, your models are having serious problems ( AUC close to 0.5 for classification, negative R² for regression). This shows that the current data, with the available features, **does not contain enough information to predict accurately** Offence Type or Offence Count using traditional supervised learning methods.
* **The underlying problem may lie in the data:** The fact that all three models (classification, regression, clustering) performed poorly (except for association rules) suggests a strong possibility that **the data provided does not contain enough predictive power** for those tasks. The categorical variables may not have linear or complex relationships for the supervised learning model to capture.
* **Value of Unsupervised Learning: However, using Unsupervised Learning** methods like **Association Rules Mining** and **Clustering** , we have found certain **relationships and structures** between Crime Type and Occurrence Location that can provide useful information.

**Action recommendations based on Association Rules:**

The association rules found are **actionable insights** and **have practical implications** for law enforcement and local governments:

* **Allocate patrol resources effectively:**
  + **For Property and deception offences : Increase** patrols, surveillance and installation of security systems in **retail areas** , **grounds/surrounding land** and **other transport** , as these are locations where this type of crime is likely to occur.
  + **For Justice procedures offences : Although** they primarily occur in judicial venues, the fact that they also involve **private dwellings** suggests that specific procedures or strategies are required when carrying out these procedures in residential settings.
* **Develop a targeted crime prevention strategy:**
  + **In retail stores:** Implement programs to prevent shoplifting and fraud, and train employees on suspicious signs.
  + **In residential areas:** Raise community awareness of property security, encourage the installation of cameras and alarm systems; and consider special security measures when there are legal activities in private homes.
  + **In public areas and transportation:** Increase surveillance, lighting, and security presence to reduce property crime.
* **Raising public awareness:** Organize media campaigns to alert the public to specific crime risks in different types of locations and provide guidance on appropriate preventive measures.
* **Aid in investigation:** When a particular type of crime occurs, association rules can provide initial clues about the type of location likely involved, helping to narrow the scope of the investigation.

**Limitations and Recommendations for the Future:**

* **Lack of detailed temporal elements:** The current data does not include detailed temporal elements (e.g., hour of day, day of week, season). Adding this data could enrich the association rules (e.g., {Theft, Night} $\rightarrow$ {Residential} ), providing more detailed information for resource allocation over time.
* **Lack of complementary data:** It is possible to incorporate additional socioeconomic data (median income, population density, unemployment rates) of areas to find macro-environmental factors related to crime. This can help explain why certain types of crime are more prevalent in certain locations and provide the potential for better predictive models.
* **Further research on supervised learning models:** Despite poor performance, further research and testing of other supervised learning models (e.g. Gradient Boosting, Neural Networks) or advanced feature engineering techniques (Feature Engineering) may yield better results if more data or deeper understanding of the relationships between variables is available.
* **Consider spatial data mining techniques:** If more precise geographic data (GPS coordinates) are available, spatial data mining techniques can be applied to identify crime hotspots and spatial patterns of crime.